Information-seeking in the rat

Andrés Ojeda Laguna

Pembroke College and the Department of Zoology
University of Oxford

Thesis submitted for the degree of Doctor of Philosophy
Michaelmas term 2018

Abstract

From a functional viewpoint, decision-makers should value information only when it can be used it to alter future events. However, experiments show that animals (including humans) value information they cannot use, as if they were simply driven by reducing uncertainty per se. I explore this phenomenon experimentally and interpret it theoretically. In two of the chapters I use protocols in which information is unusable because it is given after choices, while in the other two subjects can choose whether to obtain information that can lead to better decisions.

In Chapter 1 I show that rats do prefer options with lower uncertainty over alternatives with higher reward rate and that the choice mechanism is consistent with treating each alternative in parallel, as proposed by the Sequential Choice Model.

In Chapter 2 I test assumptions from a model that argues that costly over-preference for information is due to an unnatural restriction in the use of information. When rats can use information, they avoid paying time costs when they can predict forthcoming negative outcomes.

In Chapter 3 rats could choose between two options immediately, with equal probability of reward or failure, or pay some cost for acquiring information that could lead to picking a reward-primed option. Rats did perform the information-seeking action, but less often than predicted by rate-maximization.

In Chapter 4 I explore whether the paradoxical preference for informative options when information is unusable applies equally to information about probability, magnitude and delay to reward. Rats reproduced the known bias for unusable information toward reward probability and amount, but not for reward delay. No single existent theoretical interpretation can accommodate these results.

Taken together, these studies demonstrate that neither purely functional nor purely mechanistic interpretations are satisfactory, illustrating the importance of integration in the specific but broadly significant field of information-seeking behaviour.

Table of Contents

General Introduction 5

Chapter 1: Paradoxical Choice in rats 17

- Introduction 18
- Methods 22
- Results 31
- Discussion 36
- Supplementary Methods 44
- Supplementary Results 45

Chapter 2: Testing an Optimal Foraging model of Paradoxical Choice 47

- Introduction 48
- Methods 53
- Results 62
- Discussion 69

Chapter 3: Rats learn to work for useful information 83

- Introduction 84
- Methods 89
- Results 95
- Discussion 98
- Supplementary Results 106

Chapter 4: Preference for unusable information is sensitive to dimension 108

- Introduction 109
- Methods 116
- Results 122
- Discussion 125
- Supplementary Results 132

General Discussion 133

References 146

General Introduction

"*scientia potentia est*" ("*knowledge is power"*)
Sir Francis Bacon

"*There is much pleasure to be gained from useless knowledge*"
Bertrand Russell

Living organisms are immersed in a complex world, and have sensory mechanisms that allow them to perceive, store and utilize aspects of their surroundings (known as affordances (Gibson, 1966) or in other words, to gather information about their circumstances. The task they face however is immense: the richness of the available information (think of the movement of every leaf of a tree in the wind) far exceeds what is useful to process. This necessarily leads to the evolution of mechanisms that filter and select potentially useable information, and, of particular relevance here, of behavioural mechanisms that facilitate the organism's encounter with information that may come useful either immediately or at a later time (Dusenbery, 1996). A consequence is the emergence of what anthropomorphically could be called 'curiosity', namely behaviour specially 'designed' by natural selection because of its adaptive properties in terms of information gathering.

At a hierarchically low level lie many activities such as sniffing or looking around indiscriminately, and then filtering the sensory input according to some functionality criteria. Interesting it is, this is not, however, the level that I wish to explore presently. I am particularly interested on whether, and how, animals behave in a way that strategically shapes the information they get, and to what extent they are willing to pay a cost just 'to know'. The issue has a long history in psychology, behavioural ecology, and artificial intelligence, and it is far broader than can possibly be treated in a single thesis. Topics such as the balance between exploration and exploitation (Berger-Tal et al., 2014), the already mentioned curiosity (Gottlieb and Oudeyer, 2018; Kidd and Hayden, 2015), and

the notion of working for information (Bennett et al., 2016; Blanchard et al., 2015), are all interconnected, and each has received a degree of research effort, some from our own research group (Houston et al., 1982; Kacelnik, 1987; Krebs et al., 1978; Krebs and Kacelnik, 1984)

The experiments in this thesis explore the issue of whether information can act as a reinforcer, that is as a motivational goal at psychological level. Early literature on this issue, which can be summarised as asking whether ***animals learn to perform actions with the objective of acquiring information,*** has been reviewed and examined by Mackintosh (1974, 259), who clearly formulated the problem, and, after examining available evidence at the time, concluded that, in his own words, '*None of the studies under review has provided convincing evidence that observing responses are reinforced because they provide information or reduce uncertainty'.* After a period of relatively less intense research effort, interest in this important issue has re-emerged, probably because in the intervening years the development of behavioural ecology led to greater interest on the functional consequences (adaptive value) of behaviour, an issue that had been virtually neglected in the earlier literature (for a relevant exception see the already cited study of optimal sampling (Krebs et al., 1978). Another reason is the development of computational neuroscience, an interdisciplinary field that employs mathematical models to study neural function (Beierholm and Dayan, 2010; Iigaya et al., 2016). Given the rich and complex history of this problem, I will focus on more recent contributions, so as to frame the particular background and experimental questions of the present work. To do this, I will first elaborate on the concepts of "*learning"*, "*objective*" and "*information*" as these are complex concepts that might be in need of further clarification.

"Information" is a common term in contemporary society. Even though we live in what has been called the information age, the concept of 'information' is not as straightforward as it might seem. Information does not refer to a concrete item, amenable to a definition with sharp boundaries. Rather, it is an abstract concept that is difficult to define, and whose nature may take on different forms depending on the purpose with which it is being used. For the aims of this thesis a particularly useful conceptualization is that of Shannon Information (Shannon, 1948). Shannon developed a mathematical treatise of information in which he defined information as a reduction in uncertainty. In other words, acquiring information narrows the possibilities. One reason information exists is that there are dependencies in the world. For instance, the colour of the clouds in the sky has a relationship with the probability of rain occurring in the next hour. This is to say that cloud colour carries information about rain probability. The world is full of such correlations and animals would be well advised to able to attend to features of their environment that provide them information that can help them to survive and reproduce.

To say that animals learn to perform actions with the "*objective"* of acquiring information is to make a statement about the purpose of the behaviour. Defining the purpose of a behaviour is not straightforward. Purpose can be used at different levels of analysis and different time scales, including the metaphorical use in terms of thermodynamics (isolated systems do what it takes 'in order to' gain entropy), the trans-generational usage in evolutionary biology (the genome is modified by natural selection so as to maximise fitness) and the psychological level (choices are driven by a tendency to maximize utility).

At a proximate psychological level, the purpose of a behaviour can refer to a 'drive' that is being satisfied by performing it, in the sense used by Hull (1943) 'drive reduction theory',

which states that a drive is an "*excitatory state produced by a homeostatic disturbance*", an instinctual need that has the power of driving the behaviour of an individual.". For instance, when an animal learns to visit a certain location to obtain water, it can be said that the purpose of the behaviour is to quench thirst. At an ultimate level the purpose of a behaviour can refer to the behaviour's function in evolutionary terms, which would be akin to asking what selective pressures gave rise to the behaviour. For instance, the function of the startle response to a loud noise would be to minimize the potential harm of a threat, or, in the previous example, the purpose of thirst is to direct the individual towards eliminating a water imbalance.

The above-mentioned definitions of purpose are problematic and subject to philosophical debate, among other reasons because it is not clear how can drives be directly observed and measured, or how the evolutionary function of a behavior can be determined. Nonetheless, my approach in this thesis is to tolerate some semantic uncertainties and to treat the problem empirically, and specifically asking whether animals might be psychologically motivated to acquire information, and relating the problem with its ecological, adaptive consequences. Treating the idea that information can act as a primary reinforcer as a working hypothesis makes it possible to generate experimental predictions, and over the longer term, to clarify what exactly one is referring to when talking about information-seeking as a motivator of animal learning.

A common element of the experiments in this thesis is that they all concern "*learning"*. They were conducted under controlled laboratory circumstances, which allowed us set up scenarios that due to their arbitrary configuration must necessary involve learning. The subjects had to repeatedly interact with the experimental tasks, extract the relevant

contingencies out of this interaction and adjust their behaviour accordingly. As opposed to studying learning of information-seeking behaviour, one could also have studied innate information-seeking behaviour in animals. An example of this kind of behavior is the mechanism human babies are born with that causes them to move their gaze towards items of particular relevance, such as the faces of the people around them. In ecological scenarios it is often difficult to tease apart what is learned and what is innate, because from very early on animals are exposed to vast amounts of information from their environment, and no behavioural system is free from either component. Although the topic of innate information-seeking behaviours is a rich and relevant one, in this thesis I will limit myself to studying scenarios of information-seeking in which learning is fundamentally involved.

Background

To survive and reproduce animals must acquire valuable resources such as food, water and access to mates. However, acquiring these resources in an efficient way often entails first sampling the environment for information and using this information to guide subsequent behavior. As mentioned earlier, the tradeoff between acquiring food and acquiring information has been explored from a functional perspective trough the exploration-exploitation dilemma (Berger-Tal et al., 2014; Berger-Tal and Avgar, 2012; Eliassen et al., 2007; Sutton and Barto, 1998; Krebs et al., 1978) and from a mechanistic perspective through the potential reinforcing properties of acquiring information (Bennett et al., 2016; Blanchard et al., 2015; Bromberg-Martin and Hikosaka, 2009; Iigaya et al., 2016; Reed et al., 1996). Different lines of research provide evidence for and against the hypothesis that animals learn behaviours with the objective of acquiring information. We can classify

studies investigating the learning of information-seeking behaviour in animals as belonging to one of three paradigms.

In the first experimental paradigm, known as "*contra-freeloading",* subjects are presented with a choice between a freely available and a hidden source of food. A number of studies have shown that subjects prefer (or at least do not discard working for) the hidden food, even though it requires more time and effort to acquire than its freely available alternative (for reviews see: D'amato (1974), Inglis et al. (1997), Osborne (1977)). Because this preference has a cost, it has been hypothesised that subjects must be obtaining something else that they value in return. The "*Information-primacy*" hypothesis claims that what the animals gain is information (Bean et al., 1999; Inglis, 1987; Inglis and Shepherd, 1994). However, this is a difficult proposition to assess, as the kind of information that could be acquired in this type of experiment was not explicitly defined. Relatedly, the potential utility of acquiring said information has not been elaborated on. An alternative interpretation of the findings, which argues against the idea that the animals' behaviour is motivated by obtaining information, is that it reflects the animals' innate preference for foraging in a way that approximates their species-specific natural behaviour (Kacelnik, 1987)

A second paradigm that has been used to explore the idea that information can act as reward is the "*observing response*" reviewed in detail by Mackintosh (1974). With this procedure it has been found that animals will choose to perform an action that produces a discriminative stimulus but does not alter the reinforcement schedule. In one version of this paradigm, the animal can give either of two responses during the course of the experiment. The first response is towards an alternating schedule: it results in reward

during reinforcement periods (or trials), while it does not result in reward during extinction periods. The alternative response, also known as the observing response, produces a discriminative stimulus signalling whether the first response is in an extinction or reinforcement period. It has been shown that animals perform the observing response when it reliably signals which experimental period is in effect (Wyckoff Jr., 1952). This finding could be interpreted to mean that animals perform the observing response in order to obtain information. However, a finding that argues against this interpretation is that subjects will also perform the observing response when it exclusively signals periods of reinforcement, but not when it exclusively signals periods of extinction (Browne and Dinsmoor, 1974; Wilton and Clements, 1971). As these cues are equivalent in terms of the information they convey, this finding has been argued to be incongruent with an informational interpretation of the animals' behaviour (Dinsmoor, 1983).

The "*paradoxical choice"* procedure is a third paradigm that has been used to study information-seeking in animals more recently (reviewed in McDevitt et al. (2016) and Zentall (2016)). In it subjects are faced with a choice between two options that differ in the degree of uncertainty and (in some versions) in the probability of delivering reward. Here an initial choice is followed by a delay, or waiting time. During the delay that follows a choice the 'informative' option produces cues predicting whether a reward will be delivered or not, while in the delay that follows a choice for the 'non-informative' alternative the reward outcome is uncertain until the delay ends. The crucial finding emanating from this paradigm is that animals prefer an option that provides them with immediate information about upcoming reward outcomes, even if it has a lower probability of delivering reward (eg. Smith et al., 2016; Stagner et al., 2011; Stagner and Zentall, 2010; Vasconcelos et al., 2015; Zentall, 2011). However, not all findings derived

from this paradigm have been interpreted to support the information-seeking hypothesis (Roper and Zentall, 1999). Instead, the costly preference for information has been interpreted as a maladaptive or 'sub-optimal' property of the mechanism generating this behaviour and linked to pathological conditions in humans such as gambling addiction (Zentall, 2011).

To relate preference for the informative option in the paradoxical choice paradigm with its adaptive consequences, Vasconcelos et al. (2015) developed a model, which we will hereafter refer to as the *Neglect-Bad-News* model. The *Neglect-Bad-News* model is inspired by optimal foraging theory (Stephens and Krebs, 1986) and postulates that, rather than being '*sub-optimal'* as claimed by Zentall (2011) the preference exists in the laboratory because in natural circumstances obtaining such information can be used to guide behaviour so as to maximise reward rate. Specifically, in nature an animal that perceived a cue specifying a null outcome is forthcoming, would be able to redirect its behaviour to chase an alternative prey, thus saving valuable time. Typically, this certainty of failure occurs when the prey falls out of sight or range so that pursuing is not possible any more, as when a leopard abandons a chase when the gazelle has been able to establish some separation. The loss of rewards associated with information-seeking in experimental circumstances are therefore attributed to a discrepancy between natural and laboratory environments. Although the *Neglect-Bad-News* model appears to account for several findings from the paradoxical choice protocol, its interpretation, range of occurrence, and presence across species remains under debate.

My thesis focuses on exploring the paradoxical choice phenomenon. It does so by testing the *Neglect-Bad-News* model as well as alternative interpretations (Bennett et al., 2016;

Blanchard et al., 2015). The paradoxical choice paradigm has several advantages for the purposes of investigating information-seeking behavior. An advantage relative to the contra-freeloading or observing response procedures is that it offers a minimal contrast between the informative and non-informative options: both options expose subjects to physically equivalent stimuli and require the same type of response. The only difference between the options is in their informational content, thus avoiding potential confounds. Another advantage of this paradigm relative to the contra-freeloading or observing response procedures is it that in the paradoxical choice paradigm information-seeking has an easily quantifiable cost, namely, the reduction in reward probability associated with choosing the informative option. Lastly, the paradoxical choice phenomenon appears to demonstrate a particularly puzzling kind of information-seeking mechanism, namely preference for information in a situation in which information appears to be useless. This phenomenon has therefore been interpreted as evidence that information can be valued intrinsically (Blanchard et al., 2015; Gottlieb et al., 2013). Three of the four experimental chapters in this thesis explore information-seeking behaviour in rats using modifications of the paradoxical choice protocol, while in the remaining chapter I also investigate information-seeking behaviour, but using a schedule aimed at greater ecological realism, as it incorporates the possibility of using the information that becomes available after each choice.

Chapters' aims:

Chapter 1. As preference for information within the paradoxical choice paradigm has previously only been demonstrated in starlings and pigeons, the main aim of this chapter is to investigate whether this phenomenon is also present in rats. If this were the case, it would indicate that the phenomenon is taxonomically widespread, thus strengthening the argument that it reflects an adaptation to an ubiquitous feature of ecological environments, an argument made by the *Neglect-Bad-News* model. In addition, I use the results from this chapter to test a general-purpose decision model called the Sequential Choice Model (SCM). The SCM has a related rationale, but applies more widely (Kacelnik et al., 2010; Shapiro et al., 2008). It states that most foraging decisions in nature imply something more akin to go/no-go protocols than to simultaneous choice ones. Animals are more likely to encounter prey items one at a time, rather than multiple items simultaneously. The logic underling this assertion is that whatever the probabilities of encountering each of two prey items individually, the probability of encountering them together is necessarily smaller, as it is the multiplication of the two individual probabilities. The decision an animal would most commonly have to make is whether to pursue a particular prey item or to continue searching for alternative opportunities in the environment. For this reason, behaviour is expected to show more adaptive properties when tested in such protocols and observed behaviour in sequential encounters should have greater predictive power. The predictions of this model link response latencies in single-option trials (go/no-go) to preferences in simultaneous choice trials.

Chapter 2. This chapter aims to test several predictions of the *Neglect-Bad-News* model of paradoxical choice. In the first experiment I test whether, if the protocol is modified such

that the information can be useful within the task, then rats will use the information appropriately. Specifically, whether rats will escape from a trial in response to a signal for no-reward. In a second experiment I extend the *Neglect-Bad-News* model to test whether (as expected from the model) a controllable dead time between choices will affect preference for an informative option.

Chapter 3. In this chapter I develop an experimental procedure to investigate whether rats will learn to perform an action that provides no immediate benefit, but produces information that may later be used to act more efficiently. Unlike in the paradoxical choice paradigm, in this protocol information-seeking is separated from food-seeking as these require distinct responses.

Chapter 4. In this chapter I test two alternative interpretations of the paradoxical choice phenomenon: the explanation provided by the *Neglect-Bad-News* model, as well as a theory positing that animals have a broader intrinsic preference for information. I was able to distinguish between these two hypotheses by experimentally varying food reward outcomes across three informational dimensions: probability, magnitude and delay.

Chapter 1: Paradoxical Choice in rats

Published in: Ojeda, A., Murphy, R. A., & Kacelnik, A. (2018). Paradoxical choice in rats: Subjective valuation and mechanism of choice. *Behavioural processes*, 152, 73-80.

Introduction

Organisms frequently face tradeoffs between maximising immediate reward and gathering information. From a functional perspective this problem has been addressed in terms of the fitness contributions of exploration and exploitation (Berger-Tal et al., 2014; Krebs et al., 1978), while on a mechanistic level the focus has resided on the potential reinforcing properties of gaining information (Bennett et al., 2016; Blanchard et al., 2015; Iigaya et al., 2016). The biological significance of maximising the rate of gains in dimensions like food is obvious, but that of seeking information is not, and in particular it is not clear whether information *per se* can act as a primary reinforcer in learning processes. When information is relevant to the subject's current problem and serves to act more efficiently, for instance by increasing feeding rate, its biological relevance is derived from the increase in efficiency (Krebs et al., 1978) but in situations when the available information cannot be used to modify action, any information-seeking behaviour is hard to interpret and hence interesting. Consider a choice between two stochastic sources of delayed reward, where the outcome of each individual choice can be obtained at a price, either before or immediately after one of the options is chosen. If outcomes are signalled before each choice, it will often be worth investing in information as this allows to choose the better option. However, if the outcome is signalled after the agent makes its choice, the information would appear to be worthless, as the subject cannot change its choice. Post-choice information, however, is considered significant both theoretically (Loomes and Sugden, 1982) and empirically (Palminteri et al., 2017) in humans, and, as we discuss below, has also proven influential in experimental choice experiments with non-human

species. From a reward-maximizing perspective, for information that cannot be used to be valued is paradoxical, and hence worth exploring in some detail.

Several recent studies have found experimental conditions under which birds (pigeons and starlings) prefer a lower-reward but more informative option (Smith et al., 2016; Stagner and Zentall, 2010; Stagner et al., 2011; Vasconcelos et al., 2015; Zentall and Stagner, 2011). In these experiments (hereafter referred to as paradoxical choice) animals are given choices between two options that deliver reward probabilistically a fixed delay after being chosen. During the delay that follows the choice, in one of them (the informative option) signals are provided that anticipate whether a reward will be delivered or not, while in the alternative (non-informative) the outcome remains uncertain until the end of the delay. The critical finding is that both bird species show a strong preference towards the informative option. Remarkably, this preference for the informative option can lead birds to lose up to (and perhaps more than) 80% of available food rewards (Vasconcelos et al., 2015).

Although several studies have now firmly established the existence of such paradoxical choice (McDevitt et al., 2016; Zentall, 2016), its theoretical interpretation remains under debate. Broadly, the phenomenon can be interpreted in two ways. The first is that it reflects a maladaptive property of the psychological mechanism producing it. Indeed the phenomenon has been frequently labelled as *suboptimal behaviour* and proposed as an analogue of pathological human gambling (Zentall, 2011). It has also been claimed (Smith et al., 2016) that this phenomenon challenges Optimal Foraging Theory, which proposes that predictive models of animal behaviour can be produced by assuming that animals behave as if they were designed by natural selection to maximize Darwinian fitness

(Stephens and Krebs, 1986). When the only variable affecting fitness is feeding rate, OFT does not predict any loss of food in exchange for unusable information. However, the predictions of OFT address natural, rather than laboratory, circumstances, and this can be a factor of major significance.

The second way to interpret this phenomenon takes into account the laboratory-natural world distinction, arguing that while preference for the post-choice informative option leads to sub-maximal payoffs under artificial conditions, the mechanisms underlying such preferences could be beneficial in the environment in which the animals evolved (Vasconcelos et al., 2015). It is hard to imagine natural circumstances in which information about forthcoming food events cannot be used to modify behaviour adaptively. In particular, a forager that perceived a signal telling that no reward would follow from its last action would not pay the cost of waiting for the null outcome; it would instead divert its attention to exploit alternative opportunities, something subjects cannot do in the typical laboratory experiments. This hypothesis has been formalized by postulating that in the experiments birds edit out any time waiting for sure no-reward when computing reward rates (Vasconcelos et al., 2015).

A question of importance for the theoretical interpretation of paradoxical choice is whether this phenomenon is restricted to birds or is more taxonomically widespread. If the same phenomenon were to be found in a mammal, this would support its interpretation as an adaptation to a common characteristic of natural foraging scenarios, rather than being a maladaptive constraint on the cognitive competence of birds. Attempts to reproduce paradoxical choice in rats have so far yielded conflicting results. One research group has repeatedly failed to find evidence supporting paradoxical choice (Martínez et al., 2017;

Trujano et al., 2016; Trujano and Orduña, 2015), leading the authors to conclude that a genuine species difference exists. Another group has found positive evidence (Chow et al., 2017) but only under certain specific experimental conditions that might not be equivalent to those used with the birds (Martínez et al., 2017). Because several protocol features varied between experiments, it is not clear what is driving the different results. Here we adapt the paradoxical choice protocol to rats incorporating novel technical details aimed at resolving this conflict: 1) We use auditory stimuli to signal reward contingencies, as this modality is well suited to the rat's discriminatory capabilities (Heffner and Heffner, 1985); 2) We use different stimuli to signal the four possible contingencies: Two for the informative option (one for sure reward and another for sure no reward) and two for the non-informative option (two signals, each followed by 50% chance of reward); 3) Before testing for preferences, we ensure that the rats discriminate between the contingencies; 4) We vary the magnitude of reward loss associated with the informative option.

Another modification that we introduce here is aimed at investigating whether a well-established psychological mechanism underlying choice in starlings and pigeons (the so-called Sequential Choice Model), is also present in rats. To this effect, in addition to collecting data on choice behaviour in simultaneous encounters, we record latencies to respond (reaction times) in both single option and choice trials. While many models of choice processes treat decisions as the result of a 'tug-of-war' between the stimuli cueing for options present at the time of each choice, the Sequential Choice Model (SCM, Kacelnik et al., 2010; Shapiro et al., 2008) suggests that animals' decision mechanisms are adapted for sequential rather than simultaneous choices. According to this model, the subjective value of each option is reflected in the latency to respond towards it when encountered in isolation, and choices in simultaneous encounters result from a race

between these latencies. We examine this issue with our data, as a means to place the phenomenon of paradoxical choice in the more general framework of adaptive animal behaviour.

Methods.

Subjects

All experiments were carried out in compliance with the UK Animal (Scientific Procedures) Act (1986) and its associated guidelines. 16 male Lister Hooded rats (provider Envigo), 9 weeks old at the start of the experiment served as subjects. They were housed in groups of four. Throughout the experiment they were food deprived to 85-90% of their free-feeding weight using growth curves from the provider [Initial weight: 260g ± 12. Final Weight: 330g ±20 (Mean ± std. dev.)]. Water was provided ad-libitum in their home cages and they were maintained in a 12-hour dark/light cycle with lights on starting at 6 AM.

Apparatus

Testing was carried out in eight operant chambers (Med Associates, USA.) Each chamber contained three retractable levers: one in the back panel (centre) and two in the front panel, left and right of a central food magazine delivery tray equipped with an infrared beam and a sensor to record head entry. Each reward delivery consisted of two 45mg sucrose pellets (TestDiet, USA). A speaker (3.5" in diameter) was positioned above the back lever in the back panel. Each chamber was also equipped with a house light and a fan, which were switched on for the duration of the session. The chambers were controlled via custom-written Med-State Notation programs running on MED-PC V (Med Associates, USA).

General procedure.

The experiment was based on a trial-based chain procedure. On *choice* trials (**Figure 1**) responding on an attention manipulandum (the initial link in the response chain)) led to a choice between two other manipulanda (the options). Responding to either of these levers caused the onset of an acoustic cue that persisted for 10 seconds until an outcome occurred. Pressing the "*info*" (informative) option resulted in either the *i-plus* (sure reward) or *i-minus* (sure no-reward) cues. Responding to the "*non-info*" (non-informative) option resulted in either the *n1* or *n2* cues (both associated with 0.5 probability of reward). On *forced* trials only one of the options was available. We varied across treatments the probability of reward in *info*, with p(*i-plus*) being either 0.2, 0.4 or 0.5. As a consequence, the *info* option had a reward probability that was 60% lower, 20% lower, or the same as the *non-info* option, respectively. For clarity, we describe first the cues and then proceed to detail the training procedures and the final protocol.

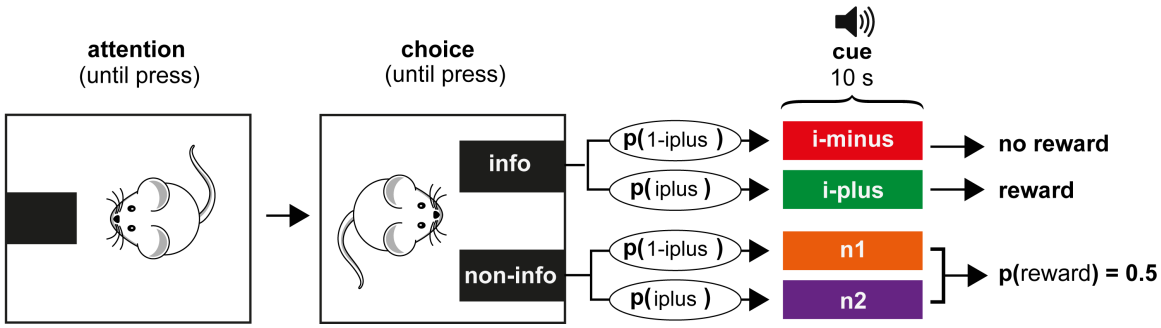


Figure 1. Choice Trial Structure

Cues

The four cues consisted of sounds, all with a duration of 10 seconds, and each associated to a reward probability. There were two cues for the informative option: *i-plus* (100% reward probability and *i-minus* (0% reward probability), and two cues for the non-informative option (*n1* and *n2*, both with 50% reward probability). The four sounds were: a pure tone (3 kHz, 78 dB), a buzzing sound (74 dB), white noise (74 dB), and a clicking sound (74 dB). Assignment of sounds to reward probabilities was counterbalanced across subjects to avoid the possibility of option preferences being generated by any intrinsic aversive or attractive properties of the sounds.

Pre-training

Magazine training:

To train the rats to collect reward in the apparatus training began with a single session with 60 reward deliveries. The interval between reward deliveries was sampled from a Poisson distribution with a mean of 30 seconds.

Lever training:

Over the next three sessions the rats were trained to press the two front levers for food. The left lever was available on the first session, the right lever on the second and either lever (with equal probability across trials) on the third. Once a lever extended into the chamber, a single press resulted in its retraction and immediate reward delivery (fixed ratio 1 schedule). In the first two sessions, the lever became available again after 10 seconds. In the third session, this delay was composed of a constant element plus a variable one. The constant one was 5 seconds, and the variable one was sampled from a Poisson distribution

with a mean of 5 seconds. All three sessions concluded after 60 reward deliveries.

Cue training

To ensure that in the main experiment the subjects were tuned to the reward probabilities of the four cues (*i-plus*, *i-minus*, *n1* and *n2*) we preceded the choice phase with a Pavlovian protocol in which the rats were passively exposed to the four cues, each with its associated reward contingency. Like in the main phase that was to follow, each cue (sound) was presented for 10s and reward was delivered (probabilistically) at the end of the stimulus, but in this phase no action was required to produce the cues. These cue-training sessions consisted of 120 trials, with 30 trials for each cue. To avoid large deviations from the expected reward probability of each cue, proportions of rewards were fixed for every 10 trials with each cue. These trials were then randomly intermixed. Thus, if for example a 50% reward probability was associated with a cue, exactly 5 out of every 10 trials with that cue ended in reward, but the order was randomised. The ITI (inter-trial interval) was composed by sampling a Poisson distribution with a mean of 15 seconds and adding 10 seconds. The subjects performed one daily session of this phase for 11 days.

Main experiment

Treatments

A within-subjects design was used, with each rat participating in the three treatments. In all treatments pressing the *info* (informative) lever resulted in the presentation of either the *i-plus* (sure reward) or *i-minus* (sure no-reward) cues, with the probability of *i-plus* differing between treatments (0.2, 0.4 and 0.5), while responding on the *non-info* (non-informative) lever caused either of the *n1* or *n2* cues, with the same probabilities as *i-plus*

and *i-minus* in that treatment. Because both *n1* and *n2* delivered reward with 0.5 probability, the probability of reward in the *non-info* option was always 0.5, and treatments differed only on the probability of reward for the *info* option. Treatments were named according to the probabilities of reward in the two options: 1) 0.2 *info* vs 0.5 *non-info*, 2) 0.4 *info* vs 0.5 *non-info* and 3) 0.5 *info* vs 0.5 *non-info*. In all cases, the probability of reward in *info* (the probability of the cue *i-plus*) was equal to or lower than 0.5, hence the info option was at best neutral.

Subjects were split into two groups that performed the treatments in two different orders: *"0.5/0.4/0.2"* and *"0.2/0.5/0.4"* according to the probability of reward in the informative option. Both groups performed the experiment concurrently. Subject assignment to treatment was counterbalanced such that there was no correlation between treatment and any of the following parameters: side of the informative option; hour of testing; cage in which the animals were housed, or sound-reward contingencies. For each treatment, the subjects performed one daily session for 14 consecutive days. Each rat was trained at the same time of day; one cohort of rats began the experiment at 7:30 A.M. while the other started at 11:30 A.M. Between treatments, the *info* and *non-info* option lever sides were reversed for half of the rats, to control for carry-over effects of the spatial position of the options in the previous treatment.

Trial structure

There were two kinds of trials: *choice* and *forced.* All trials started with the rear (attention) lever extending. Pressing this lever resulted in its retraction, and either one (forced trials) or both (choice trials) of the front levers to be presented (Figure 1). The purpose of requiring a back-lever press to start a trial was to ensure task engagement, to position the rat centrally in the chamber with respect to the options, and to enable measurement of

latency to respond starting from a central position. Pressing a front lever initiated a cue (the terminal link) and also caused that lever (forced trials) or both levers (choice trials) to retract. The auditory cue was presented for 10 seconds, after which reward delivery could occur. Trials were separated by an ITI (inter-trial interval) generated by sampling from of a Poisson distribution with a mean of 15 seconds and adding 10 seconds.

Session structure

Each session was composed of 120 trials: 80 *forced* (half *info* and half *non-info*) and 40 *choice* (randomly intermixed). A session finished after completing the 120 trials or 3 hours, whichever occurred first. Other trial sequence details were determined as previously described in the *cue discrimination training* section

Data analysis

All data processing and statistics was done in MATLAB 2017a. Before statistical analysis proportion data was arc-sine square root transformed to normalize the residuals (Grafen and Hails, 2002). When performing repeated measures analysis-of-variance the data were verified for sphericity using Mauchly's test.

Cue training phase

The focus of this study is choice between actions leading to different cues and outcomes. It is thus important to establish if subjects were tuned to the programmed contingencies of each cue. To this effect we quantified for each trial the total duration of head entries into the food magazine during cue presentation. To check for statistical significance of differences between cues we pooled the data from the last five sessions and performed a

repeated-measures ANOVA with cumulative head poking time as the dependent variable and cue as a within-subject factor. Post-Hoc tests between all cue pairs were then performed using the Bonferroni correction for multiple comparisons.

Option preferences

To examine preference in choice trials, defined as the proportion of choice trials in each session allocated to the *info* or *non-info* options, we performed a repeated-measures ANOVA with treatment (categorical) and session (continuous) as within-subject factors, and preference for *info* as the dependent variable. To establish if asymptotic preference differed from indifference we pooled preference through the last five sessions of each treatment and compared it against a 0.5 expectation of indifference, using one sample t-tests.

Latency analysis

Testing the Sequential Choice Model required examining latencies to respond and predicting proportions for each option in choice trials using latency to respond in forced trials. To do this, for each session, we compared the latencies of 40 pairs of forced trials. Pairs were composed of a forced *info* trial and a forced *non-info* trial, following the order in which they were presented. As these were unique pairs, each latency was used only once for the prediction of choice proportions in that session. Predicted preference for *info* was defined as the proportion of forced trial latency pairs in which *info* had a shorter latency than *non-info* (for an example see Table 1 in Supplementary Methods). We then analysed statistically the predicted preferences, following the same procedure as for observed preferences.

To establish the statistical significance of the match between predicted and observed preferences we performed a permutation test (Golland et al., 2005), as follows. Pooling all data, the whole experiment yielded a total of 672 sessions, each one with a predicted and observed preference (*info* or *non-info*). We shuffled the predicted preferred options for each session in 10^6 random permutations, and determined for each permutation the percentage of sessions in which the predicted and observed preferred option matched. The probability of the observed proportion under the null hypothesis was calculated as the fraction of permutations with a higher match percentage than the actual (non-shuffled) match.

A difference between Tug of War models and the Sequential Choice Model (Shapiro et al., 2008; Kacelnik et al., 2010) is that the former implies that latency to respond towards a given option in choice trials should be longer than latency towards the same option in single option trials, while the latter predicts the opposite. Further, the SCM predicts that the shortening of latencies in choice trials (relative to forced trials) should be stronger for the non-preferred than for the preferred option. This is caused by the assumption of parallel processing of each option under the SCM model. As in each choice trial the option yielding a shorter latency to act is expressed as the preferred alternative, the obtained distribution of latencies for each option in choice trials has a richer representation of shorter latencies than the distribution for the same option in the absence of competition. This bias should be greater the stronger the censorship suffered by the original (no competition) latency, and this occurs for the least preferred option. To test this, we determined the difference between single option and choice trial latencies as a function of the level of preference for that option (complete, partially preferred and partially non-

preferred). This poses statistical difficulties because latency data are right skewed and the sample size (number of trials) of each option in choice trials is determined by the preferences of each subject. Therefore, we used a metric (the *latency index*), that is robust in this respect. For a given option, this latency index takes a value of 0 when all latencies for that option in choice trials are shorter than the median latency for the same option in single option trials, 1 when all choice trials yield longer latencies than the median latency in forced trials, and 0.5 when choice and forced trials have the same median latency. The *latency index* was computed as follows. For each session of each subject the following procedure was carried out once (if all choice trials belonged to the same option) or twice (once for each option if preferences were partial): 1) determine the median latency on forced trials (the *reference*) 2) determine the number of choice trials of the option with a latency shorter than the *reference* (*shorter* trials) and longer than the *reference* (*longer* trials). For each session with partial preferences this resulted in two pairs of values (the number of *shorter* and *longer* trials for the preferred and non-preferred options) while when preference was absolute (all choice trials were allocated to the same option) one pair of values was obtained per session (the number of *shorter* and *longer* trials for the option that was completely preferred). The total number of *shorter* and *longer* trials for each of the three categories (completely preferred, partially preferred, and partiallly non-preferred) was then calculated by summing across all sessions of the same subject, and the index was computed by dividing the number of *longer* trials by the total number of trials (*shorter* plus *longer)*. The SCM predicts values below 0.5, the Tug of War models predicts values above 0.5, and, of course, under the null hypothesis the index would approximate 0.5.

Results

Cue discrimination

The average cue discrimination performance during the five last sessions of pre-training is shown in **Figure 2.** The figure shows that the rats' head poking duration during cue presentations varied appropriately according to each cue's reward contingency. A one-way repeated-measures ANOVA with cue as a within-subject factor revealed a significant effect of cue: ($F_{3,45}$=41.74, P<0.0001). Differences between cues were confirmed by post-hoc pairwise comparisons, with significant differences between all cue pairs (Bonferroni-corrected, P<0.001), except for the *n1* vs *n2* pair, as should be expected given that these two cues signal the same reward probability. In the main experimental phase, the difference between the *i-minus* cue and all other cues was maintained (Bonferroni-corrected, P<0.001), but there was no longer a significant difference between the *i-plus* cue and the *n1* or *n2* cues (P>0.05, for details see Figure S1 in Supplementary Results). This confirms that the bias towards the *info* option could not be explained by a failure to discriminate the corresponding higher incidence of no-reward.

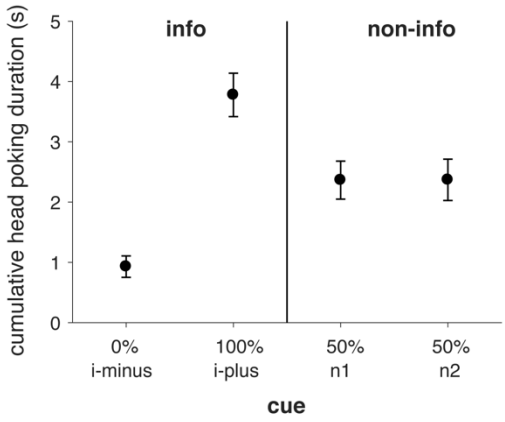


Figure 2. Cue discrimination
Mean ± S.E.M. (n=16) average cumulative duration of head poking into the food magazine (per trial) during cue presentation in the five last sessions of the pre-training phase. The rats showed good discrimination between the cues associated with the *info* (informative) and *non-info* (non-informative) options, and within the *info* signals when a reward was due and when it was not.

Option preferences (Choice trials)

Figure 3 shows that in choice trials the rats preferred *info* when its reward probability was the same or 20% lower than that in *non-info* (0.5 *info* vs 0.5 *non-info* and 0.4 *info* vs 0.5 *non-info*). However, when reward probability in *info* was 60% lower (0.2 *info* vs 0.5 *non-info*) rats preferred the *non-info* option (**Figure 3**). A two-way repeated-measures ANOVA with treatment and session as within-subject factors and preference as the dependent variable revealed a significant effect of treatment ($F_{2,30}$=27.75, P<0.0001), session ($F_{1,15}$=268.19, P<0.0001), and session by treatment interaction ($F_{2,30}$ = 44.12, P<0.0001). Pooled over the last five sessions, preference for *info* is significantly above 50% for treatments 0.5 *info* vs 0.5 *non-info* (t_{15}=11.86, P<0.0001) and 0.4 *info* vs 0.5 *non-info* (t_{15}=9.18, P<0.0001), and significantly below 50% for treatment 0.2 *info* vs 0.5 *non-info* (t_{15}=3.87, P<0.002). For individual subject data see Figure S2 in Supplementary Results.

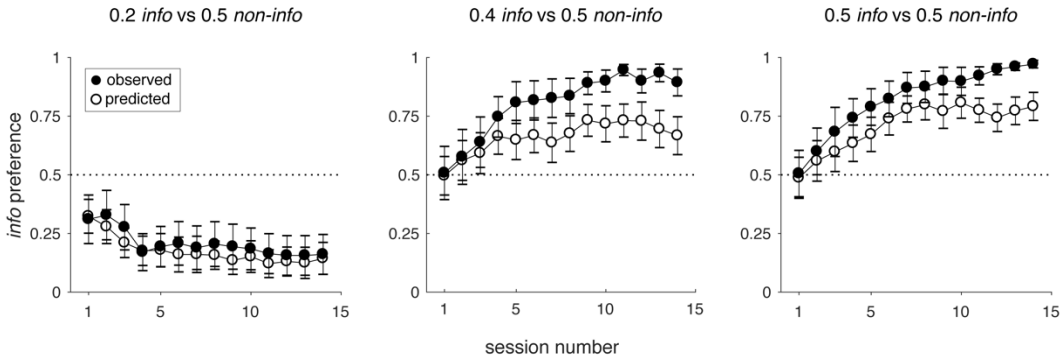


Figure 3. Observed and predicted preferences

Mean ± S.E.M. (n=16) proportion of choices for the *info* (informative) option in each treatment. Each panel is a treatment with a different reward probability for the *info* option. Observed preferences in choice trials are depicted in full symbols. Empty symbols show predicted preferences in choice trials derived from latencies in forced (single-option) trials, according to an implementation of the Sequential Choice Model. The dotted line represents indifference.

Latencies (shortening)

The Sequential Choice Model predicts differences in latencies to respond to each option in forced (single-option) and choice trials (i.e. when the option was chosen), as well as differences between preferred and non-preferred options (defined as being chosen above or below 50% of choice trials, respectively). According to the model, when preferences are partial latency to select an option on choice trials should be shorter than when the same option is encountered on its own, due to cross-censoring of the distributions of latencies. This shortening effect should be more pronounced for the option preferred less than 50% of the trials, because its latency distribution is censored more strongly, than for the preferred alternative. However, when an option is exclusively preferred, latencies in choices should equal those in single-option trials, because when the alternative is never chosen there is no censorship. To this effect we split latencies according to which was the preferred option for a given subject and a given session, and whether the degree of preference was extreme (absolute preference, hence no censorship), or partial (i.e. both options received at least some choices). **Figure 4** shows the results. Latency to respond to each option was either indistinguishable or shorter in choice trials relative to forced trials. When preference was absolute, the chosen option showed no detectable difference between the two kinds of trials (t_{15}=1.57 P=0.13, Bayes Factor for the null = 1.4054). When preference was partial, both the preferred and non-preferred options showed significantly shorter latencies in choice trials, and the shortening effect was stronger for the non-preferred option (t_{15}=2.4, P< 0.03 and t_{15}=3.3, P<0.005 for preferred and non-preferred respectively).

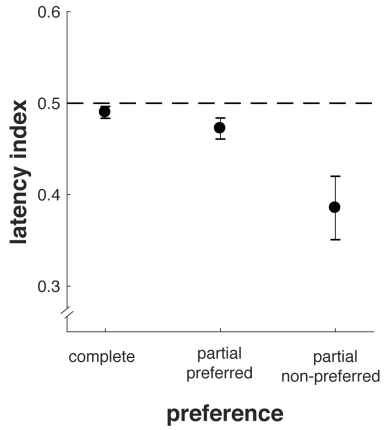


Figure 4. latencies (choice versus forced trials)
Mean ± S.E.M. (n=16) latency index as a function of option preference. Values lower than 0.5 indicate that latency to respond to an option when the alternative is present (choice trials) is shortened with respect to when the same option is presented in isolation (forced trials). When preference was absolute, the index did not differ from 0.5, but when preference was partial, both preferred and non-preferred options showed significantly shorter latencies in choice trials. The shortening effect was stronger for the non-preferred option.

Latencies (choice predictions)

In addition to predicting the shortening of latencies in choice trials, the Sequential Choice Model states that the difference in latencies between options in forced trials should predict preference in choice trials. This analysis is shown in **Figure 3** (empty symbols). Predicted preferences derived from latencies on forced (single-option) trials mirror observed preferences on choice trials (Figure 3, empty versus solid symbols), albeit with predicted preferences tending to be less extreme than observed ones. The procedure used to predict preference from latencies is described in detail in the Methods. A two-way repeated-measures ANOVA with predicted preference as the dependent variable and treatment and session as within-subject factors revealed a significant effect of treatment ($F_{2,30}$=47.38, P<0.0001), session ($F_{1,15}$=138.74, P<0.0001), and treatment by session interaction ($F_{2,30}$=64.16, P<0.0001) Pooled over the last 5 sessions, predicted preference for *info* was significantly above 50% for treatments 0.5 *info* vs 0.5 *non-info* (t_{15}=4.99, P<0.0002) and 0.4 *info* vs 0.5 *non-info* (t_{15}=2.78, P<0.02), and significantly below 50% for treatment 0.2 *info* vs 0.5 *non-info* (t_{15}=5.67, P<0.0001), just as for the experimental data. The predicted preferred option for a session matched the actual preferred option in 88% of all sessions. This match is significantly above chance (permutation test, P<0.0001).

Discussion

We discuss first the specific mechanism of paradoxical choice with respect to outcome information and then the more general problem of how subjective valuation of alternatives leads to choice behavior.

Paradoxical choice in rats

Our results show that rats, like pigeons and starlings, prefer an informative over a non-informative option, to the extent of showing systematic deviations from reward rate maximization. Rats' preference for an informative option is not absolute, but depends on how much loss it causes: rats preferred an informative option yielding a 50% or 40% reward probability against a non-informative alternative offering 50% reward probability, but when the informative option offered a 20% reward probability they switched preference towards the higher yielding non-informative option. Birds (pigeons and starlings), in contrast, seem to go as far as preferring an informative option yielding as little as 10% reward probability against a 50% non-informative alternative (Vasconcelos et al., 2015; Smith et al., 2016).

A potential critic could argue that the difference in reward probability between 0.4 info and 0.5 non-info is not large enough for the rats to be able to detect it and thus their preference for the informative option is not really paradoxical. With the present data we cannot know if the animals can discern that the informative option has a lower probability of reward and are still willing to incur the cost of selecting it, or rather if they fail to

discriminate the two reward probabilities. This ambiguity however does not make the behavior non-paradoxical. If preference for an informative option leads an animal to prefer an option with a lower reward probability, then the mechanism that generates this behavior is costly regardless of what the animals "know". Our results are 'paradoxical' because the rats were not indifferent: they systematically preferred an option causing a 20% loss of reward. The precise reasons for this paradoxical preference do of course, admit further investigation. Another potential issue is that we employed a within-subject experimental design. Because of this, carry-over effects from one treatment to the next could have affected preference for each option. To control for this possibility, we reversed between treatments and for half of the subjects the spatial position of the *info* and *non-info* options. The individual-level data (Figure S2 in Supplementary Results) show that for each treatment subjects displayed the same terminal preferences, regardless of whether the options had been reversed or not. This demonstrates that preferences are genuine, and not due to the order of presentation of the treatments.

Rats' preference for the informative option cannot be attributed to a putative inability to discriminate between the terminal link cues, as the (Pavlovian) nose-poking behaviour recorded in the presence of each cue during the pre-training phase (Figure 2) showed that subjects could discriminate between cues according to their reward contingencies. This was also the case for starlings, which, when given a choice between the terminal links showed orderly preference for the higher payoff alternative (Vasconcelos et al., 2015). Interestingly, in the rats, during the main experimental condition the cue for absence of reward continued to be associated with a suppressed Pavlovian response during the waiting time, but the three cues associated with a possibility of reward elicited similar durations of

head-poking even though one signaled sure reward and the other two partial reinforcement with equal chance of reward or no reward. Although this is curious in its own right, it cannot explain the preference for *info* as that would presumably manifest itself as a lack of discrimination between the cue for the absence of reward and rest of the cues. Also, it does not impact our conclusions as their behaviour during the pre-training phase demonstrates that the rats can discriminate between the cues, even if they do not explicitly show it during the main experiment.

The fact that the phenomenon is present in birds and mammals argues against labelling it 'sub-optimal choice' (Zentall, 2011), as its presence in taxonomically distant species suggests that the systematic failure to maximize reward in this specific laboratory situation probably reflects a psychological mechanism maintained and/or reinstated by natural selection under natural conditions. The challenge lies in identifying the critical differences between natural and experimental situations.

The comparison between rat and bird results requires some caution. Previous work with rats in similar protocols has had conflicting results, calling for a detailed analysis of the protocols used in each case. In the present experiments we used four auditory cues, one signaling sure reward, another one sure lack of reward, and 2 more associated with an equal chance of reward or lack of it. We also ensured that the animals did discriminate between the four contingencies (Figure 2). These tests serve to reject the possibility that differences between tests in birds and rats reflect differences in discrimination and not in valuation.

A previous study also found evidence of paradoxical choice in rats (Chow et al., 2017), but using a protocol that had an important difference relative to ours. Chow and co-workers used one cue (a lever) to signal partial reinforcement in the non-informative option and another cue (another lever) to signal sure reward in the informative option, signalling sure no-reward implicitly by a blackout. Such a setup could have biased results in favour of the informative option by not providing a specific stimulus signalling sure no-reward. Martínez et al. (2017) made a similar argument stating that Chow et al.'s 2-cue setup lacks a stimulus that can function as a conditioned inhibitor. To support their argument, Martinez and co-workers (2017) performed an experiment that also used levers as stimuli, but with a 4-cue setup (one lever for each contingency, including sure lack of reward) and reported that they were unable to find preference for the informative option. Given that we found paradoxical choice with a 4-cue setup, and these cues were highly salient, our result cannot be attributed to the lack of a specific stimulus signalling sure no-reward. This still begs the question of why in Martínez et al. (2017) as well as other studies in rats (Chow et al., 2017; Trujano et al., 2016; Trujano and Orduña, 2015) paradoxical choice was not evident. One likely cause of failures to show the added value of information is that in some of these experiments the reward loss associated with choosing the informative option might have been beyond the range of what rats will pay. Martínez and co-workers (2017) gave rats a choice between an informative option offering a reward probability of 0.5 and a non-informative option offering 0.75, a loss of 33% of possible rewards. In our experiments rats reverted to preferring the higher reward but non-informative option when the loss associated with preference for the informative option increased from 20% to 60% of possible rewards. Therefore, Martinez et al.'s null finding is consistent with the results of one of our treatments. Nonetheless, other possible procedural differences call for detailed analysis. Several of the previous studies used lights as cues (Chow et al., 2017;

Trujano et al., 2016; Trujano and Orduña, 2015) although the visual modality is not ideally suited to albino laboratory rats. Generalization is another potential issue. In the studies that used lights or levers as cues (Chow et al., 2017; Martínez et al., 2017; Trujano et al., 2016; Trujano and Orduña, 2015), what made the cues different from each other was their spatial position and this could have led to some degree of generalization between them.

In the present study, we used four auditory cues, one for each outcome, randomised the assignment of cues to outcomes, ensured that rats discriminated between these cues, and varied the magnitude of the loss in rewards associated with the informative option between no loss, a 20% loss, and a 60% loss. Our subjects preferred the higher probability, non-informative option for the more extreme condition, but showed a very strong and reliable preference for the informative option up to a loss of 20% (Figure 3). We surmise that the evidence for the presence of the phenomenon in rats is conclusive, and sensitive to the magnitude of the incurred loss.

The fact that rats' preference for an informative option is not strong enough to overcome large losses contrasts with the results found with pigeons and starlings (Smith et al., 2016; Vasconcelos et al., 2015), and may reflect a true species difference worth of further investigation. The argument put by Vasconcelos et al. (2015) to explain birds' extreme preference is that the '*cue that signals a sure absence of a forthcoming reward does not contribute to the animal's computation of lost opportunity'* (Vasconcelos et al., 2015; Smith et al., 2016), because in nature animals in such situations can re-direct behavior to other targets, so they would not incur the time cost of waiting for something that will with certainty not come. An associative learning account of the same phenomenon would posit that the cue for the absence of reward does not acquire inhibitory properties. This argument leads to expecting that preference for the informative option will be extreme

regardless of its probability of reward, a prediction that has been supported by bird data (Vasconcelos et al., 2015; Smith et al., 2016) but does not apply to the picture that is emerging for rats.

Mechanism of choice: the Sequential Choice Model

The value a subject attributes to an option, and the process by which this valuation translates into choices between options are distinct and separable empirically, because subjective valuation is expressed also in no-choice situations by readiness to take an option. One particular model, the Sequential Choice Model (SCM, Kacelnik et al., 2010; Shapiro et al., 2008) implements the following rationale: 1) when a subject encounters signals that anticipate the possibility of a reward in nature, it normally does so in isolation (different kinds of opportunities (such as prey species) are met sequentially); 2) An experienced animal has expectations for the potential outcome of each kind of opportunity, and for how it compares with the average payoff experienced in the environment as a whole; 3) These expectations translate into noisy distributions of latency to respond to each signal, with signals leading to more favourable outcomes resulting in shorter average latencies; 4) If 2 (or more) signals are encountered simultaneously (a 'choice'), then the said distributions are sampled, and the one yielding the shorter latency elicits a response, resulting in a choice at behavioural level. In this scenario, the observed distribution of latencies to choose results from statistical cross-censorship between two distributions, and hence has some predictable properties. The most specific, and that which separates this model from more conventional models for cognitive processes underlying behavioural choices, is that when responses to a given signal occur in the context of a choice, the observed latencies will be on average the same or shorter than when the same signal elicits

a response in a sequential encounter. An option that is always preferred faces no censorship and its latencies should remain the same. If preferences are partial the preferred and non-preferred options cross-censor each other and therefore have shorter average latencies. The predicted shortening should be more evident for the signal that elicits longer latencies when alone and as a consequence is chosen less often (the non-preferred option), as only the left tail of its distribution will be expressed in choice situations. These predictions contrast with more intuitive and widespread hypotheses, which fail to relate choices to behaviour in sequential encounters and predict, with different degree of explicitness, that choices rely on specific cognitive mechanisms of comparison which deal with the attributes of signals' outcomes in the choice scenario. This leads to the usual expectation of a cognitive evaluation time applicable to models that have been called "Tug-of-War".

Some of the different predictions of these two kinds of model have been compared empirically in birds and the SCM's assumptions linking sequential situations with simultaneous choices have on the whole been supported (Freidin et al., 2009; Kacelnik et al., 2010; Shapiro et al., 2008; Vasconcelos et al., 2010, 2015), but what has hitherto remained most elusive is the demonstration of a latency shortening, affecting primarily the non-preferred alternative. In the present experiment, latencies to accept each option in forced (single-option) trials predicted preferences in choice trials (Figure 3), supporting the SCM's assumption that the subjective value of an option is inversely related to the latency to accept it in sequential (forced) trials. Further, Figure 4 examines the results in terms of the relation between latencies in sequential and choice trials, showing that latency to accept an option in choice trials was equal or shorter than when presenting the option in isolation (forced single-option trials). When an option was preferred in exclusivity, its

latency did not differ, and when preferences were partial, both the preferred and non-preferred options had shortened latencies, with the non-preferred option experiencing the stronger shortening between choice and sequential trials. These results constitute the most direct corroboration of the predictions of the SCM reported so far.

In summary: two hypotheses derived from behavioural research in birds have been corroborated using rats. First, rats, like previously tested birds, objectively sacrifice reward rate in exchange for a reduction in uncertainty, but this does not mean that uncertainty avoidance acts as primary reinforcement. The mechanism underlying this tradeoff is not yet clear, but a strong candidate relates to the learning process by which signals are credited with the value of their outcomes, so that time costs for periods of lost opportunity due to waiting for sure no reward are somehow excluded, or down-weighted with respect to times waiting for at least some probability of reward. Second, the mechanism that leads from valuation to preference in choice situations seems well accounted for by a model that gives priority to sequential encounters without postulating any special mechanism adapted for choice situations. The fact that these two somewhat counterintuitive phenomena are present in such distant taxa suggests that both are adaptive in natural circumstances, rather than resulting from unavoidable cognitive constraints.

Supplementary Methods

trial number	1	2	3	4	5	6	7	8	9	
trial type	forced info	forced non-info	choice	forced info	forced non-info	choice	forced info	forced non-info	choice	
Latency (s)	3	2		2	3		2	4		
predicted	non-info			info			info			**total**= 2/3 info
observed			info			info			non-info	**total**= 2/3 info

Table 1. Schematic example of the prediction of preferences for a session

Supplementary Results

Figure S1 shows the head poking behavior during the different treatments of the main experiment. A two-way repeated-measures ANOVA with cue and treatment as within-subject factors revealed a significant effect of cue: ($F_{3,45}$=80.91, P<0.0001), treatment ($F_{2,30}$=3.97, P<0.0001) and their interaction ($F_{6,90}$=2.78, P<0.02). Post-hoc comparison between cues revealed a significant difference between the *i-minus* cue and all other cues, below is the table with the p-values (Bonferroni-corrected) for all pairwise comparisons.

		p-value
'iminus'	'iplus'	<0.0001
'iminus'	'n1'	<0.0001
'iminus'	'n2'	<0.0001
'iplus'	'n1'	0.0875
'iplus'	'n2'	0.2415
'n1'	'n2'	0.9234

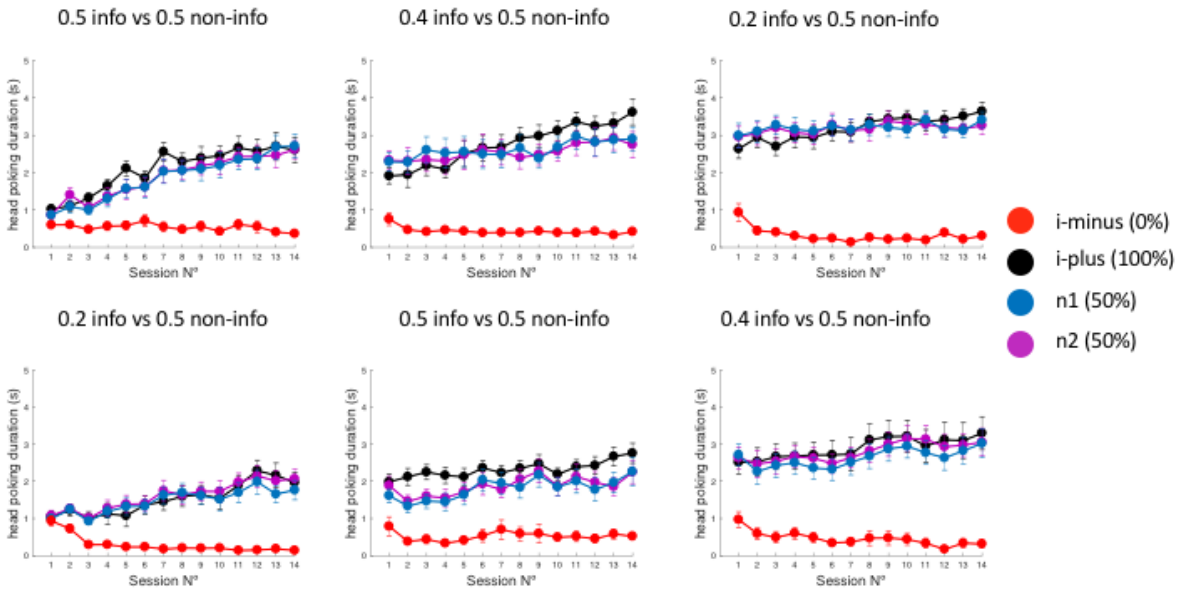


Figure S1. Cue discrimination
Mean ± S.E.M. duration of head poking into the food magazine during cue presentation in each of the conditions of the main experimental phase

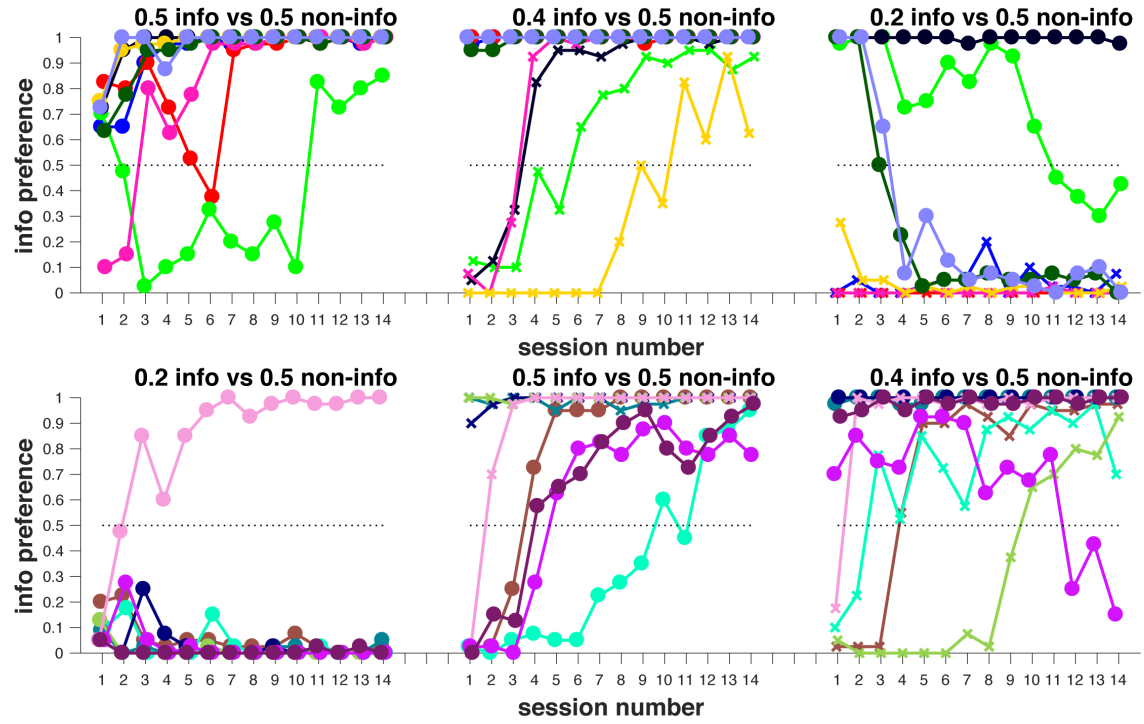


Figure S2. Info preference (individual)
Preference for the informative option in each of the treatments of the main experiment. Each colour represents a different subject. Between phases the spatial position of the options was reversed for half of the subjects, points from subjects that experienced a reversal are marked with x's.

Chapter 2: Testing an Optimal Foraging model of Paradoxical Choice

Introduction

In natural environments, animals often have to choose between obtaining immediate rewards and acquiring information. Whereas the purpose of obtaining rewards (such as food or water) is readily apparent, that of obtaining information is more difficult to determine. Sometimes information can aid in securing a future reward, but in other instances the information cannot be used to change the animal's course of action (at least not in the short term). The value of the former type of information is readily understandable, as its biologically relevance is derived from the increase in efficacy (Krebs et al., 1978). However, animals' preference for the latter type of information is intriguing (e.g. Bennett et al., 2016; Bromberg-Martin and Hikosaka, 2009; Iigaya et al., 2016), as it is not clear how information that seemingly cannot be used to change behaviour could increase the evolutionary fitness of animals.

The problem of working for information is often framed as an exploration-exploitation conflict, and is of widespread importance for many behavioural scenarios. For instance, in artificial intelligence a strategy that is frequently explored is what is called an 'epsilon-greedy strategy', namely one in which the subject selects the highest paying alternative with probability 1-epsilon and selects options at random with complementary probability, thus balancing the need to sustain payoff and to correctly track the environmental opportunities (Sutton and Barto, 1998). If epsilon is very large, then behavioural allocation approaches randomness, but always with some bias towards the higher payoff. However, this strategy was conceived as a reasonable policy when the ranking in probability of reward between alternatives is incompletely known, as in a Two-Armed Bandit. We are

interested in a related, but not identical kind of problem: when the probability of reward is fully known and stable, but the options differ in (unusable) information delivered between choice and outcome.

Several studies have described experimental conditions under which birds (pigeons, starlings) and more recently rats, facing options that differ in both expected payoff and degree of reward uncertainty prefer the lower-reward but more informative option (Smith et al., 2016; Stagner and Zentall, 2010; Stagner et al., 2011; Vasconcelos et al., 2015; Zentall and Stagner, 2011; Ojeda et al., 2018) . In these experiments (referred to as ´paradoxical choice´) subjects are given choices between two options that provide reward probabilistically a fixed delay after being chosen. During the delay that follows choice, in the informative (‘*info’)* option signals are provided that anticipate whether a reward will be delivered or not, while in the alternative non-informative (‘*non-info’*) option, the outcome remains uncertain until the delay has elapsed.

All species tested so far, pigeons (Smith et al., 2016), starlings (Vasconcelos et al., 2015) and rats (Ojeda et al., 2018), show a strong preference for the informative option. Although the existence of the paradoxical choice phenomenon has been firmly established (for review see McDevitt et al. (2016) and Zentall (2016)), its theoretical significance is not well understood. Two different interpretations of animals' preference for the informative option have been put forward.

On the one hand, it has been argued that paradoxical choice reflects a "*suboptimal"* choice mechanism that provides no advantage to the animal (Smith et al., 2016). This interpretation questions the generality of Optimal Foraging Theory (OFT), which states

that predictive models of behaviour can be generated by assuming that animals behave as if they were designed by natural selection to maximize Darwinian fitness (Stephens and Krebs, 1986). According to OFT, animals should not sacrifice food in return for useless information if food reward rate is the only variable affecting fitness.

On the other hand, an alternative interpretation is that the paradoxical choice does not reflect a "*suboptimal"* choice mechanism, but rather that its adaptive value only becomes apparent under natural circumstances. Whereas in the laboratory environment the animals cannot use the information to guide their behaviour, this might be the case in natural environments. This is to say that the preference for information may be beneficial in the environment in which the animal evolved (Vasconcelos et al., 2015). In the nomenclature used by human cognitive researchers, such mechanisms are 'ecologically rational'(Gigerenzer, 2008) being only apparently maladaptive in the arbitrary circumstances of the experiments. In general, in natural environments obtaining information about forthcoming reward events can be used to modify behaviour. Specifically, an animal that receives a signal of no-reward would not pay the time cost of waiting for a null outcome, rather it would redirect it behavior towards pursuing an alternative opportunity in the environment, which is something that the animals in the paradoxical choice experiments are unable to do. In mathematical terms this was formalized by claiming that when calculating the profitability of each option, the animals edit out (or ignore) any time waiting for sure no-reward (Vasconcelos et al., 2015). We will hereafter refer to this account of paradoxical choice as the *Neglect-Bad-News* model

The fact that the phenomenon of paradoxical choice is present in taxonomically distant species supports its interpretation as an adaptation to a common characteristic of natural

foraging scenarios, rather than as a maladaptive constraint on the cognitive competence of a particular species. There seem to be species differences since rats, although displaying the phenomenon of paradoxical choice like pigeons and starlings, do it with lesser strength. Pigeons and starlings prefer an informative option over a non-informative one even when the probability of reward of the informative one is 80% lower (Smith et al., 2016; Vasconcelos et al., 2015). Rats, however, only seem to be willing to pay a limited cost to avoid the uncertainty. They prefer the informative option when the loss is of up to 20% but switch to preferring the higher reward non-informative option, when tested with a loss of 60% or more. This preference reversal makes investigating the empirical validity of the *Neglect-Bad-News* model particularly important. In its basic form, this model is consistent with total neglect of the difference in objective payoffs, as observed in starlings and pigeons. This discrepancy may however be resolved if, rather than assuming total neglect of times waiting for sure no-reward, such times are discounted by a weight factor. This would require adding an extra parameter to the model, which in its basic form is parameter-free. Since adding parameters is preferably avoided, before engaging in some model-fitting strategy, here we test two novel predictions derived from the structure of the *Neglect-Bad-News* model.

In Experiment 1 we assess the assumption that, if the protocol is modified so that information is usable, then animals will use it appropriately to guide their behaviour. Specifically, if the protocol used in the previous chapter is modified by introducing an "escape" response that allows the termination of a trial during the interval between choice and outcome, then rats will emit such a response when they face a signal that predicts sure no-reward but not when the signal predicts sure reward. We thus modified the protocol used in our previous study (Ojeda et al., 2018) to enable such an escape response. If

subjects have at least some sensitivity to the time waiting for sure non-reward, then the addition of an escape possibility should further increase the preference for info, as selective escaping objectively eliminates such cost.

In Experiment 2 we introduce a further modification aimed at testing another prediction derived from the rationale of the *Neglect-Bad-News* Model: we add a controllable common waiting time (the *pre-choice* delay) before the rat can choose between *info* and *non-info*. In this protocol, a trial begins by pressing the *start* manipulandum. After such response the options (*info* and *non-info*) are not immediately available. They both become available once the *pre-choice* delay has elapsed. This delay is similar to an inter-trial interval (ITI) in more conventional schedules of reinforcement, in the sense that it represents a dead time between the end of a trial and the following opportunity to choose between *info* and *non-info*, but since it is a pre-requisite for the animal starting each cycle it is perhaps more accurately equated with a common initial link in a chain schedule. To avoid issues of confusing nomenclature we simply refer to it as the *pre-choice* delay.

The reason a *pre-choice* delay should have an influence on preference between *info* and *non-info* under a rate-maximising logic is presented formally in the Methods section. Verbally, the argument is the following. When escape is possible, and there is no pre-choice delay, the advantage of the *info* option resides in that the subject only has to wait (pay a time cost) when it is waiting for a forthcoming sure reward (during the post-choice cue presentation period). This occurs in the trials in which the subject chooses *info* and is lucky. When it chooses *info* and it is unlucky (i.e. it gets a signal for sure non-reward), it can escape to face a new choice immediately after choosing (i.e. without waiting to complete the post-choice delay). Thus, the subject does not invest time waiting in unlucky trials. A perfectly behaving subject, capable of escaping as soon as it is given the

opportunity to do so, will experience a higher reward rate in the *info* option, regardless of the difference in reward probability with the non-info alternative. However, if a pre-choice delay is introduced, trials with the sure no-reward cue also involve some waiting, and now the relative reward probabilities of *info* and *non-info* do affect the ranking of reward rates between *info* and *non-info.*

According to this reasoning, adding a pre-choice delay should reduce preference for *info*. If the pre-choice delay is sufficiently long, then *non-info* should be preferred. In other words, given such an extra delay, which is likely to exist in most natural foraging cycles, we should expect a preference reversal as a consequence of the interaction between pre-choice time and reward probability in the *info* option. These intuitions were formalized in a rate-maximising mathematical model (described in detail in the methods) that allowed us to estimate two durations of the pre-choice delay that predicted opposite preferences (*info* vs *non-info*). To summarize, in this experiment we tested the prediction that manipulating a *pre-choice* delay should result in a preference reversal.

Methods

Subjects

All experiments were carried out in compliance with the UK Animal (Scientific Procedures) Act (1986) and its associated guidelines. 24 male Lister Hooded rats (provider Charles River), 9 weeks old at the start of the experiment served as subjects. They were housed in groups of four. Throughout the experiment they were food deprived to 85-90% of their free-feeding weight using growth curves from the provider. Initial weight: 291g ± 14. Final Weight: 326 ±18 (Mean ± std. dev.) Water was provided ad-libitum in their

home cages and they were maintained in a 12-hour dark/light cycle with lights on starting at 6 AM.

Apparatus

Testing was carried out in eight operant chambers (Med Associates, USA.) Each chamber contained three retractable levers: one in the back panel (centre) and two in the front panel, left and right of a central food magazine delivery tray equipped with an infrared beam and a sensor to record head entry. Each reward delivery consisted of two 45mg sucrose pellets (TestDiet, USA). A speaker (3.5" in diameter) was positioned above the back lever in the back panel. Each chamber was also equipped with a house light and a fan, which were switched on for the duration of the session. The chambers were controlled via custom-written Med-State Notation programs running on MED-PC V (Med Associates, USA).

General procedure.

The experiments were based on a trial-based chain procedure. In *choice* trials (**Figure 1**) responding on a trial-start manipulandum (the initial link) led to a controllable waiting time (the *pre-choice* delay) during which the trial-start lever was retracted, and then to a choice between two other manipulanda (the options). Responding to either of the options caused the onset of an acoustic cue that persisted for 40 seconds. At the end of this interval the sound ceased and the reward/no reward outcome occurred. Choosing the "*info*" (informative) option resulted in either a 100% (sure reward) or a 0% (sure no-reward) cue. Responding to the "*non-info*" (non-informative) option resulted in either a *50%a* or a 50%b cue (both associated with 0.5 probability of reward). Importantly, the trial-start manipulandum was again available immediately after pressing any of the options' levers and pressing it during cue presentation caused the immediate offset of the current cue with

no reward and the immediate start of a new pre-choice delay (constituting an escape response and the start of a new trial). If a trial was escaped the trial counter advanced to the next trial. In *forced* trials only one of the options (either *info* or *non-info*) was available.

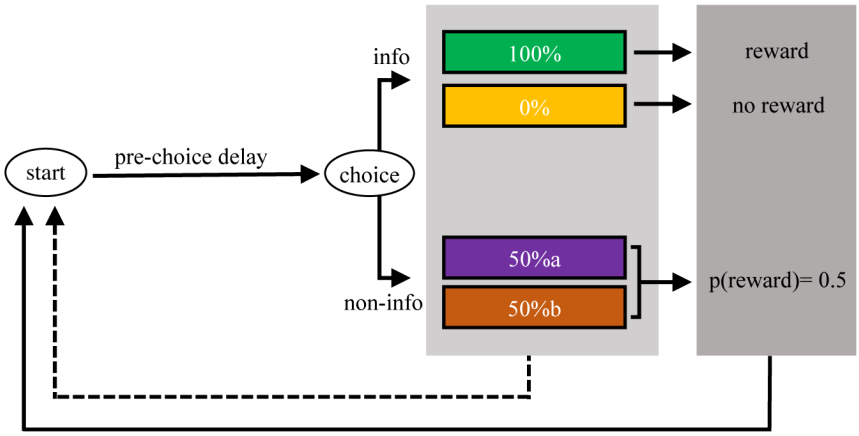


Figure 1. Trial structure
Pressing the start manipulandum during the presentation of the cues allowed subjects to escape a trial (marked with a dotted line). See the main text for a full description.

Cues

The four cues consisted of sounds, all with a duration of 40 seconds, and each associated to a reward probability. There were two cues for the informative option: *100%* (100% reward probability and *0%* (0% reward probability), and two cues for the non-informative option (*50%a* and *50%b*, both with 50% reward probability). The four sounds were: a low frequency pure tone (3 kHz, 70 dB), a buzzing sound (70 dB), a high frequency pure tone (6 kHz, 70 dB), and a clicking sound (70 dB). Assignment of sounds to reward probabilities was counterbalanced across subjects to avoid the possibility of option preferences being generated by any intrinsic aversive or attractive properties of the sounds.

Pre-training

Magazine training:

To train the rats to collect rewards in the apparatus, training began with a single session with 60 reward deliveries. The interval between reward deliveries was sampled from a Poisson distribution with a mean of 30 seconds.

Lever training:

Over the next three sessions the rats were trained to press the two front levers. The left lever was available on the first session, the right lever on the second and either lever (with equal probability across trials) on the third. Once a lever extended into the chamber, a single press resulted in its retraction and immediate reward delivery (fixed ratio 1 schedule). In the first two sessions, the lever became available again after 10 seconds. In the third session, this delay was composed of a constant element plus a variable one. The constant one was 5 seconds, and the variable one was sampled from a Poisson distribution with a mean of 5 seconds. All three sessions concluded after 60 reward deliveries.

Cue training

To ensure that in the main experiment subjects were tuned to the reward probabilities of the four cues (*100%, 0%*, 50*%a,* and *50%b*) we preceded the choice phase with a Pavlovian protocol in which the rats were exposed to pairings of the four cues, with their associated reward contingencies. Like in the main phase that was to follow, each cue (sound) was presented for 40s and reward was delivered (probabilistically) at the end of the stimulus, but in this phase no action was required to produce the cues. These cue-training sessions consisted of 120 trials, with 30 trials for each cue. To avoid large

deviations from the expected reward probability of each cue, proportions of rewards were fixed for every 10 trials with each cue. These trials were then randomly intermixed. Thus, if for example a 50% reward probability was associated with a cue, exactly 5 out of every 10 trials with that cue ended in reward, but the order was randomised. The ITI (inter-trial interval) was composed by sampling a Poisson distribution with a mean of 15 seconds and adding 10 seconds. The subjects performed one daily session of this phase for 11 days.

Main experiments

Subject assignment to group was counterbalanced such that there was no correlation between group and any of the following parameters: side of the informative option; hour of testing; cage in which the animals were housed, or sound-reward contingencies. The subjects performed one daily session for 14 consecutive days. Each rat was trained at the same time of day; the first cohort of rats began the experiment at 7:00 A.M. the next started at 11:30 A.M and the last started at 4:00 PM. A session finished after 120 trials or 4 hours, whichever occurred first.

Trial structure

There were two kinds of trials: *choice* and *forced.* All trials started by pressing the rear (trial-start) lever. Pressing this lever resulted in its retraction and initiated the *pre-choice delay*, once this delay elapsed, one (forced trials) or both (choice trials) of the front levers were presented. Pressing one of the front levers: 1) initiated a cue, 2) retracted that lever (forced trials) or both front levers (choice trials), and 3) presented the rear (trial-start) lever. The auditory cue was presented for 40 seconds, after which reward delivery could occur or not according to the respective contingency. If the trial-start lever was pressed

during cue presentation this immediately caused the termination of the cue without reward delivery and the start of the pre-choice delay of the next trial (constituting an escape response).

Session structure

Each session was composed of 120 trials: 80 *forced* (half *info* and half *non-info*) and 40 *choice* (randomly intermixed). Other trial sequence details were determined as previously described in the *cue discrimination training* section.

Experiment 1

Experiment 1 tested the prediction that rats will selectively escape from a stimulus that predicts sure no-reward, as implied in the rationale of the *Neglect-Bad-News* model. In this experiment, both *info* and *non-info* had the same probability of delivering reward (0.5 for both) and the *pre-choice delay* had a duration of 1s. A single group of eight rats participated in this experiment.

Experiment 2

Experiment 2 tested the prediction that manipulating the *pre-choice* delay (the time for the options to become available after starting a trial) can result in a reversal of the ranking of reward rates under the specified conditions, and this can translate into a preference reversal. In this experiment the reward probability of *info* was lower than that of *non-info* (0.2 *info* vs 0.5 *non-info*), but this could, in theory, be compensated by time saving if the subject used the available information appropriately. Two groups of eight rats each participated. In one group the *pre-choice* delay had a duration of 1s while in the other group the *pre-choice* delay was 25s long. These durations were chosen by means of a

foraging rate model that served to identify a duration of the *pre-choice* delay where the predicted experienced reward rate was higher for exclusive choice of *info* (1s group) and another duration where the same was true for *non-info* (25s group). In the following section this model is specified in detail.

Modelling reward rate in Experiment 2

We used a mathematical model to predict the effect of the *pre-choice* delay on preferences. The model yields the reward rate of a hypothetical subject that chooses either *info* or *non-info* exclusively, as a function of the duration of the *pre-choice* delay, and predicts preference assuming the highest rate is chosen. Forced trials are not included. We assumed the subject would be either a specialist in *info* or in *non-info*. The info specialist would escape immediately in all trials in which it gets the signal for sure no-reward, and would stay to experience the outcome when the signal is for sure reward. The non-info specialist is assumed to never use the escape lever (since it has no means of predicting the outcome). The reward rates for each policy are thus the following:

Reward rate for a specialist in *info*:

$$rr_{info} = \frac{P_i}{P_i * \left(t_p + t_c\right) + (1 - P_i) * \left(t_p\right)}$$

Reward rate for a specialist in *non-info:*

$$rr_{no-info} = \frac{P_n}{P_n * \left(t_p + t_c\right) + (1 - P_n) * \left(t_p + t_c\right)}$$

Where:

P_n = probability of reward for the *non-info* option

P_i = probability of reward for the *info* option

t_p = duration of the *pre-choice* delay

t_c = duration of the cues, i.e. the delay between choice and outcome (reward/no-reward)

In each equation the numerator is the expected reward of the option while the denominator is the time cost associated with that option. Because the reward is the same in all cases, it can be cancelled out so that the numerator reduces to the probability of reward. The time cost is composed of two terms, the first being the time cost associated with rewarded trials. This cost is the same for both options: the *pre-choice* delay plus the cue duration. The second term is the time cost associated with unrewarded trials, this term is different between options because in the non-informative it comprises both the duration of the *pre-choice* delay and the duration of the cues while in the informative option it comprises only the duration of the *pre-choice* delay (because in this option the subject can escape when receiving the sure no-reward cue).

For the instantiation of the model P_n (probability of reward of *non-info*) was set at 0.5 while P_i (probability of reward of *info*) was set at 0.2, and t_c (the duration of the *cues*) was set at 40s. The results of the model are depicted in **Figure 2**. The model shows that increasing the duration of the pre-choice delay leads to a reversal of the ranking in payoffs, and hence a potential reversal in preference. We chose two values of the pre-choice delay, such that one resulted in a predicted preference for *info* (higher reward rate for *info* than *non-info)* while the other did for *non-info* (the converse). With a *pre-choice* delay of 1s the reward rate of *info* is 1.8 times larger than that of *non-info,* while with a pre-*choice* delay of 25s the reward rate of non-*info* is 1.3 times larger than that of *info*.

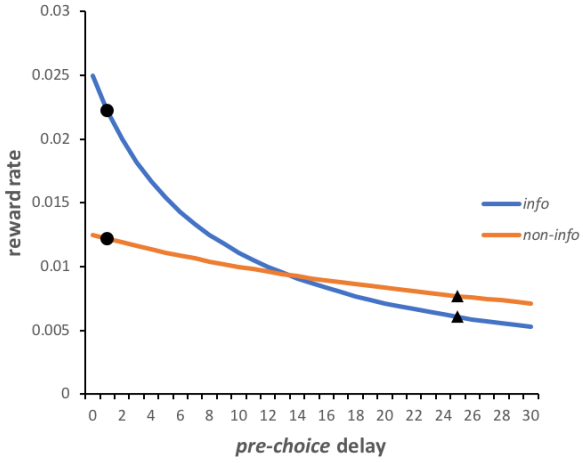


Figure 2. Modelled reward rates
Estimated reward rates for the *info* and *non-info* specialists in Experiment 2. Circles mark the delay chosen for the first group (1s) while triangles mark the delay chosen for the second group (25s). The model does not include forced trials or latencies (delayed actions), both features present in the real situation.

Data analysis

All data processing and statistics was done in MATLAB 2017a. Before statistical analysis cumulative head poking time was square-root transformed in order to normalize the residuals. Pairwise tests between cue pairs were performed using the Tukey-Kramer critical values. Analysis of preferences or escape proportions was done on the data from the last session of the main experiments.

Results

Experiment 1: Escape from a stimulus that predicts the absence of reward.

The purpose of this experiment was to test an aspect of the rationale of the foraging interpretation of so-called paradoxical choice. The *Neglect-Bad-News* model of (Vasconcelos et al., 2015) argued that the apparently irrational preference for avoiding uncertainty, in the absence of any use for the information acquired, could be that in nature animals do not face situations where information is useless, as in the laboratory protocol. A basic mechanism of preference towards being informed could then reflect a heuristic that always works in the wild but can lead to losses in contrived experimental situations. In particular, the argument is that when an animal can be certain that no benefit is forthcoming, it can abandon the present prey pursuit and avoid paying the time cost. Hence, the only time cost associated to a fully informational option is that waiting for sure reward, making it psychologically preferable. If such a mechanism for using the information exist, one may expect that in a laboratory protocol where subjects can use the information, for instance by aborting the present food pursuit, they will use it selectively to escape from a cue that predicts sure no-reward.

Pre-training

With the aim of ensuring good discrimination between the cues we preceded the main experiment with a pre-training phase in which rats were exposed to the cues and their associated-reward contingencies without any instrumental action required. The Pavlovian

head-poking response to the different cues during the pre-training phase (**Figure 3)** shows evidence of discrimination. A repeated-measures ANOVA with cue and session as within-subject factors and duration of head poking as the dependent variable revealed a significant effect of cue ($F_{2,14}$=7.8, P<0.01), session ($F_{1,7}$=, P<0.0001), and cue by session interaction ($F_{2,14}$=10, P<0.01). Please note that the two non-informative cues (*50%a* and *50%b*) were averaged given that they signal the same reward probability. Pairwise comparisons between the cue pairs revealed a difference for the 0% vs 100% pair (P<0.05) but did not reach significance for the 0% v 50% pair (P=0.12) or the 50% vs 100% pair (P=0.13). Notice that while a difference in Pavlovian head poking is evidence for discrimination, lack of difference is not at all suggestive of its absence. This is true for the usual statistical reasons associated with null results, but particularly relevant here because the rats may find that an expected reward with 50% probability is sufficient to warrant visits to the magazine. For the purpose of this study, the discrimination between sure and null probability of reward was critical and satisfied.

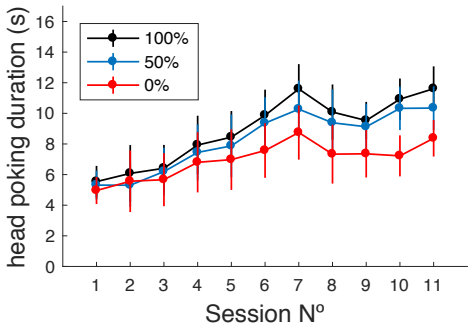


Figure 3. Cue discrimination
Mean ± S.E.M. (n=8) cumulative duration (in seconds) of head-poking into the food magazine during cue presentation in the pre-training phase. Each coloured line represents a cue's reward probability: 50% for the *non-info* option and 100% or 0% for the *info* option.

Main experiment

Rats developed a preference for the informative option even though its probability of reward was the same as the non-informative option (0.5 *info* vs 0.5 *non-info*) (**Figure 4)**. Preference for *info* in the last session is significantly above 50% (Wilcoxon signed-rank, n=8, W=34, *P<0.05* two-tailed). All subjects were sensitive to the contingency for the "escape" possibility, as shown in **Figure 5**: When responding to the *info* option, the probability that rats would actuate the escape lever and thus abort the trial to fast-forward to the next one was higher for a cue that predicted sure no-reward (*0%* cue) reward relative to one that predicted sure reward (*100%* cue) (Wilcoxon signed-rank, n=8, W=36, *P<0.01* two-tailed). It is curious, however, that the rats did escape in a high percentage of the trials when reward was sure to be delivered. This is never predicted by the rate maximizing policy and may be an expression of impulsivity (the difficulty in withholding a response even when doing so would be beneficial).

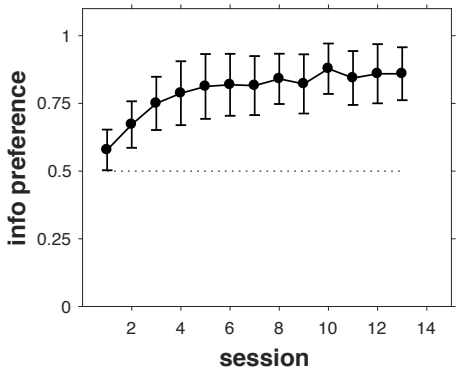


Figure 4. Preference for *info*
Mean ± S.E.M. (n=8) proportion of choices for the *info* (informative) option.

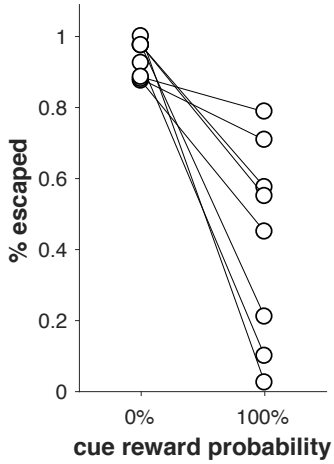


Figure 5. Escape behaviour
Proportion of trials that were escaped as function of the *info* cue presented. Each line represents a different subject (n=8).

Experiment 2: The effect of the pre-choice delay on preference for an informative option

The rate maximizing model described earlier on shows that if the probability of reward in the informational side is lower than in the non-informational side (i.e. 0.2 vs. 0.5 respectively in the example) then the relative reward rate of a specialist in info and non-info show a reversal as a function of times other than the interval between choice and outcome (in this protocol the 'pre-choice' interval). This should lead to a preference reversal. The purpose of this experiment was to test this prediction by using those reward probabilities and two values of the pre-choice interval.

Pre-training

The Pavlovian head-poking response to the different cues during the pre-training phase (**Figure 6)** shows evidence of discrimination. A repeated-measures ANOVA with cue and session as within-subject factors and duration of head-poking as the dependent variable revealed a significant effect of cue ($F_{2,28}$=11.32, P<0.001), session ($F_{1,14}$=96.6, P<0.0001),

and cue by session interaction ($F_{2,28}$=, P<0.001). Pairwise comparisons between cue pairs revealed a difference for the 0% vs 100% pair (P <*0.01*) and the 0% vs 50% pair (P<0.05) but did not reach significance for or the 50% vs 100% pair (P=0.1).

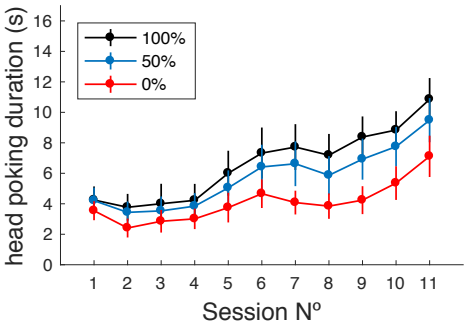


Figure 6. Cue discrimination
Mean ± S.E.M. (n=16) cumulative duration (in seconds) of head-poking into the food magazine during cue presentation in the pre-training phase. Each coloured line represents a cue's reward probability: 50% for the *non-info* option and 100% or 0% for the *info* option.

Main experiment

Figure 7 shows the mean preference for *info* in two groups that differed in the length of the *pre-choice* delay. There was no significant difference in preference between groups (Wilcoxon rank-sum, W=68, n_1=n_2=8, P=0.68 two-tailed). Also, in both groups preference was not significantly different from indifference (Wilcoxon signed-rank, n=8, 1s: W=18, P=0.68,25s: W=15.5, P=0.85). This indifference is notable because in one of the conditions of the experiment described by (Ojeda et al., 2018) when there was no escape possibility and the probabilities were 0.2 for info and 0.5 for non-info, the rats strongly preferred non-info, and because indifference implies a loss of available rewards. Comparing between experiments, this implies that adding the escape possibility did increase preference for the info side, as expected. Another aspect to consider is that Figure 5 shows interindividual averages. When observing individual trajectories (**Figure 8**) we can see a difference between the two groups: when the pre-choice interval was just 1s, all

but one of the rats showed strong preference for one of the options. In contrast, when the pre-choice interval was 25s, the individual results are much less orderly, with most of the subjects showing true intermediate preference.

Figure 9 shows that sensitivity to the escape contingency was present in the two groups. After choosing *info*, escape was more likely for the cue that predicted sure no-reward than for the cue that predicted sure-reward (Wilcoxon signed-rank, n=8, 1s: W=21, $P<0.05$, 25s: W=28, $P<0.05$

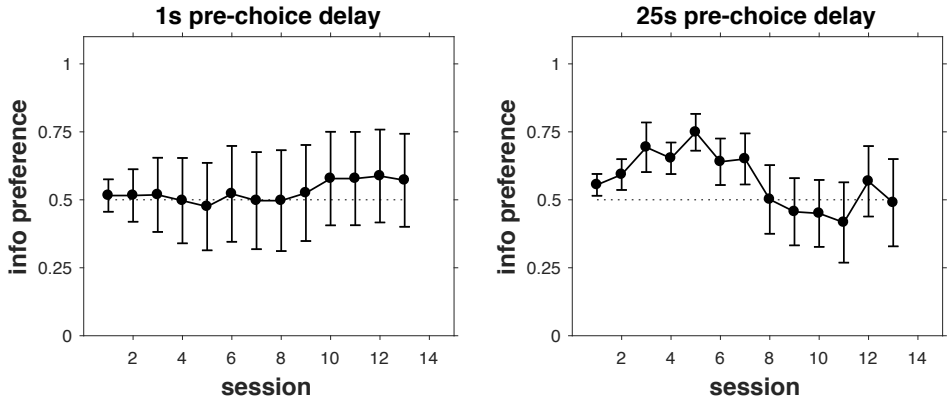


Figure 7. Preference for info
Mean ± S.E.M. proportion of choices for the *info* (informative) option.

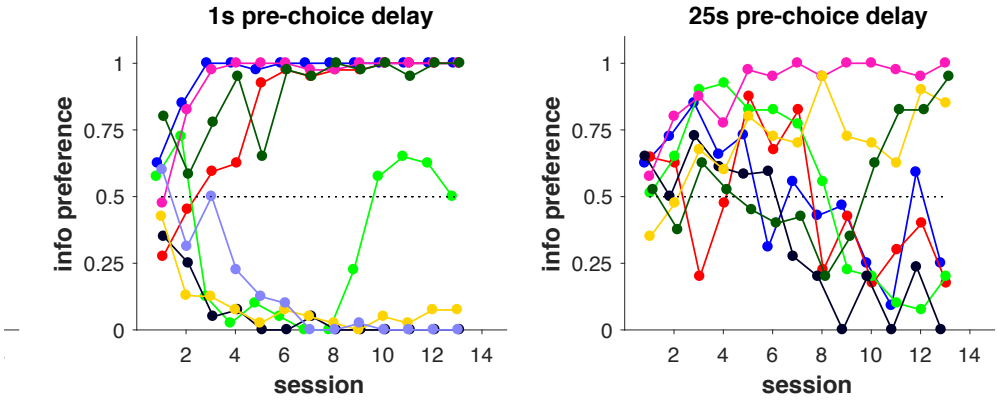


Figure 8. Preference for info (individual)
Proportion of choices for the *info* (informative) option. Each line represents a different subject.

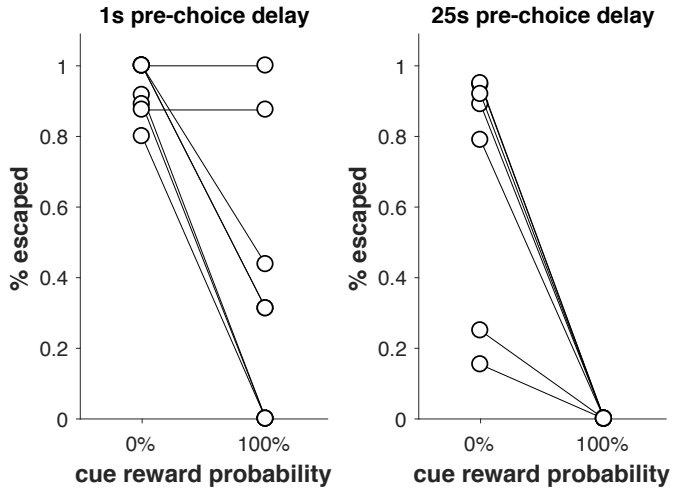


Figure 9. Escape behaviour
Proportion of trials that were escaped as function of the cue presented. Each line connects a different subject (n=8).

Discussion

The objective of both experiments in this study was to test two predictions of the *Neglect-Bad-News* (Vasconcelos et al., 2015) model of so-called paradoxical choice, namely the fact that animals prefer an informative option that signals reward outcomes in advance over an uninformative one that keeps them under uncertainty, and that they would escape from a sure no-reward waiting time when this is possible. The preference for useless information in the standard protocol is never functional, because it leads to null gains when there are no differences in reward probability or losses when the informative option has lower probability than the alternative. According to the *Neglect-Bad-News* model this paradox is explained when one takes into account a key distinction between nature and the laboratory. Specifically, in natural environments an animal that received a signal specifying that no reward is forthcoming, would not pay the time cost of waiting for this outcome, because it would instead abort the current pursuit and look for alternative opportunities in the environment. By contrast, in the paradoxical choice laboratory protocol (Ojeda et al., 2018; Smith et al., 2016; Vasconcelos et al., 2015) the animal is forced to wait until the end of the trial, even if it knows that it will not be rewarded. The behaviour observed in the experiments can thus be a consequence of the animals obeying a psychological mechanism that edits out the time spent under certainty for no-reward from the reward rate computations. The *Neglect-Bad-News* model thus predicts that if the paradoxical choice paradigm is modified such that the information about upcoming rewards is task-relevant, then animals will use this information appropriately. In Experiment 1 we tested this prediction and found that, as expected, rats were more likely to escape from a trial when presented with a cue that predicts sure no-reward relative to a

cue that predicts sure reward, but this difference was less pronounced than what a strict rate maximiser subject would do.

I also examined the possible effects of times other than those between a choice and its outcome. A mathematical model showed that the addition of such time could lead to a reversal in the ranking of reward rates for an animal that behaves strictly as a specialist in info or non-info, with the former always escaping for sure non-reward and never for sure reward, while the non-info specialist never escapes from any trial. Following this line of reasoning, in Experiment 2 we compared two groups that differed in the durations of a pre-choice delay, but the result of this experiment are not sufficiently clear to claim a definite support or rejection of the model. We discuss each of the experiments in greater detail below.

Experiment 1

In the paradoxical choice procedure animals opt between an alternative where information about forthcoming individual outcomes is absent and one in which this information is given, but presented after that animal has made its choice, so that it cannot be used to influence individual choices. Our rationale is that this preference is adaptive in nature because if the information were usable, animals would use it appropriately. Following this logic, in Experiment 1 we tested the prediction that if the paradoxical choice protocol is modified to make the information about individual outcomes potentially usable, then rats would use it appropriately. Specifically, animals given a chance to escape or "bail-out" of a trial, should always do so when receiving a signal for sure no-reward (*0%* cue) while, if they receive a signal for certain reward (*100%* cue), they should always wait for the

reward to be delivered before starting a new trial. We modified the paradoxical choice protocol to allow for these choices. In the new version, at any moment during cue presentation the rats could choose to terminate the current trial and start the next one immediately. Rats escaped almost all of the trials in which they were presented with the *0%* cue. In contrast, when presented with the *100%* cue rats escaped a significantly smaller proportion of trials. This differential escape behaviour is in accordance with the direction of predictions of the *Neglect-Bad-News* model, even though the model assumes more extreme differences than observed. This model has also been discussed in the context of data obtained from pigeons and starlings (Vasconcelos et al., 2015). The fact that novel predictions derived from the underlying rationale are corroborated in rats supports qualitatively this line of thinking, thus strengthening the argument that this phenomenon reflects an adaptation to a common characteristic of natural foraging scenarios.

Despite the broad agreement with the predictions of the *Neglect-Bad-News* model some interesting discrepancies do exist. Rats escaped in almost all trials with the 0% cue, just as the model predicted, but they also escaped a significant proportion of trials with the 100% cue. The model assumes that none of these trials would be aborted. Although it is not possible to establish functional explanations for this discrepancy, different mechanistic possibilities exist, and are discussed below.

One possible explanation for the escape responses to the *100%* cue is incomplete sensory discrimination between the *info* cues. The Pavlovian head-poking during the pre-training phase showed different levels of responding to the two *info* cues (*0%* and *100%).* However, the results of this analysis only demonstrate different mean levels of responding, and do not guarantee that in every individual trial the rats could correctly discriminate

between the two cues. Future experiments should aim at measuring the effect of discriminability of the cues on this effect.

Another possible cause for the escape responses to the *100%* cue can be sought in the associative processes involved in acquiring the escape response. The escape response (pressing the trial-start lever) had two ways of acquiring a positive association with reinforcement: Secondary reinforcement and generalization. Either way, once the start lever response was reinforced, that response needed to be inhibited in the trials in which the *100%* cue was presented.

The first way is trough secondary reinforcement (Williams, 1994). To escape a trial, rats had to press the trial-start lever (positioned at the rear of the chamber). Pressing this lever was required both to escape a trial and start a new one, and to start a new trial when the last one had concluded normally (without an escape response). On both escaped and non-escaped trials the start lever gave access to the front levers (the options), and the front levers in turn produced reward. Because of this, the start lever is likely to have acquired value through secondary reinforcement (i.e. with a reinforcement chain going backwards in time: food > option levers >trial-start lever).

The second way the start lever could have acquired a positive association with reinforcement is trough generalization (Ghirlanda and Enquist, 2003). The *info* and *non-info* options in our experiment corresponded to levers on the front panel of the chamber. The manipulandum used to escape a trial was also a lever but located in the rear panel of the chamber. This physical similarity between manipulanda could have led to

generalization from the front levers to the start (rear) lever and therefore a transfer of value from the front levers to the start lever.

Because of the abovementioned associations between reinforcement and pressing the trial-start lever, rats had to inhibit any tendency to respond to this lever to obtain reward when the *100%* cue was presented. This inhibitory control might have been difficult due to the long duration of the cues (40 seconds). It is well known that withholding a response is often very difficult, even if it leads to a loss of benefit, as in DRO (Differential Reinforcement of Other Behaviour) schedules, in which responding systematically reduces payoff (Ainslie, 1974; Williams and Williams, 1969). In psychological terms, our task implementation required complex occasion setting discrimination (Schmajuk and Holland, 1998): the same manipulandum (the trial-start lever) should have been treated in different ways according to state of the cycle: after a reward outcome (reward or no reward) an immediate response was always required; after a choice for non-info, a response should never have been present; after a choice for info, a response was expected when the acoustic cue indicated a sure reward and not expected when the cue indicated a sure reward. With hindsight, this may have posed too complex a task.

Future experiments could try to diminish the potential influence of generalization and secondary reinforcement on escape responses to the *100%* cue. To eliminate generalization physically distinct kinds of manipulanda could be used for the escape response and for the options. For instance, the options (*info* and *non-info*) would remain as levers while the escape response would be pulling a chain. The effect of secondary reinforcement could be reduced by having separate "escape" and "trial-start" events (they are one event in the present task). For instance, by requiring one response to escape a trial (e.g. pressing a

lever), and another distinct response (pressing a different lever) to start a trial. These changes may be practically complicated but they could make the task clearer for the animals.

The abovementioned potential experimental modifications highlight the importance of understanding psychological mechanisms to be able to predict behaviour. These modifications would not alter the formal structure of the task. Therefore, the rate-maximising behaviour predicted by an optimal foraging model would remain the same. These modifications could however change which psychological mechanisms are recruited when engaging with the task and could therefore impact behaviour. The mismatch between predictions derived from an optimal foraging model and those derived from knowledge of psychological mechanisms involved in the task might suggest there are important differences between the structure of the experimental task and its equivalent ecological scenario. This is a difficult proposition to asses as we have little knowledge about the natural behaviour of these animals in situations analogous to our experimental task.

Despite the incomplete differentiation of escape responses, at a broader level the results of this experiment support the *Neglect-Bad-News* model, since the rats did use the escape response in the direction predicted by the model. This result fulfils a core assumption of the model, highlighting that a rate-maximizing approach together with a focus on psychological mechanisms can be a way forward to interpret apparently paradoxical observations.

Experiment 2

The purpose of Experiment 2 was to test the prediction that a longer *pre-choice* delay should diminish preference for an informative option. In our experiments *pre-choice* delay refers to the length of time between pressing the trial-start lever (the rear lever) and the options (front levers) becoming available. The rationale for this prediction is that if the *pre-choice* delay is sufficiently short, it is best to choose the informative option even if its probability of reward is lower, because the subject can escape trials with the *0%* cue, thus avoiding to pay for the post-choice waiting time cost. A sufficiently long *pre-choice* delay eliminates this advantage, because by the moment a cue is presented and the subject can decide whether to escape a trial or not, it has already paid a significant time cost. This relationship was formally derived trough an optimal foraging (rate maximizing) model which predicts preference for *info* in the 1s *pre-choice* delay group and for *non-info* in the 25s *pre-choice* delay group. Our experimental results did not match this prediction. Both groups were indifferent between *info* and *non-info.*

Although the prediction on the effect of the pre-choice delay did not hold, it does seem to be the case that comparing with the results of (Ojeda et al., 2018) the presence of an escape possibility increased preference for the informative alternative. In that work in a condition with the same probabilities of reward for the options as Experiment 2 (0.2 for info and 0.5 for non-info) the rats strongly preferred non-info, nonetheless some caveats apply as the two studies did not have identical protocols, in the present experiment the cue period had a duration of 40s whereas in (Ojeda et al., 2018) the cues lasted for 10s. Additionally the display of indifference might just reflect that subjects did not learn the

contingencies associated with the options, rather than any bias towards the informative option.

It is not possible to conclude with certainty why the experimental results failed to match some predictions derived from the rate maximizing model, but broadly speaking there are three possibilities. 1) the rats lack a mechanism for dealing with the scenario we envisaged (i.e. variations in a pre-choice delay). Alternatively, the rats do have a mechanism for dealing with this scenario, but 2) our model lacks psychological realism, or 3) our experimental task does not approximate the scenario well enough. Although the first possibility cannot be ruled out, it seems likely that second and third options contributed to some extent to the discrepancy between the experimental results and the predictions of the model. There are several assumptions of our model that seem likely to have shown significant departures in the experiment. Also, the implementation of our experimental task seems to have resulted in a differential level of engagement between groups, and that could have caused some artefactual results.

Our model incorporated several assumptions. The model estimated the reward rate for a hypothetical subject that chose exclusively *info* or *non-info*. If the subject chose *info* then we assumed it escaped all trials with the 0% cue and stayed in all trials with any of the other cues. We also assumed that latencies to respond would be small enough to be negligible, and that there were no forced trials. We now turn to discuss how some of these assumptions might have shown significant departures in the experiment.

The first departure is the incomplete differentiation of escape responses to the *info* cues. One assumption was that subjects would escape all of the trials with the *0%* cue and none

of the trials with the *100%* cue. The results in Figure 9 show that was not the case. Although the rats escaped substantially less trials with the *100%* cue than with the *0%* cue, some of the trials with the *100%* cue were escaped, and not all of the trials with the *0%* cue were escaped. In Experiment 1 the escape behaviour showed a similar pattern, and as we discussed previously, we do not know if this was because they did not discriminate the cues perfectly or due to other details of how the task was implemented. Regardless of the cause, this pattern of escape behaviour diminishes the reward rate associated with the *info* option relative to our model, and therefore would be predicted to reduce preference for this option. Therefore, this could potentially explain why subjects in the 1s group were indifferent between the options whereas our model predicted preference for *info*. An interesting aspect of the results in Figure 9 is that it in the 1s *pre-choice* delay group rats always escaped the 0% cue (as predicted), however they also sometimes escaped the 100% cue (not matching the prediction). By contrast in the 25s *pre-choice* delay group rats never escaped the 100% cue (as predicted), but they sometimes did not escape the 0% cue either (not matching the prediction). This suggest that the length of the *pre-choice* delay influenced the escape behaviour, making it overall less likely for the 25s *pre-choice* delay group. Differences in escape behaviour between groups were not predicted by our model and are likely to have influenced preference for the informative option. The *pre-choice* delay in our task is similar to the *travel time* of the Marginal Value Theorem (Charnov, 1976b). According to this theorem an increased *travel time* predicts a longer time before abandoning a patch, this could perhaps explain the reduced propensity to escape shown in our experiment for the group with a longer *pre-choice* delay.

The second departure is the latencies to respond. These latencies were not incorporated into our model and they can change the reward rate associated with each option and therefore the predicted preferences. There were three latencies in the task: 1) to escape a trial (by pressing the trial-start lever during the presentation of a cue) 2) to begin a new trial when the previous one concluded normally (by pressing the trial-start lever after the reward outcome) and 3) to select an option after starting a trial (by pressing a front lever after having pressing the trial-start lever). Regarding this last latency type, although we manipulated the duration of the *pre-choice* delay, this value only set the minimum delay between pressing the trial-start lever and selecting an option, but there was no upper boundary. By means of our optimal foraging model we determined the parameters of this experiment such that there were large differences in estimated reward rates between the options, and therefore latencies in the range that were physically realistic should have not changed the direction of preference, nonetheless it is possible that the rats exceeded our estimates of these latencies considerably. In this respect, a study by (Shapiro et al., 2008) with starlings in which the birds had to choose between combinations of reward size and delay, the predictive accuracy increased when these self-generated latencies were included in the models.

Exploring the possible effect of latencies and the incomplete differentiation of escape responses to the *info* cues could be done by estimating these values from the experimental data and adding these values as new parameters of the model. Subsequently, we could test if this provides a closer match between predicted and actual preferences. A better fit would not however explain the origin of these departures, e.g. why animals did not further differentiate their escape behaviour between the *0%* and *100%* cues or why they had latencies to respond well above what was physically possible. Because these parameter

values would not be programmed independently but generated from the subject's own behaviour, this strategy significantly reduces the explanatory power of the model.

The other possible explanation for the discrepancy between predicted and actual preferences is that our experimental task did not approximate well enough an ecological scenario. This is difficult to assess as our task is not based on any particular example of information-seeking behaviour in nature. Nonetheless, the behaviour of the rats in the group with a *pre-choice* delay of 25 seconds suggests that this may have been a serious issue. It seems like decreasing the contiguity between pressing the back lever (the trial-start manipulandum) and the front levers becoming available (the options) made the rats much less willing to engage with the task. Specifically, subjects in the 25s group did not engage as actively with the task as those in other conditions. They were slow to complete the trials relative to the 1s group, not as an obligate consequence of the additional 25 seconds imposed by the longer *pre-choice* delay but due to a slower working rate. Two of the eight rats did not complete all the programmed trials within the available time, and one rat stopped working altogether. Another aspect of the results that suggests lack of engagement are the unstable individual-level preferences, as seen in Figure 8. Although average results seem similar between the two groups with different pre-choice delay, the subject-level data display a different pattern. In the 1s group each rat (but one) developed an absolute preference for either of the options (choice trials were allocated to one option exclusively). By contrast in the 25s group preferences were partial (choice trials were allocated to both options) and unstable.

The lack of task-engagement in the 25s group might mean that we did not obtain a reliable estimate of the preferences of the subjects in this group. This could possibly be avoided

with a different implementation of the task. The motivation for Experiment 2 was to study the effect on preference of a delay between an outcome (reward, no-reward or an escape response) and the following opportunity to choose. The task implementation we selected to study the effect of this delay required that subjects in the 25s group produce an action (pressing the trial-start lever) that was temporally distant from the actions that directly produced reward (pressing a front lever). This delayed chaining of instrumental responses might have made the initial response (the trial-start response) unattractive and reduced overall engagement in the task. Such chaining of delayed actions was not an unavoidable feature of the task. An alternative implementation of the task could have been to introduce a delay between an outcome (reward, no-reward or an escape response) and the start of a trial. A delay of this sort would conventionally be known as an ITI. This could have been implemented by having trials start automatically and requiring a response only to escape from a trial. This might have made the rats more consistently engaged with the task while leaving the equations of our rate-maximising model unchanged.

To summarize, In Experiment 2 we tested whether a longer *pre-choice* delay reduced preference for an informative option. Our results did not corroborate this prediction. This could mean that rats do not have a mechanism for dealing with the scenario we envisaged in this experiment, but alternatively it could mean we did not incorporate enough parameters into our model or it could be due to details of how the experimental task was implemented. Additional experiments and analysis would be needed to clarify this ambiguity.

General discussion

We tested two predictions of the *Neglect-Bad-News* model of paradoxical choice. The first is that animals will selectively escape from a stimulus that predicts no-reward. The results from Experiment 1 match this prediction, although not as expected from an absolute rate-maximiser. We also tested whether a longer *pre-choice* delay diminishes preference for an informative option. The results from Experiment 2 did not corroborate this prediction. Overall, support obtained for the *Neglect-Bad-News* model was mixed, but in my view the prediction tested in Experiment 1 is the most important. The assumption that subjects should be able to selectively escape from the 0% cue is at the very core of the *Neglect-Bad-News* model, whereas the effect of the pre-choice delay is a secondary prediction that can be derived from extending the *Neglect-Bad-News* model but is not integral to it. The model could still help to understand the paradoxical excessive preference for information even if there was no sensitivity to *the pre-choice* delay. The positive results of this chapter strengthen the *Neglect-Bad-News* model and suggest that similar mechanisms underlie paradoxical choice in taxa as distant as birds (pigeons and starlings) and mammals (the rat). Additional experiments are needed to further explore the significance of results from Experiment 2.

Addendum

After completing the experiments in this chapter Fortes at al. (2017) published a work with research aims overlapping those of the present chapter. They also tested a series of predictions derived from the *Neglect-Bad-News* model but did so in pigeons instead of rats. Their work tested one prediction we also assessed. This is that animals should

selectively escape when presented with a cue that predicts sure no-reward. Their results showed that pigeons did indeed escape the 0% cue and stayed with the 100% cue. However, like in our results, the proportion of escape responses was not always zero or one, but sometimes took intermediate values. Fortes et al. (2017) also manipulated the length of the ITI in in one of their experiments. The ITI is analogous to the *pre-choice* delay in our experiment, in that it reflects a waiting time that is common to both options. Therefore, according to our modelling work, lengthening the ITI should diminish preference for an informative option. This change in preferences did not take place in the work of Fortes at al. (2017). However, we do not know if the parameter values used in their experiment are adequate for our model to predict a significant change in preferences. Overall, the fact that the main result of Experiment 1 was replicated in a different species highlights the utility of rate-based models as a tool to understand the mechanisms underlying paradoxical choice across taxa.

Chapter 3: Rats learn to work for useful information

Introduction

Probably the strongest established principle of learning by experience is reinforcement learning, a derivative of the Law of Effect first articulated by Thorndike, and later formalized by Skinner (Skinner, 1938; Thorndike, 1898) More recently, reinforcement learning has turned into a central tool in artificial intelligence research, as a simple process through which artificial agents can adjust to their world (Frankenhuis et al., 2018; Sutton and Barto, 1998) Basically, the idea is that given that the outcomes of actions can be assigned a positive or negative valence and that each outcome can be 'attributed' to a preceding event (the credit attribution problem), an agent can adapt to its environment by initially emitting actions at random, with a certain probability or frequency, adjusting this probability or frequency upwards for positively valence outcomes and vice versa for negative ones. There are some unsolved issues of circularity (technically, for an event to be called 'reinforcement' it must have a positive valance, and its valence is determined by its effect on the frequency of emission of actions, and not by objective biological consequences), but the principle of reinforcement is undoubtedly extremely useful and applicable to many animal learning cases. But not to all. Cases in which reinforcement seems insufficient to explain learning processes include those in which organisms need to follow complex sequences of actions to achieve some goal, so that intermediate steps in the chain do not have opportunity to change their frequency by causing objective improvements in the circumstances of the agent. And yet, animals do learn such sequences.

This simple observation applies sharply to some very widespread problems, such as how does experience combine with pre-programmed adaptive traits to control the frequency of

intermediate steps in chains of actions leading to a physical reward, and similarly, what drives behaviour when the biological value of an action in a chain towards physical reward is to generate information rather than a substantial benefit such as food? Is information reinforcing per se, so that information-seeking behaviours are maintained because they 'inform' (i.e. reduce the uncertainty of) the agent, or it is entirely driven by its effect in producing objectively beneficial consequences?

The last problem is ubiquitous: animals generate behaviours that provide them with information about their environment, information which often, but not always, can serve to guide subsequent behaviour, but the interaction between innate action patterns and behaviours shaped by their consequences can be complex. For instance, an animal might frequently sniff as it moves around, and use its olfactory stream to assess the proximity of food or danger, approaching or avoiding a location accordingly. Sniffing is thus information-seeking, but is only operational in concert with other, environmental, factors.

In cases such as this, the whole compound of information-seeking behaviour involves the presence of two associations. One of them is between an environmental variable that may not be intrinsically valenced (an odour) and a meaningful variable (a suitable prey, or predator). This is the 'information' that ends up being acquired. The other association is between an element of the behavioural repertoire and the information it produces in a given context (e.g. the association between sniffing and a specific odour). One or both of these associations may involve learning. In our example, sniffing as a form of searching is in itself a genetically programmed behaviour by which many animals, specially mammals, map their affordances, just as other mammals (i.e primates) use their eyes to visually map their surroundings. Animals are intrinsically motivated to do such things. In other cases,

however, information-seeking may not rely on pre-programmed traits, and this is the topic of this study. We ask whether the information-seeking behaviour can be instrumentally maintained, that is whether an animal can acquire or sustain a pattern of behaviour whose consequence is to produce information. In the long term we wish to separate information seeking from enhancement of objective benefits by the information acquired, namely whether uncertainty aversion is in itself reinforcing.

In the present experimental context, we operationalize the problem by separating information seeking from food seeking, where each is achieved through pressing different levers. Because of its arbitrary configuration, this problem involves learning both associations (neither is innate). We strived to keep the task as simple as possible while still being able to speak to the question of whether animals will learn to perform an action that provides no immediate objective benefit but useful information that may or may not cause a benefit after some delay.

The paradigm we developed can be summarized as follows: In each trial there are two options available in distinct locations (left vs right levers). Only one of these options leads to food as a reward if effected (is *active*). The other leads to a timeout (is *inactive*). Thus, both options have the same upfront cost, but only one of them leads to a measurable objective benefit. Which of the options is active or inactive is determined randomly for each trial. Because of this uncertainty, if the subject directly operates either option it is rewarded in half of the trials. The subject is given another possibility however, in the form of an *information-seeking* response. When performed, this third action produces an acoustic cue that signals the identity of the active lever in the current trial. The information-seeking response also has a cost in time, but this can be compensated by

making later responses more effective. Thus, the choice faced by the subject at the start of each trial is whether to act directly on one of the options, obtaining reward in only half of the trials, or alternatively to first perform the *information-seeking* response, obtain the cue signalling the location of the active option, operate it, and obtain reward in all trials. A case of 'pay now, benefit later'.

This protocol relates to impulsiveness paradigms, in which subjects choose between actions leading to a lesser reward delivered soon or an otherwise preferable reward delivered later (Stephens and Anderson, 2001; Stevens et al., 2005). A fundamental difference, however, is that in impulsiveness protocols both alternatives compete as actions that could generate reward, while here the goal is to offer choices between a manipulandum directly linked to food and an action that modifies the former's probability of food delivery. The protocol here also relates to the observing response (Dinsmoor, 1983) as both produce discriminative stimuli that do not directly alter the reinforcement schedule. The crucial difference is that the observing response is designed to avoid any potential benefit of acquiring information, whereas in the task described here the acquisition and use of information has an easily quantifiable benefit, namely an increase in the reward rate.

There are known learning mechanisms that could account for the acquisition of the two associations this task requires in order to be solved efficiently. The association between each acoustic cue and the location of the food-producing lever is analogous to what is known as an occasion setter in learning theory, i.e. a discriminative stimulus that signals to the subject when responding will be reinforced (it sets the occasion). Quoting Schmajuk and Holland (1998): "*Over sixty years ago, B. F. Skinner proposed that a discriminative*

stimulus in an operant conditioning paradigm does not elicit a response but simply "sets the occasion" for the response to occur. More recently, investigators have claimed that, whereas in many Pavlovian conditioning paradigms a conditioned stimulus (CS) elicits a conditioned response (CR) because it signals the occurrence of an unconditioned stimulus, in other paradigms a CS elicits a CR because it sets the occasion for the responding produced by another CS.". Notice that these authors were focusing in Pavlovian associations, but it is known that rats are capable of modulating an instrumental response on the basis of such a stimulus (Holland, 1991; Trask et al., 2017). Occasion setters are classified into positive and negative on the basis of whether the stimulus predicts reinforcement or non-reinforcement (Holland, 1989; Meyer and Bucci, 2017). In our task, each cue is simultaneously a positive occasion setter (for one option) and a negative occasion setter (for the other).

The other association (between the *information-seeking* action and the different acoustic cues) could be acquired by conditioned reinforcement, in which pairings of two stimuli cause an increase of responding to the first of them only because the other has earlier acquired 'reinforcing value' by being associated to a stimulus of biological value, such as food, water etc. (for a review see Williams, (1994)), In our task the acoustic information-carrying cue may have become reinforcing because it allows for more efficient food-producing behaviour. Each of the acoustic cues could be labelled as conditioned reinforcer, but this describes rather than explains the problem we pose. A more explanatory framework may be constructed by linking these mechanistic and descriptive discussions to experiments in which the quantitative determinants of when and how information-seeking can be reinforced.

Methods

Subjects

All experiments were carried out in compliance with the UK Animal (Scientific Procedures) Act (1986) and its associated guidelines. 16 male Lister Hooded rats (provider Envigo), 11 weeks old at the start of the experiment served as subjects. They were housed in groups of four. Throughout the experiment subjects were food deprived to 85-90% of their free-feeding weight [Initial weight: 298g ± 11. Final Weight: 390g ±27 (Mean ± std. dev.)]. Water was provided ad-libitum in their home cages and they were maintained in a 12-hour dark/light cycle with lights on starting at 6 AM.

Apparatus

Testing was carried out in eight operant chambers (Med Associates, USA.) Each chamber contained three retractable levers: one in the back panel (centre) and two in the front panel, left and right of a central food magazine delivery tray equipped with an infrared beam and a sensor to record head entry. Each reward delivery consisted of two 45mg sucrose pellets (TestDiet, USA). A speaker (3.5" in diameter) was positioned above the back lever in the back panel. Each chamber was also equipped with a house light and a fan. The chambers were controlled via custom-written Med-State Notation programs running on MED-PC V (Med Associates, USA).

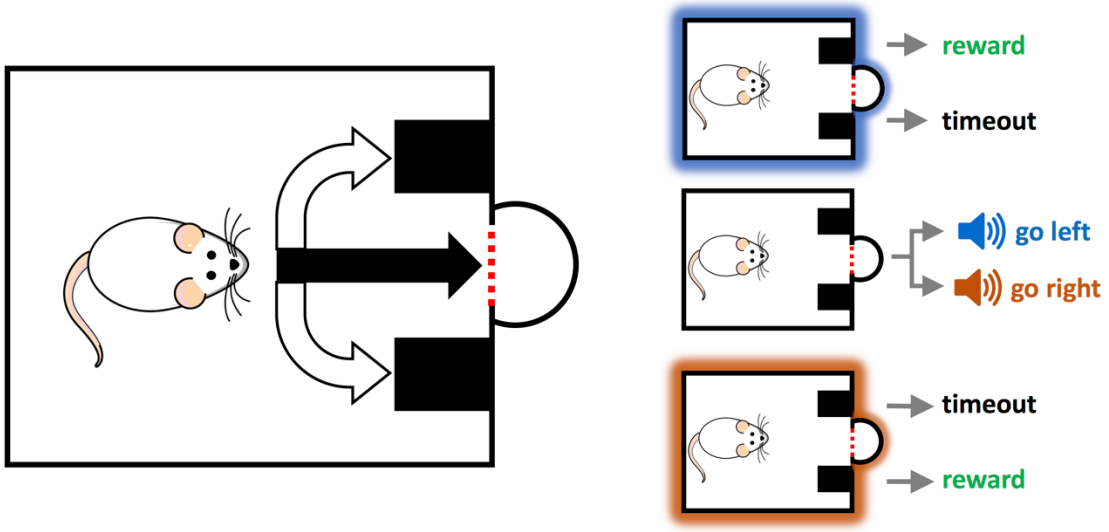


Figure 1. Probe trial structure
Probe trials assessed if rats would produce an action that provides no immediate reward, but information that can lead to better choices later on. Left panel: After starting a trial by pressing a back lever, subjects could press either of two front levers (empty arrows) or nose-poke into a central magazine (filled black arrow). Right panel, top and bottom: one of the front levers led to reward while the other led to a timeout, with locations randomized across trials. Right panel, middle: A nose-poke at the magazine (red dotted line) triggered an acoustic cue signalling the location of the active (rewarding) lever. Thus, the subjects could directly press either of the front levers and be rewarded in 50% of trials, or first nose-poke and then use the information provided to choose the active lever, thus being able to obtain reward in all trials.

Lever training Over the next three sessions the rats were trained to press the two front levers. The left lever was available on the first session, the right lever on the second and either lever (with equal probability across trials) on the third. Once a lever extended into the chamber, a single press resulted in its retraction and immediate reward delivery (fixed ratio 1 schedule). In the first two sessions, the lever became available again after 10 seconds. In the third session, this delay was composed of a constant element plus a variable one. The constant one was 5 seconds, and the variable one was sampled from a Poisson distribution with a mean of 5 seconds. All three sessions concluded after 60 reward deliveries.

Pilot experiment:

In an initial exploratory phase, the two option levers at the front of the cage were

programmed to deliver food after either 10s or 30s, with the two delays assigned randomly to each side between trials. Trials started when the subject pressed the back wall lever. After this, there were three different kinds of trial, as follows. In 'Forced-info' trials, nothing else happened until the subject poked its nose in the food magazine. Nose-poking triggered the onset of either of two acoustic cues, one when the short delay was assigned to the left and the other when it was assigned to the right lever. At the same time both front levers were offered. In such trials the subjects could anticipate which of the two levers would deliver the shorter delay. In 'Forced-side' trials, contingencies were the same except that only one front lever was made available, and the acoustic signal now anticipated whether pressing it would produce a delay of either 10s or 30s.Finally, in 'Forced-Ignorance' trials only one lever was offered, and no acoustic cue was provided even if the subject nose-poked. Subjects performed a total of 46 daily sessions in this phase, but their choice of lever in Forced-info trials showed no evidence of discrimination according to the acoustic signal (see Figure S1 in Supplementary Results). As a consequence, the task was modified to increase the incentive of the discrimination.

Main Experiment:

The protocol for the main experiment was modified with respect to the pilot in two important respects: the two outcomes were made more extreme, so that the better outcome became a 1s delay (once the lever was pressed, the first response after 1s resulted in the delivery of food) and the worse one no-reward, and a forced waiting time of 60s, after which a new trial could start. The FT1s lever is from now on called the 'active' one for that trial. Two acoustic cues (either a pure tone (78 dB, 3 kHz) or a buzzing sound (74

dB)) were correlated with the location of the FT1s. Assignment of sound cue to each side lever was counterbalanced across subjects. The trial structure was now as follows:

1)Probe trials: In probe trials, pressing the rear lever caused it to retract and activated the nose-poking sensor (as signaled by a light on top of the magazine) and at the same time caused both front levers to extend, priming them for responding. At that point, the choice faced by the subjects was between pressing either of the front levers, leading to an FT1s in 50% the trials and a time-out of 60s in the remaining 50% or 2) nose-poke into the central magazine to cause the onset of the acoustic location cue, and then use this information to choose the active lever. The latter would allow a good discriminator to obtain the FT1s reward in all trials (**Figure 1)**. The acoustic cue remained on until either front lever was pressed. To facilitate acquisition, after a front lever had been pressed a visual cue provided immediate feedback on whether the active or inactive lever had been selected (i.e. whether food was forthcoming or not in the present trial). This cue was the house light turning steadily on or blinking on/off (frequency=10hz, counterbalanced across subjects). The visual cue remained on until the selected lever had retracted.

2) Forced-info trials: The purpose of these trials was to ensure subjects were exposed to the contingency between nose-poking and the delivery of the acoustic signals. The only difference with probe trials was that upon starting a trial the front side levers were not extended, so that the subjects could only proceed by nose-poking and thus generating the discriminant acoustic signal. After nose poking the side levers were activated as in probe trials.

3) Forced-side trials. The purpose of these trials was to ensure subjects were exposed to all the contingencies associated with the acoustic signals. Pressing the trial start lever led to the immediate onset of one of the acoustic signals (without the need to nose-poke), and the extension of only one of the two front levers (with equal probability). Thus, subjects were not able to choose how to respond to the available information.

Groups

Subjects were assigned to either the Experimental or Control group (n=8 each). Both groups were exposed to the training and testing regime, except for one crucial difference in the protocol. The difference resided in the statistical properties of the acoustic cue produced by nose-poking into the central magazine. In the Experimental group, this cue was perfectly correlated with the location (left vs right) of the active lever for each trial. In the Control group no such correlation existed. The purpose of including this Control group was to ensure than any nose-poking action in the experimental group was due to the informative properties of the location cue. This is important because some basic level of magazine inspection can occur spontaneously or through Pavlovian association with the magazine. By focusing on the difference between groups this base rate can be accounted for and the instrumental component isolated.

Phases

The main experiment comprised two phases with different trial composition: The first phase included all three trial types (probe; Forced-info and Forced-side, one third of trials in a session each). Subjects performed 19 daily consecutive sessions of this phase. The behaviour at the end of this phase showed subjects performing the information-seeking response in only a small fraction of probe trials, and this was interpreted as being

potentially attributable to the presence of Forced-side trials, in which animals could not use the signals. Therefore, a second phase included only probe and Forced-info trials (one half of trials in a session each). Subjects performed 12 consecutive daily sessions of this phase.

Session structure

Each session consisted of 96 trials. The number of trials of each trial type was evenly split within each session (one third each in the first phase, and one half each in the second phase) but their order within a session was determined randomly. Trials were separated by an ITI (inter-trial interval) composed of a constant element plus a variable one. The constant one was 3 seconds, and the variable one was sampled from a Poisson distribution with a mean of 3 seconds.

Data analysis

All data processing and statistics was done in MATLAB 2017a. Statistical analyses were performed on the last session of each phase. Analysis of accuracy (in choosing the rewarding side after hearing the location cue) was performed on Forced-info trials only because including probe trials would have resulted in unequal sample sizes for each subject.

Results

The main motivation of the present experiment was determining if rats would acquire and then sustain an action whose consequence was providing information that could alter adaptively the consequences of a later action. This requires to seek evidence of two separable components of behaviour, namely their ability to use the information, and their willingness to acquire it. The structure of our protocol ensured that they experienced the contingencies, by the use of 'Forced-info' trials in which they were forced to emit the information seeking response before choosing a final link lever, and probe trials in which they could choose whether to emit the information seeking behaviour or not.

Phase 1 (with Forced-side trials):

The ability to discriminate and use the information was examined in Forced-info trials. Rats in the experimental group used the location cue effectively (**Figure 2**, panel A). Accuracy was high and significantly above the chance level. (Wilcoxon signed-rank against 0.5, n=8, W=36, $P < 0.01$). As is to be expected, the uncorrelated control did not depart from the chance expectation.

The second question, namely whether they would work for this information given that they are competent to use it, is examined by means of probe trials. In these trials, the subjects opted to seek information, although not in all trials (**Figure 2,** Panel B.) The proportion of probe trials in which rats in the experimental group performed the cue-producing action was significantly different from zero for the experimental group (Wilcoxon signed-rank,

n=8, phase 1: W=28, $P<0.05$) but not for the control group (Wilcoxon signed-rank, n=8, W=3, $P=0.5$ two-tailed), and the proportion of probe trials in which information was sought was higher for the experimental group than for the control group (Wilcoxon rank-sum, W=47.5, $n_1=n_2=8$, $P < 0.05$ two-tailed**)**.This indicates that the use of the nose-poking, information-seeking response, was conditional on its informational contingency.

Although experimental subjects did seek information, they did not do so on all possible opportunities (in probe trials). We wondered if this could be due the influence of forced-side trials, in which subjects experienced the acoustic stimulus predicting the side to be reinforced but were not always to choose freely where to respond (only one of the levers was available, which could be either the correct or the wrong one given the acoustic cue). Forced-side trials diluted the benefit associated with the location cue (because on half of these trials subjects were forced to respond on the wrong lever), and this could be diminishing the propensity to perform the information-seeking response. To explore this possibility, Forced-side trials were eliminated in Phase 2.

Phase 2 (without Forced-side trials):

Figure 2B shows that the elimination of Forced-side trials produced a sudden elevation of the information-seeking response, consistently with the hypothesis that these forced trials acted through curtailing the rats' ability of using such information. For the experimental group the proportion of probe trials in which information was sought was higher in Phase 2 than in Phase 1 (Wilcoxon signed-rank, W=3, n=8, $P< 0.05$ two-tailed). Otherwise, the pattern of results in the second phase follows that of the first: For the experimental group the proportion of probe trials in which information was sought was higher for the

experimental group than for the control group (Wilcoxon rank-sum, W=41.5, $n_1=n_2=8$, $P < 0.01$ two-tailed), for which this proportion was not significantly different from zero (Wilcoxon signed-rank, n=8, W=1, $P=1$ two-tailed).

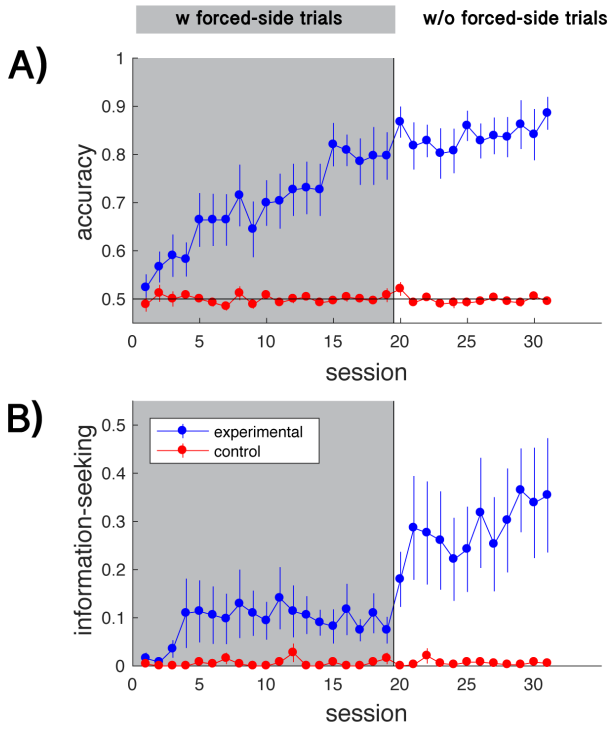


Figure 2. Information-seeking
A) Proportion of Forced-info trials in which rats correctly chose the rewarding side after hearing the location cue (Mean ± S.E.M, n=8). **B)** Proportion of probe trials in which rats performed the cue-producing action (Mean ± S.E.M, n=8). In the experimental group nose-poking into the central magazine triggered a sound that signalled the rewarding lever in the current trial, whereas for the control group the sound triggered by the same action was uncorrelated with the location of the rewarded lever.

Discussion

The main objective of the present chapter is to determine if rats would acquire and sustain an action that produces no immediate physical reward but provides information that is potentially useful to improve later choices. To test this question, it was necessary to ensure subjects learned to use the information effectively, regardless of how they came to it. This condition was satisfied: When forced to obtain the location cues (in Forced-info trials) subjects did choose the more profitable option on each trial (the lever that lead to reward). The second and main question of the chapter was whether, given that their choices did improve when they came by the information, they would emit the behaviour necessary to generate that information. Our results showed that subjects in the experimental group chose to perform the information-seeking action, while controls with non-informative cues did not. These results demonstrate that rats can learn to acquire responses that do not directly produce reward but provide information that can be used to act more efficiently on the environment.

Qualitatively, the results match the predictions of reward rate-maximization (Stephens and Krebs, 1986): Rats performed the information-seeking action when it allowed an increased reward rate compared to acting under uncertainty, as was the case in the experimental group. By contrast, in the control group, in which the cues could not be used to increase the reward rate (because they were uncorrelated with the location of the rewarding lever), this action was absent. The absence of the cue-producing action in the control group rules out alternative psychological causality for the nose-poking action, such as autoshaping or a non-specific exploratory drive.

Nonetheless, there is an important quantitative deviation from rate-maximization: subjects did not use all the available opportunities to seek information. Although the proportion of probe trials in which the information-seeking response was performed is higher for the experimental than for the control group, there was a substantial fraction of opportunities in which information was not sought (although the proportion seemed to be still improving at the end of the experiment). This represents a departure from rate-maximizing: Given the level of correct choices upon possessing the information, it would have been optimal (rate maximizing) to always seek information. Each trial in which information was not sought reduced the probability of obtaining reward from the level of accuracy achieved by the animals by about 35% (from around 85% at the end of training to the chance level of 50%). This causes an additional time cost, as each choice of the non-rewarding lever resulted in a 60s time-out. Because of these two factors, not seeking information reduced the reward-rate substantially, not just respect to what was possible according to the programmed contingencies, but according to the demonstrated discrimination capabilities of the animals.

This can be formalized as follows:

Say that the proportion of correct responses achieved by the animal after nose-poking is Pi, while responding without nose-poking leads to an even chance of reward, but implies a time cost of n seconds. Say, further, that the time cost of a correct response is c seconds while that of a wrong response is w seconds, and there is a common ITI.

The rate of reward obtained by nose-poking in every trial is then:

$$R_{info} = \frac{P_i}{ITI+n+P_i c+(1-P_i)w} = \frac{1}{\frac{ITI+n+w}{P_i}+c-w} \qquad \text{Eq 1}$$

while that obtained by responding directly at one of the front levers is:

$$R_{non-info} = \frac{1/2}{ITI+\frac{1}{2}(c+w)} = \frac{1}{\frac{ITI}{2}+c+w} \qquad \text{Eq 2}$$

The resulting rates of reward for the temporal parameters of our experiment (c=1s, w=60s, n<5s, ITI=6s) are shown in **Figure 3**

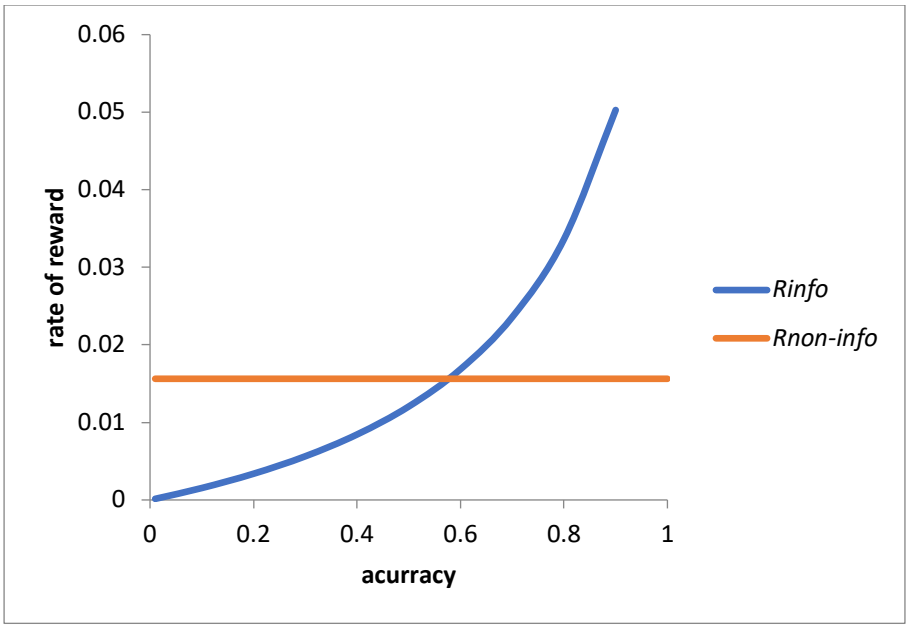


Figure 3. Reward rates

The figure shows that animals that use the information (to select the correct front lever) with an accuracy in the neighbourhood of 80%, as ours did, would benefit very substantially from seeking-information (nose-poking) in all trials. The crossing over point can be found using the rate equations, since $R_{info} > R_{noninfo}$ when

$$\frac{1}{\frac{ITI+n+w}{P_i}+c-w} > \frac{1}{\frac{ITI}{2}+c+w} \qquad \text{Eq 3}$$

Or

$$\frac{ITI}{2} + c + w > \frac{ITI+n+w}{P_i} + c - w \qquad \text{Eq 4}$$

Cancelling c and solving for the critical Pi, we get that poke nosing is beneficial when

$$P_i > \frac{ITI+n+w}{\frac{ITI}{2}+2w} \qquad \text{Eq 5}$$

In our experiment there was no programmed cost n, but if we assume a cost of, say 5s, given that w=60s and the ITI=6s, it would pay subjects to nose-poke for all values of Pi greater than 71/123=0.57.

The model makes it clear that rate maximizing rats with the discriminatory capabilities we observed would have made the information response in every trial, but they did not, raising an interesting problem for a foraging perspective. This behavior is an example of partial preferences, a well know issue in the optimal foraging literature (Krebs et al., 1977; Recer and Caraco, 1989; Stephens, 1985). One factor that we could empirically identify as reducing the proportion of trials in which the rats worked for information was possible artefactual, namely the presence of trials in which the rats did get the information but were forced to press a lever chosen randomly by us, as happened in Phase 1. The purpose of these trials was to ensure subjects experienced the information provided by the location cue, but it seems, with hindsight, to have been a design weakness, which we corrected in Phase 2 of the experiment. In this second phase, as soon as the force-side trials were eliminated, the rats increased dramatically the intensity of information-seeking responses, as shown in figure 2B. However, this does not seem to be the whole story, since nose-

poking trials were still below 50% at the end of the experiment, notwithstanding an apparent increasing trend.

Another candidate cause for the reduced information-seeking resides in the training protocol. The animals used in the main experiment were not naïve when the experiment started. They had been previously exposed to a different version of the task (during the pilot experiment). In this version the contrast between outcomes was smaller. Instead of one lever leading to almost immediate reward and the other leading to no-reward and a time-out (as it was later, in the main experiment), in the pilot experiment both levers led to reward, but following delays of different duration (10s vs 30s). With these parameters the rats failed to use the location cue even when the information was given without the nose-poking response. They chose the same lever irrespective of the cue presented. It is possible that the subjects might have acquired habit-controlled responses that were not completely extinguished in the main experiment. This possibility can be directly tested by repeating the experiment with totally naïve rats.

The level of information-seeking shows substantial variation between subjects. Some subjects never performed the information-seeking response while others did so on the majority of opportunities. (see Figure S2 in Supplementary Results). This prompted us to investigate if there was a relationship between accuracy (selecting the lever on the correct side given the cue) and information-seeking. That is, if subjects with low accuracy also had low-levels of information-seeking. This was not the case, there is little variation in accuracy and this variation is not correlated with the level of information-seeking (see Figure S3 in Supplementary Results**).** This finding is congruent with the fact that even though accuracy increased gradually during the first phase of the experiment, the level of

information-seeking remained constant. This is puzzling as it would seem reasonable that these two would have risen together, because the higher the accuracy, the larger the benefit of information-seeking.

Another possible cause for sub-maximal information-seeking is that certain aspects of the task do not closely match ecological circumstances. For instance, the location of the rewarding lever was determined stochastically on every trial. This degree of extreme "volatility" of the environment is not common in nature, where states are highly autocorrelated. Therefore, animals might have evolved a prior belief of environmental stability that caused them not to seek information on every occasion.

The information-seeking action in our task has some parallels to the so-called "observing-response|" (Kelleher et al., 1962; Wyckoff Jr., 1952), in the sense that both produce discriminative stimuli but do not directly alter the reinforcement schedule. However, in the "observing-response" paradigm only one potentially reward-producing action exists. The discriminate stimulus signals if this action will produce reinforcement or not (or in some versions of the paradigm the delay to reinforcement (Bower et al., 1966)), but crucially no alternative reward-producing action is available. This setup is designed to cancel any benefit of acquiring information, because the only potential advantage of being informed is the physical effort saved by not performing a response that will not be reinforced, or perhaps to prepare to exploit the food more efficiently, for instance by salivating. The total number of rewards and the time taken to acquire them (aka. the reward rate) cannot be increased by acting with information in this paradigm. By contrast in our task, because there are two potentially reward-producing actions available, information has an easily

quantifiable benefit, in effect an increase in reward rate resulting from making a better choice as described in our model.

A recent work has proposed a general mathematical treatment of choice behaviour (Friston, 2010; Friston et al., 2015). As part of this treatment they define a term in their equations labelled "epistemic value" that is directly relevant to the results of the present chapter. *"Epistemic value is the expected information gain under predicted outcomes. In other words, it reports the reduction in uncertainty about hidden states afforded by observations. […] Heuristically, this means valuable policies will search out observations, cues or "signs" that resolve uncertainty about the state of the world"*. "Epistemic value" could thus potentially explain how rats acquired the information-seeking response in our task, and why they only performed it on a fraction of trials. Although a formal mathematical model such as this one seems potentially very valuable, this work is entirely theoretical and was restricted to simulating behavioural and neural data, therefore its ability to predict real world behaviour is unknown. Interestingly, the task they simulated is at a formal level very similar to our task, therefore the results of the present chapter are potentially a useful dataset against which to test the predictions of this theory.

In summary, our results show that rats are willing to incur a cost to gain information that can subsequently be used to act more efficiently in the environment. However, a foraging analysis shows that rate-maximizing behaviour required a higher level of information-seeking than shown by the animals. Our model allows for an analysis of these quantitative discrepancies and helps to identify the factors that should influence the relative benefit of each pattern of responding. Some recent behavioural and brain research (Bennett et al., 2016; Blanchard et al., 2015; Bromberg-Martin and Hikosaka, 2009, 2011; Iigaya et al.,

2016) argue that animals may assign value to information per-se (namely a reduction of uncertainty independently of objective consequences). This is not contradicted by our results, but, in contrast with the present approach, most of those studies do not give animals the opportunity of combining uncertainty reduction with adaptive foraging strategies. These two approaches should ideally converge in the future.

Supplementary Results

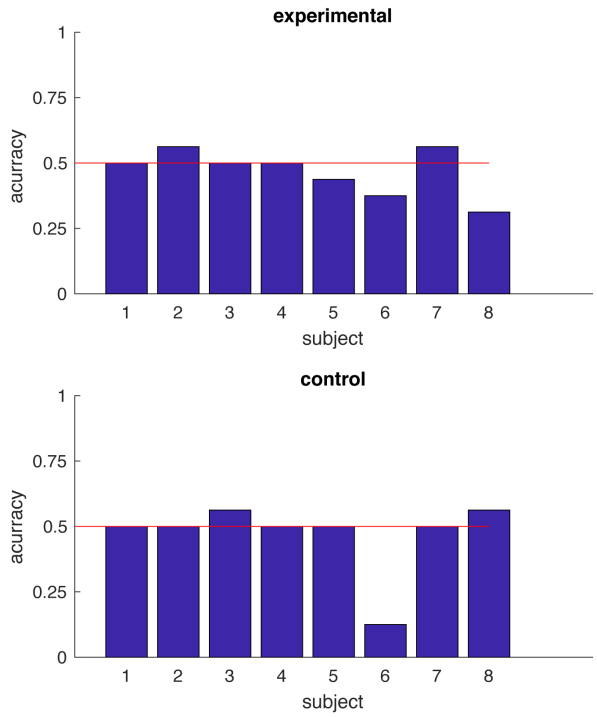


Figure S1. Pilot experiment
Proportion of Forced-info trials in which rats correctly chose the rewarding side after hearing the location cue.

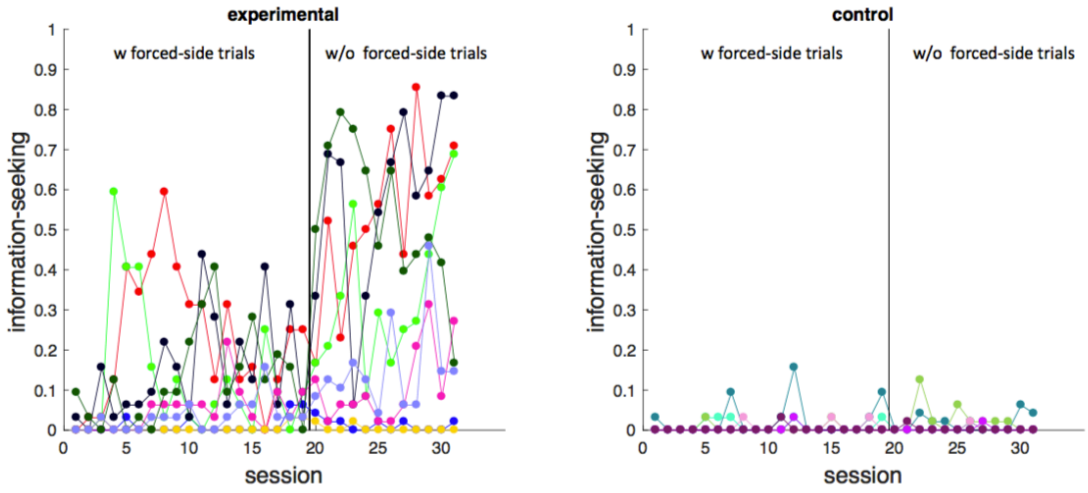


Figure S2. Information-seeking (individual results)
Proportion of probe trials in which subjects opted to perform the cue-producing action for the experimental and control groups. Each line represents a different subject (n=8).

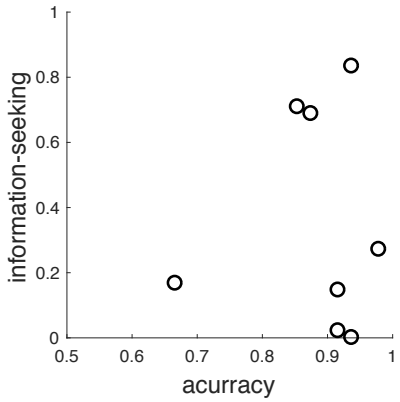


Figure S3. Relationship between accuracy and information-seeking.
Proportion of probe trials in which subjects in the experimental group performed the cue-producing action as function of accuracy in selecting the rewarding lever in forced-info trials. (Spearman's rank correlation, rho=-0.12, *P*= 0.78).

Chapter 4: Preference for unusable information is sensitive to dimension

Introduction

Classical learning theory postulates that animals learn by changing the frequency of actions that result in reinforcement. From a psychological perspective reinforcement has been defined in various operational manners, e.g. 1) The law of effect: If a given event increases the frequency of a behaviour, then it constitutes reinforcement (Skinner, 1935), 2) Premack's principle: Behaviour A can serve as reinforcement for behaviour B only if A is more likely to be freely performed (Premack, 1959), and 3) The response deprivation hypothesis : Behaviour A can serve as reinforcement for behaviour B if performing A can restore B to its free baseline level (Timberlake and Allison, 1974). From a biological perspective, reinforcing events are assumed to be those that in the species' history typically resulted in fitness gains (although fitness itself has some definitional circularity). It is thus readily understandable why items like food, water or mating have the capacity to energize the behaviour of an animal and influence learning. However, animals often perform actions that do not seem to provide an identifiable substantial benefit and yet animals will choose over food, such as the opportunity to play or socialize (Vásquez and Kacelnik, 2000). One possible interpretation of this observation is that animals might find that acquiring information in itself constitutes reinforcement (for an earlier debate on these matters see: Inglis, 1987; Inglis and Shepherd, 1994; Kacelnik, 1987).

One way to investigate the extent to which obtaining information is reinforcing is to offer experimental subjects the option of acquiring information by relinquishing the opportunity to obtain food, similarly to cases in which they faced tradeoffs between defending their territory or foraging (Kacelnik et al., 1981) or forage less profitably but closer to conspecifics (Vásquez and Kacelnik, 2000). A number of empirical studies have suggested

that human and non-human animals find acquiring information reinforcing (Bennett et al., 2016; Blanchard et al., 2015; Bromberg-Martin and Hikosaka, 2009, 2011; Iigaya et al., 2016). Crucially, in these studies the information provided cannot be used to modify subsequent behaviour in the task, thus in theory ensuring that it is information per-se that is valued. It is likely that in all these protocols a psychological mechanism evolved under circumstances when information-gathering is beneficial, and thus, while the studies serve to understand the psychology of the animals, their adaptive implications are often harder to interpret. To explore this issue in some detail, we now turn to describing one such experimental paradigm and its possible interpretations, in detail.

A number of studies have found experimental conditions under which animals, confronted between options that differ in both the degree of uncertainty and the probability of delivering reward, prefer a lower-reward but more informative option. (Ojeda et al., 2018; Smith et al., 2016; Stagner et al., 2011; Stagner and Zentall, 2010; Vasconcelos et al., 2015; Zentall, 2011). In these studies, hereafter referred to as 'paradoxical choice' (for review see (McDevitt et al., 2016; Zentall, 2016)), animals choose between two alternatives that deliver reward probabilistically a fixed delay after being chosen. During the delay that follows choice, one of the options (the informative or '*info'* option) produces cues predicting whether a reward will be delivered or not, while in the other alternative (the non-informative or '*non-info'* option) the reward outcome is uncertain until the delay has elapsed. The crucial result emerging from this paradigm is that all species tested so far: pigeons (Smith et al., 2016), starlings (Vasconcelos et al., 2015) and rats (Ojeda et al., 2018) have a strong preference for the informative option, and that this preference can lead to significant (>80%) losses in potential food rewards.

Although the existence of paradoxical choice has been firmly established, its theoretical interpretation remains unclear. A key element to understand this phenomenon appears to be the cues that signal sure no-reward. When a cue's contingency is with a null outcome, subjects appear to ignore or downweigh this signal and its associated cost of waiting for no-reward. Several findings support this interpretation. For example, the cue predicting no-reward does not acquire persistent conditioned inhibitory effects (Laude et al., 2014). Also, the relative frequency of presentation of this cue (i.e. the probability of no-reward for *info*) has little or no effect on preference for the informative option (Smith et al., 2016; Vasconcelos et al., 2015). Finally, preference for *info* is not affected by inserting a gap between choice and onset of this cue (McDevitt et al., 1997).

A theoretical model that aims to explain paradoxical choice is the "*Neglect-Bad-News*" (Vasconcelos et al., 2015). This model relates the observations to a foraging framework, and is congruent with the evidence suggesting subjects ignore the cue for sure no-reward. The model is inspired by optimal foraging theory (Stephens and Krebs, 1986) and proposes that the phenomenon of paradoxical choice can be understood by paying attention to a crucial difference between the structure of the experimental task and the structure of the natural environments to which animals are adapted. In natural environments, there are multiple potential opportunities available, thus at a given moment an animal can abort the pursuit of a prey and redirect its behaviour towards seeking another one. By contrast, in the paradoxical choice experiments, subjects' freedom is limited: trials are sequentially presented and of fixed duration. This discrepancy between the natural and experimental settings could have important consequences for interpreting behaviour in the laboratory, as animals are likely to behave as if they were in a natural environment. Specifically, they might act as if they can use the information about the

presence or absence of reward to guide subsequent behaviour. In nature, an animal that perceived a signal specifying that no reward is forthcoming would abort the current pursuit of prey and redirect its behaviour to alternative opportunities in the environment. Thus, the animal would not pay the time cost of waiting for a null outcome, something which animals in the paradoxical choice protocol are forced to do. In summary, the *Neglect-Bad-News* model postulates that animals in the experiments behave as if they could escape from the sure no-reward cue. Formalized mathematically within a reward-rate maximization framework, this is equivalent to stating that animals edit out any time waiting for sure no-reward when assessing the profitability of the informative option. This editing-out of time costs explains why subjects prefer an informative option when its probability of reward is the same, or even lower than that of a non-informative option. The *Neglect-Bad-News* model accounts for several key findings of the paradoxical choice experiments and multiple novel predictions derived from the model have been empirically confirmed (Fortes et al., 2016, 2017, 2018).

The alternative interpretation of paradoxical choice is that the phenomenon reflects an intrinsic preference for information acquisition (namely uncertainty reduction). According to this interpretation, information is a primary reinforcer in the learning process (Bennett et al., 2016; Blanchard et al., 2015; Iigaya et al., 2016). In other words, information influences learning similarly to primary outcomes known to influence fitness, such as water, food and mating. As these different types of rewards can all be appetitive, they can be traded off against each other. Thus, this interpretation of paradoxical choice can explain why subjects prefer an informative option even if the probability of obtaining food is the same as, or lower than for a non-informative option (a closely related idea is that of 'information-primacy' or 'contra-freeloading' (Bean et al., 1999; Inglis et al., 2001;

Talling et al., 2002). As the motivation to obtain the information is "cognitive or endogenous", it has been labelled as an *intrinsic*, as opposed to an *extrinsic,* preference for information (Gottlieb et al., 2013). In the latter case the information would only be valued as a means to increase the amount of truly primary rewards. It has been speculated that an intrinsic preference for information could underlie higher order cognitive processes such as curiosity and self-directed learning (Gottlieb et al., 2013; Kidd and Hayden, 2015). Nonetheless, there is little theoretical work specifying how such a mechanism would be implemented (but see Baldassarre et al., 2014; Baldassarre and Mirolli, 2013; Barto, 2004) and none that addresses a scenario like that of the paradoxical choice experiments.

Both the *Neglect-Bad-News* model and the intrinsic preference for information hypotheses are consistent with the phenomenon of paradoxical choice. Nonetheless, these hypotheses are not equivalent. According to the *Neglect-Bad-News* model, paradoxical choice is the product of a mechanism that ignores the time costs associated with waiting for null rewards, because in nature this is a cost that does not have to be paid. Thus, the apparent intrinsic preference for information is an artefact resulting from the discrepancy between natural environments and the artificial conditions of the laboratory. In other words, according to this interpretation the animals do not value information per-se but are following a drive to maximise food rewards. Such a mechanism is effective in natural environments, but neutral or detrimental within the context of the experimental protocol of paradoxical choice. Additionally, according to the *Neglect-Bad-News* model, preference for an informative option exists only in a very specific circumstance, namely when the information delivered includes as one of its possible outcomes the absence of reward. By contrast, the alternative hypothesis attributes an intrinsic motivational value to

information, with no a-priori constraints on the kind of outcomes to which information relates.

The purpose of the present chapter is to disambiguate between the *Neglect-Bad-News* and the intrinsic preference for information hypotheses in the context of the paradoxical choice protocol. We aim to distinguish between these two competing interpretations by varying food reward outcomes across several dimensions. The basic structure of the present experimental task follows that of previous implementations of the paradoxical choice protocol: it presents subjects with two options with equal food reward contingencies and different post-choice signalling properties (*info* vs *non-info*). However, here we manipulate between groups the dimensions across which reward outcomes vary between trials. In the "*Probability*" group, we vary the outcome dimension that has been manipulated in previous studies: trials differ in the presence or absence of reward. In two other groups we vary dimensions of reward that have previously not been manipulated: in the "*Magnitude*" group possible outcomes are either one or three food pellets in different trials, while in the "*Delay*" group the interval between choosing an option and receiving food varies between trials, being either five or fifteen seconds.

The two hypotheses we have previously discussed make different predictions for these three experimental groups. If the *Neglect-Bad-News* interpretation is correct, then preference for an informative option should exist exclusively in the probability group: only in this group is the absence of reward one of the possible outcomes, and hence it is to be expected that in the wild a forager would abandon the chase and not pay the waiting cost. Subjects in the other groups (magnitude and delay*)* would complete the chase and pay the time cost, so would not benefit by preferring the informative option. Alternatively, if

intrinsic preference for information drives the paradoxical choice phenomenon, then preference for the informative option should exist in all three groups, because in all of them the cues carry information about reward outcomes. Thus, this experiment will allow us to elucidate the mechanism underlying paradoxical choice and improve our understanding of information-seeking behaviour.

Methods

Subjects

All experiments were carried out in compliance with the UK Animal (Scientific Procedures) Act (1986) and its associated guidelines. 24 male Lister Hooded rats (provider Envigo), 9 weeks old at the start of the experiment served as subjects. They were housed in groups of four. Throughout the experiment they were food deprived to 85-90% of their free-feeding weight using growth curves from the provider. Initial weight: 275 ± 8, final weight: 341±15 (Mean ± std. dev.) Water was provided ad-libitum in their home cages and they were maintained in a 12-hour dark/light cycle with lights on starting at 6 AM.

Apparatus

Testing was carried out in eight operant chambers (Med Associates, USA.) Each chamber contained three retractable levers: one in the back panel (centre) and two in the front panel, left and right of a central food magazine delivery tray equipped with an infrared beam and a sensor to record head entry. Each reward delivery consisted of two 45mg sucrose pellets (TestDiet, USA). A speaker (3.5" in diameter) was positioned above the back lever in the back panel. Each chamber was also equipped with a house light and a fan, which were switched on for the duration of the session. The chambers were controlled via custom-written Med-State Notation programs running on MED-PC V (Med Associates, USA).

General procedure.

The experiment was based on a trial-based chain procedure (**Figure 1,** panel **A**). On *choice* trials responding on an attention manipulandum (the initial link in the response chain) led to a choice between two other manipulanda (the options). Responding to either of these

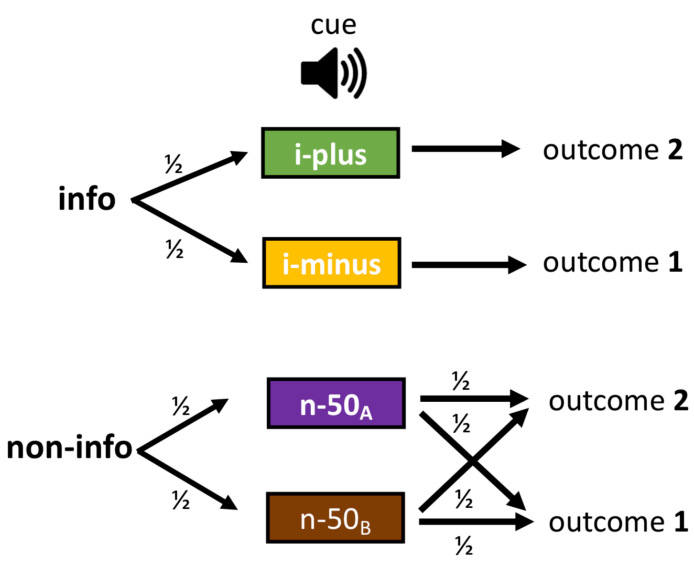
options caused the onset of an acoustic cue that persisted until a reward outcome occurred. Choosing the "*info*" (informative) option resulted with equal probability (0.5) in either the cue *i-plus* (associated with outcome one) or cue *i-minus* (associated with outcome two). Responding to the "*non-info*" (non-informative) option resulted with equal probability (0.5) in either cue $n\text{-}50_a$ or cue $n\text{-}50_b$ (both associated with 0.5 probability of either outcome). Thus, *info* and *non-info* had the same outcome probabilities (0.5 for both) but differed in their signalling properties. On *forced* trials only one of the options was made available after pressing the attention lever.

Outcomes

Between groups we varied the dimension across which outcomes differed between trials (**Figure 1**, panel **B**). In group "*probability*" we manipulated whether on a given trial the chosen option would result or not in food (zero vs. four pellets). In group "*magnitude*" all trials led to reward, but we manipulated the amount of food (trials could end in one or three pellets). Finally, in group "*Delay*" the duration of the choice-outcome delay varied between trials (food was delivered after either 5s or 15s). In groups "*probability*" and "*magnitude*" the duration of the choice-outcome delay was held constant at 10s, while in "*delay*" it was the amount of reward that was held constant (at two pellets). Thus, all three groups and both options had the same mean post-choice food over time profitability (0.2 pellet/second).

For clarity, we describe first the cues and then proceed to detail the training procedures and the final protocol.

A)



B)

	Probability	Magnitude	Delay
outcome 1	no-reward (0 pellets)	small (1 pellet)	long (15s)
outcome 2	reward (4 pellets)	big (3 pellets)	short (5s)

Figure 1. Information task

2.4 Cues

The following description applies to all three conditions. In each trial, a choice of an option was followed by one of four cues, each one consisting of an acoustic stimulus associated to a food reward contingency. There were two cues for the informative option: *i-plus* (for outcome one) and *i-minus* (for outcome two), and two cues for the non-informative option (*n-50$_a$* and *n-50$_b$*, both with 50% probability of either outcome one or two). The four sounds were: a low frequency pure tone (3 kHz, 78 dB), a high frequency pure tone (6 kHz, 78 dB), a buzzing sound (78 dB) and a clicking sound (74 dB). Assignment of sounds to outcomes was counterbalanced across subjects to avoid the

possibility of option preferences being generated by any intrinsic aversive or attractive properties of the stimuli.

Pre-training

Magazine training

To train the rats to collect rewards in the apparatus, training began with a single session with 60 reward deliveries. The interval between reward deliveries was sampled from a Poisson distribution with a mean of 60 seconds.

Lever training

Over the next three sessions the rats were trained to press the two front levers. Either lever (left or right with equal probability) was available on each trial. Once a lever extended into the chamber, a single press resulted in its retraction and immediate reward delivery (fixed ratio 1 schedule). The lever became available again after a delay composed of a constant element plus a variable one. The constant one was 5 seconds, and the variable one was sampled from a Poisson distribution with a mean of 20 seconds. All three sessions concluded after 60 reward deliveries.

Cue training

To ensure that in the main experiment subjects were tuned to the reward contingencies of the four cues (*i-plus, i-minus*, n-50$_a$, and n*50*$_b$) we preceded the choice phase with a Pavlovian protocol in which the rats were exposed to pairings of the four cues with their associated reward contingencies. In this phase no action was required to produce the cues. These cue-training sessions consisted of 40 trials, with 10 trials for each cue. To avoid

large deviations from the expected outcome probabilities of each cue, proportions of outcomes were fixed for every 10 trials with each cue. These trials were then randomly intermixed. Thus, if for example a 50% probability of outcome one was associated with a cue, exactly 5 out of every 10 trials with that cue resulted in outcome one, but the order was randomised. The ITI (inter-trial interval) was composed by sampling a Poisson distribution with a mean of 50 seconds and adding 10 seconds. Subjects performed one daily session of this phase for 20 days.

Main experiments

Groups

A between-subjects design was used, with three groups (probability, magnitude and delay) of eight rats each. Groups differed in the possible reward outcomes as previously detailed. Subject assignment to group was counterbalanced such that there was no correlation between group and any of the following parameters: side of the informative option; hour of testing; cage in which the animals were housed, or sound-reward contingencies. For each group, the subjects performed one daily session for 14 consecutive days. Each rat was trained at the same time of day; one cohort of rats began the experiment at 8:00 A.M. while the other started at 11:30 A.M. A session finished after 60 trials or 3 hours, whichever occurred first.

Trial structure

There were two kinds of trials: *choice* and *forced.* All trials started with the rear (attention) lever extending. Pressing this lever resulted in its retraction, and either one (forced trials) or both (choice trials) of the front levers to be presented. The purpose of requiring a back-

lever press to start a trial was to ensure task engagement, to position the rat centrally in the chamber with respect to the options, and to enable measurement of latency to respond starting from a central position. Pressing a front lever initiated a cue (the terminal link) and also caused that lever (forced trials) or both levers (choice trials) to retract. The auditory cue was presented for 10 seconds in groups *probability* and *magnitude,* and equiprobably for either 5 or 15 seconds in group *delay*. After the cue had been presented, reward delivery could occur (zero or four pellets in group *probability*, one or three pellets in group *magnitude*, and two pellets for group *Delay*). Trials were separated by an ITI (inter-trial interval) generated by sampling from of a Poisson distribution with a mean of 50 seconds and adding 10 seconds.

Session structure

In the main experiment each session was composed of 60 trials: 40 *forced* (half *info* and half *non-info*) and 20 *choice* (randomly intermixed). Other trial sequence details were determined as previously described in the *cue discrimination training* section.

Data analysis

All data processing and statistics was done in MATLAB 2017a. Analysis of cumulative head-poking time was performed on the whole cue period for groups Probability and Magnitude and on the first five seconds for group delay. For the statistical analysis the cumulative head poking time was square-root transformed, to normalize the residuals. Pairwise tests between cue pairs were performed using the Tukey-Kramer critical values. Statistical analysis of preferences was done on choice trials, using the last session of each experimental group.

Results

In this experiment we assessed rats' preference for an informative option across different dimensions of reward: *probability*, *magnitude*, and *delay*. In all cases the informative option signalled outcomes immediately, while the non-informative one kept the rats under uncertainty until the outcome was due. At this point it is worth remembering that the *Neglect-Bad-News* interpretation of paradoxical choice predicts a preference for the informative option exclusively in the *probability* group (where on half of trials there is no salient outcome at the end of the waiting period), the idea that there is an intrinsic preference for information (uncertainty aversion) predicts preference for the informative option in all three groups.

Pre-training

With the aim of ensuring good discrimination between the cues we preceded the main experiment with a pre-training phase, in which rats were exposed to the cues and their associated-reward contingencies without any instrumental action required. The Pavlovian head-poking response to the different cues (**Figure 2)** shows evidence of discrimination in the *probability* and *delay* groups, but not in the *magnitude* group. In the *probability* group a two-way repeated-measures ANOVA with cue and session as within-subject factors and preference as the dependent variable revealed a significant effect of cue ($F_{2,14}$=93.86, P<0.0001), session ($F_{1,7}$=462.81, P<0.0001), and cue by session interaction ($F_{2,14}$=117.51, P<0.0001). Differences between all cues pairs were confirmed by pairwise comparisons (reward vs no-reward: P<0.0001, no-reward vs 50/50: P<0.001 and 50/50 vs reward: P<0.05). A similar pattern exists in the *delay* group with a significant effect of cue

($F_{2,14}$=32.53, P<0.001), session ($F_{1,7}$=470.35, P<0.0001) and cue by session interaction ($F_{2,14}$=25.29, P<0.0001). Likewise, pairwise tests revealed differences between all cue pairs (short vs long: P<0.001, long vs 50/50: P<0.01 and 50/50 vs short: P<0.01). In contrast to the *probability* and *delay* groups, no evidence of discrimination was found in the *magnitude* group. Although there is a significant effect of session ($F_{1,724}$=243, P<0.0001), there was no effect of cue ($F_{2,14}$=0.55, P=0.58) or cue by session interaction ($F_{2,14}$=0.92, P=0.41). Pairwise comparisons between cues were not significant (big vs small: P=0.74, small vs 50/50: P=0.66 and 50/50 vs big: P=0.95). Notice that while a difference in Pavlovian head-poking is evidence for discrimination, lack of difference does not demonstrate that the animals have not learned the cues' significance. This is true for the usual statistical reasons associated with null results, but particularly relevant here, because of a Pavlovian ceiling effect: the subjects in the *magnitude* group may find that a small reward warrants visits to the magazine just as much as a big reward.

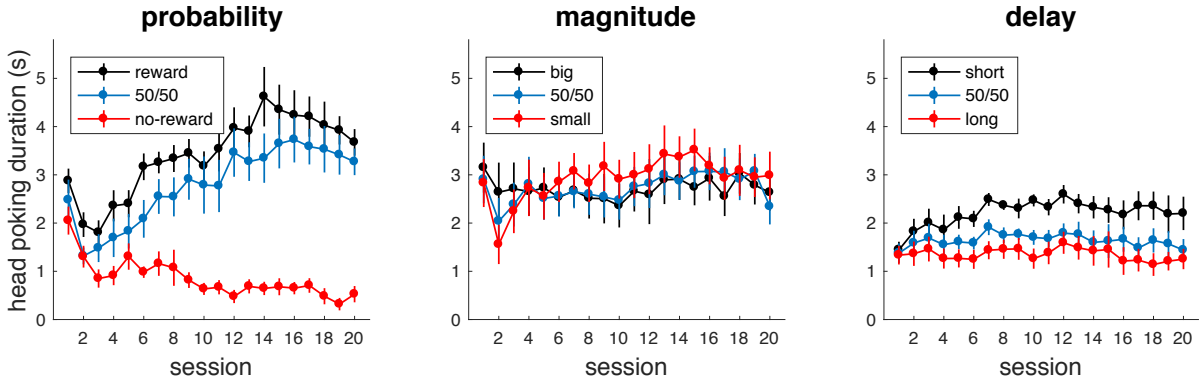


Figure 2. Cue discrimination
Mean ± S.E.M. (n=8) duration of head-poking into the food magazine during cue presentation in the pre-training phase for each of the dimensions of reward for which information preference was assed. Regarding absolute values, notice that head poking time in the delay condition is out of the first 5s of waiting time while in the other two conditions is out of 10s.

Main experiment

A strong preference for *info* developed in both the *probability* and *magnitude* groups. By contrast, in the *delay* group the rats were indifferent between the options (**Figure 3)**. For the *probability* group preference for *info* was significantly above 50% (mean=1 SD=0, Wilcoxon signed-rank, n=8, W=36, *P<0.01* two-tailed). For the *magnitude* group preference for *info* took longer to develop, but eventually was also significantly above 50% (mean= 0.93, SD=0.09, Wilcoxon signed-rank, n=8, W=36, *P<0.01* two-tailed). In the *delay* group, however, there was no difference between the options (mean=0.52, SD=0.46, Wilcoxon signed-rank, n=8, W=13, *P=1* two-tailed). For individual subject data see Figure S1 in Supplementary Results.

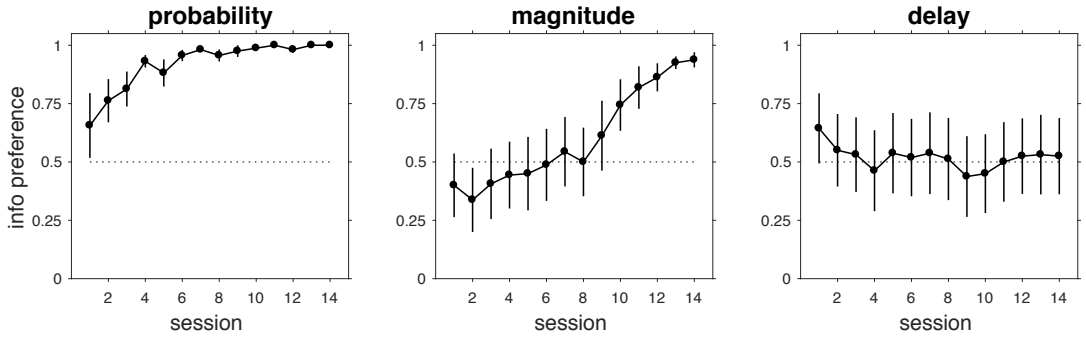


Figure 3. Preference for info

Mean ± S.E.M. (n=8) proportion of choices for the *info* (informative) option in each of the conditions. The dotted line represents indifference.

Discussion

The aim of the present chapter is to arbitrate between competing interpretations of the phenomenon of paradoxical choice. In this phenomenon, animals prefer an informative (*info*) option that signals outcomes in advance, relative to a non-informative (*non-info*) one that keeps them under uncertainty. Animals show this preference even if the probability of reward of the informative option is the same or lower than that of the non-informative option. One interpretation of this phenomenon that was proposed for the probability protocol is the *Neglect-Bad-News* model, that posits that this paradox is due to animals ignoring the time costs associated the cue signalling sure no-reward outcomes, because in nature this is a cost that does not have be to be paid. This argument should not apply (and hence the *info* preference should disappear) in the two new conditions employed here, *magnitude* and *delay*, because in both cases the subject arguably would stay to complete the cycle even after receiving the bad news, and then would pay all time costs. Another interpretation of the same phenomenon is that it reflects a broader preference for acquiring information (or reducing uncertainty) per-se. This predicts that the same preference for *info* should be present in all three conditions. Thus, our present experiment arbitrates between these interpretations, by modifying the paradoxical choice protocol and varying reward outcomes across three different dimensions of food reward (*probability*, *magnitude* and *delay*): The *Neglect-Bad-News* interpretation predicts a preference for the informative option only in the probability group while an intrinsic preference for information predicts preference for the informative option in all three groups.

The pattern of our results (especially indifference in the delay condition) cannot be due to an inability to discriminate between cues and their associated outcomes, as we have

evidence that subjects were able to perform this discrimination in all three groups. The head-poking behaviour in the *probability* and *delay* groups demonstrate this discrimination. This was shown in the pre-training phase, in which subjects were exposed to the cues and their associated outcomes without any instrumental action required. Although this was not the case for the *magnitude* group, nonetheless effective discrimination can also be inferred in this condition from the rats' preferences between options in the main experiment. The fact that subjects in this group preferred *info* demonstrates that they could discriminate between cues, even if they did not explicitly show this with their Pavlovian behaviour. The lack of differences in head-poking behaviour in the *magnitude* group could be due to a ceiling effect: the subjects in this group may find that a small reward warrants visits to the magazine just as much as a big reward. It is also possible that a metric of head-poking different from cumulative total time could produce differences between the cues, but this seems to belabour the point, a as argued above, failure to discriminate the cues is an implausible explanation.

The results of the present experiment argue against the *Neglect-Bad-News* interpretation, or at least show that it is not sufficient to account for all of our findings. According to this interpretation the preference for the informative option in the paradoxical choice experiments is the result of a mechanism that fails to take into the account the time cost associated with the sure no-reward outcome, because in nature this is a cost that does not have to be paid. This interpretation predicts preference for *info* only in the *probability* group, because only in this group is one of the possible outcomes the absence of reward, whereas we found a preference for *info* both in the *probability* and *magnitude* groups. In the *magnitude* group possible outcomes were either one or three pellets of food, therefore the *Neglect-Bad-News* model has no way of accounting for a preference for *info* in this

group. From a logical standpoint, *info* preference in the *magnitude* group does not falsify the *Neglect-Bad-News* interpretation of the conventional *probability* treatment. It is possible that this model accurately describes the mechanism operating in the probability group, and a different mechanism explains *info* preference in the *magnitude* group, Nonetheless, this possibility is not very parsimonious, as it requires multiple mechanisms where a single one might exist.

The alternative hypothesis, of an intrinsic preference for information, also does not account for all of our findings. Preference for *info* in the *probability* and *magnitude* groups is consistent with this interpretation, but indifference in the *delay* group is not. This last result is incongruent with the hypothesis because in this group, just like in the other two groups, the cues in the informative option predicted differential food reward outcomes. It is possible that only certain kinds of information are valued, but it is not readily apparent why information about delays would not be valuable, while information about probability (presence/absence) or magnitude would be. Regardless, the lack of an *info* preference in the *delay* group is a null result, and therefore some caveats apply. It is possible that the specific delay durations chosen were not adequate to show an effect, for instance if the outcomes were not sufficiently differently from each other. However, this seems unlikely as the long delay had three times the duration of the short delay, a difference in reward rate between outcomes just as large as the one in the *magnitude* group.

Evidence available at the moment is not sufficient to reconcile the different findings of the present experiment under a single mechanistic or functional hypothesis. Neither of our two pre-existing hypotheses is consistent with all the results. The *Neglect-Bad-News* model cannot explain preference for *info* in the *magnitude* group, intrinsic preference for

information cannot explain indifference towards the options in the *delay* group. An interesting aspect of results that also remains unaccounted for, is that preference for the informative option developed more rapidly in the *probability* group than in the *magnitude* group.

The Neglect-Bad-News model could perhaps be modified to account for *info* preference in the *magnitude* group, but it would then not be able to explain indifference in the *delay* group, for the reasons we develop below. Under the optimal diet selection strategy formulated by (Charnov, 1976a) instead of signalled no-reward outcomes being ignored as in the current model, signalled poor (relative to the context) outcomes would be ignored if the bad news are sufficiently bad that they would lead to aborting the chase in the wild. This could be consistent with Charnov's diet selection model, which posits that when animals are searching for prey their overall rate of capture is maximised if they reject any opportunity whose profitability (i.e. energy per unit of handling time) falls below the expected overall rate a specialist that rejected that prey type would experience (Charnov, 1976a). (Charnov, 1976a)Although this modified model needs to be developed before it can be properly assessed, it would face some strong difficulties: Crucially, it would appear to be incongruent with the finding of indifference towards *info* in the *delay* group. Given the profitability of each of the outcomes in the present experiment, the threshold for a specialist (who rejects the poor outcome) to earn a higher reward rate than a generalist (who accepts all outcomes) goes in the direction of a better premium for the specialist in the *delay* condition than in the *magnitude* condition. Additionally, Charnov's model assumes perfect knowledge of the parameters of each prey type. If a prey is ignored, it is because its magnitude and time cost (or at least their ratio, or profitability) are known,

while the learning-inspired Neglect-of-Bad-News model assumes that the time costs of the leaner outcome are simply not included in the computations.

Additionally, if the *Neglect-Bad-News* interpretation is not correct, it is not clear what would account for previous findings that appear to be specific to this interpretation. For instance, inserting a gap between choice and the start of an *info* cue affects preferences when this is done for the sure reward cue, but it does not if done for the sure no-reward cue (McDevitt et al., 1997). An intrinsic preference for information account would not predict this difference, as different outcomes carry just as much information regardless of their valence (i.e. whether they predict good or bad news). To probe this issue, future experiments could test whether such a differential effect of manipulating *info* cues can be found in a magnitude condition. Similarly, a problem with the intrinsic preference for information account is that it appears to be inconsistent with a previous study in pigeons (Roper and Zentall, 1999). This study tested (and failed to corroborate) an important prediction derived from Shannon's mathematical formalization of information content. According to Shannon's information theory (Shannon, 1948), the amount of available information is maximized when uncertainty is greatest. Therefore, preference for *info* should be greatest when outcomes are equally likely (i.e. 0.5 as in the present experiment). Instead, subjects' preference for the informative option was greatest when the probability of reward (one of the outcomes) was lowest (0.125 specifically).

In order to be able to assess the validity of hypothesis of intrinsic preference for information, the theoretical framework needs to be further refined. Models of this hypothesis should define exactly what kind of information is preferred, as this often remains unspecified (e.g. Bennett et al., 2016; Blanchard et al., 2015). A recent theoretical

work posits that it is information about the temporal properties of reward delivery that can explain the paradoxical choice phenomenon (Cunningham and Shahan, 2018; Shahan and Cunningham, 2015), see also (Balsam and Gallistel, 2009). This hypothesis would have thus predicted preference for *info* in the *delay* and *probability* groups, and indifference in the *magnitude* condition. The findings of the present chapter appear to be directly incongruent with this hypothesis, as the *delay* group was the only condition in our experiments were the rats showed indifference towards the informative option. The present experiment, as well as previous work on paradoxical choice, have been limited to information about the value of food rewards. Concerning future explorations, it must be taken into account that information could be about the value of outcomes unrelated to food. For example, avoiding danger or attracting mates could be similarly treated. Information could also be about outcomes unrelated to the value of rewards, for instance features of the environment such as its layout. It is reasonable to assume that some kinds of information are beneficial to animals, while most are not. However, this distinction has not been explored either theoretically or empirically in the context of the current paradigm. Even if at mechanistic level acquiring information were rewarding per-se, a statistical correlation ought to exist between such information acquisition and fitness-increasing events. Otherwise, this preference would have not evolved. In other words, models should be able to explain how the information acquired can eventually be used (e.g. to modify behaviour or to reduce computational costs), a link that is currently missing. There are significant hurdles in developing such a model but also potentially big payoffs.

To summarize, the results of the present experiment are comforting because they do replicate and expand previous evidence, but they are only partly consistent with the pre-existing hypotheses. On the one hand, the *Neglect-Bad-News* model cannot explain

preference for the informative option when outcomes correspond to different non-zero quantities of reward (magnitude). On the other hand, models based on a putative intrinsic motivational value of information cannot explain different responses between the content of information. This means that these theories need to be refined, and alternatively novel models need to be developed. There is work to be done, and this is encouraging!

Supplementary Results

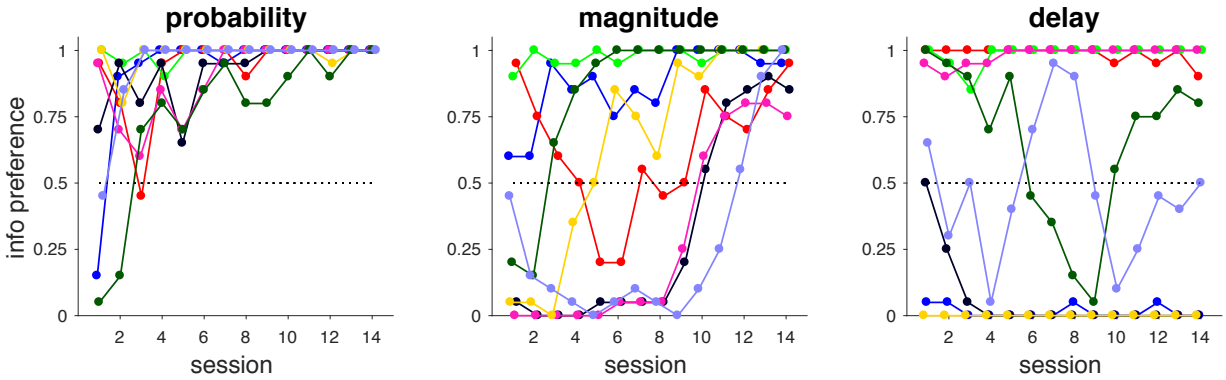


Figure S1. Preference for info (individual)
Proportion of choices for the *info* (informative) option in each of the conditions. Each line represents a different subject. The dotted line represents indifference.

General Discussion

From the very start of my research I faced several puzzling questions regarding the relation between information available in the environment and its potential to guide animal behaviour. These problems have interested behavioural scientists for a long time. For instance, (Dusenbery, 1996) was concerned with the innate mechanisms possessed by juvenile soil nematodes that allow them to move along sensory gradients so as to locate suitable food sources, (Krebs et al., 1978) studied how foraging birds allocate behavior between potential food sources so as to refine their estimates of relative value, and studies inspired by the observation that rats would rather bar press for food than eat from an open bowl (Jensen, 1963) led to the 'information primacy' hypothesis (Inglis et al., 2001) which states that animals would primarily work to reduce uncertainty, but that this tendency is diminished by hunger. In the laboratory operant chamber context this was studied using the 'observing response', summarized by Mackintosh (1974, 251–252) thus: "*Observing responses, therefore, reduce uncertainty about whether, when and where reinforcement is to be delivered. Since animals learn to make observing responses, it is to be presumed that the stimulus change contingent on the observing response is an effective conditioned reinforcer and since the stimuli provide information about the occurrence of primary reinforcement, it is possible that that it is precisely this information that is the basis of their effectiveness (Berlyne, 1960; Hendry, 1969)* ." The issue, however, is far from solved, as a variety of competing ideas and interpretations have been published, on both sides of the argument. For instance, Blanchard (1975), Steiner (1970), Silberberg and Fantino (2010) and Kacelnik (1987) have questioned the idea that it is information itself that constitutes the reinforcer.

The overarching aim of my thesis was then, to investigate whether animals can learn to produce actions with the objective of acquiring information. In chapters one, two and four I addressed this question by investigating the paradoxical choice phenomenon (McDevitt et al., 2016; Zentall, 2016), which in recent years has received considerable attention by laboratories working with starlings and pigeons and to some extent in rats (with contradictory results), while in the third chapter I tackled the same question by developing a protocol similar to that used previously in pigeons by Silberberg and Fantino (2010). In this case, information-seeking and food-seeking required separate actions, allowing me to test the motivational value of information in a context in which it can be used to increment the reward rate.

In the first chapter I showed that the paradoxical choice phenomenon (where subjects prefer a lower-reward but more informative option) and which had previously been demonstrated in pigeons and starlings (Stagner and Zentall, 2010), is also present in rats, a species in which the existence of this phenomenon had previously been questioned (Chow et al., 2017; Martínez et al., 2017; Trujano et al., 2016; Trujano and Orduña, 2015), The fact that the phenomenon is taxonomically widespread strengthens the argument that it reflects an adaptation to common characteristic of natural foraging scenarios, an assertion made by the *Neglect-Bad-News* model. It also argues against the interpretation that the phenomenon reflects a flaw in the design of the psychological mechanism producing it (what had been called 'suboptimal behavior' or 'maladaptive (Stagner et al., 2011; Stagner and Zentall, 2010; Zentall, 2011, 2011; Zentall and Stagner, 2011). Nonetheless, the strength of the preference for the informative option in rats is not as big as that found in pigeons and starlings. In order for the *Neglect-Bad-News* to account for this, an additional parameter needs to be incorporated, thus reducing the heuristic appeal of the model.

Showing that the phenomenon of paradoxical choice is taxonomically widespread, as I did in the first chapter, confirms the need to develop an explanatory framework, but does not, on its own, constitute a very specific corroboration of the *Neglect-Bad-News* model. Although the model asserts that the phenomenon represents an adaptation to a common characteristic of most natural foraging scenarios, and thus predicts that it should be taxonomically widespread, this applies equally to any other model that claims adaptiveness to a common characteristic. Therefore, while the presence of paradoxical choice in rats argues against the idea that it represents a maladaptive property, as embodied in 'suboptimal choice' label (Smith et al., 2016; Stagner et al., 2011; Stagner and Zentall, 2010) it does not support any specific adaptive interpretation of it. Also, the fact that fitting the behavior of the rats to the *Neglect-Bad-News* model requires adding a parameter limits the appeal of the model. Therefore, to further test the *Neglect-Bad-News* model, in the second chapter I set out to corroborate more specific predictions derived from the model.

In the second chapter I tested several predictions of the *Neglect-Bad-News* model. The first prediction comes directly from the model and is fundamental to it. The prediction states that if the paradoxical protocol is modified so that the information can be used, then the subjects will use the information appropriately. Specifically, if the subjects are able to terminate a trial and start the next one (that is, they can 'escape'), they will do so when presented with the cue for sure reward, but not when presented with the cue for sure no-reward. The results of the chapter corroborated this prediction. The second prediction was derived from extending the *Neglect-Bad-News* model, and states that preference for an informative option should diminish as the duration of the pre-choice delay increases. This

delay is a controllable dead time between the start of a trial and the moment when subjects can choose between the options. This second prediction was not supported by the data. In summary, the experiments of this chapter corroborated one of the predictions derived from the *Neglect-Bad-News* model, thereby supporting its viability but still leave some open issues. Importantly, the prediction that was corroborated (selective escape) had greater importance for the rationale that originates the model than the one that was not (effect of pre-choice delay). Additionally, it is not clear whether the lack of corroboration of the second prediction represents a failure of the model, or rather reflects issues in the task design or methodological implementation.

Although the findings from the first and second chapter support the *Neglect-Bad-News* model, I had only attempted to find corroborating evidence for the model. I had not, however, attempted to falsify the model by looking for evidence for an alternative hypothesis. In the fourth chapter I set out to do this. The *Neglect-Bad-News* model posits that preference for an informative option can be explained at a mechanistic level by animals ignoring the time cost associated with a signaled no-reward outcome, because in nature this cost does not have to be paid. Therefore, the model predicts preference for the informative option whenever the cues signal the availability (presence/absence) of rewards. However, an alternative interpretation of paradoxical choice exists, which is that it reflects a broader intrinsic preference for information (Bennett et al., 2016; Blanchard et al., 2015; Bromberg-Martin and Hikosaka, 2009; Iigaya et al., 2016). This hypothesis predicts that an informative option will be preferred regardless of the kind of information (e.g. viz probability, food quality, food amount, food delay or food palatability) to which it refers, as the animal's goal is assumed to be to reduce uncertainty per se.

To distinguish between these hypotheses in the fourth chapter I manipulated reward outcomes across three dimensions (probability, magnitude and delay). The *Neglect-Bad-News* model predicts preference for the informative option only in the probability condition, because only in this case there is a definite 'sure no reward' outcome, where the foraging interpretation is that the subject allocates time to some alternative pursuit, and the associative account is that there is no salient 'end of interval' signal that allows the conditioning process to attribute credit for the lost time to the earlier choice of option. In the magnitude and delay dimensions, the subject would take the time taken after bad news into account, and then would not show a preference for information. In contrast, an intrinsic preference for information predicts preference for the informative option in all three conditions.

The results from the fourth chapter are, on this light, surprising. The experiment confirmed the previous results with probability, showed a preference for information in the magnitude treatment, and total indifference in the delay treatment. On one hand these results reveal a finding that cannot be explained by the *Neglect-Bad-News* model, namely preference for the informative option in the magnitude condition. The *Neglect-Bad-News* model does not account for this finding, because in this condition all outcomes were a non-zero number of pellets, On the other hand, the lack of a preference for the informative option in the delay condition is inconsistent with the interpretation that the primary motivational objective is an acquisition of information per-se. Therefore, neither of our pre-existing hypotheses can explain the full set of results.

The *Neglect-Bad-News* model could be modified to account for *info* preference in the magnitude group, but it would then appear to be inconsistent with the finding of

indifference in the delay group, for the following reason. The model can be modified such that "*bad-news*" are not only empty trials (no-reward), but trials in which the outcome's profitability falls below some value. A reasonable threshold for this value would be the threshold in profitability identified by Charnov (1976) as worth accepting in the rate maximizing (optimal) diet. This could explain preference for information in the magnitude group, when the small reward's profitability is below Charnov's threshold, but would be inconsistent with indifference in the delay group, because the difference in profitability between outcomes in this group is even larger than in the magnitude group.

At the moment it is not clear how to reconcile all the findings from my thesis into a single theoretical framework. One possibility is to refine the hypothesis that information has an intrinsic reinforcement value, such that it specifies what kind of information is valued. This could explain why there was an effect in the magnitude and probability conditions, but not in the delay condition. Recent theoretical work (Cunningham and Shahan, 2018; Shahan and Cunningham, 2015), see also (Balsam and Gallistel, 2009) has developed a model with this kind of specificity. However, the predictions of this model are in direct contradiction to my own results. This recent model postulates that what drives the paradoxical choice phenomenon is valuation of information about the temporal properties of reward delivery. By contrast, the findings of the fourth chapter show that the delay condition was the sole dimension of reward for which the rats were indifferent towards the informative option.

In addition, the intrinsic preference for information interpretation makes predictions for the paradoxical choice protocol that have not been yet tested. One such prediction is that more information is better than less information. In a related protocol Silberberg and Fantino

(2010), found no support for this idea, by making the 'bad news' option more informative than the 'good news' option, and finding that the pigeons still preferred the good news alternative. In their own words: *"In one condition, the bad news was actually more informative about reinforcement than was the good news. Nevertheless, a preponderance of the birds' responses was made on the nominally good-news option*". In the experiments in my thesis the amount of information was not parameterized: one option was completely informative, while the alternative option was completely non-informative. In order to test whether an orderly preference exists along the information continuum, the amount of information provided to the animal could be graded. Validating this prediction would make the paradigm more representative of ecological environments, in the sense that perfect information is unlikely to be common in nature. Instead, cues that are only partly correlated with outcomes are likely to be more prevalent.

There are at least two ways to test the prediction that an orderly preference should exist between different degrees of informativeness. One way would be to vary the reliability of the cues. In the experiments of my thesis, the cues in the informative option were perfectly correlated with the outcomes, but this correlation could instead be partial. The specific prediction would be that if subjects are faced with a choice between two informative options, they will prefer the option in which the cues have a higher correlation with the outcomes. Another way to test the general prediction would be to increment the number of outcomes. Until now, the experiments have been restricted to two possible outcomes, but this number could be increased (for instance to three or four). Subjects would then be made to choose between pairs of informative options. In one option each outcome would be uniquely signaled by a cue, whereas in the alternative option, it would be subsets of outcomes that are signaled by each cue. Thus, all options would provide information about

upcoming reward outcomes, but one would do so with more specificity, thereby delivering a greater amount of information (i.e. leaving different degrees of uncertainty).

In the third chapter I explored a scenario distinct from that posed by the paradoxical choice protocol. Like in all chapters of my thesis, the experimental task aimed to assess whether animals can learn to produce an action with the objective of acquiring information. However, there was an important difference in the procedure of this chapter with respect to the other chapters. In the task used here, food-seeking and information-seeking required separate actions, while in the paradoxical choice protocol, being a simultaneous choice, a single action was both a reward-producing response and an information-producing one. Like in the second chapter (but unlike the first and fourth chapters) acquiring information can affect reward rate within the task. The potential of information to affect reward rate also distinguishes this protocol from other 'observing-response' studies (e.g. Silberberg and Fantino, 2010). Our results showed that rats did indeed perform the information-seeking action, and that performing this action was sensitive to the informational content of the cues. Thus, the behavior in this task is potentially an example of "extrinsically-motivated" information-seeking (Gottlieb et al., 2013), of a kind that had not been demonstrated to our knowledge before. Interestingly, they only performed the information-seeking action in a fraction of possible trials. This result was not predicted by rate-maximization and should be explored by future research.

The results of this thesis provide valuable findings to further our understanding of paradoxical choice and its relationship to information-seeking. Nonetheless, there is much work to be done, as none of the currently proposed theoretical interpretations can account for all the empirical findings. To progress further, it would be advantageous to generate

models that specify the psychological mechanisms involved in the acquisition of preference for an informative option. The *Neglect-Bad-News* model is connected with an ecological, foraging, interpretation, but does not succeed in accounting all available evidence. It also relates exclusively to terminal (steady-state) preferences but does not aim to explain how these preferences develop. In this respect, a reinforcement learning (Sutton and Barto, 1998) approach could prove useful. The findings from the fourth chapter constitute a particularly powerful dataset to test potential models of paradoxical choice. They provide important constraints that theoretical interpretations must satisfy, specifically preference for an informative option in magnitude and probability conditions but not in a delay condition. Another constraint could be the pattern with which preferences develop. An interesting aspect of results from the fourth chapter is that preference for the informative option developed more rapidly in the probability group than in the magnitude group.

Another potentially useful approach is to continue developing normative models of information-seeking. Some models of information-seeking describe mechanisms that result in preference for information (Bennett et al., 2016; Iigaya et al., 2016), but they do not specify the ecological, and hence evolutionary, consequences of any particular preference, thus neglecting either one or other of Tinbergen's critical four kinds of explanation (Tinbergen, 1963). By contrast, the *Neglect-Bad-News* model, in connection with the diet choice model, specifies adaptive reasons why, in natural environments, information about probability of capture may be more influential than information about magnitude of, or delay to outcome. However, this framework cannot explain all the findings of my thesis related to the paradoxical choice protocol, and in any case was not intended as a general-purpose account of information-seeking.

For an animal to value information *per se*, irrespective of its content, cannot be adaptive, because most of the information potentially available to any animal is likely to be useless.. Normative models of information-seeking should therefore specify the filters that determine what kind of information is valued. Some work has been done following an approach of this sort, but it remains mostly limited to the robotics field and its relevance to real-world animal behavior has not yet been established (Baldassarre and Mirolli, 2013; Barto, 2004; Chentanez et al., 2005; Hester and Stone, 2017; Oudeyer and Kaplan, 2009; Santucci et al., 2016; Sorrentino et al., 2014). One interesting example is the algorithm known as '*Targeted Exploration with Variance-And-Novelty-Intrinsic-Rewards'* (TEXPLORE-VENIR)' (Hester and Stone, 2017) in which a learning agent "*sets an internal goal for which the program rewards itself for comprehending something new, even if the knowledge doesn't get it closer to the ultimate goal*." While such progress is fundamental to understand what the problems are any autonomous agent faces, a specific biological problem is that normative models of information-seeking must reflect how the acquisition of certain types of information can increase the evolutionary fitness of animals under specific scenarios. In this respect, more work needs to be done to identify and describe instances of information-seeking behavior in natural environments. The long-term goal should be to develop a multi-level account of information-seeking in the four senses addressed by Tinbergen (1963) and the three levels addressed by Marr and Poggio, (1976) . The goal or problem to be solved should be specified by reflecting about the agent in the environment in which it has evolved. The algorithmic description should detail the psychological mechanisms that perform the different functions. Lastly, the implementation level should describe the neural apparatus that performs the computations in the biological system. Studying behavior in natural environments should also allow us to understand why

certain aspects of information-seeking behaviors are innate while others are learned. This distinction might be connected with features of the environment that are stable or show variability (Dunlap and Stephens, 2009)

As a last remark, a recurring finding of the experiments in this thesis is that rats do behave in ways that systematically fail to maximize their food reward-rate, a result that does not violate the principles of optimality modeling in evolutionary biology, but that undoubtedly exposes difficulties to be addressed by the behavioural biologist. The main focus of this thesis was in one such instance, namely the so-called paradoxical choice phenomenon, but other aspects of the results also failed to conform to the expectations generated from reward-rate maximization. For instance, in the second chapter subjects occasionally escaped in response to a cue that signaled sure reward when the background alternative could only be worse, while in the third chapter subjects chose to forgo seeking potentially useful information. These instances of non-maximization of reward-rate should not be taken as demonstrations that the optimal foraging framework has little predictive power and is therefore not very useful. Rather it should be taken as evidence that what is beneficial in the experimental paradigm was uncoupled in some way from what is beneficial in natural environments. This suggests the animals have prior beliefs about the structure of the world that are not completely flexible to individual scenarios (a related point is made by Versace et al. (2018)). Therefore, the particular circumstances under which animals do not maximize reward-rate in laboratory tasks are interesting, because they can potentially help us to understand us what animals assume 'by design' to be true about the environments in which they live. Thus, failures of maximization in the laboratory should not be taken as a nuisance but rather as opportunities, because they are avenues into understanding adaptive behavior in the wild.

References

Ainslie, G. W. (1974). IMPULSE CONTROL IN PIGEONS 1. *J. Exp. Anal. Behav.* 21, 485–489.

Baldassarre, G., and Mirolli, M. (2013). "Intrinsically Motivated Learning Systems: An Overview," in *Intrinsically Motivated Learning in Natural and Artificial Systems* (Springer, Berlin, Heidelberg), 1–14. doi:10.1007/978-3-642-32375-1_1.

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., and Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Front. Psychol.* 5. doi:10.3389/fpsyg.2014.00985.

Balsam, P. D., and Gallistel, C. R. (2009). Temporal maps and informativeness in associative learning. *Trends Neurosci.* 32, 73–78. doi:10.1016/j.tins.2008.10.004.

Barto, A. G. (2004). Intrinsically motivated learning of hierarchical collections of skills. Available at: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.6436 [Accessed January 26, 2018].

Bean, D., Mason, G. J., and Bateson, M. (1999). CONTRAFREELOADINGIN STARLINGS: TESTING THE INFORMATION HYPOTHESIS. *Behaviour* 136, 1267–1282.

Beierholm, U. R., and Dayan, P. (2010). Pavlovian-Instrumental Interaction in 'Observing Behavior.' *PLoS Comput. Biol.* 6. doi:10.1371/journal.pcbi.1000903.

Bennett, D., Bode, S., Brydevall, M., Warren, H., and Murawski, C. (2016). Intrinsic Valuation of Information in Decision Making under Uncertainty. *PLOS Comput. Biol.* 12, e1005020. doi:10.1371/journal.pcbi.1005020.

Berger-Tal, O., and Avgar, T. (2012). The glass is half-full: overestimating the quality of a novel environment is advantageous. *PLoS One* 7, e34578.

Berger-Tal, O., Nathan, J., Meron, E., and Saltz, D. (2014). The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLOS ONE* 9, e95693. doi:10.1371/journal.pone.0095693.

Berlyne, D. E. (1960). Conflict, arousal, and curiosity.

Blanchard, R. (1975). The effect of S− on observing behavior. *Learn. Motiv.* 6, 1–10. doi:10.1016/0023-9690(75)90031-4.

Blanchard, T. C., Hayden, B. Y., and Bromberg-Martin, E. S. (2015). Orbitofrontal Cortex Uses Distinct Codes for Different Choice Attributes in Decisions Motivated by Curiosity. *Neuron* 85, 602–614. doi:10.1016/j.neuron.2014.12.050.

Bower, G., McLean, J., and Meacham, J. (1966). Value of knowing when reinforcement is due. *J. Comp.* 62, 184–192.

Bromberg-Martin, E. S., and Hikosaka, O. (2009). Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards. *Neuron* 63, 119–126. doi:10.1016/j.neuron.2009.06.009.

Bromberg-Martin, E. S., and Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nat. Neurosci.* 14, 1209–1216. doi:10.1038/nn.2902.

Browne, M. P., and Dinsmoor, J. A. (1974). Wyckoff's observing response: Pigeons learn to observe stimuli for free food but not stimuli for extinction. *Learn. Motiv.* 5, 165–173. doi:10.1016/0023-9690(74)90023-X.

Charnov, E. L. (1976a). Optimal foraging: attack strategy of a mantid. *Am. Nat.* 110, 141–151.

Charnov, E. L. (1976b). Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.* 9, 129–136. doi:10.1016/0040-5809(76)90040-X.

Chentanez, N., Barto, A. G., and Singh, S. P. (2005). Intrinsically motivated reinforcement learning. in *Advances in neural information processing systems*, 1281–1288.

Chow, J. J., Smith, A. P., Wilson, A. G., Zentall, T. R., and Beckmann, J. S. (2017). Suboptimal choice in rats: Incentive salience attribution promotes maladaptive decision-making. *Behav. Brain Res.* 320, 244–254. doi:10.1016/j.bbr.2016.12.013.

Cunningham, P. J., and Shahan, T. A. (2018). Suboptimal choice, reward-predictive signals, and temporal information. *J. Exp. Psychol. Anim. Learn. Cogn.* 44, 1.

D'amato, M. R. (1974). Derived motives. *Annu. Rev. Psychol.* 25, 83–106.

Dinsmoor, J. A. (1983). Observing and conditioned reinforcement. *Behav. Brain Sci.* 6, 693–704. doi:10.1017/S0140525X00017969.

Dunlap, A. S., and Stephens, D. W. (2009). Components of change in the evolution of learning and unlearned preference. *Proc. R. Soc. Lond. B Biol. Sci.* 276, 3201–3208.

Dusenbery, D. B. (1996). Information is where you find it. *Biol. Bull.* 191, 124–128.

Eliassen, S., Jørgensen, C., Mangel, M., and Giske, J. (2007). Exploration or exploitation: life expectancy changes the value of learning in foraging strategies. *Oikos* 116, 513–523.

Fortes, I., Machado, A., and Vasconcelos, M. (2017). Do pigeons (Columba livia) use information about the absence of food appropriately? A further look into suboptimal choice. *J. Comp. Psychol.* 131, 277.

Fortes, I., Pinto, C., Machado, A., and Vasconcelos, M. (2018). The paradoxical effect of low reward probabilities in suboptimal choice. *J. Exp. Psychol. Anim. Learn. Cogn.* 44, 180.

Fortes, I., Vasconcelos, M., and Machado, A. (2016). Testing the boundaries of "paradoxical" predictions: Pigeons do disregard bad news. *J. Exp. Psychol. Anim. Learn. Cogn.* 42, 336.

Frankenhuis, W. E., Panchanathan, K., and Barto, A. G. (2018). Enriching behavioral ecology with reinforcement learning methods. *Behav. Processes*.

Freidin, E., Aw, J., and Kacelnik, A. (2009). Sequential and simultaneous choices: Testing the diet selection and sequential choice models. *Behav. Processes* 80, 218–223. doi:10.1016/j.beproc.2008.12.001.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi:10.1038/nrn2787.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–214. doi:10.1080/17588928.2015.1020053.

Ghirlanda, S., and Enquist, M. (2003). A century of generalization. *Anim. Behav.* 66, 15–36. doi:10.1006/anbe.2003.2174.

Gibson, J. J. (1966). The senses considered as perceptual systems.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press.

Golland, P., Liang, F., Mukherjee, S., and Panchenko, D. (2005). Permutation Tests for Classification. in *Learning Theory* Lecture Notes in Computer Science. (Springer, Berlin, Heidelberg), 501–515. doi:10.1007/11503415_34.

Gottlieb, J., and Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nat. Rev. Neurosci.*, 1.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593. doi:10.1016/j.tics.2013.09.001.

Grafen, A., and Hails, R. (2002). *Modern statistics for the life sciences*. Oxford, UK.: Oxford University Press.

Heffner, H. E., and Heffner, R. S. (1985). Hearing in Two Cricetid Rodents: Wood Rat (*neotoma floridana*) and Grasshopper Mouse (*onychomys leucogaster*). *J. Comp. Psychol.* 99, 275–288.

Hendry, D. P. (1969). Reinforcing value of information: Fixed-ratio schedules. *Cond. Reinf.*, 300–341.

Hester, T., and Stone, P. (2017). Intrinsically motivated model learning for developing curious robots. *Spec. Issue AI Robot.* 247, 170–186. doi:10.1016/j.artint.2015.05.002.

Holland, P. C. (1989). Occasion setting with simultaneous compounds in rats. *J. Exp. Psychol. Anim. Behav. Process.* 15, 183–193. doi:10.1037/0097-7403.15.3.183.

Holland, P. C. (1991). Acquisition and transfer of occasion setting in operant feature positive and feature negative discriminations. *Learn. Motiv.* 22, 366–387. doi:10.1016/0023-9690(91)90002-P.

Houston, A. I., Kacelnik, A., and McNamara, J. (1982). Some learning rules for acquiring information. *Funct. Ontog.* 1, 140–191.

Hull, C. (1943). Principles of behavior.

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

Inglis, I. R. (1987). The information-primacy approach: A reply to Kacelnik.

Inglis, I. R., Forkman, B., and Lazarus, J. (1997). Free food or earned food? A review and fuzzy model of contrafreeloading. *Anim. Behav.* 53, 1171–1191.

Inglis, I. R., Langton, S., Forkman, B., and Lazarus, J. (2001). An information primacy model of exploratory and foraging behaviour. *Anim. Behav.* 62, 543–557.

Inglis, I. R., and Shepherd, D. S. (1994). Rats work for food they then reject: support for the information-primacy approach to learned industriousness. *Ethology* 98, 154–164.

Jensen, G. D. (1963). Preference for bar pressing over" freeloading" as a function of number of rewarded presses. *J. Exp. Psychol.* 65, 451.

Kacelnik, A. (1987). Information primacy or preference for familiar foraging techniques? A critique of Inglis & Ferguson.

Kacelnik, A., Houston, A. I., and Krebs, J. R. (1981). Optimal foraging and territorial defence in the great tit (Parus major). *Behav. Ecol. Sociobiol.* 8, 35–40.

Kacelnik, A., Vasconcelos, M., Monteiro, T., and Aw, J. (2010). Darwin's "tug-of-war" vs. starlings' "horse-racing": how adaptations for sequential encounters drive simultaneous choice. *Behav. Ecol. Sociobiol.* 65, 547–558. doi:10.1007/s00265-010-1101-2.

Kelleher, R. T., Riddle, W. C., and Cook, L. (1962). OBSERVING RESPONSES IN PIGEONS. *J. Exp. Anal. Behav.* 5, 3–13. doi:10.1901/jeab.1962.5-3.

Kidd, C., and Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. *Neuron* 88, 449–460. doi:10.1016/j.neuron.2015.09.010.

Krebs, J. R., Erichsen, J. T., Webber, M. I., and Charnov, E. L. (1977). Optimal prey selection in the great tit (Parus major). *Anim. Behav.* 25, 30–38.

Krebs, J. R., and Kacelnik, A. (1984). Optimal learning rules. *Behav. Brain Sci.* 7, 109–110.

Krebs, J. R., Kacelnik, A., and Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature* 275, 27–31. doi:10.1038/275027a0.

Laude, J. R., Stagner, J. P., and Zentall, T. R. (2014). Suboptimal choice by pigeons may result from the diminishing effect of nonreinforcement. *J. Exp. Psychol. Anim. Learn. Cogn.* 40, 12.

Loomes, G., and Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *Econ. J.* 92, 805–824. doi:10.2307/2232669.

Mackintosh, N. J. (1974). *The psychology of animal learning.* Academic Press.

Marr, D., and Poggio, T. (1976). From understanding computation to understanding neural circuitry.

Martínez, M., Alba, R., Rodríguez, W., and Orduña, V. (2017). Incentive salience attribution is not the sole determinant of suboptimal choice in rats: Conditioned

inhibition matters. *Behav. Processes* 142, 99–105. doi:10.1016/j.beproc.2017.06.012.

McDevitt, M. A., Dunn, R. M., Spetch, M. L., and Ludvig, E. A. (2016). When good news leads to bad choices. *J. Exp. Anal. Behav.* 105, 23–40. doi:10.1002/jeab.192.

McDevitt, M. A., Spetch, M. L., and Dunn, R. (1997). Contiguity and Conditioned Reinforcement in Probabilistic Choice. *J. Exp. Anal. Behav.* 68, 317–327. doi:10.1901/jeab.1997.68-317.

Meyer, H. C., and Bucci, D. J. (2017). Negative occasion setting in juvenile rats. *Occas. Setting Past Present Future* 137, 33–39. doi:10.1016/j.beproc.2016.05.003.

Ojeda, A., Murphy, R. A., and Kacelnik, A. (2018). Paradoxical choice in rats: Subjective valuation and mechanism of choice. *Quant. Comp. Anal. Behav.* 152, 73–80. doi:10.1016/j.beproc.2018.03.024.

Osborne, S. R. (1977). The free food (contrafreeloading) phenomenon: A review and analysis. *Anim. Learn. Behav.* 5, 221–235.

Oudeyer, P.-Y., and Kaplan, F. (2009). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobotics* 1. doi:10.3389/neuro.12.006.2007.

Palminteri, S., Lefebvre, G., Kilford, E. J., and Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Comput. Biol.* 13, e1005684. doi:10.1371/journal.pcbi.1005684.

Premack, D. (1959). Toward empirical behavior laws: I. Positive reinforcement. *Psychol. Rev.* 66, 219.

Recer, G. M., and Caraco, T. (1989). Sequential-encounter prey choice and effects of spatial resource variability. *J. Theor. Biol.* 139, 239–249.

Reed, P., Mitchell, C., and Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Anim. Learn. Behav.* 24, 38–45. doi:10.3758/BF03198952.

Roper, K. L., and Zentall, T. R. (1999). Observing Behavior in Pigeons: The Effect of Reinforcement Probability and Response Cost Using a Symmetrical Choice Procedure. *Learn. Motiv.* 30, 201–220. doi:10.1006/lmot.1999.1030.

Santucci, V. G., Baldassarre, G., and Mirolli, M. (2016). GRAIL: A Goal-Discovering Robotic Architecture for Intrinsically-Motivated Learning. *IEEE Trans. Cogn. Dev. Syst.* 8, 214–231. doi:10.1109/TCDS.2016.2538961.

Schmajuk, N. A., and Holland, P. C. eds. (1998). Occasion setting: Associative learning and cognition in animals. *Occas. Setting Assoc. Learn. Cogn. Anim.*, xxi, 440–xxi, 440. doi:10.1037/10298-000.

Shahan, T. A., and Cunningham, P. (2015). Conditioned reinforcement and information theory reconsidered. *J. Exp. Anal. Behav.* 103, 405–418. doi:10.1002/jeab.142.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423.

Shapiro, M. S., Siller, S., and Kacelnik, A. (2008). Simultaneous and Sequential Choice as a Function of Reward Delay and Magnitude: Normative, Descriptive and Process-based Models Tested in the European Starling (*sturnus vulgaris*). *J. Exp. Psychol. Anim. Behav. Process.* 34, 75–93. doi:10.1037/0097-7403.34.1.75.

Silberberg, A., and Fantino, E. (2010). Observing responses: Maintained by good news only? *Behav. Processes* 85, 80–82. doi:10.1016/j.beproc.2010.06.002.

Skinner, B. F. (1935). The generic nature of the concepts of stimulus and response. *J. Gen. Psychol.* 12, 40–65.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis.* Oxford, England: Appleton-Century.

Smith, A. P., Bailey, A. R., Chow, J. J., Beckmann, J. S., and Zentall, T. R. (2016). Suboptimal Choice in Pigeons: Stimulus Value Predicts Choice over Frequencies. *PLOS ONE* 11, e0159336. doi:10.1371/journal.pone.0159336.

Sorrentino, E. P. di, Sabbatini, G., Truppa, V., Bordonali, A., Taffoni, F., Formica, D., et al. (2014). Exploration and learning in capuchin monkeys (<Emphasis

Type="Italic">Sapajus</Emphasis> spp.): the role of action–outcome contingencies. *Anim. Cogn.* 17, 1081–1088. doi:10.1007/s10071-014-0740-5.

Stagner, J. P., Laude, J. R., and Zentall, T. R. (2011). Sub-optimal choice in pigeons does not depend on avoidance of the stimulus associated with the absence of reinforcement. *Learn. Motiv.* 42, 282–287. doi:10.1016/j.lmot.2011.09.001.

Stagner, J. P., and Zentall, T. R. (2010). Suboptimal choice behavior by pigeons. *Psychon. Bull. Rev.* 17, 412–416. doi:10.3758/PBR.17.3.412.

Steiner, J. (1970). Observing responses and uncertainty reduction. II The effect of varying the probability of reinforcement. *Q. J. Exp. Psychol.* 22, 592–599.

Stephens, D. W. (1985). How important are partial preferences? *Anim. Behav.* 33, 667–669. doi:10.1016/S0003-3472(85)80092-0.

Stephens, D. W., and Anderson, D. (2001). The adaptive value of preference for immediacy: when shortsighted rules have farsighted consequences. *Behav. Ecol.* 12, 330–339. doi:10.1093/beheco/12.3.330.

Stephens, D. W., and Krebs, J. R. (1986). *Foraging Theory*. Princeton University Press.

Stevens, J. R., Hallinan, E. V., and Hauser, M. D. (2005). The ecology and evolution of patience in two New World monkeys. *Biol. Lett.* 1, 223–226.

Sutton, R. S., and Barto, A. G. (1998). *Introduction to reinforcement learning*. MIT press Cambridge.

Talling, J. C., INGLIS, I., and Van Driel, K. S. (2002). EFFECT OF HUNGER ON STARLINGS'PREFERENCESFOR FOOD SOURCES ASSOCIATED WITH VARIABILITY OR UNCERTAINTY. *Behaviour* 139, 1223–1235.

Thorndike, E. (1898). SOME EXPERIMENTS ON ANIMAL INTELLIGENCE. *Science* 7, 818. doi:10.1126/science.7.181.818.

Timberlake, W., and Allison, J. (1974). Response deprivation: An empirical approach to instrumental performance.

Tinbergen, N. (1963). On aims and methods of ethology. *Z. Für Tierpsychol.* 20, 410–433.

Trask, S., Thrailkill, E. A., and Bouton, M. E. (2017). Occasion setting, inhibition, and the contextual control of extinction in Pavlovian and instrumental (operant) learning. *Occas. Setting Past Present Future* 137, 64–72. doi:10.1016/j.beproc.2016.10.003.

Trujano, R. E., López, P., Rojas-Leguizamón, M., and Orduña, V. (2016). Optimal behavior by rats in a choice task is associated to a persistent conditioned inhibition effect. *Behav. Processes* 130, 65–70. doi:10.1016/j.beproc.2016.07.005.

Trujano, R. E., and Orduña, V. (2015). Rats are optimal in a choice task in which pigeons are not. *Behav. Processes* 119, 22–27. doi:10.1016/j.beproc.2015.07.010.

Vasconcelos, M., Monteiro, T., Aw, J., and Kacelnik, A. (2010). Choice in multi-alternative environments: A trial-by-trial implementation of the Sequential Choice Model. *Behav. Processes* 84, 435–439. doi:10.1016/j.beproc.2009.11.010.

Vasconcelos, M., Monteiro, T., and Kacelnik, A. (2015). Irrational choice and the value of information. *Sci. Rep.* 5, 13874. doi:10.1038/srep13874.

Vásquez, R., and Kacelnik, A. (2000). Foraging rate versus sociality in the starling Sturnus vulgaris. *Proc. Biol. Sci.* 267, 157–164. doi:10.1098/rspb.2000.0981.

Versace, E., Martinho-Truswell, A., Kacelnik, A., and Vallortigara, G. (2018). Priors in Animal and Artificial Intelligence: Where Does Learning Begin? *Trends Cogn. Sci.* 22, 963–965. doi:10.1016/j.tics.2018.07.005.

Williams, B. A. (1994). Conditioned reinforcement: Neglected or outmoded explanatory construct? *Psychon. Bull. Rev.* 1, 457–475. doi:10.3758/BF03210950.

Williams, D. R., and Williams, H. (1969). AUTO-MAINTENANCE IN THE PIGEON: SUSTAINED PECKING DESPITE CONTINGENT NON-REINFORCEMENT 2. *J. Exp. Anal. Behav.* 12, 511–520.

Wilton, R. N., and Clements, R. O. (1971). The role of information in the emission of observing responses: A test of two hypotheses. *J. Exp. Anal. Behav.* 16, 161.

Wyckoff Jr., L. B. (1952). The role of observing responses in discrimination learning. Part I. *Psychol. Rev.* 59, 431–442. doi:10.1037/h0053932.

Zentall, T. R. (2011). Maladaptive “gambling” by Pigeons. *Behav. Processes* 87, 50–56. doi:10.1016/j.beproc.2010.12.017.

Zentall, T. R. (2016). Resolving the paradox of suboptimal choice. *J. Exp. Psychol. Anim. Learn. Cogn.* 42, 1–14. doi:10.1037/xan0000085.

Zentall, T. R., and Stagner, J. (2011). Maladaptive choice behaviour by pigeons: an animal analogue and possible mechanism for gambling (sub-optimal human decision-making behaviour). *Proc. R. Soc. Lond. B Biol. Sci.* 278, 1203–1208. doi:10.1098/rspb.2010.1607.