ABC 2021, 8(2):166-189
Animal Behavior and Cognition
DOI: https://doi.org/10.26451/abc.08.02.06.2021

New Caledonian Crows and Hidden Causal Agents Revisited

Laurie O'Neill[1,2,*], Garp Linder[3], Magdalena van Buuren[4], and Auguste M.P. von Bayern[1,2,3,*]

[1]Max Planck Institute for Ornithology, Germany
[2]Max Planck Comparative Cognition Research Station, Loro Parque Fundación, Spain
[3]Department Biology II, Ludwig-Maximilians-University of Munich, Germany
[4]Department of Behavioural Biology, University of Vienna, Austria

*Corresponding authors (Email: laurencerichardoneill@gmail.com & avbayern@orn.mpg.de)

Citation – O'Neill, L., Linder, G., van Buuren, M., & von Bayern, A.M.P. (2021). New Caledonian crows and hidden causal agents revisited. *Animal Behavior and Cognition, 8*(*2*), 166-189. https://doi.org/10.26451/abc.08.02.06.2021

Abstract – A previous experiment suggested that New Caledonian crows (*Corvus moneduloides;* henceforth NCCs) can reason about hidden causal agents (Taylor et al., 2012). In that study, subjects showed greater vigilance towards an area from which they had previously witnessed a threatening "stick attack" if a hidden causal agent (a human) could still be present in that area compared to when a human person had visibly left. Interpretations of these results were challenged in two commentaries (Boogert et al., 2013; Dymond et al., 2013). We aimed to replicate this experiment with a different group of NCCs (*N = 14*) whilst also adding three additional control groups that addressed the issues raised in the two commentaries. These four experimental groups included a direct replication (*n = 4*), a counterbalance of the events of that replication (*n = 4*), a control group to see if alternative associative cues would create the same effect (*n = 3*) and finally, a counterbalanced group of these alternative associative cues (*n = 3*). The direct replication group did not replicate the effect of Taylor et al. (2012). The fact that we did not replicate the effect meant that further interpretation of our other control groups proved difficult. The low sample size of our replication group meant we could not be sure if we did not replicate the effect due to low power or due to actual differences. Our findings neither support nor refute whether NCCs reason about hidden causal agents.

Keywords – Causal agents, Causal cognition, New Caledonian crows, Corvids

A classical developmental study on human children involves presenting them with a scenario in which two objects on a screen collide with each other (Ball, 1973; Michotte, 1963; Spelke & van der Walle, 1993; Spelke et al., 1994). Children from a very young age react surprised by objects that do not follow typical cause-effect relationships. For example, if a child repeatedly sees an instant transfer of movement between two blocks when one collides into the other, they quickly become disinterested as if it were a normal event (Ball, 1973; Leslie & Keeble, 1987). However, if the normal causal structure is broken, for example if there is a transfer of force between objects that do not touch or if the transfer of movement is delayed, the children maintain an interest in this event for longer, as if observing that it is atypical (Leslie & Keeble, 1987). Various versions on this test have been used to show that children develop some basic understanding of causality and object motion from an early age.

In a similar manner, children as young as six months old show an increased interest in objects with unexpected animacy (Luo et al., 2009). By ten months, if a child sees an object moving unexpectedly, they expect to later see a hidden cause of this movement. Specifically, they express a heightened reaction to a thrown toy if it comes from the direction of an inanimate object (a toy train) rather than an animate agent (a person's hand) (Saxe et al., 2005, 2007). The development of how children perceive the causal structure of why objects move, whether it might be self-propelled, and how hidden agents may cause movement has been studied intensely (Spelke & van der Walle, 1993; Spelke et al., 1994), and scientists continue to employ these paradigms (Wu et al., 2016). However, it is important to determine if the development of these factors is similar in animals other than humans. To this end, animals' perceptions of animacy and agency cues have been studied in many taxa, including fish (Wisenden & Harter, 2001), dogs (Abdai et al., 2017), and infant monkeys (Tsutsumi et al., 2012).

Some studies have investigated how birds understand the causality of motion. A test with newborn chicks (*Gallus gallus)* showed that detection of self-propelled, animate objects might be an innate sensitivity (Mascalzoni et al., 2010). New hatchlings showed a preference to follow objects that were self-propelled rather than ones which were propelled by the force of other objects. There is also evidence that some birds are able to categorize which objects in their environment are expected to have animacy. Jackdaws (*Coloeus monedula)* showed greater wariness towards an animate stick than towards an animate model bird or snake (Greggor et al., 2018). Furthermore, New Caledonian crows (*Corvus moneduloides;* henceforth, NCCs*)* can use the movement of a stick to infer the presence of a hidden causal agent (Taylor et al., 2012). It is this last study that we will now focus our attention on as the interpretations of this paper have drawn some criticism.

In this study, a group of NCCs were first habituated to extract food from a box placed directly next to a hide. Following this, they faced two different experimental conditions after each of which they were allowed to feed from the same box again. In the first condition, they observed a human enter the hide, then a stick protrude from the hide in a repeated, poking motion towards the box and finally the human leaving the hide ("Human causal agent" condition). In the second condition, they observed only the identical movement of the poking stick but no human entering beforehand or exiting the hide afterwards ("Unseen causal agent" condition). In this second condition, the birds made significantly more visual inspections of the hide before obtaining the food from the feeding box. The interpretation of this result was that the birds must have viewed the human as the cause of the moving stick. Therefore, when they did not see a human leave the hide in the second condition, they believed the human was still in the hide and therefore might still poke them with the stick when they approached the feeding box.

The interpretations of these results have raised criticism because desirable controls were lacking. The first missing control was a counterbalance of the two conditions presented to the birds (Dymond et al., 2013). Taylor and colleagues (2012) presented all eight subjects with the same order of conditions, namely "Human causal agent" before "Unseen causal agent." A counterbalanced order would have been necessary to establish that the crows' change in vigilance behavior was due to the change in conditions only, and not any factor related to the order in which they were presented. For example, there is evidence to show that animals will sometimes sensitize to an aversive stimulus (Groves & Thompson, 1970 *from* Dymond et al., 2013). The results, therefore, may have just shown the crows' sensitizing to the potential danger of the poking stick, i.e., becoming more vigilant to a known threat after a certain amount of exposure. The counterbalance would have also added more conclusive evidence for the authors' interpretation if a similar pattern in vigilance behavior was seen; a greatly escalated level of vigilance behavior to the novel unseen causal agent condition *and* a more diminished level of vigilance to the following human causal agent condition. This could potentially be interpreted as the crows using the human as a post-explanatory factor of the sticks movement and why it would not move again, much like the reveal of a human hand explaining the movement of toy in developmental studies with children (Saxe et al., 2005, 2007).

The next set of missing controls concerned the interpretation of how the crows viewed the human in the test. The evidence does not necessarily show that the crows believed the human was a *causal* agent. Instead, they could have just used the visual cue of a human entering and exiting the hide as a signal of the start and finish of the "stick attack" (Boogert et al., 2013). The signal of a human entering and exiting the hide is missing during the unseen causal agent condition. It is possible that the crows were more vigilant during this condition because they did not have an exact cue, a human leaving the hide, of when the stick would stop moving. In this scenario, the human is not viewed as a causal agent, but simply as a conditioned stimulus that marks the timing of an unconditioned stimulus, i.e., the "stick attack." The crows had no evidence that the stick was not operated by a causal agent other than the human, or by itself a miraculously self-propelled animate object, so they showed reasonable fear of the more ambiguous situation. To control for this potential explanation of the results, an alternative non-causal agent cue could be used as an arbitrary stimulus in another set of conditions. For example, an auditory stimulus could mark the beginning and end of the stick attack in one condition, and this could then be removed in the following condition (equaling the “Unseen causal agent” condition). If the crows were to respond in the same way to this as to the human cue, then it would show that they had not necessarily viewed the human as a causal *agent*.

A similar alternative interpretation is that the birds did not perceive the “stick attack” as threatening but were afraid of the human instead. In this scenario, the birds used the stick attack as a cue for the presence of a human, whom they do not see leaving in the unseen causal agent condition. The increased vigilance in this condition is therefore not because the crows had reasoned that a hidden human was the cause of the stick attack, but that the stick attack announced the potential presence of a large, possibly scary and unpredictable animal (a human) (Boogert et al., 2013). In this instance, the moving stick is a conditioned stimulus associated with the presence of a human, which would then be the fear inducing unconditioned stimulus. Another set of controls are needed to check for this interpretation. If the potentially threatening stick attack was replaced with a completely unthreatening visual stimulus, then a similar reaction from the crows would suggest their fear is related only to assumed human presence.

Finally, there were two more issues that may have biased the results involuntarily that need to be controlled for in a replication. The stick’s movement was human controlled and it was not stated whether that human was blind to the conditions. Unconscious bias can easily create involuntary differences in movements, which may have led to the stick being moved differently in the different conditions. It is more controlled if the stick is mechanically controlled and thus has identical movements in each condition. Another factor that is likely to have been difficult to control in Taylor and colleagues’ study (2012) was the presence of a second human in the room with the crows in all conditions. This was implemented to prevent the crows from approaching the food box before the stick poking had finished. This also presents a potential problem of unconscious bias between conditions, even if this second human was facing away from the subject, so it should be changed if possible.

Taking into account the above discussion points, the experiment conducted by Taylor et al. (2012) was replicated with the following groups: A first group to directly replicate the original study, with the same conditions in the same order, a second group that completes the same conditions, but in the reverse order, then, finally, a third and fourth group that completes two new sets of conditions but in counterbalanced orders; two conditions that comprise a “stick attack” cued by a non-causal signal (zero-agent controls) and two conditions that replace the “stick-attack” with a non-threatening stimulus to cue the presence of a human (human-presence controls). These groups and the different conditions are shown in Tables 1 and 2. Throughout all of the conditions, modifications that reduce any potential unconscious bias by human experimenters have been added.

Method

Subjects

The experiments were carried out on fourteen NCCs of which six were wild-caught as adults and eight were wild-caught as juveniles in the beginning of 2010. The younger individuals were less than five years of age at the time of testing in 2014 and so were still considered “sub-adults.” Based on blood tests, seven of them were females and seven were males.

Ethics Statement

All animals were housed in accordance with British and German Law. As the experiments were strictly non-invasive and based purely on behavioral observations, they were not classified as animal experiments in accordance with the German and British Animal experiments Act (Germany: §7 Bundestierschutzgesetz; UK: Animal (Scientific Procedures) Act).

Table 1

The Order of the Protocol for the Pre-Test and the Six Different Experimental Conditions the Subjects Could Take Part in

	Condition		30 s	30 s	30 s	
	Pre-test	*Food box is baited followed by a 60 second pause*	*Cloth removed - test starts*	*N/A*	*N/A*	
Replication & Reverse Order	Human Causal Agent		Human walks *into* hide	Stick pokes 10 times	Human walks *out* of hide	*Cloth removed - test starts*
	Unseen Causal Agent		*Nothing*	Stick pokes 10 times	*Nothing*	
Human Presence Controls	Human Cued Light		Human walks *into* hide	Light is on (for 30 s)	Human walks *out* of hide	
	Ghost Cued Light		*Nothing*	Light is on (for 30 s)	*Nothing*	
Zero-agent Controls	Sound Cued Stick		Sound signal (for 30 s)	Stick pokes 10 times	Sound signal (for 30 s)	
	Ghost Cued Stick		*Nothing*	Stick pokes 10 times	*Nothing*	

Note. Two groups (both *n* = 4), the Replication and Reverse-order groups, completed the Replication & Reverse Order protocols as described below, but in a counter-balanced order. Two other groups (both *n* = 3) completed both the human presence controls and the zero-agent controls. Both groups completed the human presence controls first and then the zero-agent controls, the order of the two conditions within those were reversed between those two groups. The full protocol order is shown in Table 2.

Table 2

The Name, Age and Sex of Subjects and the Order of Experimental Conditions They Took Part in

Subjects	Age	Sex	Experimental Group	Order of testing			
Admiral Amais Jungle Tumulte	Adult Adult sub-adult sub-adult	male female male female	**Replication**	*Human Causal Agent (HCA)*	*Unseen Causal Agent (UCA)*		
Thetys Titan Liane Aigaios	Adult Adult sub-adult sub-adult	male female female male	**Reverse Order**	*Unseen Causal Agent (UCA)*	*Human Causal Agent (HCA)*		
Mermaid Papaye Tabou	Adult Adult sub-adult	female male female	**Zero agent & human presence controls**	*Human Cued Light*	*Ghost Cued Light*	*Sound Cued Stick*	*Ghost Cued Stick*
Captain Hook Mango Tortue	Adult sub-adult sub-adult	male male female	**Zero agent & human presence Reverse Order**	*Ghost Cued Light*	*Human Cued Light*	*Ghost Cued Stick*	*Sound Cued Stick*

Housing Conditions

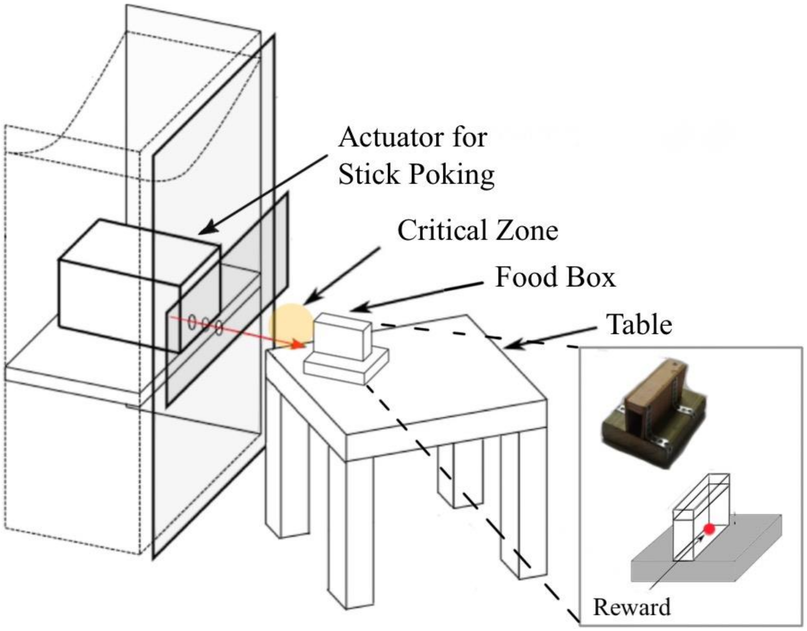
Observations and experiments took place at the Avian Cognition Research Group hosted by the Max Planck Institute for Ornithology (Seewiesen, Germany). All crows were housed in pairs (always one female and one male bird) in eight separated outdoor aviaries. The outdoor aviaries varied in size, but were all at least 17.5 m^2 in area and 2.60–3.00 m high. Each aviary consisted of an outdoor area and a heated and lit indoor area in which the birds had *ad libitum* access to food and water. The outdoor aviaries contained a variety of perches, an artificial nesting zone and were enriched with natural vegetation, pebbles, and water pools. The birds were kept on a diet consisting of a mixture of minced beef heart, curd, dried insects, rice, egg powder, vegetable oils, vitamin and mineral supplements, as well as “Versele Laga Nutribird beopearls,” soaked cat biscuits, cereals, seeds, and fruit. As rewards in the experiment, they were given larvae of *Zophobas morio*.

Experimental Setup

The experiments were conducted in the outdoor aviaries of each crow pair. A hide (ca. 2.50 m x 1 m) was constructed on the outside of one of the walls of each aviary. A table was placed in front of it inside the aviary and a food box was presented on the table (Figure 1). The hide was made of opaque green plastic and was large enough to fully hide an adult human and table with its three walls. It was designed so that there was no chance for the crows to see anything inside the hide from any position from inside their aviaries. In the experimental trials, the food box was placed on a table (55 cm x 55 cm) 20 cm from the hide in such a manner that the only way to access the food was for the bird to sit on the food box and bend their neck downwards thus displaying their sensitive nape area towards the hide while looking inside the box.

Figure 1

A Schematic of the Critical Apparatus



Note. The critical zone is the area that the crows needed to place their head in order to put a stick inside the food box for extracting the reward. This was also in the same area that the stick poked.

A remote-controlled linear actuator was used to create the protruding stick movements from the hide. The movements were therefore more standardized across conditions and there could also be no unconscious cueing by a human from a mechanical movement. The movement consisted of a 12 cm long stick emerging from the hide into the critical “danger” area near the opening of the food box and then retracting again into the hide. This was controlled remotely by the experimenter pressing a button to both make the stick emerge and retract with separate presses. In each stick attack, they made the stick probe in and out ten times spread every three seconds across 30 s in an exact and repeatable way.

The hollow food box was created so that a reward could be hidden inside (6 cm depth, 1 cm width, 3 cm height) that could be accessed only from a single end and required a stick tool for extracting the reward, which is a habitual behavior in New Caledonian crows (Hunt, 1996). Subjects were provided with two 12 cm long wooden sticks to use as tools for each trial. These were tucked into the food-box each time it was baited.

The food-box was covered with a white satin cloth to prevent birds from accessing it before the experimental manipulations of each trial had been observed. This replaced the necessity of a second experimenter being present. A thin black string was attached to the cloth so the experimenter could remove it from a distance while being out of sight.

Experimental Procedures

All the birds were housed as pairs. Before each of the habituation and experimental trials, focal subjects were separated from their partners. To do this, the non-focal subject of the pair was ushered into the indoor section of their aviary and closed inside during the trial. This prevented visual access between the two birds, which ensured that subjects did not see experimental trials they did not take part in. All tests took place between April to October of 2014.

The crows were first habituated to remove food from the food-box while the hide was in position at the side of the aviary. The table was placed in the middle of the aviary (away from the hide) and surrounded with rewards, with an especially favorable reward (giant mealworm - *Zophobas morio*) placed at the entrance of the food box. After they had eaten the reward at the entrance, rewards were placed progressively further inside the food box so the crows would learn to remove these rewards with tools. Once the birds were able to extract food from the box with a tool, the box was placed gradually closer to the hide (100 cm → 50 cm → 20 cm) in successive habituation sessions. They had to successfully extract the food twice at each distance with minimal hesitation (approaching in less than five minutes) before moving closer. After they were successful at 20 cm, the experimental phases began.

Pre-test Trials

The pre-test trials followed the procedure by Taylor and colleagues (2012) and their purpose was to establish a baseline level of the crows’ vigilance behavior that was not influenced by their neophobia of the apparatus. Thus, after they had been habituated to obtain food from the food-box at a distance of 20 cm from the hide after the satin cloth had been pulled off, the same setup was repeated and recorded three times for each subject. All the birds had to extract food from the box three times in a row but under the following conditions: landing on the table in under two minutes and then extracting the reward from the food box within one minute of landing on the table.

Subjects were then assigned to one of four experimental groups semi-randomly so that the groups had a balanced sex and age ratio. Each test group had a sequence of two or four test conditions, each condition is described below and summarized in Table 1. In total, there were six different experimental conditions, the first two groups took part in two experimental conditions, and the third and fourth groups each took part in four experimental conditions, the order of these conditions is shown in Table 2. Subjects participated in three trials per condition. These three trials were typically carried out in one session with a maximum of one session carried out per day per bird. All of the experimental conditions followed the exact same timeframe, so the human causal agent condition is described in detail, and the following conditions describe the differences to this condition only.

Human Causal Agent

The experimenter entered the aviary with the subject and set up the apparatus. This included baiting the food-box and covering the top of the table with the cloth. The experimenter then walked out of the aviary via a door on the outside of the aviary and walked to a hiding place that was out of view of the subject. The experimenter then disguised themselves with a ski mask, hat, and large red jacket (all unknown to the birds). Sixty seconds after having exited the aviary, the disguised experimenter (walking with a slight limp to camouflage any recognizable gait) walked from their initial hiding place to the hide attached to the aviary. This involved walking past one of the wire mesh walls of the aviary, which meant the subjects saw the disguised experimenter enter into the hide. Thirty seconds after the experimenter had started walking from their hiding place (by which point they were now in the hide) and 90 s after leaving the aviary for baiting, they began the stick poking protocol. This involved poking the mechanically controlled stick ten times every three seconds (within 30 s in total) into the area where the bird would need to place its head in order to retrieve the food. The disguised experimenter then left the hide and returned to their initial hiding place. Thirty seconds after having left the hide, the experimenter remotely removed the satin cloth by pulling the string it was attached to, which gave the subject access to the food box thus starting the test.

Unseen Causal Agent

The critical difference between the human causal agent and the unseen causal agent condition was that the disguised experimenter did not enter and exit the hide on either side of the stick attack. However, the timing of the condition was controlled so that it was the same as the human causal agent condition (see Table 1). Thus, the stick poked the same number of times after the same 90 s delay after the experimenter had set up the experimental apparatus and left the aviary.

Human Cued Light

In this condition, there was the same sequence of events as in the “human causal agent” condition except the stick-poking was replaced with a thirty second “blinking” light signal from a remote-controlled LED lamp mounted at the same position as the hole in the hide which the stick came out of.

Ghost Cued Light

In this condition, there was the same sequence of events as in the “unseen causal agent” condition except again, the stick poking was replaced with the thirty second blinking light signal.

Sound Cued Stick

In this condition, there was the same sequence of events as in the “human causal agent” condition except the action of a human walking to and from the hide before and after the stick poking attack was replaced by a sound signal. Specifically, it was a 30 second recording of church bells.

Ghost Cued Stick

This condition was identical to the “unseen causal agent” condition in every way except that it was paired with the ‘sound cued stick’ condition, so the different name was given to clearly separate the two conditions.

Behavioral Coding

All pre-tests and experimental conditions were filmed on HD Samsung Handycams. These videos were observed and certain behaviors were scored and coded using Solomon Coder (András Péter, solomon.andraspeter.com). The aim was to measure the degree of caution that the birds were showing in each of the conditions. This was expressed by a number of different variables.

We counted the number of “hide inspections” subjects made before obtaining the reward as defined in Taylor et al. (2012).

> "Hide inspections were defined as a crow orientating its head towards the hole (the opening of the baited food box) and then moving its head toward the hide so that one or both eyes was/were in line with the hide. Orientations to the hide were not scored if the crow was not first looking toward the baited hole or if the crow looked at an area of the cage other than the hide after looking at the hole" (p. 2).

They were counted only if the subject was standing on the experimental table; thus, from the moment the bird landed on the experimental table until they removed the reward from the food box, and not beyond two minutes. As in the original Taylor et al. (2012) experiment, the hide inspections were calculated as a rate per minute. The rate was calculated as the number of hide inspections/total time on experimental table before extraction (in minutes). Thus, the latency between the time the birds landed on the experimental table until they obtained the reward was also calculated. This differed slightly to Taylor et al.'s (2012) measure in that they measured it from the moment the crows picked up a tool from the table. In this instance, we had to differ as in their experiment, the crows were able to position the stick tools behind the feeding box freely on the surface of the experimental area. In our experiment, we had to secure the sticks by tucking them into the food box so that they would not be blown away by wind. We decided to use the cue of 'landing on the table' as it seemed more comparable to the time point in which the NCCs picked up the stick in the Taylor et al. (2012) experiment, as observable from videos of their experiment available online.

As in the Taylor et al. (2012) study, we also counted the number of times the birds "abandoned probing" during the trials. An abandoned probe was defined as a bird approaching the food box, inserting a tool, and then leaving the testing area before extracting the food. This behavior was rare in our study. It never happened in the pretest and it happened only once or twice per experimental condition. Thus, we do not discuss these data further, but they are available in the online data repository.

In contrast to Taylor et al. (2012), the final variable we decided to take to operationalize caution was the latency for the subjects to place their head in the critical "danger" zone in each trial from the moment they had landed on the experimental table. The zone is marked on Figure 1, and was defined as the area where the stick poked that the subjects needed to place their head to obtain the reward. We diverged from Taylor et al. (2012) here because we had difficulties determining the crows' "hide inspections" based on their head orientation objectively following Taylor et al.'s (2012) description and because we considered "latency to put the head in the critical zone" (also referred to as HICZ) to reflect the subjects' vigilance more objectively and more conservatively than the measure 'hide inspection rate'.

Twenty percent of the videos were then coded by a second observer to check inter-observer reliability. There was "excellent" reliability between the two observers for both hide inspection count (Intraclass correlation = 96.5% consistency, R-package "irr") and latency to place the head in critical zone (Intraclass correlation = 96.6% consistency*)*.

Data Analysis

All data were analyzed in R, version 3.6.3. Firstly, we repeated comparable frequentist statistics as used in Taylor et al. (2012). We did these tests only for the Replication and Reverse-order groups because after we finished analyzing those groups, it became clear that further statistical interpretation of the Zero-agent and Human presence control conditions was not feasible, which is discussed below. As the Replication and Reverse-Order groups had an *n* of only four, we opted to use Student's paired *t*-test for comparisons between conditions even though we could not show the data was normally distributed with such a low number of subjects. However, for such a low *n,* the *t*-test can be justifiably used for avoidance of type-2 errors (de Winter, 2013).

We additionally combined the data of the Replication and Reverse-order groups by both 'condition' (human causal agent and unseen causal agent) and order of condition faced (which of the two conditions they faced first and then second). We re-analyzed this data using Wilcoxon signed rank tests as there was now an *n* of 8 with this combined data set. This analysis however has to be taken with the following caveat in mind. As it stood, the subjects' expectancy of what might happen in the second condition of the Replication group (unseen causal agent) and in the second condition of the Reverse-order group (hidden causal agent) could have differed based on what they had observed in the first condition respectively. In order to truly judge the effect of the condition, as well as the order of condition, two further experimental groups would have been required, i.e., one in which the subjects are given the “hidden causal agent” condition twice in a row and one in which they are given the “unseen causal agent” twice in a row. Thus, although we report the comparison of those two second conditions here, we recognize they may not solely be explained by an effect of order, but additionally by the birds’ previous differing experience.

To gain further insights into the effects of the factors that affected hide inspection rate, we also ran two different models in R using the packages “lme4” and “arm.” The first was just on the two groups with the same conditions as the Taylor et al. (2012) test (Replication and Reverse-order groups), and the second included all of the experimental groups. We used Bayesian methods to draw conclusions from these models due to the low sample sizes.

For the Replication and Reverse order groups we used a generalized linear model (glm) with a Poisson distribution with the response variable of hide inspection count and the fixed effects of condition (pre-test (control), human causal agent, unseen causal agent), order of condition (first condition, second condition), an interaction between condition and order, trial number (1,2,3), age at capture (Juvenile, Adult) and finally with an offset of the experimental time in minutes (so that the dependent variable “hide inspection count” represented the hide inspection rate). Model fit was assessed by visual inspection of the residuals. Originally, “subject” was added as a random factor in a mixed model, but it was shown to have zero effect, so was removed and a simpler model was run. The natural logarithm link function was used. To obtain posterior distribution, the function ‘sim’ from the package “arm” (Gelman & Hill, 2006) was used to directly simulate 2000 values from the joint posterior distribution of the model parameters. To draw inference from the model, posterior distributions of 2000 fitted values were calculated, each with a different set of model parameters. The mean and the 2.5% and 97.5% quantiles of these 2000 values were used as the estimate and the 95% credible interval. We considered effects as statistically meaningful if there was no overlap in the credible intervals between the different factors.

The second model we ran included all the conditions from all the groups. However, as we highlighted above, we did not include the second condition that each experimental group faced as we could not be sure of the different expectations each group had going into the second condition based on the first condition they faced. We used a generalized linear mixed model (glmer) with a Poisson distribution using the response variable of hide inspection count and the fixed effects of condition (pretest (control), human causal agent, unseen causal agent, sound cued stick, ghost cued stick, human cued light, ghost cued light), age at capture (adult, juvenile), trial number (1,2,3), an offset of the experimental time in minutes and finally ‘subject’ was added as a random factor. Model fit was assessed by visual inspection of the residuals.

As we described above, we used “sim” from the package “arm” to simulate 2000 values from the joint posterior distribution of the model parameters and then extracted the means and 95% credible intervals from this simulated data. Again, we considered effects as statistically meaningful if there was no overlap in the credible intervals between the different factors.

Furthermore, the raw data of the different conditions was visualized and described using ggplot2 (Wickham, 2016). In some of the plots, the data was plotted with four different response variables. First, it was plotted with the same response variable as in Taylor et al. (2012), “hide inspection rate” (described above). Next, the two variables that make “hide inspection rate,” “hide inspection count,” and the “latency to obtain reward,” were also plotted as both also individually show subjects’ vigilance behavior. Finally, the “latency to place head in the critical zone” was also plotted.

After plotting these different response variables, it was noted that the response variable “hide inspection rate” might be a noisy variable (discussed below) and the response variable “latency to place head in the critical zone” might be a more appropriate response variable to show the subjects vigilance levels in this experiment. For this reason, we re-ran all of the statistics described above but with the new response variable of “latency to place head in critical zone” (referred to as HICZ). This included all of the comparative statistics described for use on the “Replication” and “Reverse-order” groups as well as the two models run. Nevertheless, there were some necessary differences with the models. As the HICZ data was continuous and normally distributed, both were analyzed with linear mixed effects models (lmer) with a Gaussian error distribution assumed. Otherwise, we used similar Bayesian methods, as described above, to simulate values and extract means and credible intervals and then used this simulated data to assess statistically meaningful differences.

The code for all the models is supplied as a supplementary R-markdown document.

Results

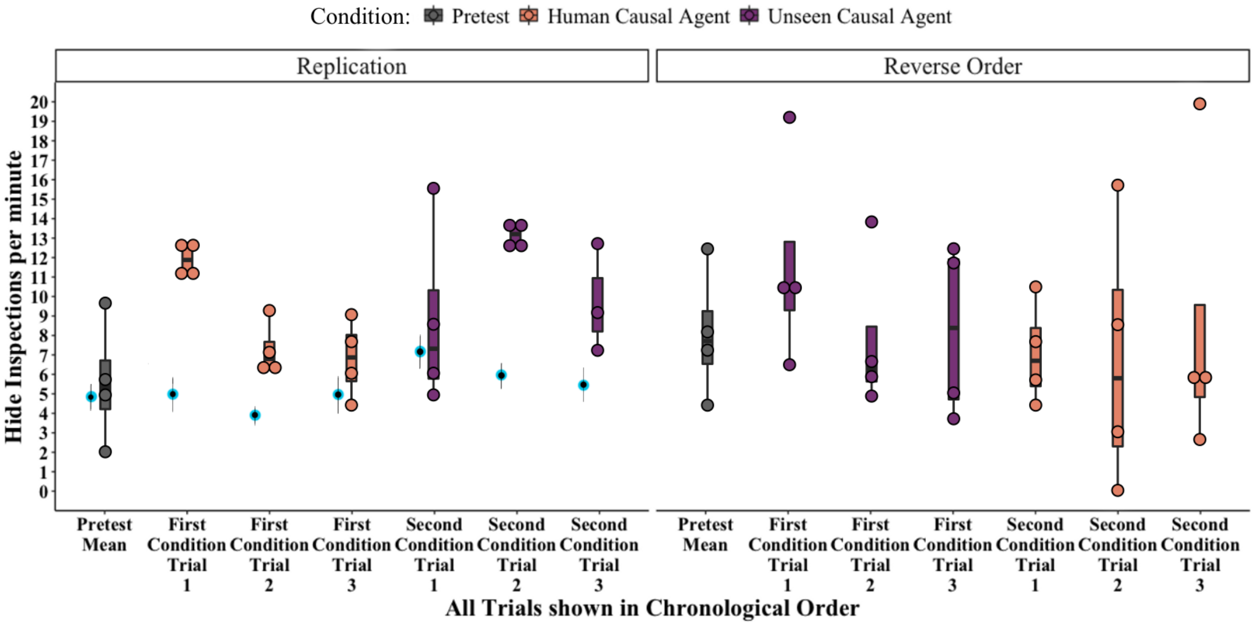
Hide Inspection Rate Data

Replication Group

From looking at the data of the Replication group, we did not appear to replicate the same effect as Taylor et al. (2012). The hide inspection rate did not increase in the unseen causal agent condition when compared to the human causal agent condition that had been tested first (Figure 2, left).

Figure 2

The “Hide Inspections Per Minute” of Both the Replication and the Reverse Order Groups



Note. The hide inspection rate of each individual is shown as colored points and additionally box-plots show the median, interquartile range, and limits. The mean of the subjects’ three pre-test trials from each group is plotted on the left of each figure. On the graph showing the replication results on the left, the hide inspections per minute from the Taylor et al. (2012) experiment are also shown (black dots highlighted in blue, showing mean and standard error of their 8 subjects). Significant differences are not shown on this figure.

There was no significant difference in hide inspection rate between the Human Causal Agent (HCA) condition *(M* = 8.69, *SD* = 1.1, *n* = 4) and the Unseen Causal Agent (UCA) condition *(M* = 10.95, *SD* = 2.31, *n* = 4; paired *t*-test; *t*(3) = -2.16, 95% CI [-5.59, 1.07], *p* = .12). There was also no significant difference between the first HCA trial (*M* = 11.93, *SD* = 0.78*)* and the first UCA trial (*M* = 8.77, *SD* = 4.72*;* paired *t*-test; *t*(3) = 1.57, 95% CI [-3.25, 9.58], *p* = *.21*). When compared with the pre-test inspection rate (*M* = 5.63, *SD* = 3.08), there was no significant difference between the HCA (paired *t*-test; *t*(3) = -1.57, 95% CI [-9.19, 3.12], *p* = .21) nor the UCA (paired t-test, *t*(3) = -2.28, 95% CI [-12.70, 2.11], *p* = .11*)* conditions. However, there was a significant difference between the pre-test inspection rate and the first HCA trial (paired t-test; *t*(3) = -3.38, 95% CI [-12.19, -0.36], *p* = .043), but not the first UCA trial (paired t-test; *t*(3) = -0.86, 95% CI [-14.64, 8.41], *p* = .45).

Reverse Order Group

There also did not appear to be significantly different effects from the HCA and UCA groups on the hide inspection rate when these conditions were done in the reverse order (Figure 2, right). There was no significant difference in hide inspection rate between the HCA (*M* = 7.52, *SD* = 4.06) and UCA (*M* = 9.26, *SD* = 3.54) conditions (paired t-test, *t*(3) = -0.93, 95% CI [-7.68, 4.20], *p* = .42). There was also no significant difference between the first HCA trial (*M* = 7.10, *SD* = 2.63) and the first UCA trial (*M* = 11.72, *SD* = 5.46; paired t-test, *t*(3) = -1.27, 95% CI [-16.22, 6.99], *p* = .29). There were also no significant differences between the pretest inspection rate (*M* = 8.05, *SD* = 3.34) and the HCA (paired t-test, *t*(3) = 0.2, 95% CI [-8.04, 9.10], *p* = .86) and UCA (paired t-test, *t*(3) = -0.72, 95% CI [-6.59, 4.17], *p* = .52) inspection rates. Finally, there was no significant difference between the pre-test inspection rate and the inspection rates of the first HCA trial (paired t-test, *t*(3) = 0.4, 95% CI [-6.47, 8.37], *p* = .71) or the first UCA trial (paired t-test, *t*(3) = -1.85, 95% CI [-9.98, 2.65], *p* = .16).

Replication and Reverse Order Groups together

Overall, when the data between the Replication and Reverse Order groups were pooled (Supplementary Figure 1), there was no significant difference between the HCA condition (*M* = 8.10, *SD* = 2.82) and the UCA condition (*M* = 10.11, *SD* = 2.91) (Wilcoxon signed rank test, *V* = 7, *p* = .15). There was also no significant difference when only the first trials of the HCA (*M* = 9.52, *SD* = 3.15) and the UCA (*M* = 10.24, *SD* = 4.98) conditions were compared (Wilcoxon signed rank test, *V* = 19, *p* = .95). There was also no significant difference between the overall pre-test inspection rate (*M* = 6.85, *SD* = 3.24) and the HCA (Wilcoxon signed rank test, *V* = 12, *p* = .46) or UCA condition (Wilcoxon signed rank test, *V* = 5, *p* = .08).

If the results of the two groups were pooled by order of conditions faced, rather than by experimental condition, then there was no significant difference between the first condition faced (HCA of replication group and UCA of reverse order group; *M* = 9.98, *SD* = 2.45) compared to the second condition faced (*M* = 9.23, *SD* = 3.56) (Wilcoxon signed rank test, *V* = 13, *p* = .55). There was also no significant difference between the pre-test inspection rate of both groups compared to the first conditions faced (Wilcoxon signed rank test, *V* = 7, *p* = .15) or the second conditions faced (Wilcoxon signed rank test, *V* = 10, *p* = .31).

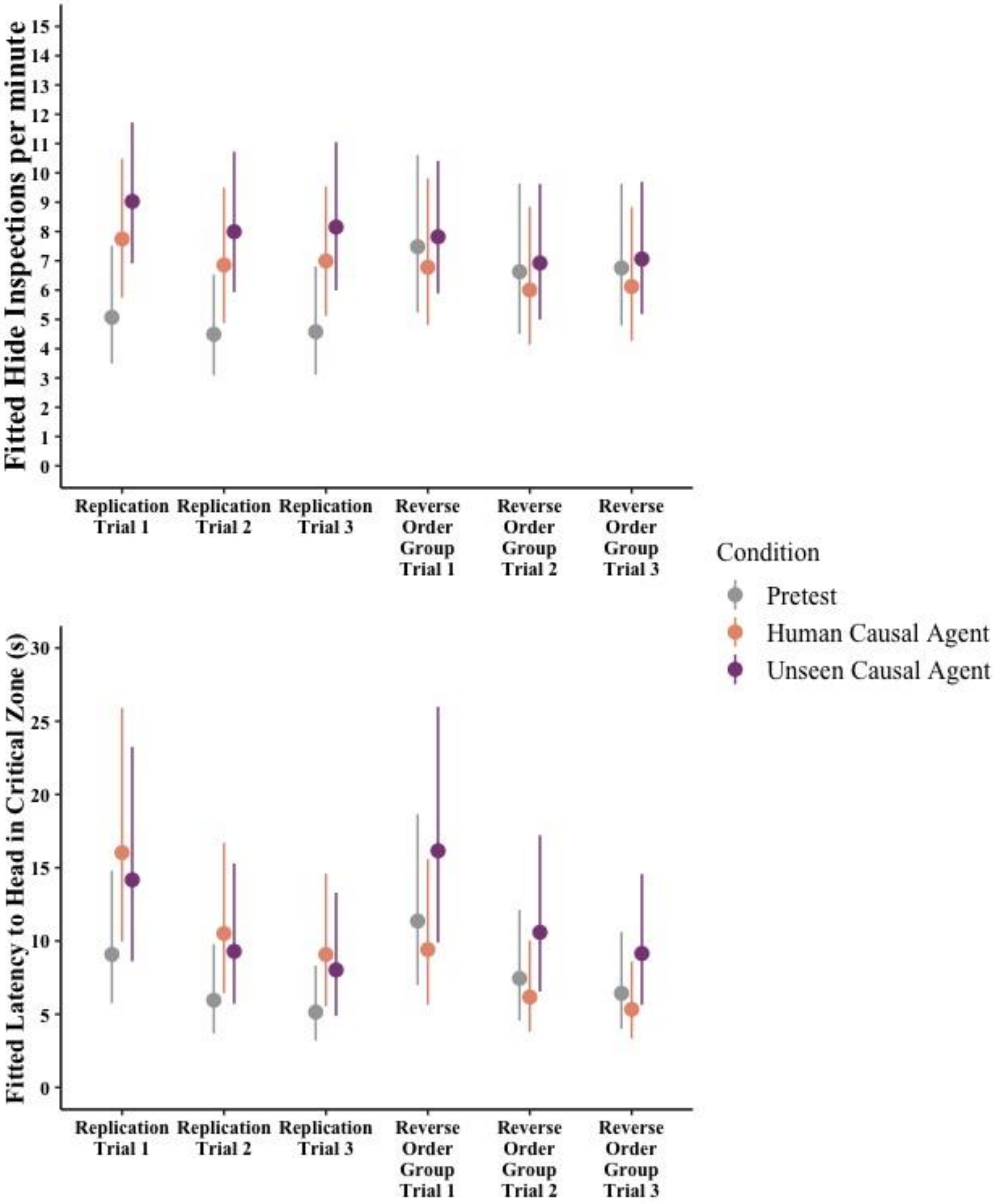
However, there was a significant difference between the pre-test inspection rate and the first trial faced (*M* = 11.83, *SD* = 3.61) by all birds in both the replication and reverse order groups (Wilcoxon signed rank test, *V* = 34, *p* = .023).

Models on Hide Inspection Rate Data

The models on hide inspection rate described in "data analysis" both showed significant effects (Supplementary Tables 1 and 2). Simulated estimates of the effects of these factors showed that none of the experimental conditions were likely to have had statistically meaningful effects on the hide inspection rate (simulated values are plotted on the upper panels of Figures 3 and 4). However, both models showed that birds that were captured as juveniles made less hide inspections than those that were captured as adults (Supplementary Tables 1 and 2).

Figure 3

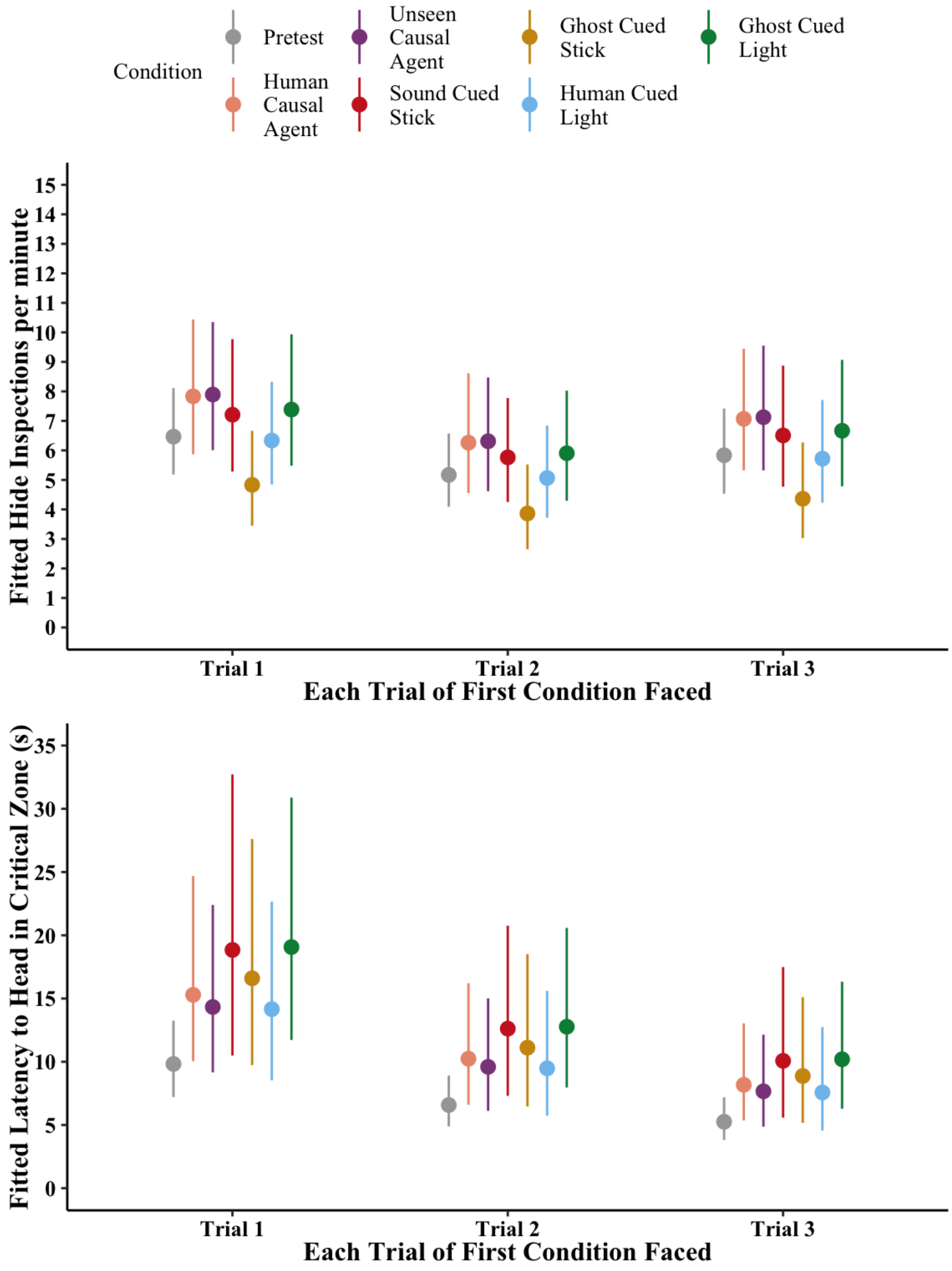
The Vigilance of New Caledonian Crows in Relation to Representations of Different Agents Causing a Stick Attack



Note. The replication and reverse order groups show the same conditions, presented in a counterbalanced order. These fitted values are drawn from 2000 simulated values based on the models shown in Supplementary Table 1 (upper figure, using "hide inspections per minute" as the dependent variable; lower figure, using latency to place head in critical zone as the dependent variable) with the effect of "age" held constant. The figures show the estimated mean and the 95% credible intervals of these simulated values, so overlapping credible intervals suggest differences are not statistically meaningful. The upper figure therefore suggests that the significant effects of the conditions shown in Supplementary Table 1 are mostly due to the lower hide inspection rates in the pre-test trials in the replication group. It is unlikely there were statistically meaningful differences between the human causal agent and the unseen causal agent in the replication group as most of the credible intervals overlap. The lower figure also shows much overlapping of credible intervals between conditions, also suggesting no statistically meaningful differences.

Figure 4

The Vigilance of New Caledonian Crows in Relation to Representations of Different Agents and Sound Cues Causing Either a Stick Attack or a Light Signal



Note. These fitted values are drawn from 2000 simulated values based on the models shown in Supplementary Table 2 (upper figure, using “hide inspections per minute” as the dependent variable; lower figure, using “latency to place head in the critical zone” as the dependent variable). The figures show the estimated mean and the 95% credible intervals of these simulated values, so overlapping credible intervals suggest differences are not statistically meaningful. In the upper figure, as most of the credible intervals overlap, it suggests that none of the different experimental conditions affected the subjects hide inspection rates in a very different way. In the lower figure, the pretest condition has a lower latency than all the other experimental conditions, which might suggest that all of the experimental conditions had some effect on the subjects’ vigilance, but again there is overlap of the credible intervals suggesting that this effect was not necessarily statistically meaningful. If there was an effect, it was always strongest in the first trial.

Other Control Conditions & Alternative Dependent Variables

To see if the different experimental conditions had affected the subjects' vigilance behavior differently, we plotted all the different conditions together to compare them (Figure 5). We checked only the first condition subjects faced from each set of conditions as we later recognized that the subjects' response to the second condition faced was likely influenced by their expectations in the first condition (order effects). This meant their expectations were different, as they had faced different first conditions.

On inspection of the data from all the different experimental groups, none of the experimental conditions appeared to have a consistent effect on the hide inspection rate (Figure 5, top left panel). To explore the data in more detail, we decided to also plot all of the different experimental conditions using alternative dependent variables, which we thought would also reflect the subjects' vigilance behavior. All four are shown in Figures 5 and 6. In Figure 5, the alternative dependent variables plotted showed that there might have been some changes in vigilance between the different conditions if measured in a different way (Figure 5, top-right and both bottom panels). In Figure 6, all three pre-test trials of each subject are also plotted next to the data from the experimental conditions. The variation within the three pre-test trials appeared to encapsulate all the variation in all the experimental conditions when using the dependent variable "hide inspection rate" (Figure 6, top left panel). This was not true for the other response variables measured (Figure 6, all other plots). Because of this, we believed it was possible that 'noise' in the dependent variable 'hide inspection rate' was obscuring effects in the data if they were present. We therefore decided to repeat the statistical analyses using one of these alternative dependent variables; in this case, we chose 'latency to place head in critical zone', which is plotted on the bottom right of Figures 5 and 6.

Latency to Place Head in Critical Zone Data

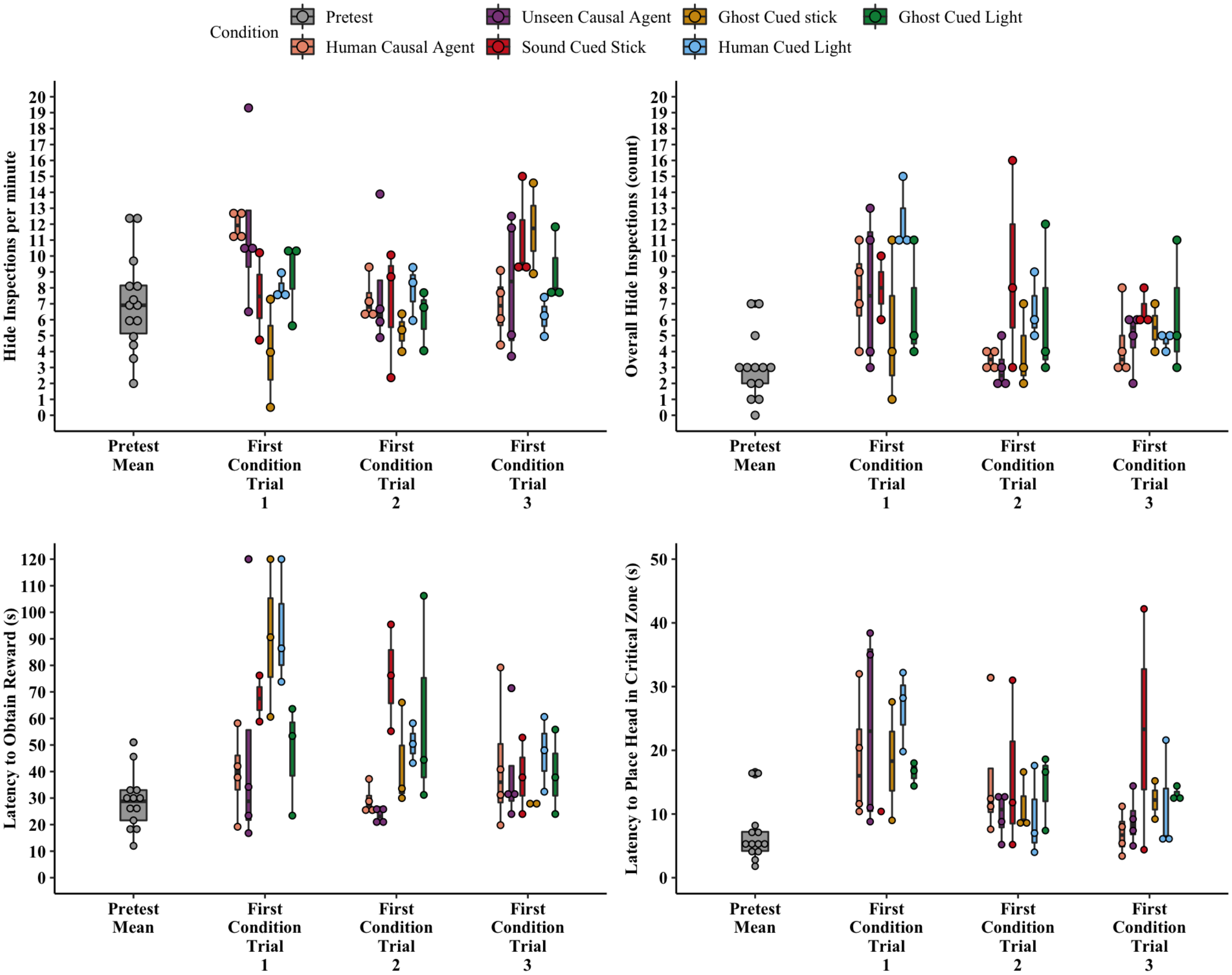
Replication Group. The latency (in seconds) for subjects to place their head in the critical zone did not appear to significantly differ between the two conditions faced (Figure 7).

In the replication group, there was no significant difference in the average latency for individuals to place their head in the critical zone between the HCA condition (*M* = 13.75 s, *SD* = 5.61, *n* = 4) and the UCA condition (*M* = 12.84 s, *SD* = 6.68, *n* = 4) (paired *t*-test, *t*(3) = 0.21, 95% CI [-12.81, 14.63], *p* = .85). There was also no significant difference between the first HCA trial (*M* = 18.60 s, *SD* = 9.98) and the first UCA trial (*M* = 15.95 s, *SD* = 13.46; paired *t*-test, *t*(3) = 0.35, 95% CI [-21.21, 26.50], *p* = .75). When compared with the pre-test latency to place the head in critical zone (*M* = 8.28 s, *SD* = 3.31), there was no significant difference between the pre-test and the HCA condition (paired *t*-test, *t*(3) = -2.27, 95% CI [-13.12, 2.19], *p* = .11) nor the UCA condition (paired *t*-test, *t*(3) = -0.97, 95% CI [-19.53, 10.41], *p* = .40). There was also no significant difference between the pre-test and the first HCA trial (paired *t*-test, *t*(3) = -2.19, 95% CI [-25.3, 4.67], *p* = .12), or the first UCA trial (paired *t*-test; *t*(3) = -0.94, 95% CI [-33.72, 18.38], *p* = .42).

Reverse Order Group. In the reverse order group, there was no significant difference in average latency to place the head in the critical zone between the HCA (*M* = 8.57 s, *SD* = 5.98) and UCA (*M* = 14.05 s, *SD* = 4.21) conditions (paired *t*-test, *t*(3) = -1.27, 95% CI [-19.22, 8.25], *p* = .29). There was also no significant difference between the first HCA trial (*M* = 13.75 s, *SD* = 14.10) and the first UCA trial (*M* = 23.30 s, *SD* = 15.56; paired t-test, *t*(3) = -0.72, 95% CI [-51.56, 32.46], *p* = .52). There were also no significant differences between the pre-test latency to place the head in the critical zone (*M* = 9.01s, *SD* = 2.72) and the HCA (paired *t*-test, *t*(3) = 0.21, 95% CI [-6.26, 7.16], *p* = .84) and UCA (paired *t*-test, *t*(3) = -2.27, 95% CI [-12.08, 2.01], *p* = .11) latencies to place the head in the critical zone. Finally, there was no significant difference between the pre-test latency to place the head in the critical zone and the latency to place the head in the critical zone in the first HCA trial (paired *t*-test, *t*(3) = -0.78, 95% CI [-24.01, 14.54], *p* = .49) or the first UCA trial (paired *t*-test, *t*(3) = -1.81, 95% CI [-39.36, 10.79], *p* = .17).

Figure 5

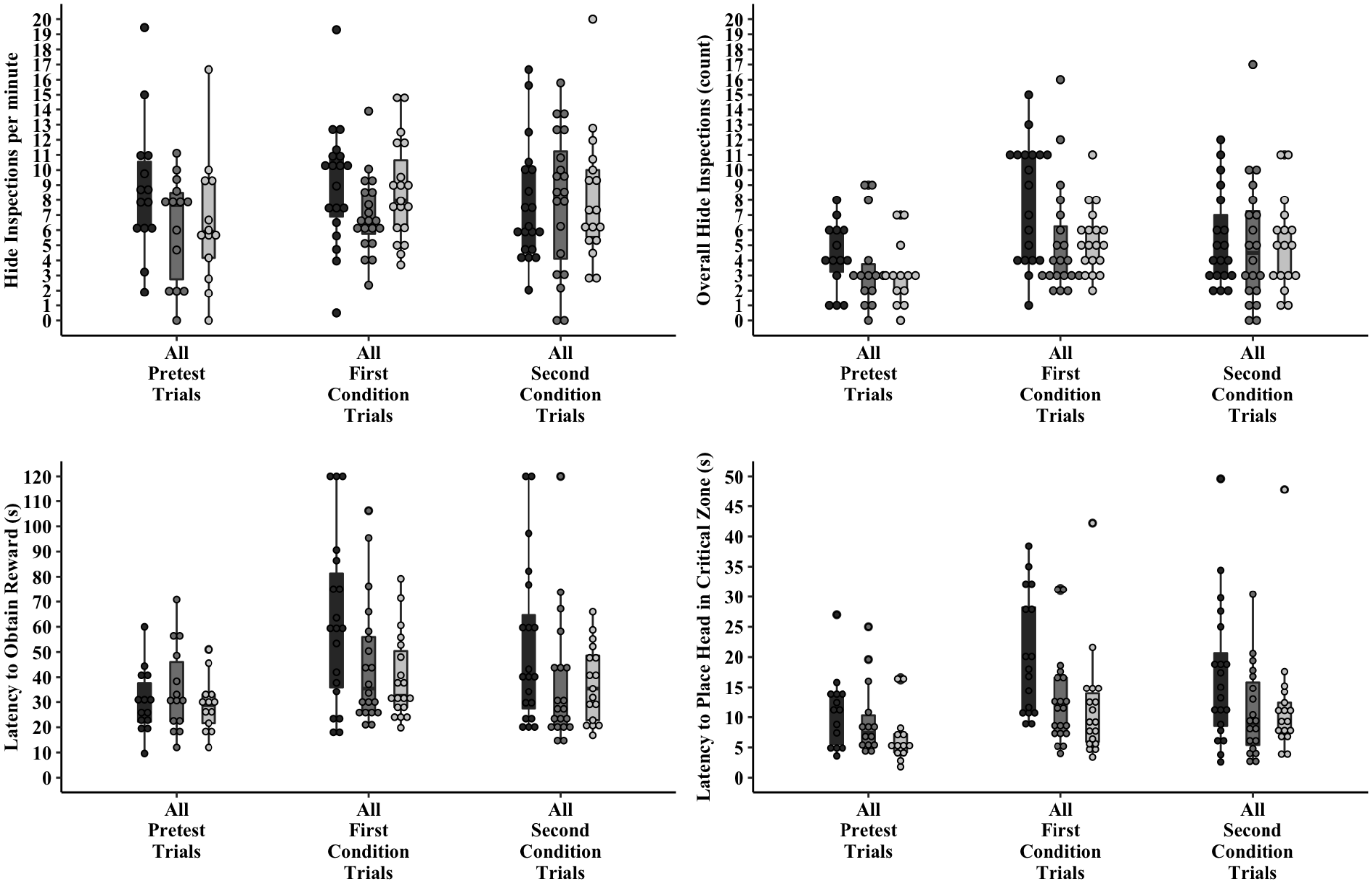
The First Condition That Each Subject Faced in the Experimental Group They Were in



Note. The four different plots show four different dependent variables, top-left: Hide inspection rate (which was obtained by divided the number of hide inspections by the latency to obtain the reward, the top right and bottom left figures), top-right: Number of hide inspections, bottom-left: The time between the start of the experiment and the subject obtaining the reward, bottom-right: The time between the start of the experiment and the subject placing their head in the “critical” zone (the danger area where the stick would poke). Most of the dependent variables show that the subjects did appear to have a response to the effects of the conditions as their vigilance increased when compared to their “baseline” vigilance level in the pretest. The human presence and zero agent control data are from the same subjects. Only the first condition is shown as the second condition each subject faced was likely affected by the subjects’ expectations from what they experienced in their first condition, which was different for each group. These “order effects” make the second condition each subject faced difficult to compare. Differences in reactions to the different conditions should still show between the groups in the first condition if it is a large difference.

Figure 6

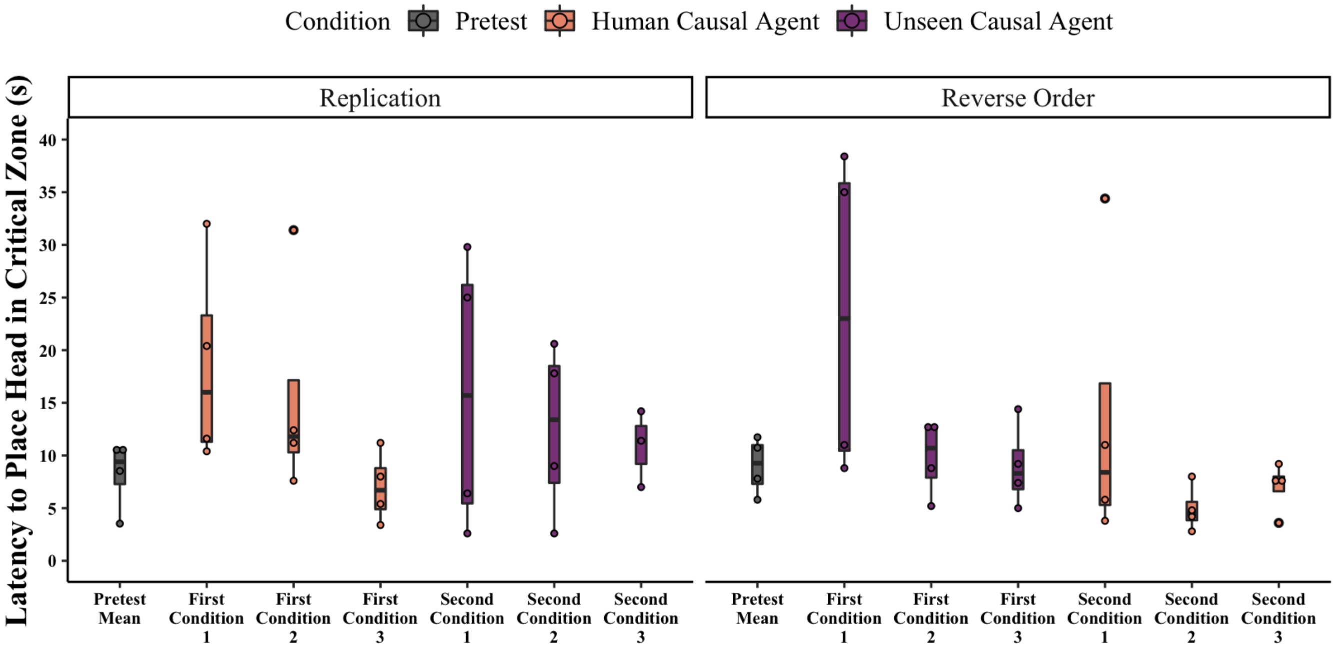
All of the Trials Subjects Faced Shown in the Order They Were Faced, Without Being Divided by Condition



Note. Each shade of grey shows the first, second and third trial of each condition. The four different plots show the data of four different dependent variables, top-left: Hide inspections per minute, top-right: Number of hide inspections, bottom-left: The time between the start of the experiment and the subject obtaining the reward, bottom-right: the time between the start of the experiment and the subject placing their head in the "critical" zone (the danger area where the stick would poke). Note that the variation in all three pre-test trials using the dependent variable "hide inspections per minute" (top-left) encapsulates almost all of the variability found in all of the other conditions, suggesting that any of the variability found in the different conditions using that measure is within the boundaries of what could be considered a "baseline" rate of inspection. It is possible that this dependent variable is too affected by noise to detect changes in inspection rates in the different conditions in relation to the factors modified in the experiment. The other three dependent variables all show a much tighter grouping and a lower value (reflecting lower rates of vigilance) in the pre-test trials when compared to the experimental trials. This suggests that the experimental factors did have an effect on their vigilance and it might be worth checking this effect using at least one of these other dependent variables.

Figure 7

The “Latency to Place the Head in the Critical Zone” of Both the Replication and the Reverse Order Groups



Note. The latency to first place the head in the critical zone of each individual is shown as points and additionally box-plots show the median, interquartile range, and limits. The mean of the subjects three pre-test trials from each group is plotted on the left of each figure. Significant differences are not shown on this figure.

Replication and Reverse Order Groups together. Overall, when the data between the replication and reverse order groups were pooled, there was no significant difference between the HCA condition (*M* = 11.16 s, *SD* = 6.04) and the UCA condition (*M* = 13.45 s, *SD* = 5.21) (Wilcoxon signed rank test, *V* = 13, *p* = .55). There was also no significant difference when only the first trials of the HCA (*M* = 16.18 s, *SD* = 11.60) and the UCA (*M* = 19.63 s, *SD* = 14.03) conditions were compared (Wilcoxon signed rank test, *V* = 16, *p* = .84). There was also no significant difference between the overall pre-test latency to place the head in the critical zone (*M* = 8.65 s, *SD* = 2.83) and the HCA (Wilcoxon signed rank test, *V* = 10, *p* = .31) or UCA condition (Wilcoxon signed rank test, *V* = 6, *p* = .11).

If the results of the two groups were pooled by order of condition faced, rather than by experimental condition, then there was no significant difference between the first condition faced (*M* = 13.90 s, *SD* = 4.60) compared to the second condition faced (*M* = 10.70 s, *SD* = 6.30; Wilcoxon signed rank test, *V* = 26, *p* = .31). Additionally, there was no significant difference between the first trial of the first condition (*M* = 20.95 s, *SD* = 12.36) and the first trial of the second condition (*M* = 14.85 s, *SD* = 12.81; Wilcoxon signed rank test, *V* = 24, *p* = .46). However, there was a significant difference between the pre-test latency to place the head in the critical zone (*M* = 8.65 s, *SD* = 2.83) compared to the first condition (Wilcoxon signed rank test, *V* = 3, *p* = .04) but not the second condition (Wilcoxon signed rank test, *V* = 15, *p* = .74). This same effect was stronger when the pre-test latency to place the head in the critical zone was compared only to the first trial of the first condition (Wilcoxon signed rank test, *V* = 3, *p* = .008), but not the first trial of the second condition (Wilcoxon signed rank test, *V* = 3, *p* = .46).

Models on Head in Critical Zone Data. The models on the latency to place head in critical data described in “data analysis” both showed significant effects (Supplementary Tables 1 and 2). Simulated estimates of the effects of these factors showed that none of the experimental conditions were likely to have had statistically meaningful effects on the hide inspection rate (simulated values are plotted on the lower panels of Figures 3 and 4). The model based on the first condition that each subject faced did provide some evidence that most of the experimental conditions did appear to have elevated the vigilance levels on all subjects in the first trial, no matter which experimental condition was faced (Figure 4, lower panel). However, the credible intervals of the simulated values still overlapped, so these differences may not have been meaningful. Unlike the models based on the hide inspection rate, the models on the latency to place the head in the critical zone data did not show that birds that were captured as juveniles had a lower vigilance than those that were captured as adults (Supplementary Tables 1 and 2).

Discussion

We did not replicate an effect that suggested that New Caledonian crows (NCCs) reason about hidden causal agents that was previously found in another group of NCCs (Taylor et al., 2012). In that previous study, subjects did not significantly increase their vigilance behavior in the human causal agent (HCA) condition compared to a baseline (= pre-tests), but following that, they did significantly increase their vigilance behavior in the unseen causal agent (UCA) condition. The logic was that the NCCs had reasoned that a hidden causal agent must have been present in the UCA condition, one that may have continued to be present throughout the test, thus explaining the necessity for increased vigilance. In the current study, when we combined the HCA and UCA results of the replication and reverse order groups, there were no significant differences in hide inspection rates between either of the pre-test, the HCA condition and the UCA condition (Supplementary Figure 1). Their hide inspection rate, as measured by the number of times the subjects looked per minute at the place where the hidden causal agent would be, increased for both conditions, but did not differ significantly between the two conditions in a way that suggested the subjects were reasoning that a human was still in the hide in only the UCA condition. Therefore, we did not find any supporting evidence to suggest that NCCs reason about hidden causal agents, as the tests on this combined data would have been the most liberal test to replicate the effect. Although a model of the data from these groups did suggest that there were significant differences between the hide inspection rates of the pre-test and both the HCA and UCA conditions (Supplementary Table 1), simulated values based on this model suggested these differences were not statistically meaningful (Figure 3, upper panel).

In the group of subjects that replicated the previous method directly (Replication group), subjects mildly increased their vigilance (as measured by hide inspection rate) in both conditions, but with no clear differences between the conditions (Figure 2). There were also no statistically meaningful differences between both of these conditions with the simulated values based on models of this data (Supplementary Table 1; Figure 3 upper panel). Further to this, in the group in which we reversed the order of conditions (Reverse order group), the subjects did not appear to drastically change their hide inspection rate in either the unseen causal agent condition or the human causal agent condition when compared to their own pre-test hide inspection rates (Figure 2, right panel). The mean hide inspection rate for this group appeared to steadily decline across all experimental trials. However, there was a large difference in baseline inspection rate between the replication and the reverse order groups. It is not clear why this was the case, but it potentially calls the accuracy of the measure "hide inspection rate" into question (which is discussed below).

At this point, it must be noted that without replicating the initial finding of Taylor et al. (2012), the zero-agent controls and the human presence controls become difficult if not impossible to interpret. The controls were designed with the underlying assumption that the difference in hide inspection rate between the human causal agent and unseen causal agent condition in the Taylor et al. (2012) study was due to a genuine effect. The uncertainty of why we could not replicate that effect means we do not know the context in which to now interpret the zero-agent and human presence controls. Although our model of all the different conditions the subjects faced suggested that there were no statistically meaningful differences between any of the experimental conditions (Supplementary Table 2; Figure 4, upper panel), this cannot be meaningfully interpreted.

Concerning the question as to why the Taylor et al. (2012) finding could not be replicated, it could be that slight variations in methodology between the two experiments had a genuine impact on the behavior of the subjects. Alternatively, it could also be that there are real differences between the two groups of birds that have been tested (Farrar et al., 2020; Forstmeier et al., 2017). For example, the birds tested in this current experiment have been in captivity for longer than the birds in Taylor et al. (2012), which may have substantially changed the subjects' vigilance behavior towards humans. It also could be that, as both experiments had small sample sizes, both experiments measured such a small sample of this species' behavior in this particular setup that the effects found are not actually representative of the population but just noise in the data (Farrar et al., 2020; Forstmeier et al., 2017). This latter point is exacerbated by the fact that the dependent variable used in this experiment ("hide inspection rate") is possibly a noisy measure; ergo, small sample sizes are more likely to capture a false-positive signal amongst unrepresentative noisy data (Forstmeier et al., 2017). All these points are discussed in more detail below.

One aspect that definitely differed between the two experiments was how we stopped the subjects approaching the feeding box while the experimental manipulations took place. In the Taylor et al. (2012) study, they used a second person with their hands crossed over their body to stand next to the feeding box so that the subjects would not approach it while the stick was poking. This person left after the stick poking had occurred. In our study, we used a cloth to cover the feeding box, which was then pulled off after the stick poking action was completed by a long piece of string that was attached to it. This change was implemented to exclude the possibility of unconscious cueing of the birds from the second experimenter, yet retrospectively, it might have caused another confounding factor. Arguably, this cloth pulling action was also the action of a "hidden causal agent", so it could be that the subjects in our experiment were already expecting "animacy violations" or the action of an "unseen causal agent" throughout the experiment, no matter what condition they were facing at that time. Because of this, they could have had reason to remain highly vigilant throughout the experiment, no matter what the differences in experimental manipulation were, as they may have reasoned that the stick could poke them at any point of the experiment even if the human causal agent had visibly left.

A second difference was from the way we made the threatening stick move. In the Taylor et al. (2012) test, they moved the stick via a string/pulley mechanism operated by an experimenter. We used a motorized stick connected to an actuator so that we could be certain that the stick moved back and forth in a standardized way and there could be no unconscious bias in the way the stick was moved by the experimenter. Newborn chicks (*Gallus gallus*) are able to spontaneously recognize (and prefer) biological motion compared to more "rigid" motion (Vallortigara et al., 2005), so it's very plausible that the NCCs recognized the mechanically moving stick as non-biological motion. This may have made the NCCs in our experiment not relate the stick movements to the human agents at all, regardless of whether they were hidden or not. Furthermore, the mechanically moving sticks may have been a source of "unexpected animacy", which has been shown experimentally to increase wariness in other *Corvidae* (Greggor et al., 2018). Considering these factors, it is therefore difficult to know what we should have expected the NCCs perceptions of the causal structure of the task were. Relating the non-biological movement of the stick to the biological human-agent would not necessarily be the most logical causal understanding of the situation.

Regarding the real differences between the subjects in ours and the Taylor et al. (2012) experiment, there are a number of factors that may have had an effect. The extended captivity of the subjects in the current experiment may have made them either habituated or sensitized to humans. This kind of erosion of anti-predator behavior in captive animals is an effect which has to be countered when captive animals are returned to the wild (Griffin et al., 2000). On a daily basis the NCCs in our study had interactions with humans such as delivering food and water, turning lights on and off and opening and closing doors from out of view. They also had occasional aversive interactions, for example, when they needed to be caught for veterinary check-ups. These extended and varying interactions with humans may have had long-term changes on their behavioral responses towards humans when compared to the more recently wild subjects in the Taylor et al. (2012) experiment. Furthermore, half of the subjects in the current experiment were caught in the wild when they were juveniles. These subjects had noticeably more positive reactions towards human experimenters compared to the subjects caught as adults. The juvenile-caught birds would often approach humans whereas the adults would keep as much distance as possible. Both our models of the hide inspection rate data suggested that these juvenile-caught birds had significantly lower hide inspection rates across all trials (Supplementary Tables 1 and 2). It would be reasonable to interpret that these birds may have felt less need to be vigilant towards a hide that may have contained a human as they did not have a particularly negative valence towards a human presence. This is suggestive evidence that there were real differences between the subjects in ours and the Taylor et al (2012) experiments.

Nevertheless, there is also evidence that the differences between the subjects in the group could just be from differences in measurement. Of particular note is the hide inspection rate the subjects showed in the pre-test trials. This could be taken as the individuals’ baseline vigilance rate. The baseline vigilance rate ranged from zero to twenty hide inspections per minute (Figure 6, top left panel, pre-test trials 1-3). All of the variation in the following test trials, regardless of condition, was encapsulated within this range of hide inspection rate. That is, any of the changes in inspection rate in the test trials is well within the baseline inspection rate of the whole group and should not therefore be considered abnormal, increased, or decreased. Our interpretation of this is that hide inspection rate is a ‘noisy’ measure. This is because other measures that also quantify the subjects’ vigilance behavior, which were measured in the exact same trials, have a much lower range of variation in the pre-test trials (Figure 6, top-right panel and both bottom panels, see pre-test trials).

Although in Taylor et al.’s (2012) experiment, the variation in hide inspection rate was shown to be much lower, it still does not necessarily mean it is a reliable measure of vigilance. The crux of the experiment relies on an increase of the birds’ looking rate at a hide from 5 times per minute to 7 times per minute. Although this difference may be statistically significant, it is not necessarily clear whether this is actually a meaningful difference of vigilance behavior. One perspective is that if the “baseline” vigilance rate is already 5 inspections per minute, then this is already more than twice as much as the rate increase between the pre-test and the Unseen causal agent condition. This might suggest that the rate increase is actually small and possibly well within a reasonable standard “baseline” inspection rate because the amount of vigilance the subjects have towards a mostly habituated object (the hide) is double the amount that they increased their vigilance towards the theoretical threat of the hidden causal agent.

This hide inspection rate issue is further compounded when one considers how this measure is actually calculated. It is made of two different measures. Firstly, one counts the number of times the subjects look at the hide between the point they picked up a stick and the point they obtained the food. Secondly, one measures the time between the point the subjects pick up the stick (or land in the experimental area in the current studies’ case) and the point they obtain the food. Arguably, both of these measures should increase if a subject is being vigilant to a purported threat; they should observe the thing they are being vigilant of more *and* they should delay approaching towards the area/thing they are being vigilant of. However, the rate is calculated by dividing the number of observations towards the hide by the time it takes the subjects to obtain the reward. Thus, an increase in hide observations increases the hide inspection rate, but an increase in time to obtain the reward actually *decreases* the hide inspection rate even though this latter measure might also be increasing if the birds are being more cautious. So hypothetically, a subject that looks at the hide 6 times in 36 s has a lower hide inspection rate (10) than a subject that looks at the hide 5 times in 25 s (12) even though both of the individual measures in the former suggest that the first subject is being more vigilant overall.

As the raw data of the two measures required to calculate the hide inspection rate are not provided in Taylor et al. (2012), it was not possible to see how each of these two measures were separately changing across the trials, i.e., were changes in their hide inspection rate due to an increased number of hide inspections, a decreased latency to obtain the reward or perhaps both. Nevertheless, the data in this current replication would suggest that calculating the hide inspection rate in this way creates a much noisier dataset. We provide the hide inspection rate, as well as the number of hide observations and the time it took the subjects to obtain the reward (Figures 5 and 6). The reduced variation in the pre-test data of the latter two measures suggests that calculating the hide inspection rate actually introduces a level of noise into the data, which possibly reduces its accuracy. It also means it is more likely to have been a false-positive result (Farrar & Ostojic, 2019; Forstmeier et al., 2017).

Alternative Measures

As we discussed, we believed the measure of hide inspection rate to have some issues. Although we did have "excellent" inter-observer agreement between our coders for counting the "hide inspections," we felt the way the hide inspections were coded ignored the possible peripheral vigilance that the NCCs possess. Although they may be more extreme in their binocular vision than other corvids, they still have large peripheries (Troscianko et al., 2012). Thus, the hide inspection measure may have missed much of the continued vigilance the subjects had towards the hide that was not solely based on binocular vision inspections. We therefore decided to also measure a behavior that did not involve assumed visual inspections.

For this measure we counted the time it took for the subjects to first place their heads in the critical "danger-zone" where the stick would be poking. We justified this as being the first point at which the individuals believed it was safe to take "a risk." We also had excellent inter-observer reliability when measuring this score. In the replication and reverse order groups, the subjects appeared to show an increase in latency to place their head in the critical zone in the first trials of either the HCA or UCA conditions (Figure 7), showing a significant increase in latency to place their head in the critical zone in the first experimental trial compared to the pre-test trials (regardless of condition). What was interesting about this measure is that it suggested that almost every single manipulation that we did led to an increased latency for the birds to place their heads in the critical zone (i.e., raised level of vigilance; Figures 5 and 6, bottom right panel). Regardless of whether this manipulation was a human cued stick, a ghost cued stick, human cued light, a ghost cued light, or a sound cued stick, the birds mostly showed an increased wariness compared to their previous baseline level of wariness in the pre-test. This suggests that the subjects in our test were possibly not discriminating between any of the different signals and were just reacting more warily to change in any form. Although the models on the "head in critical zone" data suggested that there were significant differences due to the different experimental conditions (Supplementary Tables 1 and 2), this mostly appeared to be due to all of the different conditions causing a raise in latency to place the head in critical zone compared to the baseline "pre-test" level. When fitted data was simulated from these models there was still large overlap of credible intervals between the different conditions (Figures 3 and 4, lower panels), as well as the pre-test data, suggesting no statistically meaningful difference between the subjects' behaviors.

Difficulties with Replications and an Inconclusive Conclusion

We faced a dilemma when completing this replication. We could have used all 14 of our subjects to do a direct replication of the Taylor et al. (2012) experiment to see only if the same effect could be replicated. There were only eight NCC subjects in that test, so together it would have been possible to have 22 subjects doing the same experiment, which in relative terms for the field of animal cognition is good. However, that original experiment did have issues raised in some commentaries that we felt had to be addressed (Boogert et al., 2013; Dymond et al., 2013; Taylor et al., 2013a, 2013b). The only way we could do this was to divide our limited subjects into small groups to check all the points raised in these commentaries. This left us with very low power to address any of those points. Furthermore, when we did not replicate the original effect found in Taylor et al. (2012), we could not be certain that this was because we had only four subjects in this direct replication group or whether there was an actual difference. As discussed above, this left much of the rest of our results uninterpretable.

At the same time, we lacked the sample size to run yet more experimental groups, which would have been required to accurately counterbalance for possible order effects of the different conditions on the subjects' behavior. That is, when comparing the replication and the reverse order groups, the second condition that each group of subjects faced was affected by the expectations the subjects had from the first condition they faced. In order to examine this possibility, we would have had to test two more groups, one that faced the “human causal agent” condition twice in a row and one that faced the “unseen causal agent” condition twice in a row. Only then would it be possible to know if the subjects’ increased vigilance behavior in the “unseen causal agent” condition, the second condition in the Taylor et al. (2012) experiment, was due to either the subjects “sensitizing” to the stick poking action the second time seeing it (as suggested by Dymond et al., 2013), or because of the differences between the “human causal agent” and the “unseen causal agent” conditions.

A further issue we regretted retrospectively is that we did not coordinate with the original research group at the time we conducted the study to ensure that our replication was true to their methods, and thus truly a replication. It is possible that by not reaching out to collaborate we introduced small elements of noise to the data that further caused differences between the two groups results. For example, it was not clear from the Taylor et al. (2012) manuscript if the three trials of each of the conditions were completed on the same day, one after the other, or on different days. We opted to do the two conditions on two different days, which may have been a critical difference in methods. If we had coordinated with the other group, we would have known for sure if this was the case. Furthermore, retrospectively, it turns out that it would have probably made more sense to increase the sample size of the original study and run just the replication and counterbalance with Taylor et al.’s (2012) identical method, and therefore collaborate. This would have resulted in more subjects overall together, which would further have solved many of the power issues with this current replication. Collaboration is likely to be one of the most important factors in animal cognition studies to ensure that experimental replications are both useful and reliable. It is one of our strongest recommendations to anyone considering replicating another’s work.

Overall, we did not find supporting evidence that NCCs reason about hidden causal agents. However, given our low sample size in the replication group, this does *not* mean that this result suggests that NCCs *do not* reason about hidden causal agents and it does *not* overrule the original result found by Taylor et al. (2012). Nevertheless, we feel that we did raise some general methodological concerns that call for caution when interpreting their result. Whether NCCs should be considered to reason about hidden causal agents remains an open question.

Author Contributions

The design of the experiment was initially conceived by AvB & GL. The experiments were carried out by GL, LO’N and MvB. Videos were coded by LO’N and secondary coded by Claudia Zeiträg. Data were visualized and analyzed by LO’N. The initial manuscript was written by LO’N & GL with feedback from AvB.

Acknowledgements

We thank the Province Sud de Nouvelle Calédonie and Dr. Christian Rutz for their support in 2009/2010 sourcing our subjects, and the Parque Zoologique et Forestier in Nouméa for hosting them. We are grateful to the Tierpark Hellabrunn for logistic help in 2010. We also thank Gerhard Fries and Hermann Kempf for his veterinary support and Irmgard Neumair for help in animal caretaking. We are very grateful to Claudia Zeiträg for help in the secondary coding of this experiment.

LO’N is a member of the International Max Planck Research School (IMPRS) for Organismal Biology. We are very grateful to Fränzi Korner-Nievergelt for her help producing the models in this paper as part of a statistics workshop organized by the IMPRS.

Data Accessibility

All data and the code for the models can be found at: https://osf.io/cxswd/

References

Abdai, J., Ferdinandy, B., Terencio, C. B., Pogány, Á., & Miklósi, Á. (2017). Perception of animacy in dogs and humans. *Biology Letters*, *13*(6), 20170156. https://doi.org/10.1098/rsbl.2017.0156

Ball, W. A. (1973). *The perception of causality in the infant*. Society for Research in Child Development.

Boogert, N. J., Arbilly, M., Muth, F., & Seed, A. M. (2013). Do crows reason about causes or agents? The devil is in the controls. *Proceedings of the National Academy of Sciences*, *110*(4), E273–E273. https://doi.org/10.1073/pnas.1219664110

de Winter, J. C. F. de. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, *18*(10), 1–12. https://doi.org/10.7275/e4r6-dj05

Dymond, S., Haselgrove, M., & McGregor, A. (2013). Clever crows or unbalanced birds? *Proceedings of the National Academy of Sciences*, *110*(5), E336–E336. https://doi.org/10.1073/pnas.1218931110

Farrar, B. G., Boeckle, M., & Clayton, N. S. (2020). Replications in comparative cognition: What should we expect and how can we improve? *Animal Behavior and Cognition*, *7*(1), 1–22. https://doi.org/10.26451/abc.07.01.02.2020

Farrar, B. G., & Ostojic, L. (2019). *The illusion of science in comparative cognition* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/hduyx

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Greggor, A. L., McIvor, G. E., Clayton, N. S., & Thornton, A. (2018). Wild jackdaws are wary of objects that violate expectations of animacy. *Royal Society Open Science*, *5*(10), 181070. https://doi.org/10.1098/rsos.181070

Griffin, A. S., Blumstein, D. T., & Evans, C. S. (2000). Training captive-bred or translocated animals to avoid predators. *Conservation Biology*, *14*(5), 1317–1326. https://doi.org/10.1046/j.1523-1739.2000.99326.x

Groves, P. M., & Thompson, R. F. (1970). Habituation: A dual-process theory. *Psychological Review*, *77*(5), 419–450. https://doi.org/10.1037/h0029810

Hunt, G. R. (1996). Manufacture and use of hook-tools by New Caledonian crows. *Nature*, *379*(6562), 249–251. https://doi.org/10.1038/379249a0

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288. https://doi.org/10.1016/S0010-0277(87)80006-9

Luo, Y., Kaufman, L., & Baillargeon, R. (2009). Young infants' reasoning about physical events involving inert and self-propelled objects. *Cognitive Psychology*, *58*(4), 441–486. https://doi.org/10.1016/j.cogpsych.2008.11.001

Mascalzoni, E., Regolin, L., & Vallortigara, G. (2010). Innate sensitivity for self-propelled causal agency in newly hatched chicks. *Proceedings of the National Academy of Sciences*, *107*(9), 4483–4485. https://doi.org/10.1073/pnas.0908792107

Michotte, A. (1963). *Perception of causality*. Methuen.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001. https://doi.org/10.1111/j.1467-9280.2005.01649.x

Saxe, Rebecca, Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology*, *43*(1), 149–158. https://doi.org/10.1037/0012-1649.43.1.149

Spelke, E. S., & van der Walle, G. A. (1993). Perceiving and reasoning about objects: Insights from infants. In N. Eilan, R. McCarthy, B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology* (pp. 132-161). Basil Blackwell.

Spelke, Elizabeth S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion: Continuity and inertia. *Cognition*, *51*(2), 131–176. https://doi.org/10.1016/0010-0277(94)90013-2

Taylor, A. H., Miller, R., & Gray, R. D. (2012). New Caledonian crows reason about hidden causal agents. *Proceedings of the National Academy of Sciences*, *109*(40), 16389–16391. https://doi.org/10.1073/pnas.1208724109

Taylor, A. H., Miller, R., & Gray, R. D. (2013a). Reply to Dymond et al.: Clear evidence of habituation counters counterbalancing. *Proceedings of the National Academy of Sciences*, *110*(5), E337–E337. https://doi.org/10.1073/pnas.1219586110

Taylor, Alex H., Miller, R., & Gray, R. D. (2013b). Reply to Boogert et al.: The devil is unlikely to be in association or distraction. *Proceedings of the National Academy of Sciences*, *110*(4), E274–E274. https://doi.org/10.1073/pnas.1220564110

Troscianko, J., von Bayern, A. M. P., Chappell, J., Rutz, C., & Martin, G. R. (2012). Extreme binocular vision and a straight bill facilitate tool use in New Caledonian crows. *Nature Communications*, *3*, 1110. https://doi.org/10.1038/ncomms2111

Tsutsumi, S., Ushitani, T., Tomonaga, M., & Fujita, K. (2012). Infant monkeys’ concept of animacy: The role of eyes and fluffiness. *Primates*, *53*(2), 113–119. https://doi.org/10.1007/s10329-011-0289-8

Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns. *PLoS Biology*, *3*(7), e208. https://doi.org/10.1371/journal.pbio.0030208

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Second edition). Springer.

Wisenden, B. D., & Harter, K. R. (2001). Motion, not shape, facilitates association of predation risk with novel objects by fathead minnows (*Pimephales promelas*). *Ethology*, *107*(4), 357–364. https://doi.org/10.1046/j.1439-0310.2001.00667.x

Wu, Y., Muentener, P., & Schulz, L. E. (2016). The invisible hand: Toddlers connect probabilistic events with agentive causes. *Cognitive Science*, *40*(8), 1854–1876. https://doi.org/10.1111/cogs.12309