The False Dichotomy Between Experiment and Observation: The Case of Comparative Cognition

1. **Introduction**

Comparative cognition is the interdisciplinary study of animal cognition, which brings together fields such as comparative psychology, ethology, evolutionary biology, and subfields in the neurosciences. The successful integration of these subfields has been notoriously difficult to achieve. Some of the setbacks have been attributed to the perception within some of its subfields, such as comparative psychology, that the methods of the other subfields, such as ethology and evolutionary biology, are inadequately experimental (Allen & Bekoff 1997). Prior to the synthesis, comparative psychologists conducted most of their work in a laboratory environment, while ethologists conducted the majority of their work in the field. The difference in work environments made both sides uneasy: Psychologists accused ethologists of being unable to impose the controls necessary for experimental investigation. Ethologists, in turn, accused the comparative psychologists of building elaborate experimental setups that made no contact with the natural environments of the test subjects, rendering their experimental results meaningless. Underlying this skepticism was the idea that a distinction between observation and experimentation is conceptually unproblematic. This conceptual distinction, however, had and continues to have a normative component: experimental studies are granted a greater degree of epistemic weight than observational studies. Some ethologists aimed to disprove this view, while others argued that their work was also experimental. Although comparative cognition has overcome some of the earlier divides, the reliance on the experiment/observation distinction as a measure of epistemic reliability continues to obscure similarities among its subfields, forestalling a synthesis of the field and wrongly privileging laboratory-based studies over field-based studies.

In this paper I use comparative cognition as a case study, drawing on a range of examples such as Robert Seyfarth and Dorothy Cheney’s baboon studies and Thomas Bugnyar’s studies of consolation behavior in ravens, to challenge the traditional Baconian view that experimentation is the most reliable path to secure scientific knowledge. I argue that other kinds of scientific investigation, such as observational studies, can stand on equal epistemic footing with experimentation. Toward this end, I begin with a thin conception of experiment and observation in order to examine the alleged desiderata of scientific investigation that are imputed to the former and withheld from the latter. I proceed to offer cases where an alleged desideratum is both desirable and present in observational studies. I then turn to a closer examination of the thin conception of experiment to argue that even this conception is too restrictive and rests on an overvaluation of intervention.

I conclude that, for the philosophy of science, the distinction is at best an unhelpful conceptual tool for parsing scientific activity for epistemic merit. At worst it prevents philosophy of experiment from taking as its subject a much broader range of scientific investigations. I suggest that philosophers and scientists should redraw the boundaries between types of empirical activity along epistemically informative lines, which requires a finer-grained analysis of methodological strategies drawn from a broader range of scientific investigations.

**2. The Received View: The Experimentation/Observation Distinction**

So, what is the received view of the experimentation/observation distinction? On the received view of the experimentation/ observation distinction, two things are true:

(1) it is possible to draw clear distinctions between experiment and observation, and

(2) we are right to privilege experiment over observation.

This view is so common, that official defenses of (2) are difficult to find. Meanwhile, point (1) is taken as a given in most discussions of scientific experiment.[[1]](#footnote-1) The received view of the distinction, we might say, is in the background of philosophical discussions and scientific decisions. This is in contrast with the early modern view, where, according to historians Daston and Lunbeck (2011), the sharp distinction between what we now view as experiments and all other scientific enterprises was invisible, and where it was *observation* that was considered to be the engine of scientific discovery – albeit one that required systematization and refinement.

 Although philosophy of science has had relatively little to say about the relationship between experimentation and observation this is not for lack of awareness of the subjects’ value (Cf. Okasha 2011). Hans Radder, for instance, in his (2003) anthology, *The Philosophy of Scientific Experimentation,* writes in the section on the future of the field that a full account of the relationship between observation and experimentation must feature in what he calls a “mature” philosophy of experiment (Radder 2003, 15). Such an account may, he argues, begin with a development of “a philosophical account of how observations are realized in scientific practice and to what degree they differ from experiments” (ibid). I begin a bit earlier: by examining how the definition of observational studies is constructed in the first place, in contrast with experimentation.

**2.1 Experiment and Observation: Thin conception of the distinction**

So, what is the thin conception of the experiment/observation distinction?To answer this, we must first provide rough definitions of experimentation and observation.

***Experiment***

First, let’s examine what is meant by “experimentation” on the received view. There is a variety of views on this point, but all seem to agree that experimentation is distinguished from other scientific investigation by its *intervention* in the target system and *manipulation* of the independent variables(Bogen 2010; Daston 2011; Okasha 2011). As Jim Bogen writes in the 2010 Stanford Encyclopedia of Philosophy entry on *Theory and Observation in Science*, “Contrivance and manipulation influence epistemically significant features of observable experimental results to such an extent that epistemologists ignore them at their peril” (Bogen 2010). Jim Woodward (2003) singles out intervention as crucial for establishing causal claims, equating it with experimental studies, and contrasting this practice with what he calls “*passive* observation” (Woodward 2003, 87; emphasis added). Similarly, Wendy Parker (2009) writes that “an *experiment* can be characterized as an investigative activity that involves intervening on a system in order to see how properties of interest of the system change, if at all, in light of that intervention” (Parker 2009, 487).

 Philosophers differ on the nature of intervention, however. For example, Okasha (2011) suggests that experiments take the following form: to test a generalization such as ∀x(Fx🡪 Gx), an experimental setup would first *create* an *a* with property F, and then test *a* for property G. According to Okasha, the epistemic superiority of experiment over observation rests on the experiment *bringing about* that some *a* is an F, rather than “*happening upon”* an *a* that is both F and G (Okasha 2011, 5).[[2]](#footnote-2)

Okasha’s interpretation is too narrow for my purposes, as it leaves out most studies in comparative cognition, which the scientists themselves call “experiment.” Although I agree with Radder that philosophers should not “simply take for granted the intuitions and conceptualizations of the scientists,” I want to leave the possibility open that these studies *can* count as experiments (Radder 2003, 152). Whether Okasha is correct in his assessment that studies that *create* rather than *happen upon* an instantiation of the antecedent offer the best evidence, this is not enough to establish that all and only such studies are genuine experiments. If we want our distinctions to be meaningful, I believe we should begin with accurate descriptions of the whole class of scientific investigations. For this reason I will use Parker’s (2009) much more catholic characterization of intervention, which states that “*intervention* is, roughly, an action intended to put a system into a particular state, and that does put the system into a particular state, though perhaps not the one intended” (Parker 2009, 487; original emphasis).

Whatever the definition of intervention, its utility is presumed to lie in its two functions: (1) isolation of the target from possible confounds and, (2) creation of novel scenarios that would not occur unaided. I will return to an examination of interventionism later in the paper, where I will argue that even this thin conception of experimentation faces serious theoretical challenges.

***Observation***

Zeroing in on a definition of observation within science is more difficult. This is due in part to the fact that philosophers have spent less energy considering observation on its own terms, and in part on the ambiguity of *observation* as a concept. Unlike with scientific experimentation, which is identified by its interventionism, no single feature uniquely picks out observational studies. The best we can do is to define it negatively as any empirical investigation in the sciences that does not involve intervention or manipulation. As we will see, this negative definition is unstable, since interventions are not always absent in putatively observational research.

Importantly, this definition attaches to observational *studies* and not to observation *simpliciter,* or the activity of perceiving in the course of other scientific and nonscientific activity. This is important because much of the philosophical discussion of observation in science has focused on the role of observation in the context of experiments – and on the theory-ladeness of such observations.

**3. Desiderata of Experimentation and Alleged Weaknesses of Observational Studies: Overview**

Let us now move on to an introduction of the desiderata of experimentation and their alleged absence in observational studies. These include repeatability and amenability to statistical analysis. Furthermore, experiments’ interventionism is credited with the isolation of the target from possible confounds, often with the aid of technologies. These strengths are contrasted with the weaknesses imputed to observational research. For example, secure knowledge gained through interventionism or direct manipulation is contrasted with passivity of theory-laden observation , or the reception of undifferentiated environmental signals; repeatability is contrasted with unique or rare occurrences; the quantifiability of statistical tools is contrasted with the qualitative nature of observational reports; and finally, fidelity of the technological instrumentation is contrasted with the suspicion of the human perceiver in observational studies.I turn to these now.

**3.1 The Desideratum of Precise Instrumentation**

Let us begin with the first desideratum of experimentation: the presence of specialized and finely tuned instrumentation used inter alia, togather, measure, record, and analyze data. The presence of instrumentation is desirable because, where measurement is concerned, it ensures accuracy, distilling information from a noisy environment. The *role* of instrumentation in experiment has been the subject of some philosophical discussion, though the *value* of instrumentation has not been challenged.

I do not wish to challenge it here. Instead, I wish to register that observational studies are assumed to lack the benefits of experimental studies because they are presumed to lack the proper means of parsing the noisy environment for relevant information. This however, fails to give due credit to the *trained* observer. While it is true that untrained and unfocused observation is unlikely to produce reliable conclusions about natural phenomena, this is not the sort of observation that is employed by comparative cognition researchers. Although trained observers in comparative cognition sometimes require the aid of technological enhancement, such as audio-video recording, frame-by-frame analysis software, and more, the primary instrument is the trained observer.

Consider, for instance, Robert Seyfarth and Dorothy Cheney – a psychologist and biologist, respectively – who employ both free observation and field-based experiments in their decades-long studies of a baboon troupe in the Okavango Delta of Botswana. Their form of observation requires the construction, use, and revision, of detailed ethograms – or catalogues of species-typical behaviors – as well as video-recording and frame-by-frame visual analysis of the behavior. The use and refinement of an ethogram requires astute observational skills gained through extensive experience in the field and years of academic training. These tools structure observation and allow Seyfarth and Cheney to compile a behavioral profile of their group of baboons. These profiles in turn, have both predictive and explanatory power.

One might object that this experience is precisely what makes their perception-based judgments theory-laden. In response, I would first note that although expert observers are not immune to error and bias, the theoretical knowledge that undergirds their observational abilities is unproblematic just as long as it is publically accessible (cf. Longino 2002). This is because if it is possible to articulate the framework they are employing, then it is possible to amend their modes of seeing by amending the framework. Second, although there is a common assumption that the theory-ladeness of observation is a liability, there is a positive sense in which observations may be theory laden. This is the sense in which all expert observation is theory-laden: this is what gives the expert her status *as* expert. For instance, a neurologist requires theoretical knowledge about the human brain and how it is represented by different neuroimaging technologies. This is what allows her to identify the lesion in a printout from an MRI; theoretical experience is what allows the physicist to interpret the lambda shape in *Figure A* as indicating the presence of oppositely charged particles in a cloud chamber.

**Figure A**



**3.2 Expert Observers as Instruments**

The success of Seyfarth and Cheney’s studies suggests that the expertise of scientific observers has been undervalued. Observation has been caricatured as either undisciplined perception on the one hand, or as hopelessly biased quasi-judgment on the other. I recommend a middle-ground view that recognizes the special role that expert observation plays in the scientific context. Toward this end, I recommend that we view the human observer as herself a kind of information-processing instrument. The trained observer of behavior gains increased awareness of her study subjects both through extensive training and through extensive contact with her subject. Because animal cognition is embedded in a rich ecological context, the study of animals in their natural habitats is crucial for a full understanding of their behavior. This kind of study requires expert observers who take the whole animal in its environment as their subject. Moreover, the study of putatively intentional behaviors may require instruments best suited for reading intention off behaviors, and this instrument may turn out to be the human agent. This final point is admittedly controversial, but its possibility cannot be ignored for comparative cognition.

Expert observers play a central role in observational studies, and may therefore be regarded as the primary tool for the study of animal behavior. Noting that the expert observer is different in kind from a mere *perceiver* serves as a reminder that observational studies meet the presumptive requirement of utilizing instrumentation. Insofar as experiments gain credibility due to the increased accuracy of their instruments, an observational study conducted by experts is not different in kind from an experimental study using any other appropriate tool. What matters is the fidelity of the instruments and their fit for the given task; the constitution of the instruments only matters insofar as it affects fit and fidelity. With respect to fit of the instrument to the task, it may be that human expert observers are best equipped to process and map out such complex systems as animal social networks or intentional systems. After all, for all its shortcomings, the human brain is capable of incredibly sophisticated and rapid pattern-detection abilities.[[3]](#footnote-3) Perhaps technology can be built that will match or even outstrip human abilities in this arena, but the mere fact that a better instrument *may be developed in the future* does not mean that the one we currently have is a poor performer; it only means that it is not the best possible performer. Of course, scientists do not wait until the best possible instrument is available before proceeding with their work – be it experimental or observational – nor should they.

Expert observers have their limitations, of course, and these deserve further philosophical attention. Colin Allen (2004) suggests that developing an epistemology of observers would be an important step in the direction of identifying the conditions under which field research is epistemically on a par with primarily laboratory-based research. This attention should, however, be supplemented with the reminder that technological information processing instruments have practical limitations and are not free from criticisms of theory-ladeness[[4]](#footnote-4).

**3.3. Desiderata B and C: Repeatability and Statistical Analysis versus Uniqueness and Qualitative Description**

Two other desiderata imputed to experimentation are: (1) repeatability and replicability of results, and (2) amenability to statistical analysis. I take these desiderata together, as repeatability and amenability to statistical analysis are desirable for the same reason: they ensure that the phenomenon observed in the study is a product of the hypothesized underlying structure of the system rather than a product of chance or some confound. Repeatability allows researchers to see whether the phenomenon is robust – novel conclusions can be tested by any competent researcher who wishes to replicate the original methodology. The amenability to statistical analysis is presumed to show that the conclusions of any single experimental result are not due to chance.

Let us assume for now that repeatability and statistical amenability are, in fact, desirable features of a study. Do observational studies possess these features? The answer is not prima facie obvious since one tends to think of observational studies of dynamic systems such as baboon societies, as too complex to yield repeatable results. With respect to statistical evaluation, the expectation might be that the data gathered through observational studies will be qualitative, and not, therefore amenable to statistical evaluation. Both of these assumptions are true only of a subset of observational studies, however. As just one example of a study that is both repeatable and statistically analyzable while being observational (i.e., non-interventionist), consider Orlaith N. Fraser and Thomas Bugnyar’s (2010) study of consolation behavior among ravens.

Fraser and Bugnyar (2010) tested a population of ravens for consolation behavior after aggressive bouts. They hypothesized that ravens, who live in rich social environments, should exhibit behaviors similar to those previously found in primates who also live in socially complex environments. These behaviors include, inter alia, consoling the victims of aggressive bouts and seeking out consolation from bystanders on the part of victims. The differences between primate societies, which are more stable than raven societies, suggested to the researchers that post-conflict affiliative behaviors should vary from those observed in primates. For example, they hypothesized that reconciliation between the victim and aggressor should be less likely, since for ravens “dispersal [is] a more feasible and less costly option” (Fraser and Bugnyar 2010, 2). Because ravens are social animals who also form a variety of bonds with conspecifics, Fraser and Bugnyar predicted that their population of ravens should exhibit bystander consolation behaviors following aggressive bouts and that this should be more common among ravens who shared valuable relationships.

To test this prediction, Bugnyar watched the ravens for a period of nearly two years – from August 2004 until June 2006 – recording aggressive bouts and their immediate aftermath, the identities of the victims and aggressors, and the intensity of the conflict (ibid, 3). He used in-person and video-assisted observation, classifying observed behaviors according to pre-determined definitions, such as:

bystander affiliation =def “post-conflict affiliative interaction initiated by a bystander and directed towards the conflict victim. No functional or mechanistic implications. Also known as (unsolicited) triadic or third-party affiliation.” (Fraser and Bugnyar 2010, 2)

They matched the post-conflict periods, which they called PCs, with control “MCs” or “match-control” observations of the same individuals at the same time on a different day, collecting 152 PC-MC pairs “on 11 conflict victims [with] 58 aggressor-victim dyads” (ibid). They analyzed this data using “GLMMs [generalized linear mixed models] with binomial error structures and a logic-link function … [using] Akaikes information criteria…values…to select the best (most parsimonious) model for all mixed-model analyses” (ibid). Their results corroborated their hypotheses.

Note that the study was observational insofar as it was non-interventionist: Bugnyar (who conducted all of the observations) did not prompt the aggressive bouts, nor did he set up the conditions for such bouts beyond raising the ravens and providing them with a common living area. The independent variable – in this case, the aggressive bout – was not brought about, but, in Okasha’s terminology, “happened upon.” Importantly, these happenings-upon were expected and their instances meticulously recorded and scrutinized. Despite being purely observational, the study yielded quantitative data, which produced statistically analyzable results with methodology that could easily be replicated.

Let us now return to the assumption that statistical analyzability is always desirable. Some doubt has been cast on the necessity of statistical analyses, as these don’t always license straight-forward inferences about evidence from the raw data (Dienes 2008). Zoltan Dienes, a psychologist and statistician, argues that the standard (frequentist) statistical tools in psychology do not show that the result is likely to be true and, therefore, does not provide any evidence for the truth of the result. What it does provide, is the assurance that false positives will occur very infrequently (e.g., 5% of the time) when the methodology is repeated an *infinite* number of times.

Looking at the value of statistical analysis from a different angle, we see that sometimes the requirement that data be analyzable by industry-standard statistical tools shapes the methodology and may occasionally *introduce* bias into the data. This is best illustrated with an example. Consider Irene Pepperberg’s famous studies with the African gray parrot, Alex. Pepperberg conducted many tests on Alex to prove his competency in abilities such as numerosity (comprehension of relations among members of a number line), and the comprehension of abstract concepts, such as color, shape, and texture. Each of these tests required Alex to answer a series of questions, such as “What color [is the] round [object on the tray]?” with the appropriate member of the group from the relevant conceptual category. Because getting statistically analyzable data required performing the tests many times over, Alex frequently lost interest in the tests, and would begin to give incorrect or inaudible answers toward the end of the trials. These wrong answers were not a reflection of Alex’s abilities, however, but of his willingness to perform the test. In this case, the requirement for a statistically analyzable sample necessitated the use of a methodology that was guaranteed to produce false negatives, thereby introducing bias into the very data whose amenability to statistical analysis was alleged to confer on it the status of *objective* result.[[5]](#footnote-5)

This last point should be viewed not as a rejection of statistical analysis, but as a cautionary note about the non-independence of statistical analyses from the data, by way of methodological restrictions. This caution figures into the present discussion because statistical requirements *frequently* impose methodological requirements on comparative cognition studies, which are often times inappropriate (Dienes 2008). It is possible that, under some conditions, qualitative data will be the best kind of data available because collecting quantitative data would only be possible with the use of inappropriate techniques.

**3.4 Summary**

In summary, I have argued that observational studies meet the requirements of being statistically analyzable, repeatable, and – crucially – of employing appropriate instruments for its investigations: the expert observer. I now turn back to the definition of experimentation that I have, up until now, assumed.

**4. Interventionism: Pushing on the thin conception of experimentation**

Having suggested that non-interventionist studies in comparative cognition are nevertheless epistemically reliable, let us now turn to the initial assumption: namely, that experiment requires – is defined by – intervention. This is presumed to be true across a wide range of experiments, from the lesioning of a rat’s hypothalamus to study the connection between leptin and obesity (Suga et al. 1999) to testing for neutral bosons using bursts of polarized electrons (Hacking 1983).

**4.1. “Experiments” without Intervention**

However, examples of putative experiments which do *not* require intervention abound. One example is Eddington’s 1919 expedition to test Einstein’s general relativity theory, which did not involve intervention (Okasha 2011). Despite its noninterventionism, the expedition is commonly treated as an experiment: observation of bent light corroborated the theory of general relativity. What Eddington and his aids did was go to great lengths – literally and financially – to position themselves in a way that made the necessary observations of apparent stellar displacement possible.

Another set of problematic cases includes computer simulation experiments. Simulations do not intervene on the systems that they are meant to model – in fact, it is their very distance from the target system that has presented interesting problems for philosophers of science. Perhaps the best examples come from climate science, since simulation models in that field have yielded predictions that most scientists consider to be credible. One possible response to this example is to say that simulations of climate change are experiments insofar as the variables in the model are open to manipulation and adjustment. This is Wendy Parker’s answer (2009). She points out that simulation studies *do* involve intervention: viz., intervention on the model system, though not the target system it represents. Although she holds that simulation studies *are* genuine experiments, she admits to being less interested in the question of classification than in the question of whether and in virtue of what the simulations *provide epistemically reliable results.* I wish to make the same point with respect to observational studies: classification matters, though only insofar as classification guides future epistemology.

**4.2 Case Study: False belief experiment in baboons**

With this goal in mind, let us consider another example from Cheney and Seyfarth’s baboon studies. You will notice that this study involves intervention and is therefore a putative experiment. Cheney and Seyfarth’s decades of observations yielded detailed analyses of the social hierarchy of the troupe, which they determined to be structured along dynamic, matrilineal lines – with changes in dominance relations occurring periodically. Everyday interactions are structured by the dominance hierarchy. Cheney and Seyfarth hypothesized that in order for this complex hierarchy to obtain, each member of the troupe needed to keep mental track of where she belonged relative to all other individuals and where these others stand relative to one another. In order to test this hypothesis, they devised a type of false-belief test using auditory play-back technology. They hypothesized that a baboon who *knew* where others stood relative to one another in the hierarchy would respond with surprise to an upset in this hierarchy. To see whether baboons reacted with surprise to such a situation, Cheney and Seyfarth first recorded the aggressive and submissive calls of some of the members of the tribe, and then played those back in unexpected order to an unrelated baboon. The listener baboon was being tested for surprise behavior, such as looking in the direction of the calls for longer than usual. Surprise behavior was found and Cheney and Seyfarth concluded that their hypothesis – that baboons kept track of the dominance relationships among other members of the troupe – was confirmed (Cheney and Seyfarth 2007).

This example seems to be a straight-forward example of an experiment. A few things to note here: (1) the experiment was conducted in the course of a paradigmatically observational study, and (2) that the kind of intervention involved is responsible for bringing about a situation which *may have come about without any intervention*.

The first point is a reminder that what I have been calling experimental studies are often embedded in observational frameworks. In ethology, this is more common than not. For example, in addition to the play-back experiments, Cheney and Seyfarth also surreptitiously collected the baboons feces to test them for levels of the so-called “stress hormone,” cortisol. I say “so-called” because cortisol levels measure the amount of excitement or agitation, but are silent on the *valence* of this excitement: Cortisol levels could not tell you whether your dog was excited to go for a walk or terrified of a vacuum cleaner at the time the sample was taken. In the case of baboons, this test did not interfere with the animals, but it did manipulate the collected samples to produce information that human observation alone could never produce. This suggests that the intervention need not directly manipulate or interfere with the target system (the baboon mind) for the experiment to be *about* the target system.

The second point calls into question the epistemic value of interventions that bring about unlikely scenarios. In Cheney and Seyfarth’s play-back experiment, their intervention brought about an unlikely series of scenarios whose normal occurrence is rare and therefore improbable to observe under normal conditions. In this case, the circumstances could *in principle* have obtained *absent* said intervention. We may ask why such interventions should carry any epistemic weight. Suppose that Cheney and Seyfarth had had enough time observing the baboons to witness the statistically mandated number of cases, and had written these observations in a report. Would this report count for less than the play-back experiment in the court of epistemic significance? The answer, I suggest, should be “no” – the difference between the two cases is merely pragmatic: The play-back study was more *expedient* than the wait-and-see study, but it expediency should not carry epistemic weight.[[6]](#footnote-6)

The examples thus far allow us to take our pick of the following (non-exclusive) conclusions: either (1) some experiments are non-interventionist; (2) some observational studies stand on equal epistemic footing as experiments; or (3) the distinction between experiment and observation is unhelpful when deciding what kind of results count as evidence toward a conclusion. Whichever view we accept, it is clear that intervention is not always necessary for a study to produce epistemically reliable results.

**4.3 Interventions introducing confounds while aiming to isolate mechanisms.**

Finally, some interventions in comparative cognition, although designed to eliminate confounds, can actually introduce *more* confounds into a study. For example, the generalizability of results from enculturated laboratory animals to the wild population has been challenged on the grounds that removing the animal from its ecological niche and introducing it into a human-centered environment alters the very capacities being investigated. For example, the cognitive abilities that Pepperberg found in Alex, may be attributable to the linguistic scaffolding provided for Alex, which may have shaped the development of conceptual capacities being tested.

A more problematic case of intervention introducing more confounds than it removes comes from experiments that lack what is termed ecological validity. Ecologically valid studies are those that are designed around the animal’s perceptual and physical modalities and in which the tasks are designed to be salient to the test animal. For instance, Kristin Andrews (2010) argues that testing for awareness of self by using the Mirror Self Recognition (MSR) test will produce more false negatives for gorillas than for other primates. This is because direct eye contact is considered an aggressive gesture in gorillas, making them apprehensive around reflections.[[7]](#footnote-7) What makes interventionism epistemically valuable – its ability to remove potential confounds – in these cases undermines the reliability of test results. This example appears to have a clear solution: the problem is not intervention, but the wrong *kind* of intervention. The trouble is that the wrong kinds of intervention have been employed in comparative cognition precisely because laboratory conditions were presumed to be the natural habitat of experiments, and this was presumed to be the case because laboratories allowed scientists to take the animal *out* of its ecological context. These assumptions belong to comparative psychology and have been challenged by ethologists, but the proper synthesis of these fields into comparative cognition has been slowed by the assumption that experiments required the removal of ecological context. Socially impoverished animals, who may be developmentally stunted as a result of being reared under ecologically invalid conditions, continue to be tested for cognitive capacities, with the researchers drawing conclusions about general populations.[[8]](#footnote-8)

**4.3. Intervention Summary**

In summary, we have seen that interventions in comparative cognition sometimes introduce more confounds than they eliminates and that this is a function of the types of questions raised and, as such, cannot be solved by changing the *type* of intervention. In general, the utility of interventionism depends on the questions being raised. Isolating an animal from its ecological niche will not give informative answers about its natural abilities, but it may reveal something about, e.g., the plasticity of its conceptual faculties.

**5. Conclusion**

In summary, I have argued that the received view, which casts experiment as superior to observation, does not track the epistemic reliability of the methods and instruments of the investigations performed by comparative cognition researchers. This is because, like experimental (interventionist) investigations, observational (non-interventionist) investigations:

1. Are often repeatable,
2. Are frequently statistically analyzable, though the utility of quantitative data over qualitative data remains an open question
3. Include the use of specialized instrumentation: the expert observer. Additionally, the expert observer employs technologies that enhance and refine her observational abilities

Finally, some interventional studies – such as the baboon play-back experiment – only employ intervention for reasons of expediency. These cases (i) argue for the epistemic reliability of observational studies and (ii) challenge the accepted distinction between experiment and observation in comparative cognition. Consonant with Parker’s conclusion that simulation experiments should be assessed for epistemic reliability, I conclude that observational studies in comparative cognition should be assessed on the basis of the appropriateness of their methods and instruments for the questions being asked. Often, *the best tool* for the job will be an expert observer and *the best methodology* one that does not require the manipulation of behaviors. Obviously, this will not be the case all the time: some questions require intervention and the use of technologies. Answering questions about, e.g., the hormonal regulation of bonding behavior is best achieved by measuring oxytocin levels in the blood and comparing this with behavior. The hormone analysis is both invasive and cannot be done without the aid of sophisticated machinery. However, the need for interventions such as these does not undermine the general point; it merely reinforces the simple idea that different kinds of research agendas will require their own suite of technical and conceptual resources.

**5.1. Rebranding Scientific Experimentation**

So, what does this mean for philosophy of scientific experimentation? I suggest that the case of comparative cognition recommends closer scrutiny of the concepts of experiment and observation in general, since different fields may find it more or less useful. The conceptual failure of the distinction in this case suggests that we may be better off redrawing the epistemological boundaries along different methodological lines. For example, instead of dividing scientific investigation into experimental and observational based on the presence or absence of intervention, we might focus on a broader range of scientific investigations – including studies whose primary instrument is the expert human observer or, as Parker suggests, studies whose primary methodology is a computer simulation.

Such a move would present a broader class of phenomena for those interested in such questions as: What kind of manipulation or intervention is desirable and under what conditions? How should repeatability be balanced against complexity and novelty? What kinds of quantitative analysis provides evidence for the truth (or likelihood of truth) of the hypothesis? Are qualitative studies ever preferable to quantitative studies? What counts as an instrument, and how are fit and fidelity of instruments to be determined? What criteria are required to make someone an expert observer? Are there special qualities to human perceptual discrimination abilities that make human expert observers especially well-suited to some tasks in comparative cognition, such as reading intention from nonhuman animal behavior?

Given this new approach, the old observational/experimental distinctions might be replaced through the practical step of rebranding. Both observational and experimental studies might be recast in the more inclusive terminology of “empirical investigations.”[[9]](#footnote-9) The philosophical investigation of scientific experimentation –recast as the philosophy of empirical investigation in the sciences—would be enriched by this move, as it would then include a wider variety of studies, drawn from a wider range of scientific fields.

**References**

Allen, Colin. 2004. “Is anyone a cognitive ethologist?” *Biology and Philosophy* 19:589-

607

Allen, Colin, and Marc Bekoff. 1997. *Species of Mind: The Philosophy and Biology of Cognitive Ethology.* Cambridge, MA: MIT Press.

Bogen, Jim, "Theory and Observation in Science", The Stanford Encyclopedia of

Philosophy (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/science-theory-observation/>.

Cheney, Dorothy & Robert Seyfarth. 2007. *Baboon Metaphysics.* Chicago: University of Chicago Press.

Daston, Lorraine & Elizabeth Lunback (eds). 2011. *Histories of Scientific Observation.* Chicago: University of Chicago Press.

Dienes, Zoltan. 2008. *Understanding Psychology as a Science: An Introduction to Scientific*

*and Statistical Inference*. UK: Palgrave Mcmillan

Fraser O.N., Bugnyar T. 2010. “Do Ravens Show Consolation? Responses to Distressed

Others.” *PLoS ONE* 5(5): e10605. doi:10.1371/journal.pone.0010605

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy*

*of Natural Science*. New York: Cambridge University Press

Heidelberger, Michael. 2003. “Theory-Ladeness and Scientific Instruments in

Experimentation.” In *The Philosophy of Scientific Experimentation*, edited by Hans Radder. Pittsburgh: Pittsburgh University Press, 138-151.

Longino, Helen E. 2002. *The Fate of Knowledge.* New Jersey: Princeton University

Press.

Okasha, Samir. 2011. “Experiment, observation and the confirmation of laws” *Analysis*

71: 222-232.

Parker, Wendy. 2009. “Does matter really matter? Computer simulations, experiments, and

materiality.” *Synthese* 169:483 - 496

Pepperberg, Irene M. 2002. “Cognitive and Communicative Abilities in Grey Parrots” *Current Directions in Psychological Science* 11:83–87

–––––. 2006. “Grey parrot numerical competence: a review.” *Animal Cognition* 9: 377–391

–––––. 2009. *Alex and Me.* Harper Press: New York.

Sinha, Pawan, Benjamin Balas, Yuri Ostrovsky, & Richard Russell. 2006. “Face Recognition

by Humans: Nineteen Results All Computer Vision Researchers Should Know

About.” *Proceedings of the IEEE* 94:1948-62.

Tomasello, M. and J. Call. 2006. “Do chimpanzees know what others see – or only what they are looking at?” In *Rational Animals*, edited by S. Hurley and M. Nudds.

Oxford: Oxford University Press.

Radder, Hans. 2003. “Toward a More Developed Philosophy of Scientific

Experimentation” In *The Philosophy of Scientific Experimentation,* edited by Hans Radder. Pittsburgh: University of Pittsburgh Press: 1 -18

Woodward, Jim. “Experimentation, Causal Inference, and Instrumental Realism” In *The*

*Philosophy of Scientific Experimentation*, edited by Hans Radder. Pittsburgh: University of Pittsburgh Press: 87 -118

1. Scientists as well as philosophers hold something like the received view, although the specifics of their positions will surely be different if one were to press them to elaborate. [↑](#footnote-ref-1)
2. Okasha offers a Bayesian analysis for the superiority of “learning that some object is both F and G, and learning of an object, antecedently known to be F, that it is G” (Okasha 2011, 10).
 [↑](#footnote-ref-2)
3. Just one example of the human information processing abilities that currently outpace computer programs is facial recognition. Human abilities continue to serve as models for artificial facial recognition designs (Sinha et al. 2006). [↑](#footnote-ref-3)
4. For a discussion of the relationship between theory and instrumentation, see Michael Heidelberger. 2003.

“Theory-Ladeness and Scientific Instruments in Experimentation” in *The Philosophy of Scientific Experimentation*. Edited by Hands Radder. Pittsburgh: Pittsburgh University Press, 138-151. [↑](#footnote-ref-4)
5. One might object that this kind of bias can be eliminated through further statistical tools – the most basic of them being the removal of outliers. Another reply might be that while statistical amenability introduces one kind of bias (e.g., slight increase of false negatives in the data) it eliminates a more insidious kind of problem – namely, the possibility that the results are due to *chance*. The result might be that statistically significant results would look slightly *less* significant than if the bias were not present, but that this is a minor concern. [↑](#footnote-ref-5)
6. As a matter of fact, Cheney and Seyfarth did notice upsets in the social structure of their baboon troupe. However, these observations were considered preliminary – belonging, we might say, to the context of discovery rather than justification. One reason to rightly classify such observations as belonging to the context of discovery is if this observation occurred only a handful of times, and if, in each case, it was unexpected. The reasons are straight-forward: unexpected events are more difficult to properly capture in observation – much of what preceded the event, though potentially relevant, may be left out of both written documentation and the memory of even expert observers. However, this does not undermine the general point, since such an event, should it be observed by experts when it is expected, ought to count just as much as the technologically solicited occurrence. Finally, the unexpected event in observational studies has a parallel in experimental studies: novel technologies or technologies used in novel conditions with no prior expectation of their product or behavior belong as much to the context of discovery as do unexpected observations in the field. [↑](#footnote-ref-6)
7. Similarly, a number of prominent researchers, such as Dianna Reiss, have argued that the experimental setup of MSR tests wrongly privileges those species that are predominantly visually oriented over those who use difference sensory modalities to navigate their social world. [↑](#footnote-ref-7)
8. This is the case with, for example, Daniel Povinelli’s studies of chimpanzee metacognition. [↑](#footnote-ref-8)
9. Another option is to subsume under “experiment” those studies that are currently labeled “observational.” [↑](#footnote-ref-9)