Strategies for selecting and evaluating information

Alice Liefgreen[1, *], Toby Pilditch[1, 2] and David Lagnado[1]

[1] Department of Experimental Psychology, University College London, UK

[2] University of Oxford, School of Geography and the Environment, Oxford, UK

*Contact corresponding author at: Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP, UK, and London, UK

E-mail address: alice.liefgreen.15@ucl.ac.uk

Abstract

Within the domain of psychology, Optimal Experimental Design (OED) principles have been used to model how people seek and evaluate information. Despite proving valuable as computational-level methods to account for people's behaviour, their descriptive and explanatory powers remain largely unexplored. In a series of experiments, we used a naturalistic crime investigation scenario to examine how people evaluate queries, as well as outcomes, in probabilistic contexts. We aimed to uncover the psychological strategies that people use, not just to assess whether they deviated from OED principles. In addition, we explored the adaptiveness of the identified strategies across both one-shot and stepwise information search tasks. We found that people do not always evaluate queries strictly in OED terms and use distinct strategies, such as by identifying a leading contender at the outset. Moreover, we identified aspects of zero-sum thinking and risk aversion that interact with people's information search strategies. Our findings have implications for building a descriptive account of information seeking and evaluation, accounting for factors that currently lie outside the realm of information-theoretic OED measures, such as context and the learner's own preferences.

Keywords: information search, OED framework; utility functions, inquiry, question asking, strategies, probabilistic reasoning, Bayesian Networks

1. Introduction

In everyday and professional contexts, people are often required to make judgements and decisions in environments permeated by uncertainty, whether through lack of information, unreliability of sources, or complex relationships between items. To make accurate decisions we must not only integrate and evaluate information, but also seek and acquire the *right* information in the first place. For example, after an initial examination a physician might have enough information to conjecture several possible hypotheses to explain the symptoms of a patient. However, to make an accurate diagnosis the physician needs to engage in *additional questioning* (examinations or laboratory tests) to gather information that can further distinguish between the possible diagnoses. Although certain aspects of this situation are captured by the current normative frameworks of human information acquisition (e.g., Optimal Experimental Design framework; Lindley, 1956), other aspects - such as how certain pre-inquiry "preferences" can guide people's information search strategies - are less readily modelled. For example, should the physician at the outset perform an exam to facilitate the identification of a *leading* hypothesis or one that enables the *exclusion* of a hypothesis?

Obtaining a 'frontrunner' hypothesis and 'excluding' a hypothesis are both seemingly rational motivators of inquiry, and people's preferences for these will be influenced by contextual factors. For example, the physician might prioritise a test that could *exclude* a diagnosis at the outset, if it is one that has serious implications for the patient. In contrast, a criminal investigator investigating a kidnapping case might choose an initial inquiry that facilitates the identification of a lead suspect (i.e., *frontrunner*) from a given pool of suspects, in order to maximise the chances of rapidly finding the missing person. In more dispassionate situations, such as that of an employer interviewing multiple candidates for the same position, choosing between an initial question that might help detect a lead candidate and a question that could help to exclude a candidate, is even more contentious. Although central to information search situations, these strategic preferences, as well as their determinants, have been largely overlooked within the psychological literature of human information acquisition.

1.1. Information acquisition and evaluation: Normative account

Seeking and evaluating information are recognised as central components of human cognition, capturing the attention of researchers across numerous disciplines. Within the psychological literature, Optimal Experimental Design (OED) principles, based on insights from statistics and computer science, have been used to build normative and descriptive models of people's information acquisition and evaluation behaviour (Baron, 1985; Klayman & Ha, 1997; Nelson, 2005). Part of the appeal of OED models is that they allow researchers to explore evaluation and integration processes within a probabilistic framework (Savage, 1954). As such, a Bayesian OED framework integrates i) a probabilistic belief model with a set of hypotheses (with specified prior probabilities) and a set of possible "queries"[1] to discern between these, ii) a measure to quantify the usefulness of each possible query relative to the probabilistic belief model, and iii) a (Bayesian) method of updating beliefs according to a query's outcome (Nelson, 2005). OED principles posit that people search for information with the goal of optimizing the information gained from their action. Thus, queries are selected that are anticipated to return information of utmost value or 'utility', by resulting in the greatest reduction in uncertainty (i.e., Shannon entropy of a learner's belief distribution). Equation 1 illustrates the framework utilised by all OED models to quantify the utility of a query, *eu (Q),* as the *expected* usefulness (*u*), given current knowledge, of the possible query outcomes $a_{i:}$

$$eu(Q) = \sum_{a_i} P(a_i)u(a_i) \qquad \text{Equation 1}$$

Several utility functions exist that quantify the usefulness of query *outcomes*, u (a_i) in different ways. In the present paper we utilise utility functions defined purely in information-theoretic terms and thus we will not describe situation-specific utility functions with reward structures (for a discussion of these see Coenen, Nelson & Gureckis, 2018). The same mathematical framework can nonetheless be employed to both situation-specific and information-theoretic cases (Savage, 1954). Prominent utility functions include probability gain (PG; Baron, 1985), Bayesian diagnosticity (Good, 1950), log diagnosticity, information gain (IG; Box & Hill, 1967; Lindley, 1956), Kullback-Leibler divergence

[1] We adopt the term query to represent any information-seeking action (i.e., experiment, test, or question).

(KL-D; Kullback & Liebler, 1951) and Impact (IMP; Klayman & Ha, 1987; Nickerson, 1996). We will now describe in turn the utility functions that we adopt in the present work, namely; Kullback-Liebler divergence, information gain, probability gain and Impact (for a more in-depth review of these see Nelson, 2005; 2008). Diagnosticity measures (e.g., log diagnosticity and Bayesian diagnosticity) were not included following the arguments presented in Nelson (2005) stipulating they are poor theoretical models of the utility of information and are not needed to explain empirical data of information search.

Kullback-Liebler divergence conceptualises a query's usefulness as the amount that the information provided by its outcomes is expected to change one's beliefs in the hypotheses h_i within the model. As such, KL-D computes the expected usefulness of a query outcome a_i as:

$$KL(a_i) = \sum_{h_i} P(h_i|a_i) * log_2 \frac{P(h_i|a_i)}{P(h_i)} \qquad \text{Equation 2}$$

From the equation above it follows that the usefulness of a query (Q), measured as change from prior beliefs about a true hypothesis, *H,* to posterior beliefs after a particular query outcome is observed, is computed as:

$$KL(Q) = \sum_{a_i} P(a_i) * \sum_{h_i} P(h_i|a_i) * log_2 \frac{P(h_i|a_i)}{P(h_i)} \qquad \text{Equation 3}$$

Information gain (Lindley, 1956) quantifies a query's usefulness according to how much it would reduce uncertainty with respect to the true hypothesis. The expected usefulness of a query outcome a_i would be computed as the difference between the entropy of the prior distribution and that of the posterior distribution, conditional on the status of the effect, and the expected usefulness of a query would be:

$$\mathrm{IG}(Q) = \sum_{\mathrm{h_i}} \mathrm{P}(h_i) * \log_2 \frac{1}{P(h_i)} - \sum_{\mathrm{a_i}} \mathrm{P}(\mathrm{a_i}) * \sum_{\mathrm{h_i}} P(h_i|a_i) \log_2 \frac{1}{P(h_i|a_i)} \qquad \text{Equation 4}$$

It is worth noting that KL-D and IG make identical predictions regarding a query's usefulness (Oaksford & Chater, 1994), although they have been shown to give different measures of a particular query *outcome*'s usefulness (Nelson, 2008).

Probability gain values a query in terms of its expected improvement in classification accuracy, assuming that the most probable category will always be chosen. The model's informational utility function is shown in Equation 5, where the max operators choose the leading (i.e., most likely) hypothesis given the outcome of a query and the initially leading hypothesis before any query. The difference between the two terms is the expected probability gain of a query outcome:

$$\mathrm{PG}(a_i) = \max_i P(h_i|a_i) - \max_i P(h_i) \qquad \text{Equation 5}$$

It follows that the expected usefulness of a query according to probability gain is maximises computed as:

$$\mathrm{PG}(Q) = (\sum_{\mathrm{a_i}} P(a_i) * \max_{hi} P(h_i|a_i)) - \max_{hi} P(h_i) \qquad \text{Equation 6}$$

Finally, *Impact* is a measure of absolute change, quantifying the usefulness of a query as the absolute change in beliefs (Nelson, 2005; Wells & Lindsay, 1980) from prior to posterior probability of the hypotheses conditional on a query outcome. The expected usefulness of a query according to Impact can be computed as:

$$\mathrm{IMP}(Q) = \sum_{a_i} P(a_i) * \frac{1}{n} * \sum_{h_i} abs\,[P(h_i|a_i) - P(h_i)] \qquad \text{Equation 7}$$

Note that with a binary hypothesis space with equiprobable base rates, *Impact* and *Probability gain* are identical (Nelson, 2005).

The above-mentioned utility functions are distinct in terms of how they characterise the goal of the information seeker and how they quantify the diagnosicity of information; they also differ in certain inherent properties. For example, KL-D and Impact are non-negative measures, meaning that they will always quantify the expected usefulness of an outcome as being greater than zero (usefulness $(a_i) \geq 0$). IG and PG hold the property of additivity meaning that the expected usefulness of a given outcome equals the additive expected usefulness of each outcome (a_i = usefulness a_1 + usefulness a_2…). Some properties, such as non-negativity, are arguably particularly important when trying to build a descriptive account of people's information search behaviour in naturalistic situations. Non-negative utility

functions are able to intuitively capture the notion that evidence that holds the pattern and/or probability distribution in a given model constant can still be epistemically valuable (Coenen et al., 2018; Evans & Over, 1996; Roche & Shogenji, 2016). The utility functions that do not have the property of non-negativity might, on the other hand, lead one to counterintuitive conclusions. To explain, consider the scenario in which a criminal investigator has three suspects under consideration. Imagine that Suspect A is initially the lead suspect (P (Suspect A) = 70 %) and the remaining two suspects (B and C) have an equal (lower) probability, i.e., P (Suspect B) = P (Suspect C) = 15%. Suppose now that a new piece of evidence, E_1, switches the probabilities of Suspect A and Suspect B while the probability of Suspect C remains the same, so that P (Suspect B | E_1) = 70% and P(Suspect A | E_1) = P(Suspect C | E_1) = 15%. As Suspect B has now replaced Suspect A as the leading suspect (hypothesis), clearly E_1 is of great epistemic value, even though the pattern of the probability distribution given the evidence has remained the same. Given their non-negative features, KL-D and Impact would in this instance remain true to the epistemic value of information, as they do not accrue solely as a result of a change in the 'pattern' of the probability distribution. In this scenario, KL-D and Impact quantify E_1 as having positive utility, in contrast to IG and PG, which would quantify the utility E_1 as 0 since it did not decrease the degree of uncertainty, or Shannon entropy, in the model.

A further demonstration of the potentially problematic nature of non-negative measures arises in scenarios in which a learner receives information (e.g., E_2) that actually reduces their belief in a given hypothesis (i.e., posterior probability estimate is lower than prior probability estimate), and their uncertainty in the environment therefore increases. In this scenario IG and PG would assign a negative utility value to E_2, whereas KL-D would still produce a value of positive utility given that intuitively, *something* was learned, despite leading to more uncertainty in the learner's environment (Coenen et al., 2018). In fact, KL-D and Impact will always return a positive expected utility unless the prior and posterior distributions are exactly the same, in which case it would return 0 (Nelson, 2008). In a criminal investigation scenario, finding out a suspect *is not* the culprit may be pragmatically as important as identifying the person who is. Similarly, in a medical diagnosis scenario, being able to establish that a certain disease (especially if particularly fatal) that was once thought to be the leading hypothesis is

now not very probable, is of extreme value. Therefore, adopting measures such as KL-D and Impact, that do not quantify the value of information simply in terms of reduction in uncertainty might be more appropriate and avoids counterintuitive claims, especially when considering information search behaviour in naturalistic settings (Coenen et al., 2018; Evans & Over, 1996; Roche & Shogenji, 2016). This notion however might not necessarily extend to single-hypothesis scenarios (which are not tested in the present paper) in which negative measures such as IG and PG can act as better predictors of participants' ratings of the utility of query outcomes, compared to KL-D, and have been shown to reflect more closely how participants actually conceive the utility of a given datum, e.g., at times, negatively (Rusconi et al., 2014).

Overall the above points illustrate that quantifying the expected value of an outcome (evidence) even in information-theoretic terms, is not trivial. We note that future research considering information gain measures may therefore benefit from using different types of entropy metrics, beyond Shannon. For example, Crupi and Tentori (2014) discuss information gained based on quadratic entropy. Crupi et al. (2018) further outline different entropy models, obtained from mathematics, physics and other domains, that could be extremely useful in devising a descriptive theory of human information search behaviour.

The *change* or divergence between probability distributions (i.e., prior to posterior beliefs) that utility functions, such as KL-D[2], measure, assumes a Bayesian method of belief updating, such that posterior probabilities $P(h_i|a_i)$ are calculated via Bayes' theorem:

$$P(h_i|a_i) = P(h_i) * \frac{P(a_i|h_i)}{P(a_i)} \qquad \text{Equation 8}$$

In Equation 8, the prior $P\ (h_i)$ represents how likely each hypothesis (h_i) is, and the likelihood $P(a_i|h_i)$ represents how likely it is that a query outcome a_i is observed given h_i is true. The posterior $P(h_i|a_i)$ is therefore a function of the observed outcome a_i and prior knowledge about the likelihood of the hypotheses considered.

[2] For simplicity, throughout this paper we will use KL-D when making illustrative examples regarding utility functions.

Bayesian OED models assume that people not only update their beliefs as described by Equation 8, after finding out the outcome of a query in order to inform subsequent information search decisions, but also that people follow these computations to *predict* the most informative query, before observing any outcome. As such, according to OED principles, when selecting a query people should calculate its expected usefulness by weighting each of the outcome's diagnosticities by the probability of obtaining that outcome (Coenen et al., 2018). This, in turn, depends on the prior probability of each hypothesis, and the conditional probabilities of the outcomes given each hypothesis. Despite the apparent complexity of these computations, OED models have been argued to provide the best available computational-level description of human behaviour in many probabilistic information search tasks (Gureckis & Markant, 2012; Markant & Gureckis, 2012; Nelson, McKenzie, Cottrell & Sejnowski, 2010; Wu, Meder, Filimon & Nelson, 2017).

Notwithstanding the merits of OED models, we argue that alternative information-gathering strategies should be considered in theoretical frameworks of information acquisition as they may capture some richer aspects of human behaviour currently overlooked by OED models. Identifying these alternative strategies or motivators of inquiry, such as obtaining a frontrunner hypothesis at the outset or excluding a hypothesis at the outset, may shed more light on the psychological underpinnings of people's information seeking behaviour in a variety of contexts. This would help fill important gaps in the development of realistic descriptive models of inquiry that account for the information-seekers preferences within different contexts and move beyond standard OED explanations that assume people are integrating across all possible hypotheses and always aiming to maximise the information gained from their actions when determining the most useful item of information (Markant, Settles and Gureckis, 2016).

1.2. Empirical Work and Outstanding Issues

Bayesian OED models have so far been used to describe and predict information acquisition and evaluation in various domains including causal reasoning (Bramley, Lagnado, & Speekenbrink, 2015), eye-movements in visual perception (Najemnik & Geisler, 2009), hypothesis testing (Nelson, 2005), categorization (Meder & Nelson, 2012; Nelson et al., 2010) and children's exploratory behaviour

(Ruggeri & Lombrozo, 2015; Schulz, Gopnik & Glymour, 2007). As such, these models have unified a diverse number of inquiry tasks under a single framework.

Most research that has addressed people's ability to identify useful queries has used a single utility function to calculate each query's usefulness. However, in many evidence-gathering situations, more than one utility function might reasonably apply (Klayman & Ha, 1987; Oaksford & Chater, 2003). Nelson (2005) re-analyzed the tasks in several articles (Skov & Sherman, 1986; Baron et al., 1988; Slowiaczek, Klayman, Sherman & Skov, 1992; Oaksford & Chater, 2003; McKenzie & Mikkelsen, 2007) to identify the predictions of six OED models (employing six different utility functions) of the value of information, on each task. There was high agreement between models on which questions were most (and least) useful, and KL-D made the most exact predictions of people's choices. In a later study, Nelson et al. (2010) simulated environmental probabilities designed to maximally differentiate theoretical predictions of the different utility function, and tested participants' information-seeking behaviour in these environments embedded in a binary categorization task. Results suggested that in this context, PG was the primary basis for the subjective value of information. Overall, more research is needed to disentangle the competing models (Meder & Nelson, 2012), which remains an important issue for the normative analysis of search behaviour and people's sensitivity to the diagnostic value of queries (Crupi et al.,2018; Oaksford & Chater, 1994).

Within psychology, two of the most widely employed tasks to study information search behaviour are the 20-Question game, and the Planet Vuma scenario (for a comprehensive review of these, and other, tasks see Coenen et al., 2018). The 20-Question game is a deterministic task in which there are n persons (hypotheses $h_1 \ldots h_m$) and m binary-outcome features (queries: $Q_1 \ldots Q_m$). The goal in this task is to identify a randomly drawn target person by asking questions about the binary features from a pre-defined set, each pertaining to whether some feature is present or absent in the target person. Researchers have demonstrated that in this task, both children and adults seek information in a Bayesian OED congruent manner (Navarro & Perfors, 2011; Nelson et al., 2014; Ruggeri & Lombrozo, 2015; Ruggeri, Lombrozo, Griffiths & Xu, 2015). Studies have also employed the Planet Vuma scenario, a non-deterministic task in which the goal is to categorize a fictitious alien into one of two species

(hypotheses h_1 and h_2) by querying a pre-specified set of (binary-outcome) features. These studies have similarly reported that people typically have good intuitions about what queries are more informative as quantified by Bayesian OED models (McKenzie, 2006; Nelson, 2005; Nelson et al., 2010; Skov & Sherman, 1986; Slowiaczek et al. 1992; Wu et al., 2017).

Although OED principles provide an adequate computational-level method to account for people's behaviour in these tasks, from a descriptive perspective they lack explanatory power and an ability to fully account for the cognitive underpinnings of query evaluations (Coenen et al., 2018). Even in circumstances in which selection behaviour and OED model predictions are aligned, there remains ambiguity surrounding *how* people select queries that are considered to be normatively optimal by these OED models. For instance, research has identified heuristics that people employ when judging the expected informativeness of queries and has shown that these heuristics closely approximate Bayesian OED model predictions. For example, in non-deterministic tasks, the *feature (likelihood)-difference heuristic* (Nelson, 2005; Slowiaczek et al., 1992) predicts that people select the query with the largest absolute difference in feature likelihoods for either query outcome. This heuristic has been shown to consistently select the query with the highest informative value (measured by an OED model with utility function 'Impact'; Nelson, 2005) in tasks with binary-outcome queries. Similarly, the *probability of certainty heuristic* predicts that in deterministic tasks such as the 20-Question game (again, built with binary-outcome queries), people select the query with the highest probability of an outcome that grants certainty about the true hypothesis. This heuristic has been described as a type of generalized IG model, making analogous predictions to Bayesian OED models (Nelson et al., 2010). Finally, in tasks with large hypothesis spaces, such as the 20-Question game, the *split-half heuristic* agrees with OED principles by identifying the feature that comes closest to being true in half of the hypotheses, as the most informative feature (Navarro & Perfors, 2011; Nelson et al., 2014). More generally, a recent theoretical algorithmic demonstration was given by Nelson, Meder and Jones (2018) illustrating how heuristics may successfully identify queries with maximal informative value, as quantified by different Bayesian OED models, in both one-shot, stepwise and sequential planning tasks.

Given the psychological complexity of OED principles, it seems plausible that people use heuristics to evaluate the utility of queries. In real-world situations involving information acquisition it would be computationally intractable and psychologically implausible to simulate the impact of all possible outcomes on each hypothesis, assuming all possible outcomes are even known (Bramley, Dayan, Griffiths & Lagnado, 2017; Coenen et al., 2018; Huys et al., 2012). The fact that these heuristics have been shown to make predictions corresponding to those of Bayesian OED models ultimately raises concerns about the descriptive abilities of OED models. However, given that these heuristic strategies (i.e., split-half and probability of certainty) are only valid in probabilistic contexts with either binary-outcome queries (i.e., 20 Question game) or binary-hypotheses (i.e., Planet Vuma scenario), more empirical work is needed to investigate the possible strategies that people may employ in differentially motivated tasks and in different probabilistic contexts. As such, the widespread usage of tasks comprising of a binary-hypothesis space or of binary-features may have left an array of heuristics and strategies undetected. In addition, it is crucial to explore the psychological processes and motivations underlying the use of heuristic strategies in order to build a theoretical framework that has both descriptive and predictive value.

As the majority of preceding work has focused on determining whether information search behaviour matches the core predictions of optimal information search models, there is a need to investigate not only what inquiry strategies people use, but also how these are selected in different environments. Other than "OED friendly" heuristics, it is also possible that people use an entirely different set of strategies in order to balance the trade-off between computation, accuracy and processing limits when selecting and evaluating information. Gureckis and Markant (2009) demonstrated that people adopt specific strategies when searching for information in a variation of the task 'Battleship'. These strategies were adapted as they progressed throughout the task, with participants starting with an 'exploratory' strategy that deviated from OED predications, before moving onto a more 'exploitative' strategy at later stages (which followed OED principles more closely). Similar findings were reported by Ruggeri and Lombrozo (2015), who showed that children's question-asking behaviour in a 20-Question game could be accounted for by particular strategies (hypothesis-

scanning and constraint-seeking; Mosher et al., 1966) and that these were adaptively implemented throughout the task.

In addition to discriminatory strategies, people have been found to employ confirmatory strategies, including both integrative and selection (positive testing) biases in favour of a specific leading hypothesis (see e.g., Hahn & Harris, 2014). For example, during sequential learning people often only maintain a single hypothesis, which is adapted, given new evidence (Bramley et al., 2015; 2017; Markant & Gureckis, 2014). Moreover, when choosing interventions to learn about a causal system, people were found to adaptively alter their behaviour between adopting a discriminatory and a confirmatory strategy in order to balance their expected performance and cognitive effort (Coenen, Rehder & Gureckis, 2015). Adopting confirmatory strategies, may come into conflict with the discriminatory nature of OED principles. It is worth nothing however, that certain Bayesian inductive confirmation measures such as L and Z[3], have recently been proposed as quantifiers of confirmation assessments in human reasoning, though further empirical work is still needed to determine whether these models are psychologically plausible (see e.g., Crupi, Tentori & Gonzalez, 2007; Mastropasqua, Crupi & Tentori, 2010; Rusconi et al., 2014). Ultimately, given that information seeking does not occur in a vacuum, confirmatory strategies might be sensible strategies to employ if the single hypothesis addresses a learner's cogitated *goal.* Arguably, what behaviour is considered optimal should depend on the belief-system and goals of the agent. Researching how certain factors including task context, difficulty and framing impact strategy selection during inquiry is crucial in order to help explain and predict inquiry behaviour in a range of different environments, accounting for particular contextual factors and circumstances of the learner.

Many learning problems and information-seeking situations involve a tiered structure of super-ordinate goals as well as subordinate-goals. It is therefore possible that confirmatory and discriminatory strategies may be selectively employed in order to reach different sub-goals, nested under the same super-goal. This fits with the notion that during self-directed learning people divide a problem into

[3]Measure L is connected with the log likelihood ratio measure first conceived by Alan Turing (as reported by Good, 1950, pp. 62–63). Measure Z has been recently advocated by Crupi, Tentori and Gonzalez (2007). For formal definition of these measures see Crupi et al. (2007) and Mastropasqua, Crupi & Tentori (2010).

individual sub-components. For example, in the Battleship task, a learner's super-goal is to find out which ships are hidden. They might break this down by first approximating the ships' locations, and then subsequently determining their sizes. Markant, Settles and Gureckis (2016) carried out an empirical task that resulted in the majority of people decomposing a 3-way categorization task into a series of 2-way classification tasks (sub-goals) despite the super-goal being to learn all three categories. Whereas OED principles can make predictions about how to address each individual sub-goal, they do not naturally capture the process of partitioning a space into subsets of goals, and do not account for the determinants of these sub-goals.

Consider the analogous situations outlined at the beginning of this paper: a physician trying to discern between multiple plausible diagnoses, a crime investigator trying to discern between multiple plausible suspects and an employer trying to discern between multiple plausible interviewees. Although the super-ordinate goal in each case is apparent (i.e., correctly identifying the diagnosis, suspect or candidate), people may introduce different sub-goals and adopt different strategies to achieve the super-goal, such as to initially narrow the hypothesis space down from three to two. For example, one crime investigator might prefer to initially exclude a suspect, whereas another one might prefer to identify a frontrunner suspect at the outset of the investigation. These differential pre-inquiry preferences (e.g., exclude hypothesis at outset or obtain frontrunner) would determine how queries and outcomes are evaluated in ways that, in some cases, could diverge from OED principles. To explain further, within the same probabilistic context a person motivated by 'exclusion' would value the query whose outcomes are more likely to decrease the probability of *one* hypothesis as being more useful or 'informative'. In contrast, someone driven by obtaining a 'frontrunner' would rate that same query as being of less informative value. Identifying the presence of these motivated strategies and establishing whether they could be accounted for within an OED framework merits investigation given that they are at the very core of understanding *how* people select and evaluate queries in information seeking paradigms. Moreover, they are likely to influence how subsequent information is sought and evaluated.

1.3. Current Experiments

In a series of experiments, we investigate how people acquire and evaluate information in a variety of probabilistic contexts, focusing on unearthing the reasoning that underlies people's information seeking behaviour in both one-off and stepwise paradigms. This allows us to move beyond simple demonstrations of OED principles, and help explain and predict information acquisition behaviour in different environments, given the particular strategic preferences of the information seeker. To further our understanding of these processes, we explore not only how people evaluate queries, but also query *outcomes,* an approach that is often neglected in the psychological literature of human inquiry (one exception is Rusconi et al., 2014). Within the OED framework, a query's expected value is a weighted average of the value of each of its possible outcomes, therefore the value of outcomes may be seen as more basic than the value of a query. Exploring how people evaluate outcomes may thus shed light on how they are evaluating queries. For example, the space of outcomes that people consider might strongly influence the value assigned to a query. Moreover, this approach will allow us to explore, in our final experiment (adopting a stepwise paradigm), how receiving unexpected as well as expected outcomes affects belief updating. In all experiments we adopt OED principles to generate statistical environments in which the expected utility of the queries varies within Bayesian OED models with different built-in utility functions (KL-D, IG, PG and Impact) – both negative and non-negative. This allows us to explore people's sensitivity to diverse probabilistic contexts when evaluating queries and outcomes, and to see how well different OED models agree with one another as well as with participants' behaviour.

In addition, to obtain a descriptive account of people's queries and outcome evaluations, we use think-a-loud methods to extract the reasoning explanations attached to their query selections. These methods provide a solid basis for identifying the mental processes underlying complex tasks and can provide rich data on such cognitive processes (Salkind, 2010). Ultimately, they allow us to identify the principal strategies and motivators that underlie participants' information acquisition behaviour, such as obtaining a "frontrunner" hypothesis, and assess factors that influence these strategies.

Although the vast majority of information search tasks in psychology are abstract (e.g., Planet Vuma scenario and variants), we instead embed our models within a more naturalistic crime investigation task. This realism is engaging enough to motivate participants even without the use of specific reward functions. Using a crime investigation task allows us to naturally extract the different motivated strategies that might underlie participants' selections, such as obtaining a 'lead' suspect versus eliminating a suspect, whilst holding the same super-goal of carrying out an effective investigation. Moreover, identifying people's strategic preferences when searching for information in this context could have useful implications for real-world crime investigation, for example, when confirmatory search strategies have been associated with biased case construction and ultimately miscarriages of justice (Eady, 2009; Ormerod, Barrett & Taylor, 2008).

Given the critical role of the first inquiry in stepwise and sequential information-seeking tasks (Nelson et al., 2014; Wu et al., 2017), Experiments 1-3 focused on participants' first (and only) search action in a 'one-shot' paradigm. In Experiment 4 we address additional questions relating to the influence of strategies cognizant at the outset of subsequent search decisions and belief updating, through the use of a stepwise paradigm.

Using a criminal investigation task allows us to render the problem tractable in experiments featuring the one-shot paradigm. As such, being tasked as an investigator who is trying to solve a crime - but has only a single opportunity at collecting evidence - makes the optimal solution to select the query that is most likely to maximise the posterior probability across suspects, given that this equates to the probability of choosing the suspect who is most likely to be the true culprit. The set-up of our one-shot experiments (Experiments 1-3) therefore allows us to directly evaluate the optimality of participants' behaviour in environments in which this optimality is less contentious and tractable. Given the nature of our task, and the use of uniform priors in all of our probabilistic environments, the optimal strategy described above actually equates to PG (as defined in Equation 5). This set-up therefore enables us to assess whether participants' information acquisition behaviour is reflective of any OED measure, and more specifically whether it is 'optimal' when compared to predictions of a PG model. Typically, in tasks employed in the extant psychological literature of human information acquisition it is not always clear which OED measure *should* be employed in a given context. This would depend on how the

measures characterise the goal of the information seeker, for example increasing classification accuracy, as well as how they quantify informational value.

Given the computational burden imposed by the utility functions, which might be infeasible in naturalistic information search situations, we include an additional simplified 'heuristic' version of the PG model[4] for comparison purposes, which we call the Probability Gain Heuristic (PG_H). Including this model allows us to establish whether participants might be reasoning within the realm of OED frameworks, in that they are rationally following the principles of wanting to maximise the chance of obtaining a high posterior probability in the suspect pool, but are doing so via a simplified version of the underlying model. Our PG_H model is defined in the same way as the PG model (see Equations 5 and 6), bar the fact that when calculating a query's expected utility, $P(a_i)$ is defined as $1/n$ where n is the number of outcomes of a given query. This simplifies the computation people have to make significantly compared to the standard way of computing $P(a_i)$ following the law of total probability:

$$P(a_i) = P(a_i|h_i) * (h_i) + P(a_i|\neg h_i) * P(\neg h_i) \qquad \text{Equation 9}$$

Overall, in all our experiments we compare participants' behaviour to the predictions of four different OED models, fitted with different utility functions (KL-D, IG, PG and Impact) and one model fitted with a PG utility function but assuming equal outcome priors (PG_H). All models are parameterized using participants' own beliefs to increase the informativeness of our normative comparisons.

2. Experiment 1

In Experiment 1 we explored people's information-seeking behaviour in four different probabilistic contexts. This experiment was primarily exploratory as we aimed to identify people's search strategies (i.e., obtain a frontrunner vs. eliminate a suspect) and determine how these fit with OED principles. We introduced more complex probabilistic models that were used in previous research, with a ternary hypothesis space and both binary- and ternary-outcome queries. All probabilistic models were based on

[4] We thank an anonymous reviewer for this suggestion.

a three-node Bayesian Network model (BN; Pearl, 1988). BNs are graphical models of uncertainty that model dependencies between hypotheses and items of evidence and can be used to calculate posterior probability beliefs given new states of evidence, utilising Bayes theorem. Our BN comprised of one hypothesis node (identity of burglar) and two query nodes (Burglary Time, Primary Item Stolen), connected in a common cause structure (see Figure 1).

We built four models with different sets of conditional probability tables capturing the prior probability of each query outcome conditioned on each combination of states of the hypothesis node. Each model was integrated into a one-shot information-seeking crime investigation paradigm, described in section 2.2. Informed Bayesian OED (IB-OED) models, parameterized with participants' own stated priors of causes were used as normative benchmarks against which to assess the accuracy of participants' evaluation of queries and outcomes. Additionally, participants' query selection behaviour was classified in relation to different strategies identified through participants' own think-a-loud responses and these strategies were subsequently related back to IB-OED model predictions.

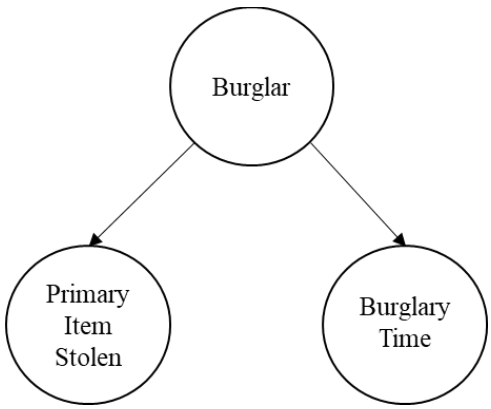

Figure 1. Graphical representation of Bayesian Network

2.1. Bayesian OED Models

Our BNs were built in R using the package gRain (Højsgaard, Edwards & Lauritzen, 2012). Each network had a three-node structure (see Figure 1) with a ternary-state hypothesis node, 'Burglar' (hypotheses: Suspect 1, Suspect 2, and Suspect 3), one binary-outcome query node, 'Burglary Time'

(outcomes: day and night) and one ternary-outcome query node, 'Primary Item Stolen' (outcomes: jewellery, electronics and money).

To fully parameterize the network, we used uniform priors for the hypothesis (Burglar) node: in all models, *P* (Suspect 1) = *P* (Suspect 2) = *P* (Suspect 3) = $\frac{1}{3}$. The conditional probabilities of each state of each query node (Burglary Time and Primary Item Stolen) given each state of the parent node (Burglar) were specified for each Model *i* where $i \in \{1, 2, 3, 4\}$ (see Table 1). In all models the hypotheses were mutually exclusive and exhaustive: one and only one of the suspects committed the burglary. Modelling our probabilistic model as a BN allowed us to uphold the condition of conditional independence ensuring that the evidence in our model was probabilistically independent given the hypotheses (Jarecki, Meder & Nelson, 2013). This is important in the present case as without this assumption the informational utility (i.e., KL-D) of different queries would not be computable from the individual likelihoods.

Table 1

Experiment 1: Conditional Probability Table with parameters employed in each model.

	Model 1			Model 2			Model 3			Model 4		
	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
P (Day| S_i)	0.1	0.9	0.25	0.05	0.95	0.05	0.05	0.95	0.05	0.1	0.9	0.25
P (Night| S_i)	0.9	0.1	0.75	0.95	0.05	0.95	0.95	0.05	0.95	0.9	0.1	0.75
P(Jewellery| S_i)	0.8	0.1	0.1	0.8	0.1	0.1	0.7	0.15	0.15	0.7	0.15	0.15
P(Electronics|S_i)	0.1	0.8	0.1	0.1	0.8	0.1	0.15	0.7	0.15	0.15	0.7	0.15
P(Money| S_i)	0.1	0.1	0.8	0.1	0.1	0.8	0.15	0.15	0.70	0.15	0.15	0.70

N.B. for S_i, *i* is a suspect $\in \{1, 2, 3\}$

Once a probabilistic BN model was built, we added a function that computed the expected utility of each query relative to the probabilistic models specified in Table 1. As such, for each of the four models parameterised as presented in Table 1, we created five versions, each measuring the expected utility of each query and outcome with a different built-in utility function computation (KL-D, Impact, PG, PG_H and IG). As can be seen from Table 2 below, this means that the query predicted to be 'optimal' differed both across utility functions, and across models.

Table 2

Experiment 1: Expected utility value of each query outcome (a_i) and each query (Q_i) predicted by each utility function in each probabilistic model

	Utility Function	a_1 Day	a_2 Night	Q_1 Burglary Time	a_3 Jewellery	a_4 Electronics	a_5 Money	Q_2 Primary Item Stolen
Model 1	KL	0.49	0.33	0.4	0.66	0.66	0.66	0.66
	IG	1.1	1.25	0.4	0.92	0.92	0.92	0.66
	PG	0.72	0.51	0.27	0.8	0.8	0.8	0.47
	PG_H	0.72	0.51	0.28	0.8	0.8	0.8	0.47
	Impact	0.26	0.18	0.21	0.31	0.31	0.31	0.31
Model 2	KL	1.03	0.44	0.65	0.66	0.66	0.66	0.66
	IG	0.55	0.55	0.5	0.92	0.92	0.92	0.66
	PG	0.9	0.49	0.3	0.8	0.8	0.8	0.47
	PG_H	0.9	0.49	0.36	0.8	0.8	0.8	0.47
	Impact	0.38	0.21	0.27	0.31	0.31	0.31	0.31
Model 3	KL	1.03	0.44	0.65	0.4	0.4	0.4	0.4
	IG	0.55	1.14	0.65	1.18	1.18	1.18	0.4
	PG	0.9	0.49	0.3	0.7	0.7	0.7	0.37
	PG_H	0.9	0.49	0.36	0.7	0.7	0.7	0.37
	Impact	0.38	0.21	0.27	0.24	0.24	0.24	0.24
Model 4	KL	0.49	0.33	0.4	0.4	0.4	0.4	0.4
	IG	1.1	1.25	0.4	1.18	1.18	1.18	0.4
	PG	0.72	0.51	0.27	0.7	0.7	0.7	0.37
	PG_H	0.72	0.51	0.28	0.7	0.7	0.7	0.37
	Impact	0.26	0.18	0.21	0.24	0.24	0.24	0.24

For example, in Model 1 KL-D predicts 'primary item stolen' to be the most informative query and in Model 3 KL-D predicts 'burglary time' to be the most informative query. In fact, we selected parameters seen in Table 1 so that according to two utility functions (KL-D and IG), in one model the query 'burglary time' would be more informative than the alternative query (e.g., for KL-D by about 0.25 bits[5]), in another model the query 'primary item stolen' would be more informative than the alternative query (again for KL-D by about 0.25 information bits), and in two models the queries would be of approximately equal informative values (both high or both low). Contrastingly, the prediction of utility functions Impact, PG and PG_H were largely the same across the probabilistic environments, with 'primary item stolen' being of greater informative value compared to 'burglary time' in three scenarios,

[5] This is arguably a 'noticeable' difference and one congruent to the difference in informativeness of features reported by previous studies in the literature (e.g., see Baron et al., 1988; Nelson et al., 2005 and Wu et al., 2017). In Skov & Sherman (1986), a 'low informativeness' feature had KL-D value of 0.001, a 'medium' informativeness feature of 0.08 bits and 'high' informativeness feature of 0.15. Our informative value differences exceeded this significantly.

and of equal informative value to 'burglary time' in one scenario. As previously discussed, given the investigative nature of our task and the parameterisation of our networks, PG and PG_H predictions would be considered to be the optimal solutions in all of the probabilistic environments adopted in the present experiment. This is due to the fact that these measures are motivated by maximising the probability of increasing a suspect's probability of being the culprit as close to 1 as possible, which is intuitively the optimal strategy to employ in a one-shot investigation task.

Overall, our set-up allowed us to explore: a) how the predictions of the most informative query and outcome differed between utility functions; b) people's sensitivity to different probability contexts when evaluating queries and outcomes, and how this relates to the predictions of the various IB-OED models; c) the optimality of participants' decisions when considering PG-based models to be the optimal solutions in the probabilistic environments embedded in the present task; d) the adaptiveness of their search strategies across these contexts; and e) how the choices stemming from their search strategies related to the different IB-OED model predictions. Moreover, it allowed us to explore people's preferences for a 'frontrunner' strategy versus an 'elimination' strategy given that, for example, according to KL-D and IG one query would guarantee the identification of a frontrunner (primary item stolen), and the alternative query (burglary time) could, given a certain outcome, lead to the identification of a higher frontrunner, but, given a different outcome, it mainly helped eliminate a suspect.

The values in Table 2 were computed as described in equations 2-7. To illustrate, in Model 1, using KL-D the expected utility of the query according to 'burglary time 'was computed by first computing the expected utility of outcome 'day' as:

$$KL-D\ (day) = P(Suspect\ 1\ |day)\ log_2\ \frac{P(Suspect\ 1\ |day)}{P(Suspect\ 1)} + P(Suspect\ 2\ |day)\ log_2\ \frac{P(Suspect\ 2\ |day)}{P(Suspect\ 2)} + P(Suspect\ 3\ |day)\ log_2\ \frac{P(Suspect\ 3\ |day)}{P(Suspect\ 3)} = 0.49$$

Subsequently computing the expected utility of outcome 'night' as:

$$KL - D\ (night) = P(Suspect\ 1\ |night)\ log_2\ \frac{P(Suspect\ 1\ |night)}{P(Suspect\ 1)} + P(Suspect\ 2\ |night)\ log_2\ \frac{P(Suspect\ 2\ |night)}{P(Suspect\ 2)} + P(Suspect\ 3\ |night)\ log_2\ \frac{P(Suspect\ 3\ |night)}{P(Suspect\ 3)} = 0.33$$

And finally utilising these values to compute the expected utility of the *query* as:

$$KL - D(burglary\ time) = P(day) * KL(day) + P(night) + KL(night) = (0.42 * 0.49) + (0.58 * 0.33) = 0.40$$

This could then be compared against the computed KL-D for primary item stolen to evaluate queries. Similar steps were carried out to compute the usefulness of the queries and outcomes according to IG, PG and Impact, utilising the pertinent equations (equations 4-7) previously outlined. As mentioned above, we included a fifth model, PG_H, that used a PG utility function and assumed all outcomes had equal priors and thus were equally likely to occur given that query's selection. This measure was defined in the same way as PG (defined in Equations 5 and 6) except for the fact that when computing a query's utility, the probability of an outcome, $P(a_i)$ was defined as $1/n$ where n was the number of outcomes of a given query.

2.2. Method

Here we present the general methods used in Experiments 1-3.

2.2.1. Participants

We tested 264 participants (n_{males} =88 males, M_{age} =34.8 years; SD = 11.9) who were recruited from Prolific Academic (www.prolific.ac.uk) and completed the study online on the Prolific Academic platform. All participants were native English speakers, who gave informed consent, and were compensated $1.20 for taking part in this experiment, which took on average (median) 13.2 minutes to complete.

2.2.2. Design and Materials

A between-subject design was adopted. Participants were randomly allocated to one of four conditions ($n_{\text{Condition 1}} = 66$, $n_{\text{Condition 2}} = 67$, $n_{\text{Condition 3}} = 64$, $n_{\text{Condition 4}} = 64$). All participants were presented with the same cover story in which they acted as crime investigators in a specialized burglary division. Participants in each Condition i (C_i) were required to reason with a Model i, where $i \in \{1, 2, 3, 4\}$, parameterized as outlined in section 2.1 so that the expected informative value of the two queries differed across OED models with different built-in utility functions and within some of these, the expected utility differed across probabilistic environments (see Table 1 and Table 2), and completed the same one-shot task described in the subsequent section (for an example of task materials see osf.io/tkr4v).

2.2.3. Procedure

Participants in each condition were initially presented with a cover story within which they were asked to imagine they were crime investigators. They were told that they were being transferred to the burglary division of a different neighbourhood and, before being involved in any new investigation, they were required to review the neighbourhood's burglary statistics and the criminal records of the (three) burglars known to operate in the area. The criminal records of the burglars contained information on the 'modus operandi' they utilised in past burglaries, in relation to the time of day they operated in and the items they primarily stole. As such, participants were provided with the percentages (likelihoods) that each burglar operated during the night (10 pm to 10 am), or during the day (10 am to 10 pm), primarily stole electronics, money or jewellery. They were told these percentages were based on all the burglaries that each burglar had ever committed in the area and that each burglar had committed an equal number of burglaries. In this manner, participants in each condition were given information on their respective model (i.e., variables present, causal relationships between these, uniform priors of hypotheses and conditional probabilities within the model). This information was presented to participants in both textual and tabular format in an accessible manner and was made available to them throughout the task. Participants were also provided with explicit instructions on the mutually exclusive and exhaustive nature of the hypotheses.

After having reviewed this information, participants were told that a new burglary had occurred in their neighbourhood and they were asked to investigate the new case. At this point, prior probabilities of each burglar being the culprit of this burglary were elicited from participants to see if the uniform priors had been accepted. Subsequently, participants were asked to select one of two investigative queries: 'burglary time' (to find out whether the burglary occurred during the day or night) and 'primary item stolen' (to find out whether electronics, money or jewellery were primarily stolen), keeping in mind they were able to make only one investigative inquiry throughout the task. The query selection question was asked in a manner that would not prime participants to adopt a particular strategy: "Please choose the query that you believe will be most useful for this investigation". Following the query selection, participants provided a textual explanation for their choice in response to the question: "Please explain the reasoning behind your choice in as much detail as you can in the text box below". No word limit was imposed. Subsequently, participants proceeded to indicate on a Likert scale ranging from (0 - not useful at all to 10 - extremely useful), the usefulness of each query ('burglary time' and 'primary item stolen') as well as of each query outcome ('day', 'night', 'money', 'jewellery' and 'electronics'). Participants did not find out the outcome of the query they selected and were not required to make any judgments on the culpability of the suspects, when selecting a query; they were thus required to evaluate its *expected* value. The task ended once query and outcome ratings were elicited.

2.3. Results

2.3.1. Prior Probabilities

The percentage of participants who correctly[6] estimated the prior probabilities of all three suspects was 76% in Condition 1, 76% in Condition 2, 73% in Condition 3 and 81% in Condition 4. These high percentages allow us to conclude that participants overall accepted the uniform priors given to them. Nonetheless, as previously mentioned, all analyses will evaluate participants' behaviour against individually fitted models parameterized with their own stated priors.

[6] In all experiments, an estimate was considered to be correct if it fell within ± 2.5% of the normative estimate (in this case ~33.3%)

2.3.2. Query Selections

Within each condition, we obtained the proportion of participants who selected each query (see Figure 2). Pearson's Chi-Square test of independence indicated that these proportions did not differ between conditions, χ^2 (3) = 2.45, p = 0.48. As can be seen from Figure 2, the majority of participants in each condition preferred querying 'primary item stolen', thus suggesting that people may not be sensitive to the change in a query's informative value within different probabilistic environments (according to all measures the two queries varied in informative value across conditions – see Table 2).

For details on the percentage of 'correct' query selections within each condition according to each utility function see supplementary materials (S1).

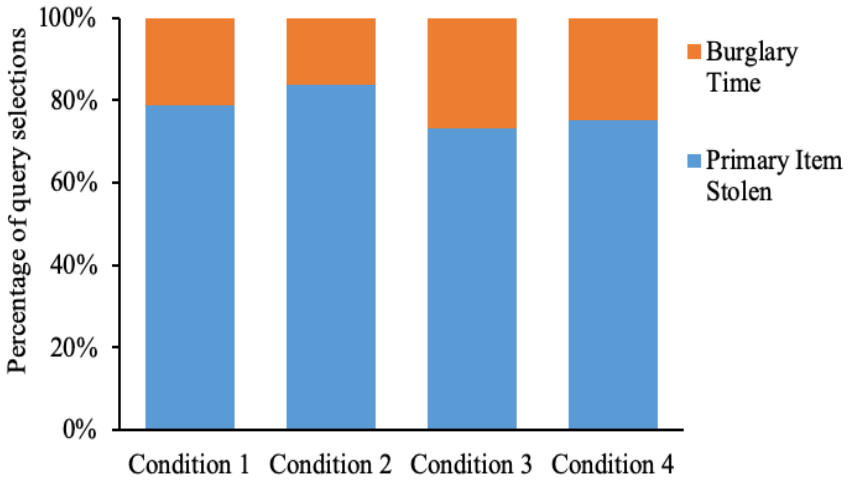

Figure 2. Experiment 1: Percentage of participants who selected each query in each condition.

2.3.3. Utility Function Model Comparisons

The breakdown of the percentage of participants for whom each utility function predicted each query to be the most informative, or for them to be equally informative[7] in each condition can be seen in Table 3.

[7] Two queries were deemed to be of pragmatically equally informative value if they were within 0.05 bits of each other. This was done to increase the fairness of our comparisons by not expecting participants to notice a difference in the expected informative value of queries if they were within a certain range of each other.

To ascertain how well each utility function is able to predict people's choice proportions (seen in Table 3), we built mixed-effects logistic regression models for each utility function, using the package *lme4* in R (Bates et al., 2012). All models were fit by maximum likelihood estimation and had an underlying binomial distribution. Our model-fitting procedure started by initially building a null model (M_0) including a random-effect with intercept for 'Subject' only and 'Participant Choice' as our outcome variable. In addition, we built as a model (M_1) that included 'Scenario' as a sole fixed-effect predicting our outcome variable, in order to ascertain whether the distribution of participants' choices varied across scenarios (M_1). A likelihood ratio test between M_1 and M_0 confirmed the findings we presented in section 2.3.2 illustrating that participants' query selections did not significantly vary across scenarios, χ^2 (3) = 2.45, p = 0.48.

Table 3

Experiment 1: Percentage of predictions made by each utility function in each condition favouring burglary time, primary item stolen, or evaluating them as equal, and percentage of participants who queried Burglary Time (represented in Burglary Time > Item Stolen column) and Item Stolen (represented in Item Stolen > Burglary Time column).

	Utility Function	Burglary Time > Item Stolen	Item Stolen > Burglary Time	Item Stolen = Burglary Time
Condition 1	KL-D	0%	100%	0%
	IG	0%	100%	0%
	PG	0%	94%	6%
	PG_H	1.5%	92.5%	6%
	Impact	0%	89.4%	10.6%
	Participant Choice	21.2%	78.8%	-
Condition 2	KL-D	0%	12%	88%
	IG	0%	12%	88%
	PG	0%	94%	6%
	PG_H	3%	92.5%	4.5%
	Impact	0%	13.5%	86.5%
	Participant Choice	16.4%	83.6%	-
Condition 3	KL-D	95.5%	1.5%	3%
	IG	95.5%	1.5%	3%
	PG	3%	82%	15%
	PG_H	10.5%	4.5%	85%
	Impact	8%	10%	82%
	Participant Choice	26.9%	73.1%	-
Condition 4	KL-D	0%	9.4%	90.6%
	IG	0%	9.4%	90.6%
	PG	0%	94%	6%
	PG_H	1.5%	95%	3.5
	Impact	0%	9.4%	90.6%
	Participant Choice	25%	75%	-

After building our intercept-only model, we iteratively increased model complexity by including the pertinent 'Utility Function' as the only predictor (M_2), both 'Scenario' and 'Utility Function' as predictors (M_3) and finally 'Scenario', 'Utility Function' and the interaction 'Scenario * Utility Function' as predictors (M_4) of our outcome variable ('Participant Choice'). All models included a random effect with intercept for 'Subject' to account for within-subject correlations. The iterative process was stopped, and a maximal model was chosen, when the likelihood ratio test showed no improvement from the preceding model. For all utility functions, the maximal model was M_2; adding 'Scenario' as a predictor did not improve any model's fit. All maximal models were checked for overdispersion and under dispersion and no issues were noted.

The outputs of the mixed-effect logistic regression analyses used to assess the predictive abilities of each utility function can be seen in Table 4 below. Through these analyses we found PG and PG_H to be significant predictors of 'Participant Choice': PG, F (2, 261) = 8, p = 0.001; PG_H, F (2, 261) = 5.4, p = 0.005. In contrast, KL-D/IG, F (2, 261) = 0.7, p = 0.5; and Impact, F (2, 261) = 1.8, p = 0.17; were not significant predictors of 'Participant Choice'.

These findings can be contextualised within the information presented in Table 3. As can be seen, PG and PG_H more closely approximate the distribution of participants' query choices by predicting 'primary item stolen' to be of greater (or equal in the case of PG_H, condition 3) value than 'burglary time', thus reflecting participants' persistent majority preference for 'primary item stolen' across conditions. Comparatively, KL-D and IG models predict 'burglary time' to be more informative in condition 2 and predict the two queries to be of equal value in two other conditions, thereby not reproducing the distribution of participants' preferences in these conditions. Finally, Impact displayed an overall lack of discriminative capacity by evaluating the two queries to be of equal value in three out of four conditions – ultimately also not reflecting participants' query preferences.

By looking at the odds ratio (OR) values in Table 4 below we can see that in the PG model, a prediction of 'primary item stolen' made an equivalent participant choice of 'primary item stolen' 6.26 times more likely than a participant choice of 'burglary time'– these odds are significantly higher than

those of a PG prediction of 'item time', and 'burglary time'. Similarly, in the PG_H model, a prediction of 'primary item stolen' made an equivalent participant choice of 'primary item stolen', 7.9 times more likely than a participant choice of 'burglary time'. Similarly, a PG_H prediction of 'burglary time' made a participant choice of 'burglary time' 4.9 times more likely than a participant choice of 'primary item stolen'. Comparatively, the OR values of KL/IG and Impact predictions are noticeably smaller, intimating they are worse predictors of participants' query choices.

Table 4

Experiment 1: Parameters of the fixed effects estimated via logistic mixed-effects models, their statistical significance, and odds ratio for the competing models.

Participant Choice = [8]Item; Reference category 'Participant Choice'= Time								
Model[1]	Parameter	Estimate ß	Std. Error ß	*t*	Sig.	Odds Ratio	OR 95%CI Lower	OR 95%CI Upper
Probability Gain	(Intercept)	-0.37	1.45	-0.25	0.8	-	-	-
	'Item'	1.83	0.47	3.95	< 0.0001	6.26	2.5	15.8
	'Time'	0.37	1.49	0.25	0.81	1.44	0.008	27.5
	'ItemTime'[a]	0[b]	-					
Probability Gain Heuristic	(Intercept)	-0.56	0.63	-0.36	0.72		-	-
	'Item'	2.06	0.65	3.1	0.002	7.9	2.14	28.9
	'Time'	1.54	0.69	2.2	0.027	4.7	1.2	18.2
	'ItemTime'	0[b]	-			-		
KL-D /IG	(Intercept)	1.42	1.46	0.97	0.33	-	-	-
	'Item'	-0.24	0.35	-0.7	0.48	0.78	0.39	1.6
	'Time'	-0.4	0.37	-1.1	0.26	0.66	0.32	1.4
	'ItemTime'	0[b]	-	-	-	-	-	-
Impact	(Intercept)	1.2	1.4	0.86	0.39	-	-	-
	'Item'	0.16	0.33	0.49	0.62	1.18	0.61	2.27
	'Time'	-1.6	0.93	-1.7	0.08	0.19	0.03	1.2
	'ItemTime'	0[b]	-	-	-	-	-	

[a] 'ItemTime' reflects a prediction of the two queries having equal value defined as an abs.diff < 0.05)
[b] Parameter is set to zero due to redundancy.
[1] Participant Choice ~ Utility Function Prediction + (1 | Subject)

This notion is additionally confirmed by the likelihood ratio results between each maximal model and M_0 presented in Table 5 below, which indicate that the only two models that significantly improved the null model are PG and PG_H.

[8] In our model comparisons we shortened the variable name 'primary item stolen' to 'Item' and 'burglary time' to 'Time'

In order to compare the competing utility functions models and select the best approximating models, we used derivatives of Akaike's Information Criterion (AIC) measure. The individual AIC values are not interpretable in absolute terms given that they contain arbitrary constants and are affected by sample size.

Table 5

Experiment 1: Likelihood ratio test results, AIC, Deviance, Akaike Weights (w) and Evidence Ratio (ER) values of the competing models.

Model	df	AIC	ΔAIC_i	w_i	ER_i	Deviance	χ^2	df	*p*-value
M_0[1]	2	284.5	-	-	-	280.5			
M_2 PG	4	272.2	0	0.92	1	264.2	16.34	2	< 0.0001
M_2 PG_H	4	277.1	4.9	0.08	11.6	269.1	11.4	2	0.003
M_2 KL/IG	4	287.2	15	0.0005	1808	279.2	1.3	2	0.51
M_2 Impact	4	284.9	12.7	0.0016	572.5	276.9	3.6	2	0.17

[1] Participant Choice ~ 1 + (1 | Subject)

In order to compare the different models and measure how much better the best approximating model is compared to the next best/alternative models, the first step therefore involved rescaling the AIC and compute ΔAIC_i by subtracting from the AIC of each model the AIC of the model with the smallest AIC value:

$$\Delta AIC_i = AIC_i - AIC_{min}$$

This transformation forces the best model to have ΔAIC=0 while the rest of the models have positive values. Although not a definitive rule, a coarse guide is that models with ΔAIC values less than 2 are considered to be essentially as good as the best model, M_r, and models with ΔAIC values of up to 6 should not be discounted (Richards, 2005). Above this, model rejection might be considered, and models with ΔAIC greater than 10 are considered implausible (Burnham and Anderson, 2004). By consulting Table 5 above, we can deduce that PG was the best model (M_r), PG_H was a contender and should not be discounted, and KL/IG and Impact models should be discounted and can be regarded as implausible models of participants' query choices. Importantly, ΔAIC can be used to calculate two additional measures used to assess the relative strengths of each candidate model (Burnham and Anderson 2004). The first measure follows an information theoretic approach because it is based on KL

divergence, which is used to represent the information lost when model M_i is used to approximate full reality (f). From the differences in AIC values, ΔAIC, an estimate of the relative likelihood L of model i can be obtained by the simple transform:

$$L(M_i|data \propto exp\{-0.5\,\Delta AIC_i\}$$

Where ∝ stands for 'is proportional to'. In the last step, the relative model likelihoods are normalized (i.e., divided by the sum of the likelihoods of all models) to obtain Akaike weights (w_i) (e.g., Burnham & Anderson, 2004), where:

$$w_i = \frac{\exp(-0.5\,\Delta AIC_i)}{\sum_{r=1}^{R}\exp\,(-0.5\;\Delta AIC_r)}$$

The Akaike weight is a value between 0 and 1, with the sum of Akaike weights of all models in the candidate set being 1, and can be considered as analogous to the probability that a given model is the best approximating model (although there are some who disagree with this, see e.g., Bolker, 2008; Link and Barker 2006; Richards, 2005). From looking at Table 5 we can see that the PG model has a 92% chance of being the correct model. Given that almost all of the weight lies in one model, we can conclude that we have low model selection uncertainty and can be confident of PG's predictive abilities.

The 'evidence ratio' (ER) can be computed as a measure of how much more likely the best model (Δ_r) is to be the best approximated model, than model i:

$$ER = \frac{\exp(-0.5\Delta\,AIC_r)}{\exp(-0.5\,\Delta\,AIC_i)}$$

According to ER, our reference model PG is 11.6 times more likely than our next best model, PG_H. This is likely due to the fact that, as seen in Table 3, PG correctly predicted a majority of participant choices to be 'primary item stolen' in all scenarios, whereas PG_H demonstrated less discriminative capacity by predicting the queries to be of equal value in one scenario. In the probabilistic environments adopted in the present experiment, it therefore seems that participants are choosing queries in line with the optimal

task strategy of maximizing their chances of maximizing the posterior of one of the suspects, as dictated by a PG measure.

2.3.4. Query and Outcome Ratings

In the task participants rated both queries and outcomes on a scale ranging from 0 - not useful at all to 10 - extremely useful. Participants' average ratings of the usefulness of each query as well as each query outcome can be seen in Table 6 below. A one-way ANOVA showed no significant between-condition difference in the average usefulness ratings of the queries 'burglary time', F (3, 263) = 0.4, p = 0.76 η_p^2 = 0.04, or 'primary item stolen', F (3, 263) = 0.9, p = 0.45, η_p^2 = 0.01.

Table 6

Experiment1: Mean participant ratings of the usefulness of each query and query outcome per condition on scale ranging from 0 to 10.

		Condition 1	Condition 2	Condition 3	Condition 4
		M (SD)	*M* (SD)	*M* (SD)	*M* (SD)
Query	Burglary Time	6.08 (2.3)	5.7 (2.2)	5.9 (2.2)	6 (2.3)
	Primary Item Stolen	7.8 (1.9)	8.1(1.8)	8.1 (1.8)	7.6 (1.9)
Query Outcomes	Night	5.7 (2.3)	5.3 (1.8)	5.4 (2.1)	5.6 (2.5)
	Day	7.2 (2.1)	7.8 (2)	8.1 (2.4)	7.3 (2.3)
	Jewellery	7.8 (1.7)	7.7 (1.7)	7.8 (2)	7.4 (2)
	Electronics	7.8 (1.7)	7.8 (1.8)	7.9 (1.8)	7.2 (2)
	Money	7.8 (1.7)	7.8 (1.7)	7.7 (2.1)	7.3 (2)

Arguably, participants' actual ratings of the utility of queries and their outcomes did not reflect those computed by any of the utility functions. For example, in contrast to participants' ratings, KL-D and IG predicted that query 'primary item stolen' to be most useful in Condition 1, and 'burglary time' to be most useful in Condition 3. Furthermore, although PG's and PG_H's higher expected utility for 'primary item stolen' in Condition 2 and 4 is comparable to participants' ratings of the usefulness of this query in these conditions, both of these models predicted the two queries to be of approximately equal value in Condition 3, which is not mirrored in participants' ratings. Finally, Impact predicted the two queries to be of approximately equal value in Conditions 2-4, which again is not reflective of

participants' consistently higher rating of the usefulness of 'primary item stolen' across conditions. Despite this however, as shown in Section 2.3.3, PG-based models were able to predict the qualitative direction of participants' query selections better than the alternative models.

In terms of participants' evaluation of query outcomes, no between-condition differences were found in the usefulness ratings of query outcome 'night', F (3, 263) = 0.5, p = 0.67, η_p^2 = 0.006; query outcome 'day', F (3, 263) = 2.4, p = 0.07, η_p^2 = 0.03; query outcome 'jewellery', F (3, 263) = 0.76, p = 0.52, η_p^2 = 0.009; query outcome 'electronics', F (3, 263) = 2.2, p = 0.08, η_p^2 = 0.025; or query outcome 'money', F (3, 263) = 0.9, p = 0.41, η_p^2 = 0.01. When comparing participants' ratings of outcomes to those predicted by the utility functions (see Table 2 and Table 6), all utility functions except IG reflect participants' evaluation of an outcome 'night' being less informative than a 'day' outcome for a 'burglary time' query in all conditions. All utility functions captured participants' evaluation of the three 'primary item stolen' outcomes as having equal utility in all conditions. Despite this, as proven by the above analysis, participants' ratings of query outcomes did not vary across conditions, which is not reflective of the computations of any of the utility functions.

To confirm that participants' usefulness ratings were representative of how they actually evaluated a query by either selecting it or not selecting it, we computed the percentage of "rating congruent" responses in each condition. A query choice was coded as congruent (1) if the participant selected the query that they also rated as being most useful on the 0-10 Likert scale. If not, a query choice was coded as incongruent (0). If a participant gave equal ratings to the two queries, their query choice was coded as congruent regardless of what query was selected. The percentage of congruent query selections was: 97% in Condition 1; 100% in Condition 2; 95.5 % in Condition 3 and 98.4% in Condition 4. These high percentages allow us to take participants' ratings as reliable representations of their evaluation of how useful they believe a query to be.

Overall, these findings suggest that, in the probabilistic environments that we embedded in this one-shot criminal investigation task, participants' query selections are mostly aligned with models based on probability gain (PG and PG_H), which we acknowledged as the optimal solutions to the task. Despite

this, participants' actual ratings of the informative value of outcomes was found to mostly deviate from those computed by our utility functions of interest.

2.3.5. Strategies: Think-a-loud responses

In order to obtain an understanding of the reasoning underlying participants' query selections and evaluate whether they aligned with the goals of any utility function, we analysed participants' think-a-loud responses. Each participants' think-a-loud response, explaining their reasoning for selecting a given query (and thus anticipating it to be more informative), were initially qualitatively analysed and coded with a single code that simultaneously categorized, summarised and accounted for the response (Charmaz, 2006) by a primary rater. Each think-a-loud response was therefore attributed a code, drawn directly from the response and not a pre-existing set, which acted as a descriptive label of an identified strategy. These strategy codes were derived from explicit statements indicating a motivation of obtaining a desired outcome as well as explanations of a systematic form of reasoning or motivation. The list of strategies (with a criteria description of each) was used to finalise a coding scheme that was agreed upon by a second independent rater who subsequently coded 50% of the total sample of responses (n = 134) being blind to condition and the query selection attached to a reasoning response. The second rater was a post-doctoral researcher familiar with qualitative methods but with minimal information on the scope of the present experiment. Cohen's weighted kappa was utilised to determine a high inter-rater agreement between the two raters, $\kappa_w = 0.81$, $p < 0.001$ in the strategy codes over imposed to participants' responses. The strategy codes we drew from our participant sample (with a description of each and frequency across conditions) can be seen in Table 7. For a graphical representation of this information broken down by condition see Figure 3 below. Responses of 41 participants (15%) out of the total sample were given a code of "n/a" as they did not provide an elaborate enough think-a-loud response for us to attribute it a specific code[9]. Subsequent analysis is carried out on the total sample (264 responses), although the "n/a" code will not be described further.

[9] In all experiments, responses that were attributed a code of "n/a" typically comprised of nonsensical letters, did not state an underlying reason for their selection, e.g., "it was easier", or stated one that did not relate to the present set-up, e.g., "could trace merchandise through pawn shops".

Table 7

List of strategies extracted from think-a-loud responses with a description of these, an example response coded with each strategy, and the percentage of participants each strategy accounted for across conditions.

Strategy Code	Description	Example	Frequency
Frontrunner	Explicitly indicating a preference for identifying a lead suspect and thus obtaining a frontrunner at the outset.	P8: "querying primary item stolen will narrow down my search to one main suspect".	26.5%
Symmetry	Preference for query with 'symmetric' parameters, i.e., in which a different burglar primarily accounts for each feature/outcome.	P134: "each burglar has a preferred 'main category' of items they like to steal".	21.2%
Differentiation	Preference for query with the most 'percentage difference' in outcomes across hypotheses and interest in maximally differentiating or disambiguating the hypotheses.	P107: "this was the most differentiating fact between the three burglars. I wouldn't choose the time of day, because 2 of the 3 burglars perform burglaries during the night".	17.4%
Frontrunner + Zero-sum/Risk Aversion	Preference for the query that is less 'risky' given that regardless of the outcome, it increases the probability of one suspect over the others.	P55: "[Suspect 1] and [Suspect 3] prefer night-time so if the robbery took place at night it would be difficult to distinguish between who did it. The items that the robbers took is more likely to point at one single culprit"	7.2%
Elimination	Preference for eliminating or excluding a suspect from the hypothesis space at the outset.	P103: "will help eliminate the likelihood of one of the suspects committing the crime."	5.3%
Highest Percentage	Preference for the query that has the *outcome* with the highest percentage for any given suspect.	P181: "[Suspect 2] commits 95% of his crimes at night, it will provide me with the best evidence".	3.4%
Zero-sum/ Risk Aversion	Avoiding the query whose outcomes would be almost equally diagnostic towards two suspects ("zero-sum reasoning" [10]) and selecting the query whose outcome has lower evidential value but was more likely to occur ("risk aversion"[11]).	P.145: "although the percentages for night/day are more severe, if it's day, there is no way of telling which of the two it is".	3.4%

[10] In game theory, 'zero-sum' describes a game where one player's gain is a loss to other players. The zero-sum fallacy in evidence evaluation occurs when evidence is dismissed as non-probative if it lends equal support to two competing hypotheses (see Pilditch, Fenton & Lagnado, 2019).

[11] We define "risk aversion" as behaviour that occurs when in the trade-off between choosing a query that holds an outcome of highest evidential value (but lower probability of occurrence) and a query that holds an outcome of lesser evidential value (but higher probability of occurrence), participants prefer the latter.

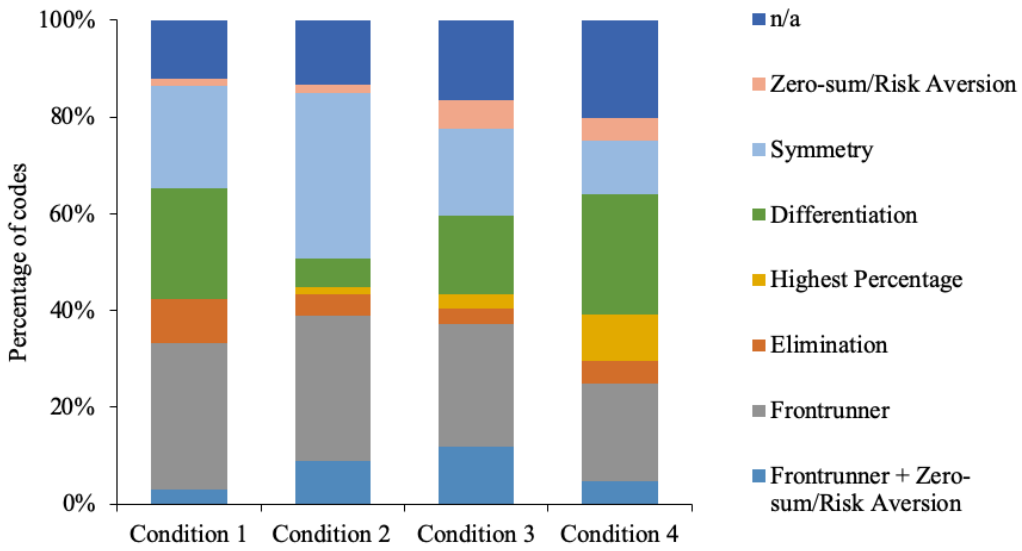

Figure 3. Experiment 1: Percentage of strategies within each condition.

2.3.5.1. Adaptability of Strategy Use

In order to explore the adaptability of participants' strategies across the different probabilistic contexts, we conducted a Chi-square test of independence on the percentage use of each strategy. Results showed a significant difference in the percentage of participants who adopted the different strategies between the four conditions, $\chi^2(18) = 34.1$, $p = 0.01$, $V = 0.23$. Bonferroni-corrected post-hoc comparisons[12] however, illustrated that the only strategy whose usage varied across conditions was 'symmetry', $\chi^2(3) = 21.1$, $p < 0.0001$. This could be due to the fact that more participants utilised a 'differentiation' strategy in this condition, and 'symmetry' and 'differentiation' are both concerned with similar features of a query (i.e., each outcome being diagnostic of a different suspect) and might be used somewhat interchangeably by participants. The extent to which participants adopted a frontrunner, differentiation, highest percentage, and frontrunner + zero-sum/risk aversion or elimination strategy did not vary across the four conditions. Post-hoc analysis could not be carried out on the strategy 'zero-sum-risk aversion' given the low numbers present within each cell.

[12] In all experiments, all post-hoc comparisons utilised Bonferroni corrections at the level of α/m where α is 0.05 and m is the number of hypotheses being tested.

Overall, it therefore seems that similarly to participants' query selection choices, the strategies participants employed also remained mostly fixed across the different probabilistic environments, with 'frontrunner', 'symmetry' and 'differentiation' accounting for approximately 65% of participants in each condition.

2.3.5.2. Strategies and Query Selection

Next, we investigated whether specific strategies systematically underlie specific query selections, in order to ascertain whether the observed query selection preferences, could be accounted for by these additional strategies. When collapsing across conditions, 86% of participants who utilised an 'elimination' strategy, and 89% of those who utilised a 'highest outcome' strategy, selected the query 'Burglary Time'. Comparatively, 86% of participants who utilised a 'frontrunner' strategy, 96% of participants who adopted a 'risk aversion' strategy (including frontrunner + risk aversion), 94% of participants who adopted a 'differentiation' strategy and 95% of participants who adopted a 'symmetry' strategy', selected the query 'Primary Item Stolen'. This suggests that the vast majority of participants' query selection preferences can be accounted for by the strategies we identified.

An omnibus test (Chi-Square test of Independence) revealed no significant difference in the overall distribution of strategies underlying the two query choices between conditions, $\chi^2(25) = 22.14$, $p = 0.63$. This suggests that certain strategies are associated with particular queries (i.e., 'frontrunner', 'differentiation', 'symmetry' and 'zero-sum/risk aversion' to 'Primary Item Stolen' and 'elimination' and 'highest percentage' to 'Burglary Time'). Similar to how participants' preference to select the query 'Primary Item Stolen' was irrespective of probabilistic contexts, the association between strategy and query selection also did not vary across contexts.

For details on the extent to which these strategies related to accuracy of query selections according to each of the utility functions see supplementary materials (S2).

2.4. Discussion

Experiment 1 explored people's information search behaviour in four different probabilistic contexts, with differing informative values for the queries, as measured by four utility functions (KL-D, IG, PG

and Impact) and a simplified heuristic PG model dubbed PG_H that assumed equal outcome priors for each query. Utilising a between-subject design we ascertained that, despite the fact that participants in each condition reasoned with a different probabilistic model, the majority of participants in all conditions preferred querying 'primary item stolen'.

Using logistic mixed-effect models, we showed that probability gain based models (PG and PG_H) best approximated the distribution of participants' query choices across probabilistic environments. PG was the best predictor of participants' choices, predicting 'primary item stolen' to be of greater informative value in each condition. Simplifying the PG model by assigning equal priors of outcomes to the queries did not improve the predictive abilities of the model, thus suggesting that in the probabilistic environments adopted in the present experiment, participants might not be suffering from integration errors at the level of estimating outcome priors when evaluating the informativeness of queries. Overall, our model comparisons suggest that participants' information search behaviour aligns with the optimal solution to this one-shot investigative task which, as represented by a PG model, entails maximising one's chance of obtaining a correct suspect classification. This makes intuitive sense given that when tasked as a criminal investigator with only one opportunity of obtaining evidence to disambiguate between suspects, one should favour maximising the posterior probability of a suspect and therefore the chances of selecting the culpable suspect.

Our analysis of think-a-loud responses suggested that underlying participants' modal querying of 'primary item stolen' was a common preference for maximally differentiating between suspects and obtaining a frontrunner hypothesis. However, adopting a strategy purely motivated by the identification of a frontrunner with the highest posterior should have led participants to correctly prefer, and rate as more informative, the query 'burglary time' in Condition 3, which instead was not what we observed. This showed that participants were willing to obtain a frontrunner, though only if this was also the 'safer' and more probable option. This strategy is arguably pragmatically reflective of the optimal strategy of maximising the *probability* of making a correct guess, akin to PG, rather than wanting to maximise absolute belief change as the Impact model would dictate.

Although only a minority of participants explicitly stated this in their think-a-loud responses, selecting the query 'primary item stolen' even in conditions in which querying 'burglary time' could have led to the identification of a frontrunner with a higher probability of being the culprit could also be the product of a frontrunner preference mitigated by risk aversion. This finding fits with work by Poletiek and Berndsen (2000), who conceptualised hypothesis-testing behaviour as risk-taking behaviour and illustrated that people displayed certain biases (though in the risk-taking direction) when carrying out a pre-posterior analysis of the probability of obtaining supporting evidence and the evidential value of this evidence. Though not in line with the findings of Poletiek and Berndsen (2000), the 'frontrunner risk aversion' tendency described above echoes the results of Skov and Sherman (1986) who frame this tendency as being a form of confirmatory strategy behaviour (we will explore this more directly in our stepwise information search paradigm – Experiment 4). In the present experiment, we found this inclination to be additionally associated with a form of zero-sum reasoning. As such, in our experiment participants perceived the outcome 'night' as being less useful than all other query outcomes, in all conditions, given that this outcome was equally diagnostic of two suspects, though drastically lowering the probability of the third suspect (to a point in which they could be pragmatically eliminated). This suggests that a) participants are not significantly driven by elimination strategies, a finding supported by our analysis of think-a-loud responses, and b) outcomes that are equally diagnostic of two suspects (e.g., that would lead to equal posteriors), were deemed to be significantly less useful. This may be related to a form of zero-sum reasoning, defined as the dismissal of potentially probative evidence because it cannot differentiate between two competing hypotheses (Pilditch, Fenton & Lagnado, 2019). In our set-up, with three hypotheses, this form of thinking led to the dismissal of outcome 'night' as probative, despite it lowering the probability of one suspect being guilty to a level whereby he can almost be eliminated.

Overall Experiment 1 illustrated that: a) participants are selecting queries mostly in line with PG within the bounds of the probabilistic environments employed in the present study, b) the strategies we unearthed from participants' own think-a-loud responses were found to underlie specific query selections and were additionally able to account for participants' preferences across probabilistic

contexts, c) participants prefer obtaining a frontrunner rather than eliminating a suspect in a ternary-hypothesis scenario, and d) these strategic preferences can be influenced by factors such as risk aversion and zero-sum thinking, that have rarely been identified in information search paradigms. Our next experiment aimed to replicate these findings in different probabilistic models to corroborate these findings as independent of particular chosen parameter sets and test the predictive abilities of the OED measures in diverse probabilistic environments.

3. Experiment 2

The primary aims of Experiment 2 were to extend the findings of Experiment 1 by: 1) determining whether participants' preferences and motivated strategies remained robust in different probabilistic contexts with a *binary* hypothesis space and 2) illustrating that with a binary hypothesis space, identifying strategies such as 'elimination' and 'frontrunner' becomes conceptually impossible. As in Experiment 1, we created models with four different parameter sets, such that the informative value of each query varied across conditions. Building on the findings of Experiment 1, we anticipated that people would once again select the query that could lead to greater hypothesis disambiguation *and* was less risky (i.e., safe frontrunner strategy). This would be with the aim of maximising the chances of obtaining a suspect with a high-enough posterior to minimise choice inaccuracy in a one-shot paradigm – ultimately reflecting the motivations of a probability gain measure. As such, we predicted that in all the probabilistic environments employed in this experiment, the majority of participants would select (and evaluate as more informative) the query 'burglary time', in all conditions.

3.1 Bayesian OED Models

Our BNs were built as described in section 2.1, except that the parent node 'Burglar' (see Figure 1) was a binary variable (states: Suspect 1, Suspect 2). Again, we used uniform priors so that in all models P (Suspect 1) = P (Suspect 2) = $\frac{1}{2}$. For each Model i where $i \in \{1, 2, 3, 4\}$ the conditional probabilities of each state of each query node (burglary time and Primary Item stolen) given each state of the common cause node (Burglar) can be seen in Table 8.

Table 8

Experiment 2: Conditional Probability Table with parameters employed in each model.

	Model 1		Model 2		Model 3		Model 4	
	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
P (Day| S_i)	0.1	0.95	0.1	0.85	0.1	0.85	0.1	0.95
P (Night| S_i)	0.9	0.05	0.9	0.15	0.9	0.15	0.9	0.05
P(Jewellery| S_i)	0.8	0.1	0.9	0.05	0.8	0.1	0.9	0.05
P(Electronics|S_i)	0.1	0.8	0.05	0.9	0.1	0.8	0.05	0.9
P(Money| S_i)	0.1	0.1	0.05	0.05	0.1	0.1	0.05	0.05

N.B. for $S_{i,}$ i is a suspect ∈ {1, 2}

Given each probabilistic model outlined in Table 8, the expected informative value of each query and each query outcome, computed through KL-D, IG, PG, PG_H and Impact, can be seen in Table 9 below.

Table 9

Experiment 2: Expected value of each query outcome (a_i) and each query (Q_i) predicted by each utility function in each probabilistic model

	Utility Function	a_1 Day	a_2 Night	Q_1 Burglary Time	a_3 Jewellery	a_4 Electronics	a_5 Money	Q_2 Primary Item Stolen
Model 1	KL-D	0.55	0.70	0.6	0.50	0.50	0	0.45
	IG	0.45	0.3	0.6	0.50	0.50	1	0.45
	PG	0.90	0.95	0.43	0.89	0.89	0.5	0.35
	PG_H	0.90	0.95	0.43	0.89	0.89	0.5	0.26
	Impact	0.40	0.45	0.43	0.39	0.39	0	0.35
Model 2	KL-D	0.51	0.40	0.45	0.70	0.70	0	0.6
	IG	0.48	0.59	0.45	0.30	0.30	1	0.6
	PG	0.89	0.86	0.38	0.95	0.95	0.5	0.43
	PG_H	0.89	0.86	0.38	0.95	0.95	0.5	0.3
	Impact	0.40	0.36	0.38	0.45	0.45	0	0.43
Model 3	KL-D	0.51	0.40	0.45	0.50	0.50	0	0.45
	IG	0.48	0.59	0.45	0.50	0.50	1	0.45
	PG	0.89	0.86	0.38	0.89	0.89	0.5	0.35
	PG_H	0.89	0.86	0.38	0.89	0.89	0.5	0.26
	Impact	0.40	0.36	0.38	0.39	0.39	0	0.35
Model 4	KL-D	0.55	0.70	0.6	0.70	0.70	0	0.6
	IG	0.45	0.3	0.6	0.30	0.30	1	0.6
	PG	0.90	0.95	0.43	0.95	0.95	0.5	0.43
	PG_H	0.90	0.95	0.43	0.95	0.95	0.5	0.3
	Impact	0.40	0.45	0.43	0.45	0.45	0	0.43

Once again, the model parameters were selected so as to yield different expected informative values of queries across utility functions and across different model parameterisations, allowing us to explore people's sensitivity and strategic adaptiveness across contexts and the abilities of utility functions to account for participants' information search behaviour. Although the binary hypothesis space does not formally allow us to tease apart frontrunner vs. elimination strategies, as seen in Table 8, our parameters rendered one query (burglary time) perceptively 'safer' given that each outcome was diagnostic of a different suspect, and one query (primary item stolen), albeit also leading to a lead suspect, somewhat riskier by including the low possibility of obtaining an outcome (money) that would not allow the disambiguation of the hypotheses. We predicted people would favour maximising their chances of identifying a leading suspect with less risk given the one-shot nature of the paradigm, by preferring the former option across probabilistic contexts.

3.2. Method

3.2.1. Participants

We tested 236 participants (104 males, M_{age} =34 years) who were recruited from Prolific Academic and completed the study online utilising the Prolific Academic platform. All participants were native English speakers, who gave informed consent, and were compensated $1.20 for partaking in the present experiment, which took on average (median) 12.7 minutes to complete.

3.2.2. Design and Materials

A between-subject design was adopted. Participants were randomly allocated to one of four conditions ($n_{\text{Condition 1}} = 58$, $n_{\text{Condition 2}} = 58$, $n_{\text{Condition 3}} = 59$, $n_{\text{Condition 4}} = 61$). All participants were presented with the same cover story in which they acted as criminal investigators. Participants in each Condition i (C_i) were required to reason with a Model i, where $i \in \{1, 2, 3, 4\}$, parameterized as outlined in Section 3.1 (see Tables 8 and 9), and completed the same one-shot task (for example task materials see osf.io/tkr4v).

3.2.3. Procedure

In Experiment 2 we followed an identical procedure as that employed in Experiment 1, outlined in section 2.2.3. This included a stage in which participants in each condition were given information on the pertinent model (parameters and relationships within the model) and a stage in which participants were required to make a query selection based on the information given to them, as well as justify their selection utilising a think-a-loud response. Finally, participants were required to provide ratings on the usefulness of the available queries and their outcomes. Once again, participants were never informed of the outcome of the query they selected, making all their evaluations pertinent to the *expected* informative values queries would yield.

3.3. Results

3.3.1. Prior Probabilities

The percentage of participants who correctly estimated the prior probabilities of all three suspects was 81% in Condition 1, 88% in Condition 2, 89% in Condition 3 and 89% in Condition 4. Given these high percentages we concluded that participants overall accepted the uniform priors given to them. Nonetheless, once again, to increase the validity of our normative comparisons, all subsequent analyses will once again evaluate participants' behaviour against informed Bayesian OED models (IB-OED), fitted with participants' own stated priors.

3.3.2. Query Selection

Within each condition, we obtained the percentage of participants who selected each query (see Figure 4). Pearson's Chi-Square test of independence indicated that these percentages did not differ between conditions, χ^2 (3) = 3.32, p = 0.34, V = 0.12. In line with our predictions, the majority of participants queried 'burglary time' in all conditions. These findings strengthen those of Experiment 1 in demonstrating that people may not be sensitive to the varying utility of queries determined by the probabilistic models to the degree of our manipulation. According to all utility functions the two queries did at least to an extent vary in informative value across conditions, except when these were computed by a PG_H model, that evaluated the query 'burglary time' as being more informative in all conditions (though only marginally in Condition 2).

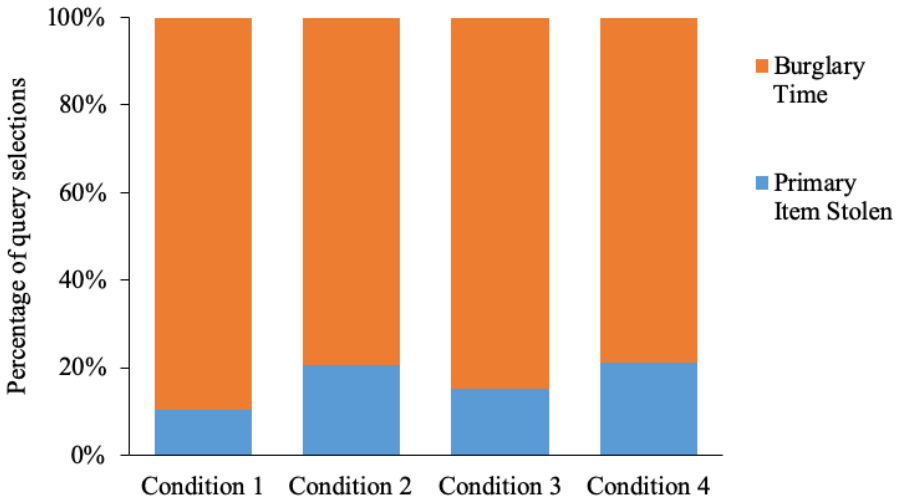

Figure 4. Experiment 2: Percentage of participants who selected each query in each condition.

3.3.3. Utility Function Model Comparisons

The breakdown of the predictions each utility functions made regarding the informativeness of each query, as well as the distribution of participants' query choices, can be seen in Table 10.

To determine how well each utility function predicted people's query choice proportions we once again built mixed-effects logistic regression models following the same model fitting procedure outlined in section 2.3.3. A comparison via likelihood ratio test of our null model M_0 (that included a random-effect of 'Subject' with intercept and outcome variable 'Participant Choice') to a model, M_1, with an added fixed-effect of 'Condition' illustrated that participants' query selections did not vary across conditions, χ^2 (3) = 3.33, p = 0.32, and that adding 'Condition' as a fixed-effect did not significantly improve the null model. Following the iterative process outlined in section 2.3.3 for Experiment 1, we found that for all utility functions the maximal model was M_2 ('Utility Function' as fixed-effect, 'Participant Choice' as outcome variable and random effect with intercept for 'Subject') – including 'Condition' as a fixed effect did not improve the fit of any model. All maximal models were checked for overdispersion and under dispersion and no issues were noted.

Table 10

Experiment 2: Percentage of predictions made by each utility function in each condition favouring burglary time, primary item stolen, or evaluating them as equal, and percentage of participants who queried Burglary Time (represented in Burglary Time > Item Stolen column) and Item Stolen (represented in Item Stolen > Burglary Time column).

	Utility Function	Burglary Time > Item Stolen	Item Stolen > Burglary Time	Item Stolen = Burglary Time
Condition 1	KL-D	100%	0%	0%
	IG	100%	0%	0%
	PG	91%	0%	9%
	PG_H	93.1%	1.7%	5.2%
	Impact	91%	0%	9%
	Participant Choice	89.7%	10.3%	-
Condition 2	KL-D	0%	100%	0%
	IG	0%	100%	0%
	PG	0%	100%	0%
	PG_H	91.4%	0%	8.6%
	Impact	0%	100%	0%
	Participant Choice	79.3%	20.7%	-
Condition 3	KL-D	0%	0%	100%
	IG	0%	0%	100%
	PG	0%	0%	100%
	PG_H	96.6%	0%	3.4%
	Impact	0%	0%	100%
	Participant Choice	84.7%	15.3%	-
Condition 4	KL-D	0%	2%	98%
	IG	0%	2%	98%
	PG	0%	0%	100%
	PG_H	95.1	3.3%	1.6%
	Impact	0%	0%	100%
	Participant Choice	83.1%	16.9%	-

The outputs of the mixed-effect logistic regression analyses used to assess the predictive abilities of each utility function can be seen in Table 11 below. Through these analyses we found no main effect of any utility function on the outcome variable 'Participant Choice': PG, F (2, 233) = 1.4, p = 0.26; PG_H, F (2, 233) = 1.7, p = 0.19; KL/IG, F (2, 233) = 1.2, p = 0.3; and Impact, F (2, 233) = 1.45, p = 0.24.

Despite these findings, by consulting the OR values in Table 11, we can see that in the PG_H model a prediction of 'burglary time' (participants' preferred query across all scenarios– see Figure 4) made a participant choice of 'burglary time' 3 times more likely than a participant choice of 'primary item stolen' Comparatively, a prediction of 'burglary time' by any of the other models, made a

participant choice of 'burglary time' *less* likely than a choice of 'primary item stolen'. The interpretations of these findings can be additionally informed by consulting Table 10. As can be seen in Condition 2, all models performed poorly (predicting primary item stolen > burglary time) except for the PG_H model, which was able to account for the directional preference of participants' query selections in all conditions (evaluating 'burglary time' to be the most informative query). As such, in Condition 3 and 4, all utility functions except PG_H predicted the two queries to be of approximately equal value, thereby failing to accurately represent participants' preference for one query (burglary time) over the other and displaying a lack of discriminative capacity.

Table 11

Experiment 2: Parameters of the fixed effects estimated via logistic mixed-effects models, their statistical significance, and odds ratio for the competing models.

Participant Choice = 'Time'
Reference category 'Participant Choice' = Item.

Model[1]	Parameter (Model Prediction)	Estimate ß	Std. Error ß	*t*	Sig.	Odds Ratio	OR 95%CI Lower	OR 95%CI Upper
Probability Gain	(Intercept)	-1.49	1.32	-1.1	0.26			
	'Item'	0.15	0.40	0.36	0.72	1.2	0.52	2.5
	'Time'	-0.78	0.53	-1.47	0.14	0.46	0.16	1.3
	'ItemTime'[b]	0[a]						
Probability Gain Heuristic	(Intercept)	0.56	0.63	-0.36	0.72			-
	'Item'	0.13	1.37	3.1	0.002	1.14	0.08	16.9
	'Time'	1.1	0.65	2.2	0.027	3.05	0.85	10.9
	'ItemTime'	0[a]						
KL-D /IG	(Intercept)	-1.48	1.3	-1.1	0.26			
	'Item'	0.12	0.4	0.29	0.77	1.13	0.51	2.49
	'Time'	-0.7	0.49	-1.37	0.17	0.51	0.19	1.35
	'ItemTime'	0[a]						
Impact	(Intercept)	-1.5	1.3	-1.1	0.26			
	'Item'	0.14	0.40	0.34	0.73	1.15	0.52	2.5
	'Time'	-0.8	0.52	-1.5	0.13	0.45	0.16	1.3
	'ItemTime'	0[a]						

[a] Parameter is set to zero due to redundancy.
[b] ItemTime refers to a model prediction of the two utilities having approximately equal value (abs diff. of two queries < 5%).
[1] Participant Choice ~ Utility Function Prediction + (1 | Subject)

The likelihood ratio test results between each maximal model and the null model are displayed in Table 12 below. As can be seen, none of the maximal models including the utility functions as predictors significantly improved the null model fit in predicting the outcome variable 'Participant Choice'.

Table 12

Experiment 2: Likelihood ratio test results, AIC, Deviance, Akaike Weights (w) and Evidence Ratio (ER) values of the competing models.

Model	df	AIC	ΔAIC_i	w_i	ER_i	Deviance	χ^2	df	*p*-value
M_0[1]	2	218.8	-	-	-	214.8		-	
M_2 PG	4	219.6	0.6	0.29	1.35	211.6	3.2	2	0.20
M_2 PG_H	4	219	0	0.39	1	211	3.1	2	0.21
M_2 KL/IG	4	220.8	1.8	0.16	2.46	214.8	0.001	2	0.97
M_2 Impact	4	220.8	1.8	0.16	2.46	214.8	0.002	2	0.97

[1] Participant Choice ~ 1 + (1 | Subject)

In order to compare the competing utility functions models, we again used derivatives of Akaike's Information Criterion (AIC) measure and followed the same procedure outlined in section 2.2.3 for Experiment 1. The computed ΔAIC values (see Table 12) suggest no model should be discounted and that there is no singular model that significantly approximates participants' choices more than the others given that $\Delta AIC < 2$ in all models. Comparing the four candidate models, PG_H has the smallest AIC value and therefore acts as the 'best' reference model. The computed Akaike weights showed that PG_H has 39% chance of being the correct model, and the next-best model, PG, has 29% chance of being the correct model. From the ER values (see Table 12) we can conclude that a PG_H model is 1.35 times more likely than our next-best model, PG, and 2.46 times more likely than the KL/IG and Impact models.

Overall, considering the percentages presented in Table 10, the OR values and coefficients presented in Table 11 and the ΔAIC, Akaike weights and ER values presented in Table 12, we can conclude that out of the candidate models, one with a built-in PG_H utility function is able to best approximate the distribution of participants' query choices. However, although faring better than its competitors, PG_H was nonetheless not found to be a significant predictor of participants' query selections. This could be due to the fact that this model was not able to account for the 15-20% of participants who on average selected the query 'primary item stolen' in each condition. Qualitatively,

however, this model was representative of the direction of the majority of participants' query selections (i.e., for 'burglary time') in all conditions suggesting that, similar to Experiment 1, participants' query choices were most in line with evaluations made by a PG-based model. Contrary to our Experiment 1 findings, in the probabilistic environments adopted in the present experiment (i.e., using a binary hypothesis space), the simplified model assuming equal priors was a better fit than the original PG model – this is discussed further in section 3.4.

3.3.4. Query and Outcome Ratings

Participants' average ratings of the usefulness of each query and query outcome can be seen in Table 13 below. A one-way ANOVA showed no significant effect of condition on the usefulness ratings of the queries 'primary item stolen', F (3, 235) = 2.38, p = 0.07, η_p^2= 0.03, and 'burglary time', F (3, 235) = 2.11, p = 0.09, η_p^2 = 0.027. As such, participants in all conditions rated the query 'burglary time' as being of higher utility than the query 'primary item stolen'. This strengthens the notion that participants may be evaluating queries using criteria that lie outside of the principles dictated by the utility functions we compared them to, given that none of our tested utility functions foretold this unvaried preference for the query 'burglary time'. In terms of query outcome ratings, whereas no between-condition differences were found in the usefulness ratings of query outcome 'night', F (3, 234) = 2.2, p = 0.09, η_p^2 = 0.09, we found a significant between-condition difference in the ratings of outcome 'day', F (3, 234) = 5.3, p = 0.001, η_p^2 = 0.065.

Post-hoc pairwise comparisons showed the significant differences to be between participants in Conditions 1 and 2, p = 0.018 and between participants in Conditions 2 and 4, p = 0.009. As such, participants in Condition 2 rated the usefulness of the outcome 'night' to be significantly lower than participants in Conditions 1 and 4. No between-condition differences were found in the usefulness ratings of query outcome 'money', F (3,234) = 0.74, p = 0.53, η_p^2 = 0.009, which mirrors the predictions of all utility functions regarding this query outcome across conditions (lowest utility - see Table 9).

Table 13

Experiment 2: Mean participant ratings of each query and each query outcome in each condition

		Condition 1	Condition 2	Condition 3	Condition 4
		M (SD)	*M* (SD)	*M* (SD)	*M* (SD)
Query	Burglary Time	8.72(1.3)	8.05 (1.5)	8.2 (1.4)	8.3 (1.7)
	Primary Item Stolen	6.2 (2.1)	6.3(1.9)	6.8 (1.8)	7 (2.1)
Query Outcomes	Night	8.74 (1.3)	8.2 (1.5)	8.2 (1.4)	8.5 (1.5)
	Day	8.6 (1.5)	7.8 (1.6)	7.9 (1.2)	8.6 (1.4)
	Jewellery	7.7 (1.7)	8.5 (1.6)	7.9 (1.2)	8.4 (1.3)
	Electronics	7.9 (1.3)	8.6 (1.2)	7.9 (1.2)	8.4 (1.3)
	Money	1.6 (2.3)	1.6 (2.7)	2.2 (2.8)	2 (2)

A significant between-condition difference was found in the usefulness ratings of query outcome 'jewellery', F (3, 234) = 3.9, p = 0.01, η_p^2 = 0.05, with post-hoc pairwise comparisons showing the difference to be between Conditions 1 and 2, p = 0.03. Participants in Condition 2 rated the outcome 'jewellery' as being significantly more useful than those in Condition 1. Finally, a significant between-condition difference was found in the usefulness ratings of query outcome 'electronics', F (3, 234) = 5.05, p = 0.002, η_p^2 = 0.06, with post-hoc pairwise comparisons showing the difference to be between Conditions 1 and 2, p = 0.01 and between Condition 2 and 3, p = 0.01. Participants in Condition 2 valued the query outcome 'electronics' as more useful than participants in Conditions 1 and 3.

Whereas some of these between-condition differences were comparable to normative value ratings of utility functions detailed in Table 9, participants failed to detect the varying informative value of the majority of outcomes across conditions. As such, all utility functions except IG predicted outcome 'night' to be less informative in Condition 2 and 3 than in the remaining conditions, and outcome 'jewellery' to be more useful in Condition 4 than in Conditions 1 and 3. In this respect our findings were in line with findings of Rusconi et al. (2014) showing IG was a better predictor of participants' query *outcome* evaluations compared to alternative OED utility functions.

To confirm that participants' usefulness ratings were representative of how they actually evaluated a query by either selecting it or not selecting it, we once again computed the percentage of "rating congruent" responses in each condition. As a reminder, a query choice was coded as congruent

(1) if the participant selected the query that they also rated as being most useful on the 0-10 Likert scale - if not, a query choice was coded as incongruent (0). If a participant gave equal ratings to the two queries, their query choice was coded as congruent regardless of what query was selected. The percentage of congruent query selections was: 94.7% in Condition 1; 100% in Condition 2; 94.8 % in Condition 3; and 93.3 % in Condition 4. These high percentages allow us to take participants' ratings as reliable representations of their evaluation of how useful they believe a query to be. Overall, it appears that participants displayed some degree of sensitivity to the different probabilistic contexts when rating the usefulness of query outcomes, though arguably insufficiently so. Moreover, this sensitivity to probabilistic context was not reflected in their ratings of the queries themselves, which remain unvaried across conditions and consistently favoured the query 'burglary time' over 'primary item stolen', aligning only with qualitative directional predictions of a PG_H model.

Given that in all conditions participants evaluated query outcomes 'jewellery' and 'electronics' as being equal to and at times superior to the outcomes 'day' and 'night', even in Condition 2, and yet displayed a modal preference for the query 'burglary time' across conditions, this suggests that pitfalls in optimally evaluating queries likely occurred at the level of integrating information (i.e., weighting the probability of outcomes occurring), rather than from bottom-up insensitivity to probabilistic contexts. Given that participants recognised the extremely low utility of outcome 'money' in all conditions, it is possible that they overweighed the probability of this outcome occurring and therefore excessively de-valued the query 'primary item stolen' and resorted to selecting the less 'risky' query of 'burglary time', even in conditions in which this would lead to a lesser (or equal) gain in information according to all utility functions. This view is supported by the analysis described in section 3.3.3, showing that the PG_H model, which assumed equal priors of outcomes, was the model that performed best in approximating the distribution of participants' query selections.

3.3.5. Strategies: Think-a-loud responses

Given the consistency in query selection and evaluation by participants across conditions, irrespective of changes in the probabilistic environment, it is important to once again assess the strategies underlying participants' evaluation of the expected value of queries and outcomes. To do this, we analysed the

think-a-loud responses related to participants' query selections. Given that one of the outcomes of 'primary item stolen' had an expected utility of zero in all contexts, we expected that more participants would employ a frontrunner strategy moderated by risk aversion, and thus prefer the query 'burglary time' in all conditions.

As all probabilistic models in Experiment 2 had binary hypotheses the states were modelled as mutually exclusive and exhaustive, discriminating between a frontrunner and an elimination strategy was not technically possible. Only two participants explicitly stated a preference for 'eliminating' a suspect, and the remainder utilised language indicating a preference for increasing the probability of one suspect or identifying the culprit. For this reason, and due to the constraints of utilising a binary hypothesis space, think-a-loud responses that voiced a preference to identify a 'frontrunner' or 'lead' suspect and those that voiced a preference for 'eliminating' a suspect were collapsed under a single code dubbed 'identify the culprit'.

The coding procedure followed that outlined in section 2.3.5 of Experiment 1. A primary rater coded all responses (236) and 50.8% of responses were randomly selected from the total sample (N = 120) and coded by a second independent rater. Cohen's weighted kappa was utilised to determine a moderately high inter-rater agreement between the two raters, $\kappa_w = 0.88$, $p < 0.001$. Using the principles outlined in Experiment 1, 36 participant responses were attributed a code of "n/a" as they did not provide relevant or elaborate enough explanations. Subsequent analysis was carried out on the total sample (including "n/a") of 236 responses. The strategy codes extracted from all responses and their prevalence in accounting for the total sample will now be presented in turn (see Table 14 for descriptives and Figure 5 for graphical representation).

Table 14

Experiment 2: Percentage use of strategy codes across conditions (collapsing scenarios).

Strategy Code	Percentage Use (across conditions)
Symmetry	11%
Differentiation	20.3 %
Highest Percentage	13.6 %
Zero-sum/Risk aversion	13.1%
Identify culprit + Zero-sum/Risk aversion	5.5%
Identify culprit	21.2%

As suspected, the extent to which participants displayed "zero-sum/risk aversion" thinking was noticeably higher (~19%) than in Experiment 1 (~10%), strengthening the notion that the modal preference in 'Burglary Time' was partly the result of a risk aversion towards the query 'Primary Item Stolen', whose outcome 'money' was not informative in any condition, and a possible failure to integrate information (i.e., weighting the probability of outcomes occurring or ignoring priors of outcomes altogether) given that this query was actually more informative or of equal informative value according to all utility functions in Condition 2 (except PG_H, which attributes equal priors to query outcomes), where participants still failed to concede this.

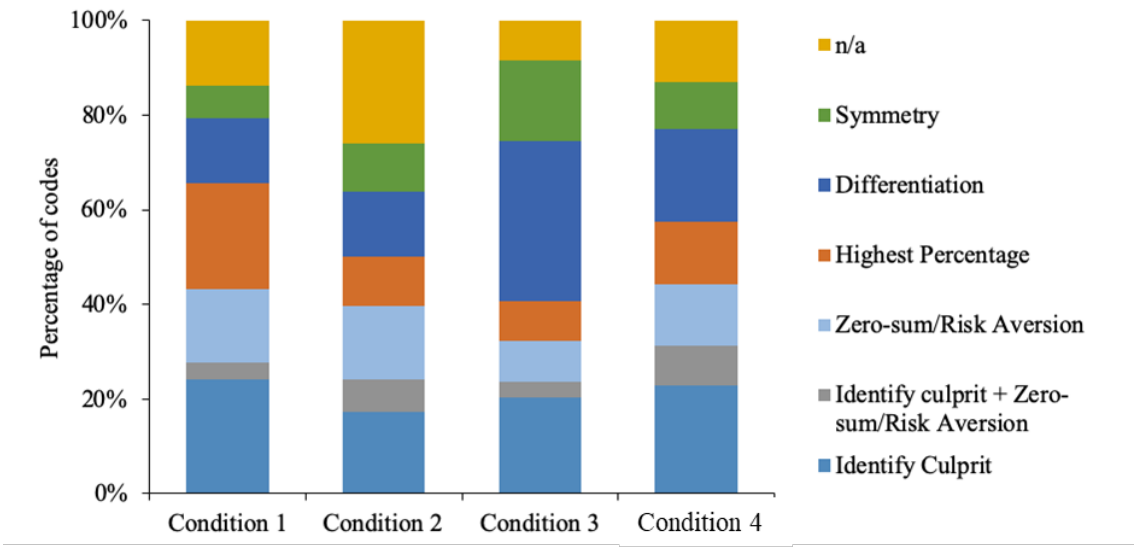

Figure 5. Experiment 2: Percentage of strategy codes within each condition

3.3.5.1. Adaptability of Strategies

In order to explore the adaptability of participants' strategies across the different probabilistic contexts, we conducted a Chi-square test of independence. We found no significant difference in the percentage of participants who adopted the different strategies between the four conditions, $\chi^2(18) = 26.1$ $p = 0.097$, $V = 0.19$. This means that the extent to which participants adopted each strategy, similarly to their query selection preferences, did not differ across the four probabilistic environments we employed. As in Experiment 1, more than half of the participants could be accounted for by 'identify suspect', 'symmetry' and 'differentiation' strategies across all probabilistic contexts.

3.3.5.2. Strategies and Query Selection

Next, we once again explored whether certain strategies systematically underlie certain query selections. Given that we found no significant variation of either strategy or query selection across conditions, this was done by collapsing across conditions. Of the participants who used an 'identify culprit' strategy (n = 50), 80% queried 'Burglary Time', as did 100% of participants who used an 'identify culprit + zero/sum risk aversion strategy' (n = 13) and 78% of participants who utilised a differentiation strategy (n = 48). Similarly, 84% of participants who utilised a zero/sum risk aversion strategy (n = 31), 92% of participants who used a symmetry strategy (n = 26), and 84% of participants who used a highest percentage strategy (n = 32), also queried 'Burglary Time'. A Chi Square test of independence illustrated that the extent to which these strategies underlay certain query selections did not vary across conditions, $\chi^2(6) = 5.85$ $p = 0.44$, $V = 0.16$.

Taken together, our findings suggest that both people's queries and their underlying strategies, remain largely unvaried across conditions. In contrast to Experiment 1, the impossibility of explicitly discerning between frontrunner and elimination strategies meant that in this experiment it was not possible to determine whether distinct strategies related to obtaining a frontrunner vs. eliminating a suspect and the majority of participants who employed each of the strategies queried 'burglary time'. For particulars on how the strategies related to accuracy of query selections according to the different utility functions see supplementary materials (S4).

3.4. Discussion

In Experiment 2 we explored people's information acquisition and evaluation behaviour in four different probabilistic contexts. We replicated findings from Experiment 1 by showing that participants across conditions displayed a strong preference for the same query (in this case 'burglary time') regardless of probabilistic environment. This switch in modal preference of 'burglary time', compared to the preference for 'primary item stolen' observed in Experiment 1 additionally allowed us to conclude that participants were not guided by the content of the queries (e.g., type of evidence per se) when judging their informativeness. Here we also note that, compared to Experiment 1, the utility functions made similar predictions, due to the binary hypothesis state (e.g., PG and Impact make identical predictions now). This highlights the importance of comparing the predictions of different utility functions across different probabilistic contexts in order to identify the environments which lead to differential versus concurrent evaluations of a query's expected utility.

Echoing the findings of Experiment 1 we showed that probability gain based models (PG and PG_H) best-approximated the distribution of participants' query choices across probabilistic environments, though neither of these models was a significant predictor of participant choice. In contrast to Experiment 1, rather than the original PG model, PG_H was best able to qualitatively predict the direction of participants' choices by evaluating 'burglary time' to be the most informative query in all conditions. In the binary hypotheses spaces adopted in Experiment 2 it therefore appears that a simplified OED model that assumes equal outcome priors best accounts for the distribution of participants' query evaluations. The adoption of a simplified PG model compared to Experiment 1, is likely due to the differences in conditional probability (CPT) values adopted in the two experiments. As such, in Experiment 1, the less preferred outcome 'night' would still result in the probabilities of two suspects increasing - albeit by the same amount thus not aiding disambiguation. In contrast, in Experiment 2, the less preferred outcome 'money' would decrease the probability of both suspects thereby going against the intuitive goal in a one-shot investigation task of increasing the probability of a suspect as much as possible. As such, it is likely that participants overweighed the probability of the outcome they rated as less favourable ('money') occurring, and therefore conformed to a model that

assumes equal outcome priors. This finding is in line with predictions of the dominant model of descriptive choice in Prospect theory (Kahneman and Tversky, 1979 – directly tested in e.g., Bleichrodt, 2001) which predicts that small probabilities (as attached to the 'money' outcome in query 'burglary time' Experiment 2) are overweighed. Therefore, the findings from Experiment 2 corroborate Experiment 1, showing that participants evaluate queries in line with the predictions of a rational model that maximises the chance of obtaining a correct suspect classification and is motivated by maximising accuracy of choice. However, participants followed the rational model less than in Experiment 1, most probably due to the above-explained increase in risk-averse behaviour.

With regard to *why* people are selecting and evaluating queries in this manner, an analysis of think-aloud responses revealed that the majority of participants across conditions were driven to identify the culprit by selecting the query that would guarantee disambiguation, such that regardless of its outcome the probability of one suspect would increase over the other. In conditions in which querying 'primary item stolen' could lead to greater disambiguation of the hypotheses and increase the probability of a suspect *higher* than any outcome of query 'burglary time' could, an analysis of participants' think-aloud responses corroborated risk-averse thinking, akin to that observed in Experiment 1. This risk aversion was possibly underlain by overweighing the probability of the least informative outcome (money) occurring, and thus resorting to choosing the alternative query, despite it resulting (in some conditions) to a lead suspect with lower probability of being the true culprit. Once again this suggests that participants, when considering the trade-off between the perceived evidential value of outcomes and the perceived probability of obtaining outcomes, trade some evidential value in favour of a query that will allow them to obtain *a* frontrunner with a higher probability (albeit with a lower chance of this frontrunner being the true culprit). These findings therefore demonstrated that known biases in decision making under uncertainty, such as risk aversion, also play a role in seeking information under uncertainty, at times leading participants to ultimately lose information (Poletiek & Berndsen, 2000).

Of note is the fact that due to the binary hypothesis space it was not possible to technically distinguish between 'frontrunner' and 'elimination' strategies in this experiment (although only 2

participants explicitly mentioned elimination in their responses), suggesting that the strategies and heuristics employed by participants are in part contingent on the probabilistic model employed. This emphasises the importance of exploring information acquisition in probabilistic contexts that extrapolate beyond binary-hypotheses and binary-outcome feature models in order to identify the strategies and heuristics people may adopt more complex probabilistic environments.

4. Experiment 3

In Experiments 1 and 2 we demonstrated that participants' deviation from Bayesian OED model predictions when evaluating queries may stem from errors when integrating an outcome's diagnosticity with the probability of its occurrence. Moreover, our results suggested that certain common motivators of inquiry underlie people's query selection and evaluation behaviour. In order to narrow the space of explanations for our findings in Experiment 3 we built probabilistic models with a ternary hypothesis space and ternary outcome queries. Moreover, to explore the adaptiveness of participants' motivated strategies and explore participants' sensitivity to changes in probabilistic models at the individual level, we introduced a within-subject design factor requiring participants to reason with multiple models.

As previously discussed, despite at times deviating from the predictions of OED utility functions, participants exhibited seemingly rational information search behaviour given the one-shot design of the task. More precisely, in the context of a one-shot criminal investigation it was rational to prefer to select the query that aided hypothesis disambiguation with minimal risk, allowing for the identification of a 'guaranteed' frontrunner to pursue and make progress in the investigation. This behaviour aligned with the motivations of a PG measure that aims to maximise classification accuracy when identifying a lead suspect.

In the present experiment we included a between-subject manipulation to investigate the effect of task framing - one-shot investigation versus perceived multiple possible inquiries - on strategy adoption and query evaluations. We predicted that participants who were under the impression the task comprised of multiple query selections would employ a different strategy than their counterparts who were told they had only one chance at obtaining information. In addition we expected participants in

the ‘multiple inquiries’ condition to display less risk aversion than participants who were told the task was a one-shot investigation, given that the former might adopt an ‘elimination’ strategy to a greater extent at the outset of the investigation, given the belief that the investigation involves multiple inquiries. This would depart from predictions made by PG, which is no longer guaranteed to be the optimal solution to a task with multiple enquiries. Overall, we predicted that in this experiment none of the utility functions would be able to account for any differences between the one-shot and multiple enquiries conditions, given that task framing is not considered by the family of information-theoretic OED measures (Coenen et al., 2018).

To further reduce the computational burden imposed on participants, and more directly assess their inquiry preferences *in the absence of uncertainty*, we also required participants to select one outcome from each query that they would prefer to receive. In this manner we directly probed their preferences by requiring them to choose between evidence that would lead to a frontrunner and evidence that would help to eliminate a suspect. Finally, we also asked participants to update their probabilistic beliefs given outcomes. This allowed us to explore whether biases in information integration, at the level of calculating the impact of different query outcomes on the various hypotheses, underlie sub-optimal query evaluation.

4.1. Bayesian OED Models

Our BNs were once again built as described in Section 2.1, except that the query node ‘burglary time’ also had ternary outcomes (‘day’, 8am to 4pm; ‘evening’, 4pm to 12am; and ‘night’, 12am to 8am). Once again, we used uniform priors so that in all models P (Suspect 1) = P (Suspect 2) = P (Suspect 3) $=\frac{1}{3}$. For each Model i where $i \in \{1, 2, 3, 4,5\}$ the conditional probabilities of each state of each query node (Burglary Time and Primary Item stolen) given each state of the common cause node (Burglar) can be seen in Table 15. Given each probabilistic model outlined in Table 15, the expected informative value of each query and each query outcome, computed through KL-D, IG, PG, PG_H and Impact can be seen in Table 16 below. Once again, the model parameters were selected so as to yield different expected informative values of queries across utility functions and across models.

Table 15

Experiment 3: Conditional Probability Table with parameters employed in each model.

	Model 1			Model 2			Model 3			Model 4			Model 5		
	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
P (Day| S_i)	0.1	0.8	0.1	0.15	0.70	0.15	0.2	0.6	0.2	0.25	0.50	0.25	0.1	0.8	0.1
P (Evening| S_i)	0.1	0.1	0.8	0.15	0.15	0.70	0.2	0.2	0.6	0.25	0.25	0.50	0.1	0.1	0.8
P (Night| S_i)	0.8	0.1	0.1	0.70	0.15	0.15	0.6	0.2	0.2	0.50	0.25	0.25	0.8	0.1	0.1
P(Jewellery| S_i)	0.9	0.1	0.1	0.9	0.9	0.1	0.9	0.9	0.1	0.9	0.1	0.1	0.98	0.1	0.1
P(Electronics|S_i)	0.05	0.5	0.35	0.05	0.5	0.35	0.05	0.5	0.35	0.05	0.5	0.35	0.01	0.6	0.35
P(Money| S_i)	0.05	0.4	0.55	0.05	0.4	0.55	0.05	0.4	0.55	0.05	0.4	0.55	0.01	0.3	0.55

N.B. for $S_{i,}$ i is a suspect $\in \{1, 2, 3\}$

In contrast to previous experiments however, given the within-subject factor, we mostly held the conditional probabilities of 'primary item stolen' outcomes given each suspect constant across probabilistic models and varied those of 'burglary time' to facilitate the identification of a stage at which seeking a safe frontrunner would cease to be the favoured strategy. As such, querying 'burglary time' in Model 1 would allow participants to identify a safe frontrunner, regardless of the outcome, with 80% probability of being the true culprit, Model 2 with 70% chance of being the true culprit, Model 3 with 60% chance and Model 4 with only 50% chance. Model 5 was included as it would allow participants to obtain a safe frontrunner with high probability but querying primary item stolen could lead to an almost certainly guilty suspect (98%) and almost certain elimination of the competing hypotheses. This set-up allowed us to determine how much information participants were willing to lose by being 'risk averse', as well as how deep-rooted their preference was for a frontrunner versus eliminating a suspect.

Table 16

Experiment 3: Expected value of each query outcome (a_i) and each query (Q_i) predicted by each utility function in each model..

	Utility Function	a_1 Day	a_2 Evening	a_3 Night	Q_1 Burglary Time	a_4 Jewellery	a_5 Electronics	a_6 Money	Q_2 Primary Item Stolen
Model 1	KL-D	0.66	0.66	0.66	0.66	0.72	0.36	0.35	0.49
	IG	0.92	0.92	0.92	0.66	0.87	1.22	1.23	0.49
	PG	0.8	0.8	0.8	0.47	0.82	0.55	0.56	0.32
	PG_H	0.8	0.8	0.8	0.47	0.82	0.55	0.56	0.31
	Impact	0.31	0.31	0.31	0.31	0.32	0.19	0.18	0.24
Model 2	KL-D	0.40	0.40	0.40	0.40	0.72	0.36	0.35	0.49
	IG	1.18	1.18	1.18	0.40	0.87	1.22	1.23	0.49
	PG	0.7	0.7	0.7	0.37	0.82	0.55	0.56	0.32
	PG_H	0.7	0.7	0.7	0.37	0.82	0.55	0.56	0.31
	Impact	0.24	0.24	0.24	0.24	0.32	0.19	0.18	0.24
Model 3	KL-D	0.21	0.21	0.21	0.21	0.72	0.36	0.35	0.49
	IG	1.37	1.37	1.37	0.21	0.87	1.22	1.23	0.49
	PG	0.6	0.6	0.6	0.27	0.82	0.55	0.56	0.32
	PG_H	0.6	0.6	0.6	0.27	0.82	0.55	0.56	0.31
	Impact	0.18	0.18	0.18	0.18	0.32	0.19	0.18	0.24
Model 4	KL-D	0.08	0.08	0.08	0.08	0.72	0.36	0.35	0.49
	IG	1.5	1.5	1.5	0.08	0.87	1.22	1.23	0.49
	PG	0.5	0.5	0.5	0.17	0.82	0.55	0.56	0.32
	PG_H	0.5	0.5	0.5	0.17	0.82	0.55	0.56	0.31
	Impact	0.11	0.11	0.11	0.11	0.32	0.19	0.18	0.24
Model 5	KL-D	0.66	0.66	0.66	0.66	0.76	0.57	0.56	0.64
	IG	0.92	0.92	0.92	0.66	0.83	1.02	1.02	0.64
	PG	0.8	0.8	0.8	0.47	0.83	0.64	0.63	0.38
	PG_H	0.8	0.8	0.8	0.47	0.83	0.64	0.63	0.37
	Impact	0.31	0.31	0.31	0.31	0.33	0.21	0.21	0.26

4.2. Methods

4.2.1. Participants

We tested 136 participants (n_{male} = 49 males, M_{age} = 33.9 years, SD = 11.8) who were recruited from Prolific Academic and completed the study online utilising the Prolific Academic platform. All participants were native English speakers, who gave informed consent, and were compensated £2.50 for partaking in the present experiment, which took on average (median) 26.4 minutes to complete.

4.2.2. Design and Materials

A mixed-subjects design was adopted. Participants were randomly allocated to one of two between-subject conditions ($n_{\text{Condition 1}} = 66$, $n_{\text{Condition 2}} = 70$) that differed in the framing of the instructions (see section 4.2.3. below). All participants were presented with the same cover story in which they were tasked as criminal investigators asked to solve various burglary cases. Each burglary case represented a Scenario *i* embedded with a Model *i* where $i \in \{1, 2, 3, 4, 5\}$ parameterised as in Table 15. By using a within-subject factor, participants in both conditions were therefore required to reason with all five models. For an example of our task materials see osf.io/tkr4v.

4.2.3. Procedure

Participants in each condition were presented with a cover story that tasked them as criminal investigators. They were told that they were being transferred on rotation to five burglary divisions in different neighbourhoods. In this manner they were instructed they would have to complete five scenarios (presented in randomized order).

In each scenario participants were, as in Experiments 1 and 2, initially required to review the neighbourhood's burglary statistics and the criminal records of the (three) burglars known to operate in the area. After having reviewed information on the model, participants were told that a new burglary had occurred in their neighbourhood and they were asked to investigate the new case. Participants in Condition 1 (one-shot inquiry condition) were explicitly told that each investigation would comprise of only one inquiry. Participants in Condition 2 (perceived multiple inquiries condition) were told that each investigation would comprise of multiple inquiries, and they would be required to make the first one. The framing of the task thus differed between conditions, but all participants were ultimately able to make only one query selection per scenario.

In both conditions, at the outset, prior probabilities of each burglar being the culprit were elicited from participants to ensure the uniform priors (as stated) had been accepted. Subsequently, participants were asked to select one of two investigative queries: 'burglary time' (to find out whether the burglary occurred during the day, evening or night) or 'primary item stolen' (to find out whether electronics, money or jewellery were primarily stolen). The query selection question was asked in a manner that

would not be leading participants to adopt a particular strategy: "Please choose the query that you believe will be most useful for the *whole* investigation". Participants were required to provide a think-a-loud response explaining the reasoning underlying their choice.

Participants did not find out the outcome of their query. Rather, regardless of their query choice, all participants were additionally asked to select the item of evidence (outcome) they would prefer to receive between an outcome of the query item stolen (randomly selected between 'electronics' and 'money') and an outcome of the query burglary time (randomly selected between all three possible outcomes).So, participants were asked to choose between receiving either electronics or money (as they both directly entailed the elimination of a suspect) or receiving any of the burglary time evidence, given that any of the outcomes of this query entailed a frontrunner (increase of probability of only one suspect). This binary choice question allowed us to directly gather participants' preferences for obtaining a frontrunner versus eliminating a suspect in a simple and direct manner that does not require participants to reason under uncertainty. Participants were again required to provide a textual explanation for their choice.

Finally, participants were required to update the probabilities of each suspect (using a slider ranging from 0 – 100 %), given that they hypothetically found out each of the aforementioned outcomes (i.e., 'electronics' and 'evening'). After giving their probabilistic estimates participants moved on to the next scenario. This procedure was repeated until a participant had completed all 5 scenarios.

4.3. Results

4.3.1. Prior Probabilities

The percentage of participants who correctly estimated the prior probabilities of all three suspects across the five scenarios was 81.8% in Condition 1 and 81.4 % in Condition 2. Given these high percentages we concluded that the uniform priors were generally acceptable to participants. Nonetheless, once again, to increase the validity of our normative comparisons, all subsequent analyses will evaluate participants' behaviour against informed B-OED models parameterized with participants' own stated priors.

4.3.2. Query Selection

The percentage of participants who selected each query in each scenario and per condition is graphically represented in Figure 6 below. For a table detailing the percentage accuracy of participants' query selections according to each utility function and within each scenario/condition see supplementary materials (S5a). This will more directly be explored in the model comparison section of the results (4.3.3).

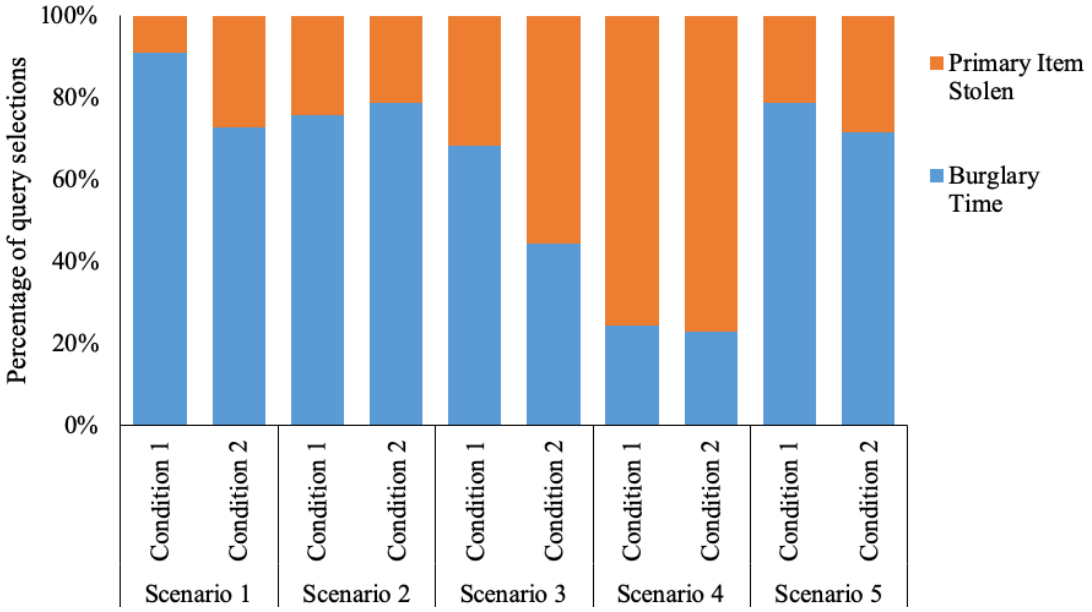

Figure 6. Experiment 3: Proportion of query selections within each condition per scenario.

To investigate whether the framing of the task (between-subjects 'Condition') and the probabilistic environments utilised in each scenario (within-subjects 'Scenario') affected participants' query selections, we built a General Log-Linear Mixed Effects Model with a binomial distribution. Our model had two fixed effects (Scenario and Condition) and a random effect (Subjects) with intercept, to account for individual differences. Results illustrated a main effect of 'Scenario' on participants' query selections, F (4, 670) = 42.1, $p < 0.001$, $\eta_p^2 = 0.20$, and a main effect of 'Condition', F (1, 670) = 8.4, $p = 0.004$, $\eta_p^2 = 0.01$. A significant interaction effect was also found, F (4, 670) = 2.6, $p = 0.036$, $\eta_p^2 = 0.015$. The interaction effect is driven by the fact that, although participants in each of the two conditions behaved comparably in scenarios 1, 2, 4 and 5, in scenario 3 they displayed markedly

different query preferences. More specifically, at this stage, as can be seen from Figure 6, participants who were in the perceived multiple inquiries condition (Condition 2) 'switched' preference at an earlier stage and queried 'primary item stolen' significantly more than participants in Condition 1, t (670) = 2.42, p = 0.015. In regard to the main effect of 'Scenario', participants queried 'burglary time' more than the alternative query, significantly more in scenario 1 than in scenarios 3 ($p < 0.001$) and 4 ($p < 0.001$) where the majority displayed a preference to query 'primary item stolen.' Moreover, participants in both conditions queried 'primary item stolen' significantly more in scenario 3 and 4 than in the other scenarios (all Bonferroni-corrected pairwise comparisons were significant at $p < 0.05$ level).

Compared to previous experiments, participants demonstrated some degree of sensitivity to probabilistic contexts. Moreover, our above analysis shows that participants' change in query preference was significantly influenced by the framing of the task; an effect that cannot be accounted for by our family of OED measures given that predictions of all utility functions remain unvaried regardless of task framing.

4.3.3. Utility Function model Comparison

The breakdown of the percentage of participants for whom each utility function predicted primary item stolen to be more informative, burglary time to be more informative, or for these to be equally informative in each condition can be seen in Table 17. From this table it appears that probability-based models best predicted the distribution of participants' query selections at least in terms of qualitative direction compared to the alterative models. As such, in Scenario 2 PG_H and PG (though marginally) are the only models that predicted the majority of participants would prefer the query 'burglary time'. Despite this however, even these measures showed poor discriminative ability by attributing equal value to the queries in scenario 3 despite a participant majority query preference of 'burglary time'.

To determine how well each utility function predicted people's query choice proportions we once again built mixed-effects logistic regression models for each utility functions using the package *lme4* in R (Bates, Maechler & Bolker, 2012). Our null model (M_0) included a random-effect with intercept for 'Subject', as well as for 'Scenario', and 'Participant Choice' as binary outcome variable.

M_1 in this case built upon M_0 by including a fixed effect of 'Condition' (task framing). A comparison via likelihood ratio test of our null model to M_1 illustrated that participants' query selections significantly varied across conditions, χ^2 (1) = 35.8, p = 0.016, echoing the findings presented in section 4.3.2. Following the iterative process outlined in section 2.3.3 for Experiment 2, we found that for all utility functions the maximal model was M_3 ('Condition' and 'Utility Function' as fixed-effects, 'Participant Choice' as outcome variable and random effects with intercepts for 'Subject' and for 'Scenario'). All maximal models were checked for overdispersion and under dispersion and no issues were noted.

The outputs of the mixed-effect logistic regression analyses used to assess the predictive abilities of each utility function can be seen in Table 18 below. Through these analyses we found no main effect of any utility function on our outcome variable 'Participant Choice': PG, F (2, 676) = 0.1, p = 0.92 ; PG_H, F (2, 676) = 0.9, p = 0.42; KL/IG, F (2, 676) = 0.01, p = 0.99; and Impact, F (2, 676) = 2.7, p = 0.07. However, in each model, we found a main effect of 'Condition', F (1, 676) = 5.45, p = 0.02. This suggests that task framing condition was a better predictor of participants' query selection preferences than any of the utility functions. Table 18 shows that the only significant individual parameter was that of an Impact prediction of query 'primary item stolen'. By looking at the corresponding coefficient value however, one can note that this prediction is actually inversely related to a participant choice of 'primary item stolen', e.g., a prediction of 'primary item stolen' increases the odds of a participant choosing 'burglary time' and decreases the odds of a participant choosing 'primary item stolen'.

Table 17

Experiment 3: Percentage of predictions made by each utility function in each scenario favouring burglary time, primary item stolen, or evaluating them as equal, and percentage of participants who selected each of the queries. Burglary Time

	Utility Function	Burglary Time > Item Stolen	Item Stolen > Burglary Time	Item Stolen = Burglary Time
Scenario 1	KL-D	98.5%	0%	1.5%
	IG	98.5%	0%	1.5%
	PG	99%	0%	1%
	PG_H	99.3%	0%	0.7%
	Impact	97%	0%	3%
	Participant	81%	19%	-
Scenario 2	KL-D	0%	96%	4%
	IG	0%	96%	4%
	PG	57%	1%	42%
	PG_H	98.5%	0.7%	0.7%
	Impact	2%	0%	98%
	Participant	78%	22%	-
Scenario 3	KL-D	0%	100%	0%
	IG	0%	100%	0%
	PG	0%	37%	63%
	PG_H	0%	4.4%	95.6%
	Impact	0%	98%	2%
	Participant	56%	44%	-
Scenario 4	KL-D	0%	100%	0%
	IG	0%	100%	0%
	PG	0%	100%	0%
	PG_H	0%	100%	0%
	Impact	0%	100%	0%
	Participant	24%	76%	-
Scenario 5	KL-D	3%	0%	97%
	IG	3%	0%	97%
	PG	95%	1%	4%
	PG_H	96.3%	1.5%	2.2%
	Impact	54%	0%	46%
	Participant	75%	25%	-

Table 18

Experiment 3: Parameters of the fixed effects estimated via logistic mixed-effects models, their statistical significance, and odds ratio for the competing models.

Reference category = 'Time' Participant Choice = 'Item'								
Model[1]	Parameter	Estimate ß	Std. Error ß	*t*	Sig.	Odds Ratio	OR 95%CI Lower	OR 95%CI Upper
PG	(Intercept)	-0.35	1.98	-0.18	0.86			
	PG Prediction							
	'Item'	0.13	0.44	0.3	0.76	1.14	0.48	2.7
	'Time'	-0.09	0.33	-0.26	0.79	0.92	0.48	1.8
	'ItemTime'	0[a]						
	Condition							
	Condition 1	-0.5	0.2	-2.3	0.02	0.6	0.40	0.9
	Condition 2	0[a]						
PG_H	(Intercept)	-0.86	2	-0.43	0.67			
	PG_H Prediction							
	'Item'	0.01	0.79	0.01	0.98	1.01	0.21	4.75
	'Time'	0.84	0.72	1.18	0.24	2.32	0.57	9.5
	'ItemTime'	0[a]						
	Condition							
	Condition 1	-0.5	0.2	-2.3	0.02	0.61	0.40	0.9
	Condition 2	0[a]						
KL-D /IG	(Intercept)	-0.43	1.9	-0.22	0.82			
	KL/IG Prediction							
	'Item'	0.09	0.73	0.13	0.89	1.1	0.26	4.6
	'Time'	0.07	0.85	0.09	0.93	1.1	0.20	5.7
	'ItemTime'	0[a]						
	Condition							
	Condition 1	-0.5	0.2	-2.3	0.02	0.6	0.4	0.9
	Condition 2	0[a]						
Impact	(Intercept)	-1.5	1.3	-1.1	0.26			
	Impact Prediction							
	'Item'	-1.6	0.79	-2.01	0.045	0.2	0.04	0.96
	'Time'	0.32	0.37	0.86	0.39	1.4	0.66	2.86
	'ItemTime'	0 [a]	-	-	-	-	-	
	Condition							
	Condition 1	-0.5	0.2	-2.3	0.02	0.6	0.39	0.92
	Condition 2	0[a]						

[a] Parameter is set to zero due to redundancy.
[1] Participant Choice ~ Utility Function Prediction + Condition + (1 | Scenario) + (1 | Subject)

The likelihood ratio results comparing each maximal model, M_3, to M_1 are displayed in Table 19 below.

Table 19

Experiment 3: Likelihood ratio test results, AIC, Deviance, Akaike Weights (w) and Evidence Ratio (ER) values of the competing models.

Model	df	AIC	ΔAIC_i	w_i	ER_i	Deviance	χ^2	df	*p*-value
M_1[1]	4	773.5				765.52			
M_3 PG	6	777.4	5.94	0.04	19.5	765.4	0.1	2	0.94
M_3 PG_H	6	775.3	3.84	0.12	6.82	763.3	2.18	2	0.34
M_3 KL/IG	6	777.48	6.02	0.04	20.28	765.48	0.04	2	0.98
M_3 Impact	6	771.46	0	0.80	1	759.46	6.1	2	0.048*

[1] Participant Choice ~ Condition + (1|Scenario) + (1 | Subject)

In order to compare the competing utility function models and select the best approximating model from our available ones, we once again used derivatives of Akaike's Information Criterion (AIC) measure. Given that the Impact model had the lowest AIC value we selected this as the 'best' reference model. The ΔAIC values presented in the above table suggest that PG_H was the next best model, and, alongside the PG model, should not be discounted given that $\Delta AIC < 6$. The KL/IG model could be discounted given that $\Delta AIC > 6$. The computed Akaike weights showed that Impact has 80% chance of being the correct model, and our next best model PG_H has 12% chance of being the correct model. Utilising the ER values, we concluded that an Impact model was 20 times more likely than a KL/IG model, 19 times more likely than a PG model and 6.8 times more likely than a PG_H model to be the correct model.

Despite these findings, we must note that none of the utility functions were significant predictors of 'Participant Choice' and that Impact's 'superior' predictive abilities compared to the alternative models were driven by the inverse significant relationship between a model prediction of 'primary item stolen' and a participant query selection of 'burglary time'. As such, this undermines the notion that Impact is a good predictive model of people's information search behaviour and suggests that none of the utility functions are truly able to capture the distribution of participants' query selections in these environments. Looking at Table 17, it appears that probability-based models best approximate the distribution of participants' query selections at least in terms of qualitative direction compared to

the alterative models. As such, in Scenario 2 PG_H and PG (though marginally) are the only models that predict the majority of participants would prefer the query 'burglary time'. This is reflected in the OR and coefficient values of a 'burglary time' prediction for the PG_H model (see Table 18) which, despite not reaching significance, increased the odds of a participant choice being time by 2.3 times compared to a choice of 'primary item stolen'. Similarly, in Scenario 3, PG is able to account for the approximate 40% of participants who prefer 'primary item stolen', although both PG_H and PG predominantly predicted the informative value of the two queries to be equal in this probabilistic environment.

Overall, however, our analysis showed that that task-framing condition is a more significant driver and predictor of participants' query selections. By consulting the coefficient and OR values in Table 18 we can see that a participant being in Condition 1 making a query selection of 'primary item stolen' is less likely than a query selection of 'burglary time' compared to participants in Condition 2. This finding is echoed by the information presented in Figure 6, showing that participants in Condition 1 selected the query 'primary item stolen' significantly less than their counterparts in Condition 2 within e.g., Scenario 3. As such it appears that contextual factors such as task framing are more important determinants of participants' query evaluations compared to the computations dictated by information-theoretic utility functions.

4.3.4. Query Outcome (Evidence) Selection

Next, we analysed participants' choices when asked to directly select which outcome they would like to receive between one of the 'primary item stolen' (either electronics or money) outcomes and one of the 'burglary time' outcomes (randomly selected between all three). Any of the 'burglary time' outcomes would have led to the identification of a 'frontrunner' albeit with different certainties of being the true culprit. In contrast, electronics and money would have led to the almost certain elimination of a suspect, thus narrowing the scope to two now approximately equally likely suspects. The proportion of participants who selected each piece of evidence (query outcome) is graphically represented in Figure 7 below, split by 'Scenario' and 'Condition'.

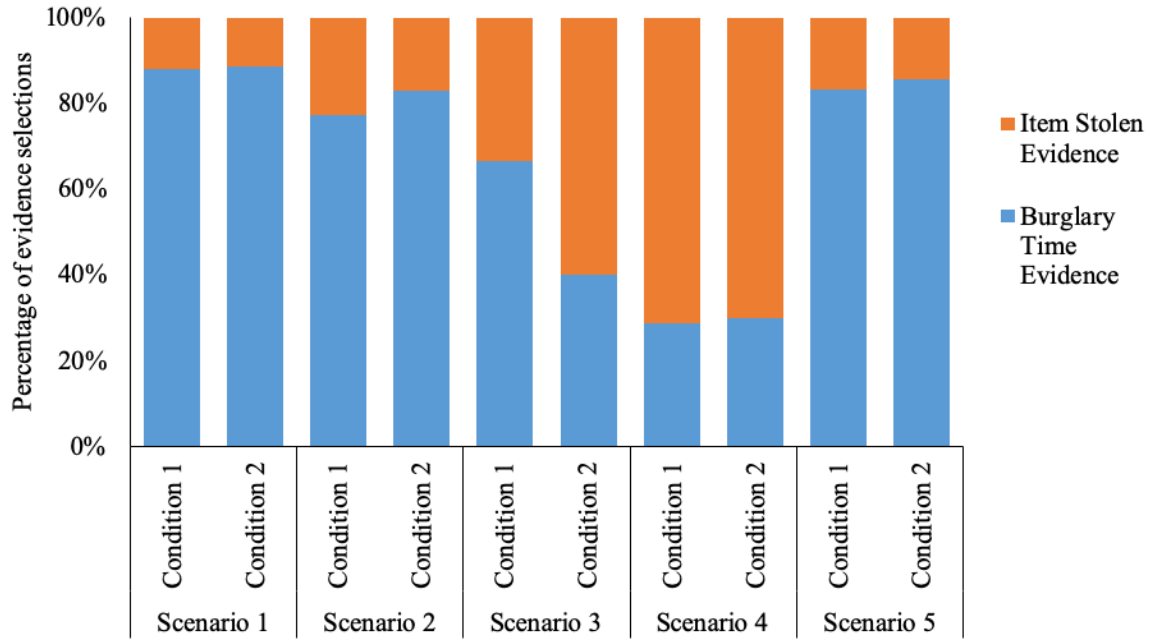

Figure 7. Experiment 3: Proportion of query outcome (evidence) selections when asked to choose between learning that electronics/money was stolen or one of the burglary time outcomes within each condition per scenario.

To investigate whether the framing of the task (between-subject 'Condition' factor) and the probabilistic environments (within-subject 'Scenario' factor) impacted participants' query outcome (evidence) selections, we again built a Log-Linear Mixed Effects Model with a binomial distribution. The model had two fixed effects (Scenario and Condition) and a random effect intercept (Subjects) to account for individual variance. This yielded no main effect of 'Condition' on participants' query outcome preferences, F (1, 670) = 0.12, p = 0.73, η_p^2 = 0.002, but a main effect of 'Scenario', F (4, 670) = 35.6, p < 0.001, η_p^2 = 0.23. A small interaction effect was found, F (4, 670) = 2.5, p = 0.036, η_p^2 = 0.02, explored below. The non-significant main effect of condition suggests that participants' preferences to obtain a frontrunner are exacerbated when uncertainty is removed and are robust across task framing conditions.

Regarding the main effect of scenario, post-hoc pairwise comparisons illustrated a significant difference in evidence preferences between scenario 4 and scenario 1, p = 0.001, as well as scenario 2, p < 0.001, scenario 3, p = 0.04, and scenario 5, p < 0.001. As such, compared to all other scenarios, in scenario 4 participants in both conditions selected the query outcome (money or electronics) that

enables the elimination of a suspect significantly more than the evidence (day, evening or night) that enables the identification of a frontrunner.

These findings suggest that, when removing uncertainty by asking participants to select which outcome they would like to directly receive (in contrast to asking them about what query they'd like to make), the majority of participants, regardless of what task framing condition they were in, "switch strategy" in scenario 4 by choosing a 'primary item stolen' outcome over viewing 'burglary time' query outcomes that would, in this scenario, only provide participants with a "lead" suspect with a 50% probability of being the true culprit (the highest posterior probability of each suspect being the culprit given the outcomes of query burglary time would be 0.5 given the parameters outlined in Table 16).

In regard to the significant interaction effect, post-hoc pairwise comparisons found the term to be driven by the significant difference across Conditions 1 and 2 within scenario 3, $p = 0.003$. As such, within this scenario, participants in the 'perceived multiple inquiries' condition (Condition 2) mostly preferred receiving evidence that would allow the elimination of a suspect (money or electronics) whereas participants in Condition 1 ('single inquiry') still preferred receiving the evidence that would enable the identification of a frontrunner (day, evening or night). This suggests that more participants in the perceived multiple inquiries condition 'switched' strategy earlier than those in Condition 1. In the latter case, in scenario 3 the majority still preferred to receive evidence that would identify a frontrunner, although only with a 60% chance of being the true culprit, rather than eliminating a suspect. We will directly assess whether participants are specifically acknowledging these strategies (i.e., frontrunner and elimination) in later analyses.

These between-condition differences do not align with the maximising information goals of the utility functions we utilised. More broadly, purely information-theoretic OED models do not take into account factors such as task framing and context when computing the expected utility of a query or a query outcome.

4.3.5. Belief Updating (Posterior Probability Estimates)

To explore whether errors of belief updating were related to participants' biased selection of queries, we compared participants' posterior probability estimates given the hypothetical observation of each of the two outcomes (i.e., the probabilistic estimates they gave when asked about the probability of guilt of each suspect given they observed certain outcomes e.g., money or electronics) to the normative posteriors given this evidence computed using the IB-OED models (see Figure 8).

As such we built a Linear Mixed Effects Model with 'probability estimate given item outcomes' as a dependent variable and 'Suspect' (indicating whether the estimate was for Suspect 1, 2 or 3), 'Condition', 'Scenario', and 'Data Type' (whether the estimate was obtained from participants or the normative model). A random effect of subject was included to account for individual differences.

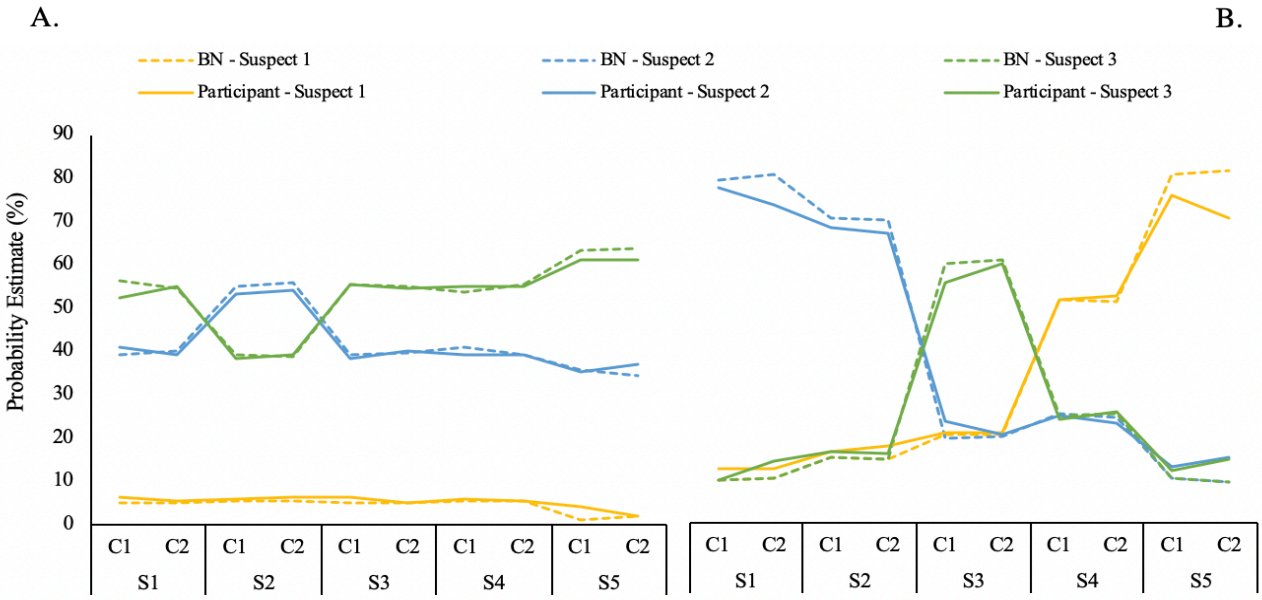

Figure 8. Experiment 3: Average probability estimates given 'primary item stolen' outcomes (A) and 'burglary time' outcomes (B) of participants and informed BN's for each suspect, in each scenario and condition.

We found no main effect of 'Condition', F (1, 4020) = 0.3, p = 0.59, $\eta_p^2 < 0.0001$, 'Scenario', F (4, 4020) = 0.45, p = 0.78, $\eta_p^2 < 0.0001$, or 'Data Type', F (1, 4020) = 0.26, p = 0.61, $\eta_p^2 < 0.0001$. The only main effect was of 'Suspect', F (2, 4020) = 37292, $p < 0.0001$, $\eta_p^2 = 0.95$, given that 'primary item

stolen' outcomes differentially affected the probability of guilt of different suspects. We found no significant interaction effects between 'Data Type', 'Condition' and 'Scenario'. The same procedure was carried out for probability estimates given 'burglary time' outcomes. Once again, we found no main effect of 'Condition', F (1, 4020) = 0.003, p = 0.9, $\eta_p^2 < 0.0001$, 'Scenario', F (4, 4020) = 0.25, p = 0.91, $\eta_p^2 < 0.0001$, or 'Data Type', F (1, 4020) = 0.03, p = 0.86, $\eta_p^2 < 0.0001$, on participants' estimates. The only main effect we found was once again of 'Suspect', F (2, 4020) = 1490, $p < 0.0001$, $\eta_p^2 = 0.43$. We found no significant interaction effect between 'Data Type', 'Condition' and 'Scenario'. Overall, it appears that participants' estimates given both 'burglary time' and 'primary item stolen' outcomes, closely approximated normative predictions.

4.3.6. Strategies: Think-aloud responses

In order to explore the motivated strategies that underlie participants' query selections, as well as their adaptiveness across conditions and probabilistic contexts (scenarios), we once again analysed participants' think-a-loud responses associated with their query selections. The coding procedure followed that outlined in previous experiments. A primary rater coded all responses and a second independent rater coded 25% of the total sample of responses (given the larger number of responses in this experiment, a smaller percentage of these was second-coded compared to the previous experiments), randomly selected (N = 170 out of 680). Cohen's weighted kappa determined a moderately high inter-rater agreement between the two raters, $\kappa_w = 0.71$, $p < 0.001$. In condition 1, 47 responses out of the total responses (N = 328), and in Condition 2, 44 responses out of the total responses (N = 349) were coded as "n/a" following the same criteria used in previous experiments. All responses were included in subsequent analysis. Below are the proportion of responses that were assigned each code throughout the task overall (Table 20) and within each scenario and condition (Figure 9).

Table 20

Experiment 3: Percentage use of strategy codes across conditions (collapsing scenarios).

Strategy Code	Percentage Use (across scenarios) Condition 1	Percentage Use (across scenarios) Condition 2
Frontrunner	37%	36%
Elimination	1.8%	11%
Symmetry	11.5%	13%
Differentiation	24%	18%
Frontrunner + Zero-sum/Risk Aversion	3.6%	1.4%
Highest Percentage	4.5%	6%
Zero-sum/ Risk Aversion	3%	0.3%

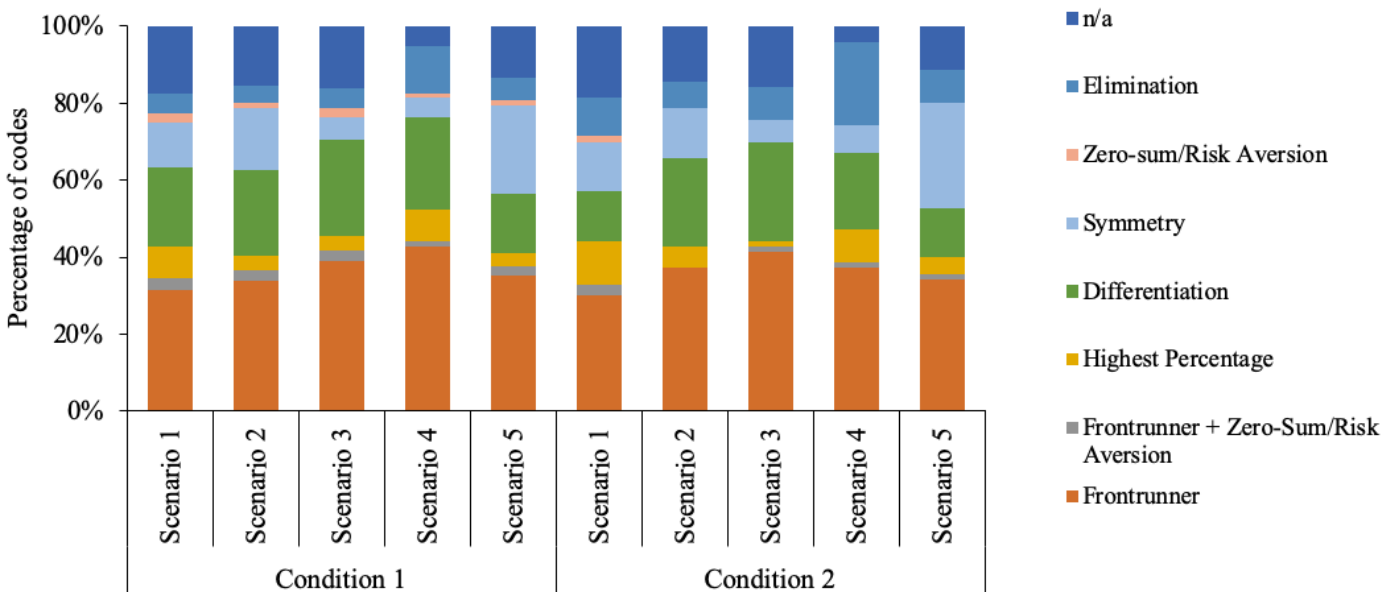

Figure 9. Experiment 3: Percentage of strategies per scenario and condition

4.3.6.1. Adaptability of Strategies

In order to explore whether participants' strategy use varied across scenarios within each condition, we initially carried out two Friedman tests (within each condition). No significant difference in the adoption of different strategies was found within condition 1, $\chi^2(4) = 6.04$, $p = 0.2$. A significant difference was found within condition 2, $\chi^2(4) = 11.2$, $p = 0.024$. This can be attributed to participants in condition 2 utilising an 'elimination' strategy significantly more in scenario 4 than in scenario 1, $p < 0.014$.

Subsequently we collapsed the scenarios in order to explore between-condition differences in the adoption of different strategies between conditions. This allowed us to test the hypothesis that participants in the 'perceived multiple inquiries' condition might employ an 'elimination' strategy significantly more than participants in the 'single inquiry' condition who comparatively would utilise a 'frontrunner' strategy significantly more and would display more risk-averse behaviour (given the one-shot nature of the task framing). A Chi-Square test of Independence illustrated that the distribution of strategies overall varied across the two conditions, $\chi^2(7) = 37.4$, $p < 0.0001$, $V = 0.24$. Post-hoc contrasts with a threshold of $\alpha = 0.05$ determined the significant between-condition differences were attributable to the differences in use of an 'elimination' strategy and a 'zero-sum/risk aversion' strategy. In line with our predictions, more participants in condition 2 adopted an 'elimination' strategy than participants in condition 1 and conversely more participants in condition 1 adopted a 'zero-sum/risk aversion' strategy than participants in condition 2 (see Table 20).

Whereas in previous experiments we noted that participants' strategies remain largely unvaried across probabilistic contexts, this was not the case here. Moreover, strategies seem to be sensitive to task framing. The finding that more participants adopted an elimination strategy in condition 2 intuitively makes sense given that they believe they have longer in the investigative process to determine an individual suspect. Thus, our findings suggest that strategies and query selections are both sensitive to task framing and probability contexts. However, such sensitivity is contingent on cases where fulfilling the strategic preferences of obtaining a frontrunner is less feasible (i.e., scenario 5 in which the frontrunner would only have a 50% chance of being the true culprit and scenario 4 to an extent, for a similar reason).

4.3.6.2. Strategies and Query Selection

Next, we explored whether, within each condition (varying in terms of how the task was framed e.g., 'perceived multiple inquiries' versus 'single inquiry') certain strategies systematically underlie different query selections. In condition 1, 63% of participants employing a 'frontrunner' strategy, 92% of participants employing a 'frontrunner + zero-sum/risk aversion' strategy, 66% of participants employing a 'differentiation' strategy, 92% of participants employing a 'symmetry' strategy, and 90%

of participants employing a 'zero-sum/risk aversion' strategy selected the query 'burglary time'. In this same condition, 83% of participants who used an 'elimination' strategy and 60% of those who employed a 'highest percentage' strategy queried 'primary item stolen'.

In comparison, in condition 2, 51% of participants employing a 'frontrunner' strategy, 60% of participants employing a 'frontrunner + zero-sum risk aversion' strategy, 65% of participants employing a 'differentiation' strategy, 93% of participants employing a 'symmetry' strategy, and 100% of participants employing a zero-sum/risk aversion strategy (though here $n = 1$) selected the query 'burglary time'. Meanwhile 75% of participants who used an 'elimination' strategy and 60% of those who employed a 'highest percentage' strategy queried 'primary item stolen'. Once again, certain strategies dictated different query selections with an 'elimination' strategy being related to querying 'primary item stolen' and the remaining strategies being predominantly related to querying 'burglary time'.

To explore the extent to which strategies differentially underlie query selections and vary both across scenarios (within-participants) and across conditions (between-participants), we built a General Linear Mixed Effects Model with binomial distribution and log link function. Our model had three fixed effects: Condition, Scenario and Strategy; one random effect: Subjects, with intercept to account for individual variability; and one dependent factor: query selection (count). This analysis showed no main effect of 'Condition', $F(1, 606) = 1.97$, $p = 0.16$, $\eta_p^2 = 0.003$, but a main effect of 'Scenario', $F(4, 606) = 7.7$, $p < 0.0001$, $\eta_p^2 = 0.05$, and a main effect of 'Strategy', $F(7, 606) = 7.8$, $p < 0.0001$, $\eta_p^2 = 0.08$. The main effect of 'Scenario' is addressed in previous analyses reported in section 4.3.2 (participants' query selection preferences did vary across scenarios).

In terms of 'Strategy', the significant pairwise comparisons in predicting query selection was between the 'elimination' strategy and: 'frontrunner', $p = 0.024$, 'frontrunner + zero-sum/risk aversion', $p = 0.028$, 'differentiation', $p < 0.0001$, 'symmetry', $p < 0.0001$, and 'zero-sum/risk aversion', $p < 0.002$. As can be seen in the descriptive statistics mentioned above, an elimination strategy was associated with querying 'primary item stolen' significantly more than the other strategies that predominantly were underlying 'burglary time' query choices. In addition, significant pairwise

comparisons were found between 'symmetry' and 'frontrunner', $p < 0.0001$, 'highest percentage', $p < 0.0001$, and 'differentiation', $p = 0.02$. As such, a symmetry strategy was more strongly associated with querying 'burglary time' than the remaining strategies (though these also were predominantly underlying the same query).

Using a full factorial design, the only significant interaction effect was between 'Scenario' and 'Strategy', $F(28, 606) = 2.1$, $p = 0.001$, $\eta_p^2 = 0.09$. This is due to the fact that although participants' query selections varied in some scenarios (e.g., between scenario 1 and 2 and scenarios 3, 4, 5), the extent to which certain strategies were related to certain queries (i.e., elimination strategy and querying 'primary item stolen') did not vary across scenarios. As such, our findings suggest that strategies dictate different query selections and they potentially indicate that the nature of the condition shifts query selections independently of strategy use.

For details on how strategies related to query selection accuracy according to each utility function, see supplementary materials (S6).

4.4. Discussion

In Experiment 3 we explored people's information acquisition and evaluation behaviours across five different probabilistic contexts using a within-subjects design. Moreover, we explored the effect of task framing on participants' query and query outcome preferences as well as on the adoption of a frontrunner vs. elimination strategy. None of the utility functions were sensitive to the different strategies that participants employed (e.g., frontrunner vs. elimination). In terms of participants' information search decisions, our mixed-effect logistic regression analyses showed that none of the models were significant predictors of participants' query preferences. Descriptively however, probability gain based models were arguably able to account for the directional majority preference of participants' choices in scenarios in which alternative models failed. In this set-up however, PG was not considered to be the optimal strategy given that half of the participants assumed that the task would involve multiple inquiries and therefore adopting a PG strategy at the outset is not necessarily the most rational choice. Overall, in the present experiment, the purely information-theoretic models we employed were not able to account for the observed differences in query evaluations between

participants who believed the task involved multiple inquiries, and those who believed it was a one-shot task.

Our results illustrated that participants in condition 1 who were told the whole investigation entailed only one inquiry, had a modal preference for querying 'burglary time' in all scenarios except scenario 4. As per the previous experiments, this was underlain by an adoption of strategies that would enable the safe identification of a frontrunner, and disambiguation of suspects. In scenario 4, querying 'burglary time' would only increase the posterior probability of any given suspect to 50%, triggering participants to switch preference and select the query 'primary item stolen'. At this point their choices aligned with the predictions of all five model predictions. On the other hand, participants in condition 2 who were told the investigation comprised of multiple inquiries (although they were required to only make the first one), evaluated queries slightly differently than their counterparts in condition 1. As such, their preference switched given a different probabilistic model, preferring to query 'primary item stolen' over 'burglary time' (which could only lead to frontrunner with only 60% probability of being the true culprit) as early as in scenario 3. Nonetheless, participants in this condition were also primarily driven by a frontrunner strategy throughout the task, although we found they adopted an elimination strategy more than participants in condition 1. We note that although some differences were found in the adoption of frontrunner vs. elimination strategies between conditions, our manipulation might not have been strong enough to induce a significant change in strategy adoption between participants. This may be due to limiting the selection of queries to one, regardless of task framing condition. The instructions might not have been enough to instil in participants the knowledge that the investigation allowed subsequent inquiries. Experiment 4 adopts a stepwise paradigm to address this issue. Nevertheless, our results still suggest that when evaluating queries, the context and task framing are significant factors that both descriptive and normative frameworks of information seeking should be able to account for.

In the current experiment, in order to specifically investigate frontrunner versus elimination driven strategies, we asked participants to select one piece of evidence between either money/electronics or a burglary time outcome. Here, participants did not reason under uncertainty, instead simulating the impact of each of the two pieces of evidence on the probability of each suspect

being the culprit. Normatively, one piece of evidence would lead to an almost certain 'elimination' of a suspect and one piece of evidence would lead to a leading hypothesis (although the extent of this differed across scenarios). Participants' outcome selections mirrored their query selections, as participants in condition 1 preferred receiving evidence relating to 'burglary time' than evidence relating to the 'primary item stolen' in all scenarios, excluding scenario 4. Participants in condition 2 switched preference earlier, at scenario 3, at which point they preferred receiving either money or electronics as evidence (and thus eliminate a suspect). The fact that participants' posterior estimates were accurate (given each item of evidence) indicates that participants could accurately predict how much the items of evidence would change the probability of guilt of each suspect, which further suggests that when they chose a certain piece of evidence it was because they could correctly anticipate its effect. This allows us to conclude that the observed preference for 'burglary time' outcomes was due to a cognizant preference for a frontrunner, and not due to belief updating errors.

This overall preference for a frontrunner is especially relevant when exploring information seeking in real-world instances such as criminal investigation, in which information itself acts as a reward and the accuracy of a judgement is of great importance (i.e., having the wrong lead suspect could have deleterious consequences on subsequent evidence gathering opportunities, potentially leading to erroneous convictions), and eliminating a suspect might be as valuable as catching the culprit. The "frontrunner bias" we observe in our studies thus far, i.e., preferring to identify a frontrunner at the outset of an investigation, could therefore prove problematic if it translates to naturalistic environments in which tunnel-vision in investigations and confirmatory information search strategies have led to miscarriages of justice (Eady, 2009). In a subsequent experiment, we will address whether adopting the frontrunner strategy at the outset leads to selective hypothesis-testing and confirmatory search strategies.

Findings from Experiment 3 yield optimistic insights, as, despite mostly selecting queries that were not optimal according to our OED utility functions, participants were able to accurately estimate the impact of evidence on each hypothesis. This is in contrast to previous literature illustrating participants' errors in integrating information within a Bayesian framework (i.e., Bar-Hillel, 1980;

Fischoff & Beyth-Marom, 1983; Tversky, 1982). In addition, it suggests that errors in estimating the change in probability distributions of outcomes do not underlie participants' deviations from OED principles when evaluating queries. Rather, participants may be committing errors in integrating the weighted probabilities of outcomes occurring when choosing what query would provide them with the most informative evidence, or not taking these probabilities into account at all. For example, similar to behaviour observed in Experiment 2, by overweighting the probability of the outcome 'money' occurring in all scenarios participants might have resorted to the safer 'frontrunner-guaranteed' query option.

Overall, Experiment 3 further substantiates the findings from previous experiments, showing that the adoption of common motivated strategies, such as the identification of a frontrunner, can account for most of participants' query selections. However, the utility function predictions were insensitive to these differences. Moreover, findings from this experiment strengthened the notion that although basic evaluations at the level of outcomes are sound, participants' strategies induce a myopic focus on lower level cues, as well as an inability to accurately integrate the prior probability of the evidence occurring with its diagnosticity. This behaviour is possibly what underlies query selections that deviate from standard Bayesian OED predictions. Finally, we have shown that purely information-theoretic utility functions do not adequately describe information acquisition when task framing is varied.

5. Experiment 4

To further explore the effect of task framing on participants' strategies, we extended the task to a stepwise paradigm. We allowed participants to make multiple sequential query selections, observe the outcome of each query, and update their probabilistic beliefs after each observation. This final experiment will allow us to explore the following points:

1) Whether people's information seeking differs at the first decision point of a stepwise paradigm compared to that observed in our previous one-shot experiments.
2) The effect of adopting frontrunner vs. elimination strategies at the outset of the task on subsequent search decisions and belief updating.

5.1. Global Bayesian OED Model

All participants were presented with materials underpinned by the same probabilistic model built as a BN (see Figure 10).

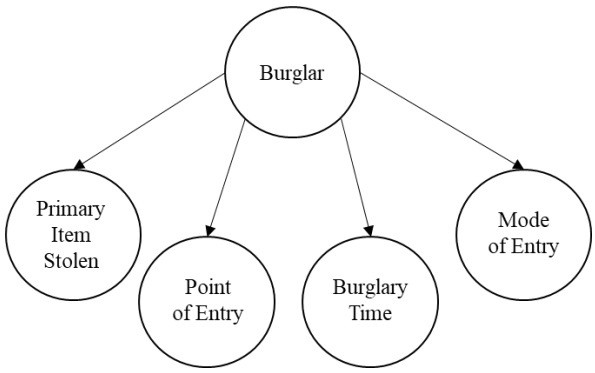

Figure 10. Experiment 4: Graphical representation of Bayesian Network

This model was more complex than that used in previous experiments. We used a ternary hypothesis node 'Burglar' (states: Suspect 1, Suspect 2, Suspect 3) with uniform priors so that in all models P (Suspect 1) = P (Suspect 2) = P (Suspect 3) = $\frac{1}{3}$, three binary-outcome query nodes: 'burglary time' (outcomes: day, night), 'point of entry' (outcomes: door, window), 'mode of entry' (outcomes: forced, not forced) and a ternary-outcome query node 'primary item stolen' (outcomes: jewellery, electronics, money). The parameters of the probabilistic model in the BN, and given to participants, can be viewed in Table 21 below. These parameters were selected as they allowed the identification of a clear order of queries in terms of their informative value, though to a different extent across utility functions, and the informative values of the queries differed enough that this difference should once again be noticeable. We kept parameters of two queries similar to our previous experiments such that one would facilitate the identification of a safe 'frontrunner' given its symmetric properties, and one might facilitate the 'elimination' of a suspect. Moreover, as will be explained below, varying the informative value of the queries allowed us to compute an optimal stepwise information search strategy.

Table 21

Experiment 4: Conditional Probability Table with parameters employed in the global CBN

Hypothesis	Primary Item Stolen (Jewellery, money, electronics)	Point of Entry (door, window)	Mode of Entry (forced, not forced)	Burglary Time (day, night)
Suspect 1	0.1, 0.1, 0.8	0.3, 0.7	0.5, 0.5	0.9, 0.1
Suspect 2	0.8, 0.1, 0.1	0.3, 0.7	0.8, 0.2	0.1, 0.9
Suspect 3	0.1, 0.8, 0.1	0.7, 0.3	0.1, 0.9	0.6, 0.4

The expected utility of each investigative query relative to the probabilistic belief model (specified in Table 21) was calculated as in the previous experiments utilising KL-D, IG, PG and Impact as utility functions. Assuming an agent selected the most informative investigative query and normatively (as dictated by Bayes theorem) updated the probabilities at each decision stage, the expected utility of each investigative query according to each utility function can be seen in Table 22 below. Given the change in paradigm from one-shot to a stepwise investigative task, probability gain based models are not necessarily the optimal strategies in this task given that maximising the suspect's posterior probability might not be the best strategy to employ when the participant knows they have multiple opportunities of obtaining information. For this reason, a simplified PG model (PG_H) was not included in the present experiment.

Although participants could freely select an investigative query at each decision stage, the outcomes of these queries were kept constant so that throughout the task, despite the changing order, all participants would have observed the same outcomes by the end of the task, making within-group comparisons more tractable. As such, if *primary item stolen* was queried the evidence 'jewellery' would be observed, if *mode of entry* was queried the evidence 'door' would be observed, if *point of entry* was queried the associated evidence was 'non-forced entry' and finally if *burglary time* was queried the evidence 'day' would be observed. We chose these outcomes to increase the complexity of the task and test participants' ability to integrate discordant evidence, as the evidence was sometimes diagnostic towards different suspects. Moreover, this allowed us to explore the adaptiveness of people's strategies and belief updating given unexpected evidence.

Table 22

Experiment 4: Expected value of each query at each decision stage predicted by each utility function.

	Utility Function	Primary Item Stolen	Burglary Time	Mode of Entry	Point of Entry
Decision 1	KL-D	0.66*	0.36	0.26	0.10
	IG	0.66*	0.36	0.26	0.10
	PG	0.47*	0.27	0.23	0.13
	Impact	0.31*	0.19	0.16	0.12
Decision 2	KL-D	-	0.26*	0.16	0.04
	IG	-	0.26*	0.16	0.04
	PG	-	0.01*	0	0
	Impact	-	0.14*	0.11	0.05
Decision 3	KL-D	-	-	0.23*	0.09
	IG	-	-	0.23*	0.09
	PG	-	-	0.12*	0.07
	Impact			0.14*	0.1
Decision 4	KL-D	-	-	-	0.03
	IG	-	-	-	0. 03
	PG	-	-	-	0.13
	Impact	-	-	-	0.13

N.B. The most informative query at each decision stage (according to each utility function) is marked with '*'

5.1.1. Informed B-OED Models

As the parameters assumed by the probabilistic belief model impact the computation of the expected utility of queries and outcomes, we did not assume participants would simply assume the stated parameters, and therefore built individually fitted (informed) Bayesian OED models (IB-OED) for each participant. These were parametrised according to each participant's stated hypothesis priors, and posterior beliefs after each decision point (having observed the query outcome), incorporating their query selection throughout the task. This allowed us to evaluate each participant's information acquisition and evaluation against the "fitted" normative model describing an individualised optimal sequential search and updating strategy (computed according to each utility function). In this way, participants were not "damned" by one initially sub-optimal query selection, allowing for meaningful assessment of normativity at later stages of the task. Expected utility was not computed at decision stage 4, as all participants by default chose the last query remaining.

5.2. Methods

5.2.1. Participants and Design

A total of 117 participants (N_{male} = 37; M_{age} = 36.1 years, SD = 7.3) were recruited from Prolific Academic. All participants were native English speakers who gave informed consent and were paid £1.5 for partaking in the present study that took on average 14 (median) minutes to complete online. The experimental task was designed in Qualtrics and powered through the online platform Prolific Academic. A within-subjects design was employed. For task materials see osf.io/tkr4v.

5.2.2. Procedure

As part of the cover story, participants were once again asked to imagine they were criminal investigators. As per the previous experiments, they were foremost asked to review the neighbourhood's burglary statistics and the criminal records of the (three) burglars known to operate in the area and were therefore provided with information on the (probabilistic) model (i.e., variables present, causal relationships between these, prior probabilities of the burglars and conditional probabilities within the model).

Participants were then told that a new burglary had occurred in their neighbourhood, and that they were to investigate the new case. Prior probabilities of each burglar being the culprit were elicited from participants to ensure the uniform priors had been accepted. Subsequently, participants were asked to select one of four investigative queries: 'burglary time' (to find out whether the burglary occurred during the day or night), 'primary item stolen' (to find out whether electronics, money or jewellery were primarily stolen), 'point of entry' (to find out whether they entered through door or window) and 'mode of entry' (to find out whether they used force entry or not). Participants were not told explicitly that they would be able to select all queries throughout the task, but this became apparent as they progressed through the task. Moreover, they were asked to carefully consider each query selection so as to maximise the effectiveness of the whole investigation. After selecting a query, participants were required to provide a think-aloud response explaining their choice. Subsequently, participants would observe the evidence associated with their selected query (query outcome). For example, if participants

selected the query 'point of entry' they would observe the following message: "You investigate the point of entry and find out the burglar came in through the door".

After learning the outcome of a query, participants were required to provide updated posterior beliefs for each suspect (as probability estimates using sliders ranging from 0% to 100%). Due to the principles of mutual exclusivity and exhaustiveness, the estimates were required to sum to 100. After providing the probability estimates participants were told to select the next query. This procedure was repeated until all investigative queries had been exhausted.

Throughout the task, all participants made three active queries in total (the final query chosen was the one left outstanding) and provided four posterior belief updates for each suspect (and the initial hypothesis priors estimates). Participants had access to an 'information review' section throughout the task, comprising of the information represented by Table 21, as well as a list of what evidence (query outcomes) had already been observed. At the end of the sequential task, given all evidence observations, participants were required to select which suspect they would like to bring in for questioning.

5.3. Results

5.3.1. Query Selection

The proportion of participants who selected each investigative query at each decision stage can be visualised in Figure 11 below. For details on query selection accuracy as defined by each of the utility functions see supplementary materials (S6). We directly test the predictive abilities of these models in the subsequent analyses in section 5.3.2.

In order to compare the predictive abilities of the utility functions, we computed Kendall-tau's correlations between the queries predicted to be most informative by each of the OED models with different utility functions (fitted to participants' own priors) at each decision stage, and participants' actual query choices. As can be seen from Table 23 below, at the first decision stage, although all models are correlated, they had virtually no correlation with participant choices at decision stage 1. This is because all utility functions predicted the query 'primary item stolen' as the most informative query at this stage for at least 95% of participants, whereas participants displayed a split preference between

two queries (see Figure 11). In subsequent stages, all utility functions displayed a moderate correlation between each other as well as participants' responses, except the KL-D and IG, which displayed an extremely strong correlation with Impact across all decision stages.

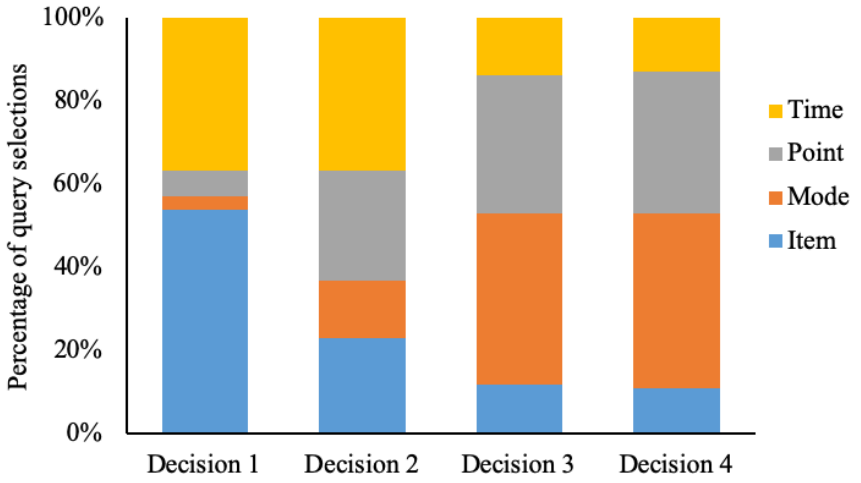

Figure 11. Proportion of participants' query selections at each decision stage

Table 23

Experiment 4: Kendall tau-b correlations between participant query selections and predictions of each utility function at each decision stage.

Decision Stage		KL-D/IG	PG	Impact
1	Participant	-0.08	-0.08	-0.05
	KL-D/IG	-	0.4**	1.00**
	PG	-	-	0.4**
2	Participant	0.61**	0.55**	0.61**
	KL-D/IG	-	0.9**	1.00**
	PG	-	-	0.9**
3	Participant	0.6**	0.36**	0.6**
	KL-D/IG	-	0.7**	0.95**
	PG	-	-	0.66**

** Significant at $\alpha = 0.05$

5.3.2. Utility Function Model Comparison

At each of the first three decision stages, we built multinomial mixed effects regression models[13] with 'Participant Choice' as our multinomial outcome variable and 'Utility Function Prediction' as our fixed effect categorical predictor, for each utility function. Each model additionally contained a random effect with intercept of 'Subject'.

5.3.2.1. Decision Stage 1

The outputs of the multinomial logistic regressions can be seen in Table 1a of Appendix 1. Our analyses yielded no main effect of 'Utility Function Prediction' on 'Participant Choice' in the PG model, F (3, 111) = 0.99, p = 0.39, the KL-D/IG model, F(3, 111) = 0.01, p = 0.99, or in the Impact model, F (3, 111) = 0.01, p = 0.99. At the first decision stage none of the utility functions were significant predictors of participants' query selections. The likelihood ratio results comparing each model to an intercept-only model are displayed in Table 24 below.

Table 24

Experiment 4: Likelihood ratio test results, AIC, Deviance, Akaike Weights (w) and Evidence Ratio (ER) values of the competing models.

Model	AIC	ΔAIC_i	w_i	ER_i	Deviance	χ^2	df	p-value
PG	239.3	0	0.56	1	227.3	3.2	3	0.36
KL/IG	241.3	1.9	0.22	2.6	229.3	1.24	3	0.74
Impact	241.3	1.9	0.22	2.6	229.3	1.24	3	0.74

In order to compare the competing utility functions models and select the best approximating model, we once again used derivatives of Akaike's Information Criterion (AIC) measure and followed the same procedure outlined in section 2.2.3. Given that the PG model had the lowest AIC value we selected this as the 'best' reference model. The ΔAIC values presented in the above table suggest that both the KL/IG and the Impact model could not be discounted given that $\Delta AIC < 2$. The computed

[13] Given the lack of implementations of repeated measures factors in mixed effects regression models for a multinomial outcome, and the fact that we are purely interested in the predictive abilities of the utility functions, the present analyses were carried out separately for each decision stage.

Akaike weights showed that PG has a 56% chance of being the correct model amongst these, with the remaining weight being equally distributed amongst the KL/IG and Impact models. These findings intimate that none of the models at decisions stage 1 are able to accurately predict participants' choices. As such, at this decision stage, despite every utility function predicting 'primary item stolen' to be the most informative query for the vast majority of participants, almost 30% of participants selected 'burglary time'.

5.3.2.2. Decision Stage 2

The outputs of the mixed-effect logistic regressions carried out on decision stage 2 can be seen in Table 1b of Appendix 1. A main effect of 'Utility Function Prediction' on 'Participant Choice' was found in the KL-D/IG model, $F(6, 108) = 2.9$, $p = 0.01$, and in the Impact model, $F(6, 108) = 2.9$, $p = 0.01$, but not in the PG model, $F(12, 102) = 1.7$, $p = 0.078$. The likelihood ratio results comparing each model to an intercept-only model are displayed in Table 25below.

Table 25

Experiment 4: Likelihood ratio test results, AIC, Deviance, Akaike Weights (w) and Evidence Ratio (ER) values of the competing models.

Model	AIC	ΔAIC_i	w_i	ER_i	Deviance	χ^2	df	*p*-value
PG	275.1	21.36	1E-05	43477	221.1	74.5	12	< 0.0001
KL/IG	253.7	0	49.99	1	235.7	75.5	6	< 0.0001
Impact	253.7	0	49.99	1	235.7	75.5	6	< 0.0001

As can be seen from Table 25, given that the KL/IG models had the lowest AIC value we selected these as the 'best' reference models. The ΔAIC values presented in the above table suggest that the PG model could be discounted as it is implausible given that ΔAIC > 10. The computed Akaike weights showed that KL/IG and Impact each have approximately a 50% chance of being the correct models amongst these. In combination with the findings presented above, this strengthens the notion that none of the utility functions are accurate predictors of participant choice behaviour in this sequential information search task.

5.3.2.3. Decision Stage 3

The outputs of the multinomial logistic regression analyses carried out in Decision Stage 3 can be seen in Table 1c of Appendix 1. A main effect of 'Utility Function Prediction' was found on 'Participant Choice' in the KL-D/IG model, F (6, 108) = 4.7, $p < 0.0001$, and in the Impact model, F (6, 108) = 4.6, $p < 0.0001$, and in PG, F (12, 102) = 2.3, $p = 0.01$. The likelihood ratio results comparing each model to an intercept-only model are displayed in Table 26 below.

Table 26

Experiment 4: Likelihood ratio test results, AIC, Deviance, Akaike Weights (w) and Evidence Ratio (ER) values of the competing models.

Model	AIC	ΔAIC_i	w_i	ER_i	Deviance	χ^2	df	p-value
PG	254.5	38	5.6E-09	178482301	224.5	69.8	12	< 0.0001
KL/IG	229.5	13	0.001	665	211.5	82.8	6	< 0.0001
Impact	216.5	0	99.8	1	198.5	95.9	6	< 0.0001

As can be seen from Table 26, given that the Impact model had the lowest AIC value we selected it as the 'best' reference model. The ΔAIC values presented in the above table suggest that both the PG model and the KL/IG model can be discounted and are implausible given that ΔAIC > 10. The computed Akaike weights showed that Impact model has an almost 100% chance of being the correct model compared to the alternative models. However, when considering the findings presented in the supplementary materials (S6), showing that at this decision stage all measures performed worse than chance level, we can conclude that although Impact is a better fit relative to the alternative models, overall none of them are particularly good fits of the distribution of participants' query selections. This is corroborated by the Table 1c in the Appendices, showing that, in the Impact model, a prediction of 'mode of entry' compared to one of 'burglary rime' significantly increased the odds of a participant choosing 'mode of entry' or 'point of entry'. It therefore seems that, despite the above-mentioned being significant prediction terms, the reference model is not able to differentiate between participants' query

preferences (e.g., mode of entry and point of entry) which were equally prevalent at this decision stage (see Figure 11).

5.3.3. Belief Updating

In order to assess participants' belief updating accuracy against IB-OED models, the absolute difference between the observed (participants' empirical estimates) and predicted (by IB-OED models) posterior belief estimates was computed for each suspect at each decision stage (see Figure 12).

In order to assess whether updating accuracy (indicated by a low absolute mean difference) differed across decision stages or suspects we ran a repeated-measures ANOVA. We found a main effect of decision stage on belief updating accuracy, F (2.9, 339) = 10.9, $p < 0.001$, $\eta_p^{2.} = 0.086$. As can be seen from Figure 12, participants were significantly less accurate at decision stages two and three.

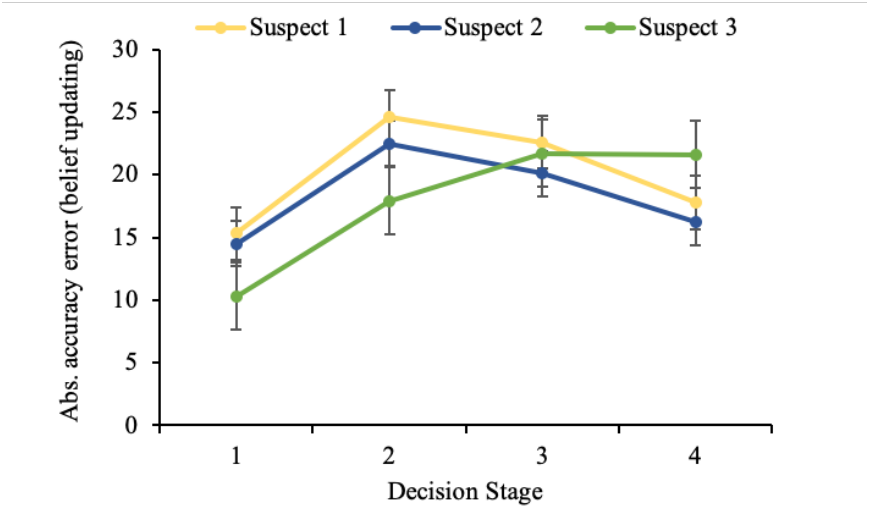

Figure 12. Experiment 4: Participants' average absolute belief updating error at each decision stage (according to IB-OED model predictions).

Additionally, a main effect of suspect on updating accuracy was not found, F (1.9, 221) = 2.14, $p = 0.12$, $\eta_p^{2.} = 0.018$. Finally, a significant interaction effect was found between suspect and decision stage, F (5.5, 638) = 4.8, $p < 0.001$, $\eta_p^{2.} = 0.04$. People's accuracy error increases throughout the decision stages for Suspect 3. As such, participants' updating error regarding suspects 1 and 2 increases noticeably from decision stage 1 to decision stage 2, before decreasing to its original level. In contrast,

participants' updating error increases throughout the decision stages for Suspect 3. We will discuss this further in relation to participants' strategies.

Given that we found no main effect of suspect on participants' accuracy, we averaged participants' absolute error of the three suspects so as to obtain a single measure of probability updating accuracy per participant at each decision stage. A repeated measures ANOVA with Greenhouse-Geisser correction illustrated a significant difference in participants' updating accuracy between decision stages, F (3, 348) = 10.9, $p < 0.001$, $\eta_p^{2.} = 0.086$. Post-hoc pairwise comparisons with Bonferroni correction, illustrated that participants' updates were more accurate at decision stage 1 (M = 13.43, SD = 14.16) compared to decision stage 2 (M = 21.7, SD = 12.1), t (116) = -5.23, $p < 0.001$, decision stage 3 (M = 21.5, SD = 13.5), t (116) = - 4.54, $p < 0.001$, and decision stage 4 (M = 18.5, SD = 13.9), t = -2.9, p = 0.02. No significant difference was found between decisions stage 2 and 3, or between stages 2 and 4, or between stages 3 and 4. This suggests that participants' updating accuracy significantly decreased after decision stage one, and then plateaued.

For details on how belief updating related to query selection accuracy as computed by each utility function, see supplementary materials (S7).

5.3.4. Strategies: Think-aloud

In order to explore the strategies that underlie participants' query and query outcome evaluations, as well as their adaptiveness across the task, we once again analysed participants' thought-a-loud responses associated with their query selections. The coding procedure followed that outlined in previous experiments. A primary rater coded all responses. A second independent rater then coded the responses of 47% of participants (randomly selected) from the first three query selections. Cohen's weighted kappa was utilised to determine a moderately high inter-rater agreement between the two raters, $\kappa_w = 0.86$ $p < 0.001$. 60 responses out of the total 351 (responses of 117 participants on three decisions) were attributed the primary code of "n/a" given the dearth of information provided. As in previous experiments, all responses were included in subsequent analyses. Figure 13 illustrates the proportion of responses accounted for by each strategy code at the first three decision stages. Table 27 shows the percentage use of each strategy throughout the task overall (collapsed by decision stage).

Table 27

Experiment 4: Participant strategy usage (percentage), collapsed across decision stages

Strategy Code	Percentage Use (across decisions stages)
Frontrunner	25%
Elimination	16%
Symmetry	5%
Differentiation	13%
Frontrunner^{+} (frontrunner-focused)	22%
Highest Percentage	3%
Zero-sum/ Risk Aversion	0%

Figure 13. Experiment 4*:* Percentages of participant strategy usage per decision stage.

The strategies were the same as ones obtained in previous experiments, with the addition of a new strategy that we termed frontrunner^{+} (frontrunner-focused). Rather than wanting to determine *any* frontrunner suspect (i.e., the standard frontrunner strategy discussed previously), the frontrunner-focussed strategy differs in that is seeks to confirm a *specific* frontrunner (i.e., the leading suspect obtained from the previous decision stage). As such it represents a more confirmatory/selective hypothesis testing strategy than simply wanting to obtain *a* frontrunner at the initial stage, at which point all suspects had equal priors. For example, one participant stated, "*Having seen the last piece of evidence, I am interested to see if the burglary time matches Suspect 2's MO*" and similarly P31 explained, "*If it was night it would add to the likelihood it's the Nightingale (Suspect 2)*".

5.3.4.1. Adaptability of Strategies

In order to determine whether participants' strategy choice varied throughout the task (i.e., across the different decision stages), we conducted a non-parametric Friedman test. Our analysis displayed a, albeit barely, significant difference in the use of strategies across the decision stages, χ^2 (2) = 5.8, p = 0.05. As such, Bonferroni corrected post-hoc comparisons of proportions illustrated the significant difference to be in the adoption of an elimination strategy across stages with participants adopting it

significantly more at decisions stage one than at both subsequent stages, $p < 0.05$. Moreover, participants utilised a frontrunner strategy significantly more at decision stage 1 and 3 than decision stage 2. At decision stage 2 participants adopted a frontrunner + (frontrunner-focused) strategy significantly more than at the first decision stage (at this stage it was conceptually impossible to utilise this strategy). Finally, participants utilised a 'highest percentage' strategy significantly more at the first decision stage than either of the subsequent two.

This seems to suggest that, in contrast to our previous one-shot experiments, participants' strategy use was responsive to the situation (e.g., probabilistic context at a certain decision stage). For example, the multiple inquiry nature of this experiment seems to have led to a large cluster of participants adopting an elimination strategy at the outset (behaviour comparable to that of participants in the 'perceived multiple inquiries' condition in Experiment 3).

5.3.4.2. Strategies and Query Selection

Next, we once again explored whether, within each condition, certain strategies systematically underlie different query selections. For details on how the strategies related to query selection accuracy as computed by each utility function, see supplementary materials (S8).

At decision stage one, 76% of participants who employed a 'frontrunner' strategy queried 'Primary Item Stolen' (the remaining participants selecting 'point of entry' or 'mode of entry'). Moreover, 77% of participants who utilised a 'differentiation' strategy, 82.1% of participants who utilised an 'elimination' strategy, and 50% of participants who utilised a 'highest percentage' strategy, selected the query 'primary item stolen'.

At decision stage 2, it was harder to detect systematic relations between query chosen and strategy employed (as previous analysis illustrated). Nonetheless, 57% of participants who utilised a 'differentiation' strategy and 50% of participants who utilised a 'highest percentage' strategy selected the query 'primary item stolen' at this stage. Comparatively, 58% of participants who utilised a frontrunner^{+} (frontrunner-focused) strategy and 42% of participants who adopted a frontrunner strategy

selected the query 'burglary time'. At decision stage three, given the smaller number of queries remaining, most of the participants' strategies underlie queries 'mode of entry' and 'point of entry'.

A general linear model with multinomial distribution and logit link function illustrated no main effect of decision point on participants' query selections $\chi^2(2) = 1.75$, $p = 0.42$, but a main effect of strategy, $\chi^2(6) = 40.2$, $p < 0.0001$. Moreover, a significant interaction effect was found between decision stage and strategy, $\chi^2(9) = 80$, $p < 0.0001$. As such, a significant difference in query preference was found within Decision Stage 1 (as previously discussed), $\chi^2(1) = 7.15$, $p = 0.007$. Moreover, at Decision stage 1 an 'elimination' strategy, $\chi^2(1) = 16.8$, $p < 0.0001$, and a 'highest percentage' strategy, $\chi^2(1) = 8.3$, $p = 0.004$, were most heavily associated with querying 'burglary time'. Comparatively, at decision stage 2 a frontrunner strategy, $\chi^2(1) = 7.2$, $p = 0.007$, and a frontrunner-focused strategy, $\chi^2(1) = 8.4$, $p = 0.004$, were both significantly associated with querying 'burglary time'.

This seems to suggest that, as in previous experiments, certain strategies dictate the observed query preferences, although in contrast to our previous one-shot studies, in a stepwise inquiry situation, strategy adoption seems to be dependent on decision stage.

5.3.5. Consequences of frontrunner-focused thinking

At the first decision stage the content of participants' think-aloud responses were reflective of the split observed in their query selections wherein a cluster of participants selected burglary time (driven by an elimination strategy) whilst the majority chose primary item stolen (driven by a frontrunner strategy). In this case, a frontrunner strategy led to an optimal choice as computed by all IB-OED models. However, this latter strategy also led many participants to less accurate belief updating at decisions stages two and three. As such, participants might have under-adjusted their beliefs for their leading hypothesis in light of contradictory evidence in order to maintain the same 'prime suspect' or frontrunner. Our previous analysis showing that a frontrunner-focused strategy led to more accurate query selections at decision stage 2 but inaccurate selections at decision stage 3 bolsters this notion.

Investigating further, we found that 53% of participants held the same suspect as lead across decision stages 1 and 2, despite the majority of participants having viewed contradictory evidence at

these stages (selecting 'primary item stolen' and 'burglary time' would show evidence diagnostic towards different suspects). Moreover, 30% of participants held the same suspect as lead across decision stages 1-3. As will be subsequently discussed, a frontrunner strategy seemed to trigger the use of a confirmatory strategy, which may have had conservative influences on belief updating, thereby explaining the increased belief updating error following decision stage 1.

It is worth nothing, however, that 57% of participants selected the correct suspect to bring in for questioning at the end of the investigation. Many of the remaining participants can be accounted for by frontrunner-focused thinking, since 33.3 % of participants ranked a given suspect as lead at decision stage 1 and subsequently reported him as the most likely culprit at the end of the task. Out of these, only 8 participants had ranked Suspect 3 as lead at decision stage 1 and thus made the correct final judgement at the end of the task. The remaining 31 participants did not update their beliefs appropriately and kept the same suspect as leading explanation for the evidence, regardless of whether the evidence could have been better explained by alternative suspects.

5.4. Discussion

In Experiment 4 we employed a more complex probabilistic context to explore people's information search and integration behaviour, exploring the adaptiveness of their strategies in a stepwise paradigm. In the extant literature, the majority of studies have relied on a variety of non-trivial assumptions when comparing human behaviour to an OED model, including the assumption that people accept and reason with the probabilistic parameters given to them by the experimenter (Nelson, 2005; Coenen et al., 2018). We increased the validity of our normative comparison by building individual models for each participant based on their own probabilistic beliefs and observed evidence. This approach allowed us to make meaningful comparisons about the optimality of their belief updating and information acquisition behaviour.

Overall, when comparing observed behaviour to IB-OED model predictions, participants tended to perform "sub-optimally" on the task – both in terms of belief updating and information seeking strategy, when compared to models with different utility functions. Of particular interest were participants' query preferences at the first decision stage, when they reasoned solely with base-rate

information. At this point, all IB-OED models (with each utility function) predicted the most informative query to be 'primary item stolen'. Although 53% of participants selected this query, a significant share of participants queried 'burglary time' (37%). According to the Bayesian formalism, the outcome of querying 'primary item stolen' substantially increased the probability of one suspect (Suspect 2), to 80%, and reduced the probability of the other two suspects to 10%. Comparatively, the outcome of querying 'burglary time' increased the probability of both Suspect 1 (to 57%) and Suspect 3 (to 37%) and decreased the probability of Suspect 2 (to 6%). All utility functions therefore computed 'primary item stolen' to be the most informative query at this stage.

In the present set-up, following a 'frontrunner' strategy at decision stage 1 led a significant number of participants to select the query that was also predicted by all IB-OED models as being most informative ('primary item stolen'). However, a large portion of participants might have reasoned under the assumption that in a multiple inquiry investigation, it might be more informative to decrease the probability of one suspect at the outset. This behaviour, consistent with an 'elimination' strategy, led participants to select the query 'burglary time', that was considered sub-optimal by all utility functions at this stage. Despite the 'sub-optimality' of this choice however, it may be seen as somewhat rational if we entertain the possibility of participants having different preferences and employing simplifying strategies in choices and integration. An analysis of participants' think-a-loud responses revealed that at decision stage 1 participants who employed either 'frontrunner' or 'elimination' strategies accounted for the observed split in query selections. We note that the large proportion of participants who selected 'burglary time' at decision stage 1 somewhat drove the inaccurate belief updating behaviour that was significantly more pronounced at later stages in the task. This may be due to the fact that selecting 'burglary time' or 'primary item stolen' at decision stage 1 leads to two different suspects being the most probable culprits given the query outcomes. When contradictory evidence was then observed at decision stage 2, it is possible that participants were unable to accurately integrate this and therefore misadjusted the posterior beliefs.

Similar to the results of Wu et al. (2017) however, we found that errors in belief updating were not correlated with accuracy of search decisions. Therefore, the claim that the strategies we identified

are used to make search decisions without explicitly using Bayesian inference is not supported. Rather, our findings suggest that participant's query selections and evaluations are consistent with certain strategies, primarily driven by obtaining a frontrunner at the outset. Although this was consistent with query selections in line with OED principles at the first decision stage, consistently adopting a frontrunner strategy across the task dovetailed notions of confirmatory heuristics stemming from overconfidence in a given focal hypothesis, akin to findings within the psychological literature (e.g., McKenzie, 2006; Skov & Sherman, 1986) as well as forensic science literature on confirmation bias (e.g., 'tunnel vision'; Findley & Scott, 2006). As such, we found that a significant proportion of participants seemed to reason with a hypothesis reduced to one leading hypothesis, which they attempted to test across the decision stages. As such, 47% of participants adopted a 'frontrunner-focused' (frontrunner+) strategy at decision stage 2 and 25% at decision stage 3. These participants maintained a single hypothesis, seemingly ignoring its alternatives and conservatively updating their belief estimates relating to their leading hypothesis in light of contradictory evidence. Moreover, they conservatively updated their beliefs regarding the alternative suspects. This finding ties together our evidence suggesting that participants' updating of Suspect 3 was increasingly inaccurate throughout the task. Given that the majority of participants had Suspect 1 or Suspect 2 as leads at the first two decision stages, they conservatively updated their probabilities regarding Suspect 3 at subsequent decision stages even though he was the most probable culprit by the end of the task. Although this suspect appeared less likely at the outset of the task, he was supported by upcoming data. Overall, this serves to demonstrate that certain strategies can lead to confirmatory information seeking and thus lead to conservative belief updating.

6. General Discussion

Within the domain of psychology, OED principles have been used to model how people seek and evaluate information. Despite proving themselves as appropriate computational-level methods to account for people's behaviour in many information search tasks (i.e., Planet Vuma and 20-Q game), their descriptive and explanatory powers are challenged by heuristic models that make the same predictions (Navarro & Perfors, 2011; Oaksford & Chater, 1994), and alternative models which deviate

from OED model predictions, but are able to account for people's behaviour (Bramley et al., 2015; Coenen, Rehder & Gureckis, 2015; Markant & Gureckis, 2014). Given that most studies have so far used tasks with simple probabilistic contexts (i.e., binary hypothesis spaces and/or binary outcome queries), it is possible that other strategies have gone undetected.

In a series of experiments, we investigated how people select and evaluate queries in diverse probabilistic contexts. Critically, these were embedded in more naturalistic crime investigation scenarios that included both binary and ternary hypothesis spaces and query outcomes. The focus of our work was not just to ascertain whether people's evaluations aligned or deviated from information seeking norms, but also to uncover the motivated strategies that might explain their behaviour. In addition, we explored the adaptiveness of the identified strategies across probabilistic contexts using both within and between-subject designs, and across one-shot and stepwise information search tasks. In all four experiments, participants' behaviour was evaluated against IB-OED models parameterized with participants' own priors and with different built in utility functions (KL-D, IG, PG and Impact). In the first three experiments we also included a heuristic model with a built-in PG function (PG_H).

Results from Experiments 1-3, which employed a one-shot task, revealed a number of noteworthy findings. Firstly, participants selected queries that coincided with those predicted by IB-OED models when these aligned with their personal strategies. This suggests that utility functions that are independent of the preferences of the learner, as is the case with information-theoretic OED measures (Coenen et al., 2018), might not be appropriate descriptors of people's information acquisition behaviour. As such, participants evaluated information as being more informative given their personal strategies of either identifying a frontrunner suspect or eliminating one. Adopting a ternary hypothesis space allowed us to disentangle these strategies, leading to the identification of a modal preference for obtaining a frontrunner in Experiments 1 and 3. Crucially, although the probabilistic model used in Experiment 2 (a binary hypothesis space) did not allow for the differentiation between a frontrunner and an elimination strategy, participants still voiced a preference for obtaining a lead hypothesis. This speaks to moving beyond the binary-feature and binary-hypothesis models frequently adopted by

researchers in the psychology domain and encourage the adoption of diverse probabilistic models that allow one to identify and discriminate between these underlying strategies.

Overall, participants' chosen queries and outcome evaluations in the first two experiments seemed to align with the intuitively optimal strategy (given our parameter sets and the investigative nature of the task) of maximising the chances of increasing the probability of a suspect as close to 1 as possible thereby reflecting the assumptions of probability gain based models. However, our mixed-effect regression model analyses illustrated that PG and PG_H were significant predictors of participants' choices only in Experiment 1 and were restricted to best approximating the qualitative direction of the distribution of participants' query selections in Experiment 2. As such, the query they predicted to be most informative, was typically chosen by the majority of participants across our experiments. Whereas in Experiment 1, a PG model outperformed a PG_H model, the opposite was true in Experiment 2. In the latter study, we found that a simplified utility function that assumes equal outcome priors best approximated participants' information search behaviour. This suggests that in some probabilistic environments, participants simplify the assumptions of OED models when evaluating the utility of information and that the computational complexity of OED measures might not be a realistic descriptor of information search behaviour especially in more naturalistic settings (Coenen et al., 2018). In Experiment 3 and 4, all utility functions, even the simplified PG_H model, were unable to account for participants' query preferences, given the introduction of task framing manipulations and stepwise information seeking.

Across experiments, the majority of participants selected queries that were consistent with their preference of obtaining a frontrunner and adopted strategies that would enable this (i.e., frontrunner, differentiation, symmetry, highest outcome). As such, participants largely displayed a preference for queries whose outcomes were most differentially probable under each hypothesis, as it allowed for the identification of a frontrunner. This is conceptually related to the feature-difference heuristic, identified by Slowiaczek et al. (1992) and subsequently tested by Nelson et al. (2010), which entails maximising the difference between the likelihoods under the competing hypotheses. The feature-heuristic strategy, however, only applies to categorization tasks with two categories and "two-value features" (i.e., two-

outcome features), whereas in in the present work we expand this concept by employing probabilistic models with both binary and ternary hypothesis spaces and binary as well as ternary-outcome features. Although prior research has illustrated how in probabilistic environments with two hypotheses and binary-outcome features, a feature-difference heuristic equates to a normative OED model with 'impact' as utility function (Nelson, 2005; 2008), we were unable to make this direct comparison given the different nature of our probabilistic environments.

We were able to conclude however, that although a 'frontrunner' strategy aligned with some OED model predictions (e.g., Experiment 1), this held true only when these predictions coincided with participants' own strategic preferences. For example, in Experiment 3 Scenario 2, participants in both task framing conditions mostly employed a 'frontrunner' and 'differentiation' strategy, which resulted in the selection of queries that were not deemed to be most informative by any of the utility functions. In addition, we found evidence for the use of an 'elimination' strategy, although this was only mentioned by a minority of participants across experiments. Those who did employ an elimination strategy evaluated (and selected) queries consistent with this strategy as being most informative.

Across the present studies we also found that the use of these strategies was sensitive to task context and demands. As such, although a frontrunner strategy dominated in most contexts, Experiments 2 and 3 showed that framing the task as involving multiple inquiries and adopting a stepwise paradigm increased the number of people who adopted an 'elimination' strategy. This serves to show that task framing, and context, may act as strategy determinants, principles currently extraneous to the purely information-theoretic OED measures. In addition, it tentatively suggests that participants' information seeking may be rational given the task framing and context. For example, it may appear sensible to seek a frontrunner at the outset in a criminal investigation case comprising of a 'one-shot' inquiry – as reflected by a PG model. In comparison, eliminating a subject at the outset may seem like a more rational strategy to follow in a context in which there are sequential inquiries. Poletiek and Berndsen (2000), though utilising a different methodology, similarly showed that altering task features like context and content affected participants' testing strategy.

Across all of our experiments, no utility function was able to consistently account for the strategies we extracted from participants' think-aloud responses. Our findings suggested that they were not adaptive across the probabilistic contexts per se, as OED principles would predict, but instead were responsive to factors such as task framing (Experiment 3). Our analysis further showed that the observed adaptiveness due to task framing is not accounted for by any of the utility functions we employed. Moreover, although the majority of participants across contexts employed a 'frontrunner' or 'differentiation' strategy, we found a variety of strategies employed by participants, illustrating that strategy choice can vary from individual to individual. In our one-shot scenarios that involved no task framing manipulations, probability gain based models best approximated the direction of participants' choices (PG in Experiment 1 and PG_H in Experiment 2). In order to account for the effect of task framing and for factors such as risk aversion which we recognised as significant determinants of participants' query evaluations, further research could investigate whether these could be formalised under different conceptualisations, for instance, in terms of risk-taking behaviour, following the work of Polietiek and Berndsen (2000).

Although preferentially adopting a certain strategy in the first three experiments (e.g., frontrunner) was not shown to be detrimental, in Experiment 4, whilst adopting a 'frontrunner' strategy at the first decision stage aligned with query selections in line with informed OED model predictions (querying primary item stolen), we found that a continued use of this strategy led to deviations from IB-OED model predictions. More specifically, adopting a frontrunner strategy at the first decision stage translated to a significant number of participants adopting a frontrunner-focused (frontrunner*) strategy in subsequent stages. This was a confirmatory strategy that entails repeatedly testing a single leading hypothesis and largely ignoring alternative hypotheses, despite their increasing plausibility. This is consistent with literature on selective exposure that finds that people with strong beliefs prefer information that they expect will confirm their beliefs and past choices (Schulz-Hardt, Frey, Lüthgens & Moscovici, 2000; Svenson, 2003). It is also consistent with selective hypothesis testing (Sanbonmatsu, Posavac, Kardes & Mantel, 1998) and positive testing strategies (Klayman & Ha, 1987; McKenzie, 2004). These search strategies are akin to those observed in forensic science whereby

investigators search for information in order to confirm their existing beliefs (Findley & Scott, 2006) which can lead to miscarriages of justice (Eady, 2009; Ormerod et al., 2008)

Despite not always deviating from OED predictions, those adopting a frontrunner-focused strategy also made updating errors, and were less likely to entertain an alternative hypothesis in light of evidence incongruent with their current leading hypothesis. Importantly, requiring participants to update their beliefs in the hypotheses after each item of evidence was viewed could have exacerbated this effect. For example, a study on professionals illustrated that asking people to state hypotheses early during a mock police investigation led to more biased information-seeking strategies (O'Brien & Ellsworth, 2006). Moreover, requiring participants to state beliefs early in a sequential task has been associated with assigning more weight on initial beliefs and conservatively updating these in light of new evidence (Phillips & Edwards, 1966). A study comparing participants required to update hypotheses in a step-wise manner versus participants who are only required to formulate a hypothesis after viewing all evidence could elucidate this matter further.

Overall, however, given the known detriments of adopting confirmatory strategies in real world settings (Kassin, Dror, Kukucka, 2013; Rassin, Eerland & Kuijpers, 2010; Van den Eeden, de Poot & Van Koppen, 2016) further work should explore the extent to which these are used in information-seeking paradigms. Arguably, evaluating queries in the real world in relation to their ability to meet certain goals (e.g., eliminating a hypothesis at the outset) and adopting strategies that facilitate this across different probabilistic contexts seems more psychologically plausible than carrying out the computations posited by a Bayesian OED framework. By testing a heuristic model that assumes equal priors (PG_H) we were able to determine that participants might be failing to integrate the prior of the outcome with the diagnosticity of the outcomes when evaluating the informative value of queries, rather than assuming query outcomes to have equal outcome priors. Given the real-world pragmatics of evidence search, mentally simulating the impact of all possible outcomes of an action on each hypothesis would be computationally intractable, assuming all possible outcomes can even be known. In step-wise or sequential information search situations, exhaustive (not goal-directed) sequential selections and information integrations would be similarly psychologically implausible.

It is also perhaps unsurprising that some participants were falling victim to well-known biases and reasoning fallacies when evaluating items of information. For example, across our experiments we noted traces of risk aversion and zero-sum thinking. Participants' preference for obtaining a frontrunner at the outset was in some cases mitigated by a form of risk aversion that led them to query the feature that was most differentially probable under each hypothesis, even if this query was not expected to yield the frontrunner with the highest probability. Through an analysis of participants' think-aloud responses this also seemed to be the product of overweighting the value of a 'safe' (information) gain, underweighting the value of an outcome that leads to the exclusion of a hypothesis (e.g., outcome 'night' in Experiment 1) and overweighting the probability of unlikely but uninformative outcomes occurring (e.g., outcome 'money' in Experiment 2). Moreover, our findings directly showed that participants, especially in Experiment 2, were averse to selecting a query that was perceived as risky since it could produce an outcome with the smallest benefit (in this case, 0 information). This suggests that some people may therefore be also evaluating information by the perceived risk associated with obtaining that value (which often does not coincide with the normative probability of obtaining the information). Although this could mean that people are reasoning systematically in respect to some utility function, it is not one among the functions we considered in the present paper. Rather, it could be a situation-specific function that captures risk-aversion for gains in information. This will be the focus of future research.

This risk-based information search finding is related to Poletiek and Berndsen's (2000) conceptualisation of hypothesis-testing behaviour as risk-taking behaviour. The authors discriminate between maximising the probability of a confirming outcome (in line with classical definitions of confirmation strategies; see Klayman & Ha, 1987) and maximising the *evidential value* of the confirming outcome. Across two experiments the authors reported a preference to maximise the confirming value of the test outcome, therefore choosing the "riskier" and taking the chance of finding no support evidence at all, with the benefit of high-value evidence if obtained, over the "safer" test that would have allowed them to obtain *some* evidence supporting the hypothesis of interest at the expense of the low evidential value of the outcome. Interestingly, although we found instances of confirmatory

information seeking behaviour (e.g., frontrunner-focused strategy in Experiment 4), we also found that, especially in Experiment 2 and 3, this was mitigated by a form of risk aversion by which participants preferred to identify a "safe" frontrunner with lower probability of being the true culprit over the riskier query that could provide them with a frontrunner with higher probability of being the true culprit, but also with an outcome that would decrease the probability of other hypotheses. One noticeable difference between our work and that of Poletiek and Berdsen (2000) that could explain the different direction of our findings, is that they provided participants with a statement indicating the verbal probability of the outcomes occurring (e.g., "there is a high probability you will obtain X outcome that will lead to Y"), whereas we left participants to infer the probability of outcomes occurring using the probabilistic information they received in the scenario. Thus, it is possible that our participants, had they not overweighed the probability of certain outcomes occurring (e.g., 'money' in Experiment 2), would have similarly been biased towards the strength of the evidence rather than the probability of obtaining that evidence. Though this renders our work more comprehensive by beginning to address these issues, further work investigating under what circumstances people adopt risk-seeking and risk-averse information search strategies is still needed.

In Experiments 1-3 we found that participants' risk aversion was interlinked with a form of 'zero-sum' thinking. Zero-sum thinking describes instances in which evidence that is equally predicted by two competing hypotheses is perceived as offering no support for either hypothesis (Pilditch et al., 2019). For this assumption to be valid the hypotheses must be mutually exclusive and exhaustive. In Experiment 2 most participants correctly evaluated the outcome money as being of little informative value across the probabilistic models (despite overweighting the probability of it occurring). In Experiment 1 however, following the same 'zero-sum' thinking led some participants to misperceive a query as being uninformative (even if it was normatively more informative) when one of its outcomes would increase the probability of two of the three hypotheses and decrease the probability of the other one. Here, these participants believed they would receive relatively no useful information given that the evidence could be highly (and at times equally) predicted by two hypotheses and overlooked the fact that it could lead to a reduction in the hypothesis space by being able to almost exclude one suspect.

Our findings, illustrating that this type of reasoning contributes to a misevaluation of the value of queries adds to previous work which showed how this type of reasoning fallacy leads to significant amounts of information (quantified by KL-D) being overlooked (Pilditch, Liefgreen & Lagnado, 2019).

Overall, the identification of zero-sum thinking and risk averse behaviour in an information seeking paradigm contributes to the existing literature by bridging the gap between known reasoning fallacies in Bayesian probabilistic reasoning tasks and information-seeking principles, two factors that are rarely considered in conjunction (Coenen & Gureckis, 2015; Coenen et al., 2018). To our knowledge only two studies identified risk aversion in information seeking (Poletiek & Berndsen, 2000; Wakabe, Sato, Watamura &Takano, 2012) although one of these was not done within a strictly Bayesian framework (Wakabe et al., 2012). Future work should therefore further investigate the presence of this phenomenon in information seeking paradigms using a purely information-theoretic set-up given it allows one to naturally capture confirmatory strategies both as search preferences aimed to maximise either the probability of a confirming outcome and/or the value of that outcome. Moreover, future work should be carried out in the pursuit of weaving risk-taking principles into current frameworks of human information seeking.

Taken together, findings from our experiments seem to counter the notion that people strictly rely on information within the probabilistic model in order to compute the informative value of a query. Rather, they suggest that people's evaluations of the value of information also depend on their own strategic preferences, task demands, and in some cases on the perceived risk of obtaining the information. Whereas in some environments these can be accounted for by a probability gain utility function, in others they could not. Nonetheless, our findings do not paint a negative picture – in one-shot investigative tasks when suspects had equal priors, participants predominantly adopted the optimal strategy of maximising their chances of finding the true culprit, as dictated by a probability gain model. This model was not able to approximate the distribution of participants' query selections in variations of the task in which this model was not actually considered to be the optimal strategy to adopt, e.g., when there were sequential query selections to be made or when the participant believed the task to be comprised of multiple enquiries. In these paradigms, adopting situation-specific rather than

information-theoretic utility functions that are additionally able to account for factors such as risk aversion, might be best. These would likely be able to account for participants' strategic preferences given different task framings and would arguably be able to fit our identified strategies (e.g., frontrunner) naturally within an OED framework.

In addition, in our sequential paradigm, we found deviations from the utility functions which were probably due to information integration errors. An additional stream of research could explore in more detail how deviations from the predictions of various utility functions influence subsequent search. In our study, we found that although participants' belief updating accuracy decreased throughout the task, this had no effect on the accuracy of their query selections. This echoes the finding of Wu et al. (2017) who reported no correlation between probability judgment error and proportion of correct search decisions. Although it appears that participants were using the probabilities given to them further work could compare the use of different methods of presenting information about the search environment. This would extend the work of Wu et al. (2017), who illustrated that presenting information as posterior probabilities and visualizing natural frequencies was helpful in guiding decisions.

A final note should be made regarding the possibility that underlying the different strategies we have observed in these experiments are different interpretations of the value of information. To model our tasks, and as a normative benchmark, we utilised a measure that quantifies the value of information in terms of divergence, whereby the amount of information proposition e_i provides to partition X is measured by the amount of divergence between the two probability distributions over X due to e_i (Roche & Shogenji, 2016). However, in circumstances such as criminal investigations people might value information in terms of how much it reduces doubt or expected inaccuracy, given that an erroneous decision can carry seriously damaging consequences. Elimination driven (as opposed to frontrunner) strategies in step-wise information search instances fits with a motivation to reduce inaccuracy, as the prospect of making an inaccurate judgment might outweigh the drive to obtain a leading hypothesis. Further empirical work should thus address if alternative measures that value information in terms of e.g., inaccuracy reduction, coincide with people's interpretation of the value of information.

7. Final Remarks

Overriding qualitative theories with principled quantitative models such as those using OED principles has allowed researchers in the past decade to successfully model information seeking in a variety of domains. However, it has also led researchers away from the question of how people are actually evaluating and selecting information, instead focusing on identifying violations of norms in information-seeking. Our work demonstrates that, although in some environments people do seek information in a manner that aligns with a PG measure, they are driven by additional strategies that cannot be entirely accounted for by an information-theoretic OED framework, which are sensitive to the framing and demands of a task. These strategies are accompanied by various well-known reasoning fallacies across a range of probabilistic contexts, and in both one-shot and stepwise information seeking tasks. This paper calls for further work to build on formalisms able to describe the richness of human inquiry, ideally by conceptualising information evaluation as a holistic form of sense making that is dependent both on context and on the seeker's own preferences, motivations, and risk-taking tendencies.

Funding

This work was partly supported by the Leverhulme Trust under Grant RPG-2016-118 CAUSAL-DYNAMICS. The authors declare no competing interests.

Open Practices

Data and example materials have been made publicly available via the Open Science Framework at https://osf.io/tkr4v/.

Acknowledgements

A thank you to all members of the Causal Cognition, Harris and Shanks labs at UCL who provided feedback on the present work, especially Dr. Stephen Dewitt.

References

Baron, J. (1985). *Rationality and Intelligence.* Cambridge: Cambridge University Press.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211-233.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H.B., Singmann, H.,…Scheipl, F. (2012) Package 'lme4'. *CRAN. R Foundation for Statistical Computing, Vienna, Austria.*

Bleichrodt, H. (2001). Probability weighting in choice under risk: an empirical test. *Journal of Risk and Uncertainty*, *23*(2), 185-198.

Bolker, B. M. (2008). *Ecological models and data in R*. Princeton University Press.

Box, G., & Hill, W. (1967). Discrimination among mechanistic models. *Technometrics,* 9, 57-71.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 708.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Charmaz, K. (2006). Coding in grounded theory practice. *Constructing grounded theory: A practical guide through qualitative analysis,* 42-71.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges*. Psychonomic bulletin & review*, 1-41. Doi: 10.3758/s13423-018-1470-5.

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? *In Proceedings of the 37th annual conference of the Cognitive Science Society*, Austin, TX.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology, 79*, 102-133.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive science, 42*(5), 1410-1456.

Crupi, V., & Tentori, K. (2014). State of the field: Measuring information and confirmation. *Studies in History and Philosophy of Science Part A, 47*, 81-90.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. Philosophy of Science, 74(2), 229-252.

Eady, D. (2009). *Miscarriages of justice: the uncertainty principle.* Doctoral Thesis submitted to: Cardiff University.

Evans J., Over D. (1996). Rationality in the Selection Task: Epistemic Utility versus Uncertainty Reduction. *Psychological Review, 103,* 356–63.

Findley, K. A., & Scott, M. S. (2006). Multiple Dimensions of Tunnel Vision in Criminal Cases. *The Wisconsin Law Review, 291*.

Fischoff, B., & Beyth-Marom,R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90* (3), 239.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. New York: Charles Griffin

Gureckis, T., & Markant, D. (2009). Active learning strategies in a spatial concept learning game. *In Proceedings of the Annual Meeting of the Cognitive Science Society* (31).

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7*(5), 464-481.

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of learning and motivation, 61,* 41-102. Academic Press.

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). Graphical models with R. Springer Science & Business Media.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology, 8*(3).

Jarecki, J., Meder, B., & Nelson, J. D. (2013). The assumption of class-conditional independence in category learning. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 35, No. 35).

Kahneman, D., & Tversky, A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of applied research in memory and cognition, 2*(1), 42-52.

Klayman, J. & Ha, Y. (1987). Confirmation, disconfirmation, and information. *Psychological Review, 94*, 211-228.

Kullback, S., & Liebler, R. A. (1951). Information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*, 986-1005.

Link, W.A., & Barker, R.J. (2006). Model weights and the foundations of multimodel inference. *Ecology, 87* (10), 2626-2635.

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behaviour? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34.

Markant, D., & Gureckis, T. (2014). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Austin; Texas.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science, 40*(1), 100-120.

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory & cognition, 38*(7), 941-950.

McKenzie, C. R. (2004). Hypothesis testing and evaluation. *Blackwell handbook of judgment and decision making,* 200-219.

McKenzie, C. R. (2006). Increased sensitivity to differentially diagnostic answers using familiar materials: Implications for confirmation bias. *Memory & Cognition, 34*(3), 577-588.

McKenzie, C. R., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. Cognitive Psychology, 54(1), 33-61.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making, 7*(2), 119-148.

Mosher, F. A., Hornsby, J. R., Bruner, J., Oliver, R., & Greenfield, P. (1966). *Studies in cognitive growth.* In (chap. On asking questions). Wiley New York, NY.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. *Vision research, 49*(10), 1286-1294.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological review, 118*(1), 120.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological review, 112(4),* 979.

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. *The probabilistic mind: Prospects for rational models of cognition,* 143-163.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological science, 21*(7), 960-969.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. Cognition, 130(1), 74-80.

Nelson, J. D., Meder, B., & Jones, M. (2018). Towards a theory of heuristic and optimal planning for sequential information search. Manuscript submitted for publication, https://doi.org/10.31234/osf.io/bxdf4.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 2*, 1-32.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*(4), 608.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. Psychonomic Bulletin & Review, 10(2), 289-318.

O'Brien, B., & Ellsworth, P. C. (2006). Confirmation bias in criminal investigations. Available at SSRN, 913357.

Ormerod, T. C., Barrett, E., & Taylor, P. J. (2008). Investigative sense-making in criminal contexts. Naturalistic decision making and macrocognition, 81-102.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology, 72*(3), 346.

Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological science, 30*(2), 250-260.

Pilditch,T.D., Liefgreen,A., & Lagnado,D. (2019). Zero-sum reasoning in information selection. *In Proceedings of the 41st annual conference of the Cognitive Science Society*, Austin, TX. In press.

Poletiek, F. H., & Berndsen, M. (2000). Hypothesis testing as risk behaviour with regard to beliefs. *Journal of Behavioral Decision Making, 13*(1), 107-123.

Rassin, E., Eerland, A., & Kuijpers, I. (2010). Let's find the evidence: An analogue study of confirmation bias in criminal investigations. *Journal of Investigative Psychology and Offender Profiling*, 7(3), 231-246.

Richards, S. A. (2005). Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology*, *86*(10), 2805-2814.

Roche, W., & Shogenji, T. (2016). Information and inaccuracy. *The British Journal for the Philosophy of Science*, *69*(2), 577-604.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition, 143*, 203-216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. *In Proceedings of the 37th annual conference of the Cognitive Science Society*, Austin, TX.

Rusconi, P., Marelli, M., D'Addario, M., Russo, S., & Cherubini, P. (2014). Evidence evaluation: Measure Z corresponds to human utility judgments better than measure L and optimal-experimental-design models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(3), 703.

Salkind, N. J. (2010). *Encyclopedia of research design* (Vol. 1). Sage.

Sanbonmatsu, D. M., Posavac, S. S., Kardes, F. R., & Mantel, S. P. (1998). Selective hypothesis testing. *Psychonomic Bulletin & Review, 5*(2), 197-220.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental science, 10*(3), 322-332.

Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of personality and social psychology,* 78(4), 655.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22*(2), 93-121.

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20(*4), 392-405.

Svenson, O. (2003). Values, affect and processes in human decision making: A differentiation and consolidation theory perspective. In *Emerging perspectives on judgment and decision research,* ed. SL Schenider, J Shanteau, 287-326. Cambridge University Press: London, UK.

Tversky, A. (1982). Evidental impact of base rate. *Judgment under uncertainty: Heuristics and biases,* 153-160.

Van Den Eeden, C. A., de Poot, C. J., & Van Koppen, P. J. (2016). Forensic expectations: Investigating a crime scene with prior information. *Science & justice, 56*(6), 475-481.

Wakebe, T., Sato, T., Watamura, E., & Takano, Y. (2012). Risk aversion in information seeking. *Journal of Cognitive Psychology,* 24(2), 125-133.

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*, 776–784

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(8), 1274.

Appendices

Appendix 1

Table 1a

Experiment 4: multinomial logistic regression output decision stage 1

Reference Category: 'Time'							OR 95 % CI		
Model[1]	Participant Choice	Prediction Parameter	ß	SE	*t*	Sig.	OR	Lower	Upper
PG	Item	Intercept	0.1	3.1	0.03	0.97			
		Item	0.4	1	0.39	0.69	1.49	0.19	11.3
		Time	0[a]						
	Mode	Intercept	-0.69	3.2	-0.2	0.83			
		Item	-1.9	1.3	-1.4	0.16	0.15	0.01	2.2
		Time	0[a]						
	Point	Intercept	-9.5	83.5	-1.1	0.9			
		Item	7.7	83	0.09	0.93	2376	3.7E-69	1.5E+75
		Time	0[a]						
KL-D /IG	Item	Intercept	8.4	66.5	0.13	0.89			
		Item	-8	66.4	0.12	0.9	0.0003	2 E-61	4.7E+53
		Time	0[a]						
	Mode	Intercept	1.7E-6	93.9	0	1			
		Item	-2.4	93.9	-0.03	0.98	0.09	1.4E-80	6.2E+79
		Time	0[a]						
	Point	Intercept	1.1E-9	93.9	0	1			
		Item	-1.8	93.9	-0.02	0.98	0.16	2.4E-82	1.1E+80
		Time	0[a]						
Impact	Item	Intercept	8.4	66.5	0.13	0.89			
		Item	-8	66.4	0.12	0.9	0.0003	2E-61	4.7E+53
		Time	0[a]						
	Mode	Intercept	1.7E-6	93.9	0	1			
		Item	-2.4	93.9	-0.03	0.98	0.09	1.4E-80	6.2E+79
		Time	0[a]						
	Point	Intercept	1.1E-9	93.9	0	1			
		Item	-1.8	93.9	-0.02	0.98	0.16	2.4E-82	1.1E+80
		Time	0[a]						

[a] Parameter is set to zero due to redundancy.

[1] Participant Choice ~ Utility Function Choice Prediction + (1 | Subject)

Table 1b

Experiment 4: multinomial logistic regression output decision stage 2

Reference Category: 'Time'							OR 95% CI		
Model	Participant Choice	Prediction Term	ß	SE	*t*	Sig.	OR	Lower	Upper
PG	Item	*Intercept*	-0.3	2.8	-0.08	0.93			
		Item	1.8	1	1.8	0.07	6.4	0.8	50.3
		Mode	-16.9	2785	-0.01	0.99	4.4E-8	.00	.
		Point	17.6	5978	0.003	0.99	42558514	.00	.
		Time	-17.3	875.6	-0.02	0.98	3.1E-8	.00	.
		ItemTime	0[a]						
	Mode	*Intercept*	-8.5	2.8	.000	1			
		Item	.85	1.1	0.79	0.43	2.3	0.28	19.4
		Mode	-16.9	2788	-0.01	0.99	4.2E-8	.00	
		Point	-2.6E-10	8455	.000	1	1	.00	
		Time	-1.7	0.92	-1.8	0.06	0.18	0.03	1.1
		ItemTime	0[a]						
	Point	Intercept	-1.1	2.9	-0.4	0.7			
		Item	2.8	1.3	2.1	0.03	16.9	1.2	230
		Mode	0.7	1.5	0.5	0.6	2	0.1	37
		Point	1.1	8455	.000	1	3	.00	.
		Time	-0.03	1.2	-0.03	0.98	0.9	0.09	10.6
		ItemTime	0[a]						
KL /IG'	Item	Intercept	-8	9.4	-0.8	0.39			
		Item	10.2	9.1	1.1	0.26	28243	0	1.9E+12
		Mode	0.28	35.7	0.008	0.99	1.3	2.6E-31	6.7E+30
		Time	0[a]						
	Mode	Intercept	-1.5	2.6	-0.6	0.5			
		Item	2.4	0.79	3.03	0.003	11.1	2.3	53.3
		Mode	0.87	1.3	0.67	0.5	2.4	0.19	30.3
		Time	0[a]						
	Point	Intercept	-1.1	2.6	-0.4	0.68			
		Item	2.8	0.7	3.9	<0.0001	16.6	4.1	66.8
		Mode	0.38	1.3	0.3	0.76	1.4	0.12	17.9
		Time	0[a]						
Impact	Item	Intercept	-8	9.4	-0.8	0.39			
		Item	10.2	9.1	1.1	0.26	28243	0	1.9E+12
		Mode	0.28	35.7	0.008	0.99	1.3	2.6E-31	6.7E+30
		Time	0[a]						
	Mode	Intercept	-1.5	2.6	-0.6	0.5			
		Item	2.4	0.79	3.03	0.003	11.1	2.3	53.3
		Mode	0.87	1.3	0.67	0.5	2.4	0.19	30.3
		Time	0[a]						
	Point	Intercept	-1.1	2.6	-0.4	0.68			
		Item	2.8	0.7	3.9	<0.0001	16.6	4.1	66.8
		Mode	0.38	1.3	0.3	0.76	1.4	0.12	17.9
		Time	0[a]						

[a] Parameter is set to zero due to redundancy.

Table 1c

Experiment 4: multinomial logistic regression output decision stage 3

Reference Category = 'Time'							OR 95% CI		
Model	Participant Choice	Prediction Term	ß	SE	*t*	Sig.	OR	Lower	Upper
PG	Item	Intercept	-3.3E-8	2.8	.00	1			
		Item	2.5	1.7	1.5	0.15	13	0.4	421.7
		Mode	-8.5	26.1	-0.3	0.75	.00	.6.5E-27	6.7E+18
		Point	-8.5E-9	92.4	0.00	1	1	.2.2E-80	4.6E+79
		Time	-10.1	59.6	-0.1.7	0.86	4E-5	1.9E-56	8.3E+46
		ModePoint	0[a]						
	Mode	Intercept	1.9	2.6	0.75	0.45			
		Item	-0.03	1.5	-0.02	0.98	0.97	0.05	18.5
		Mode	-1.2	1.1	-1.1	0.28	0.3	0.03	2.8
		Point	7.8	65.4	0.12	0.91	2443.4	1.1E-53	5.2E+59
		Time	-4.2	1.5	-2.8	0.006	0.01	0.001	0.29
		ModePoint	0[a]						
	Point	Intercept	2.3	2.6	0.89	0.37			
		Item	-1.6	1.6	-0.99	0.32	0.2	0.008	4.9
		Mode	-1.1	1.1	-0.99	0.32	0.3	0.03	3.1
		Point	6.7	65.4	0.1	0.92	854.9	3.9E-54	1.8E+59
		Time	-3.5	1.3	-2.7	0.008	0.03	0.002	0.39
		ModePoint	0[a]						
KL /IG	Item	Intercept	-11	75	-0.1	0.88			
		Item	13.7	75	0.18	0.86	86962	1.7E-59	4.3E+70
		Mode	2.5	83	0.03	0.97	12	2.3E-71	6.1E+72
		Time	0[a]						
	Mode	Intercept	-2.2	2.4	-0.9	0.35			
		Item	4	1.3	3.08	0.003	55.3	4.2	732
		Mode	3.9	0.9	4.2	<0.0001	50.1	7.9	326
		Time	0[a]						
	Point	Intercept	-1.1	2.3	-0.4	0.66			
		Item	1.7	1.3	1.2	0.21	5.5	0.37	80.9
		Mode	3.1	0.8	3.9	<0.0001	22.7	4.7	108.2
		Time	0[a]						
Impact	Item	Intercept	-12	108.8	-0.1	0.91			
		Item	14.6	108.8	0.13	0.89	2318831	5.3E-88	1E+100
		Mode	3.7	124.9	0.03	0.97	43.9	1E-106	1.5E+109
		Time	0[a]						
	Mode	Intercept	-1.8	2.2	-0.8	0.41			
		Item	3.6	1.2	3	0.003	36.2	3.3	387
		Mode	4.8	1.2	4.2	<0.0001	128.2	12.6	1294
		Time	0[a]						
	Point	Intercept	-1.12	2.2	-0.6	0.6			
		Item	1.9	1.3	1.4	0.15	7	0.5	101.6
		Mode	4.75	1.2	4.1	<0.0001	115.5	11.5	1157.3
		Time	0[a]						

[a] Parameter is set to zero due to redundancy.

Supplementary materials

S1. Experiment 1: Query selection Accuracy

We computed the percentage of correct choices within each condition according to each utility function (see Table S1a below). A query choice was coded as correct '1' if it was the one that the utility function quantified as being most informative. In addition, if a utility function quantified the two queries as having equal informative value, regardless of what query they choice, participants' selection was coded as correct.

Table S1a

Experiment 1: Percentage of correct query selections within each condition according to KL-D, IG, PG, PG_H and Impact.

Utility Function	Condition 1	Condition 2	Condition 3	Condition 4
KL-D	78.8%	98.5%	30%	94%
IG	78.8%	98.5%	30%	94%
PG	82%	85%	85 %	79.6%
PG_H	83.3%	85.1%	95.5%	78.1%
Impact	86%	98.5%	94%	94%

A Chi-Square test of independence illustrated no significant difference in the percentage of correct/incorrect choices between the different utility functions within Condition 1, $\chi^2(3) = 3.47$, $p = 0.32$. A significant difference was found within Condition 2, $\chi^2(3) = 19.6$, $p = 0.0002$; within Condition 3, $\chi^2(3) = 100.4$, $p < 0.0001$; and within Condition 4, $\chi^2(3) = 10.7$, $p = 0.01$. As such, within Condition 2 and Condition 4, PG and PG_H predicted significantly less participant query selections than the other utility functions (though still performing well), whereas within Condition 3, KL-D and IG accurately predicted significantly less participant query selections than the remaining three models.

S2. Experiment 1: Strategies and Query Selection Accuracy

Given our findings that strategies and query selections remain largely unvaried across conditions, when exploring their relationship to query selection accuracy (as defined by each utility function), we collapsed the conditions. Next, we computed the percentage of participants who chose the correct query according to each utility function within each strategy sub-set.

Out of the participants who utilised an 'elimination' strategy (n = 14), 79% chose the correct query according to Impact, 93% according to KL-D/IG and PG_H and 86% according to PG. Out of the participants who utilised a 'frontrunner' strategy (n =70), 83% chose the correct query according to Impact, 73% according to KL-D/IG, 79% according to PG_H and 74% according to PG. Out of those who employed a 'frontrunner + zero-sum/risk aversion' strategy (n = 19), 100% chose the correct query according to Impact, 84% according to KL-D/IG, 95% according to PG_H, and 89.5% according to PG. Out of those who employed a 'highest percentage' strategy (n = 9), 78% chose the correct query according to Impact, PG and PG_H and 67% according to KL-D/IG. Out of those who employed a 'differentiation' strategy (n = 46), 94% chose the correct query according to Impact, 78% according to KL-D/IG, 87% according to PG_H and 89% according to PG. Out of the participants who utilised a 'symmetry' strategy (n = 56), 88% selected the correct query according to Impact, 61% according to KL-D/IG, 80% according to PG_H and 78% according to PG. Finally, out of the participants who utilised a 'zero-sum/risk aversion' strategy (n = 9), 67% selected the correct query according to Impact, 89% according to KL-D/IG and 78% according to PG and PG_H.

Utilising Chi-Square tests of independence we found no significant difference in the distribution of correct and incorrect query selections across strategies, when correctness was computed utilising KL-D/IG, $\chi^2(6) = 10.4$, $p = 0.11$, $V = 0.22$; PG, $\chi^2(6) = 5.4$, $p = 0.49$, $V = 0.15$; PG_H, $\chi^2(6) = 14.2$, $p = 0.08$, $V = 0.23$; and Impact, $\chi^2(6) = 10.1$, $p = 0.12$, $V = 0.21$.

The above findings suggest that none of the utility functions are truly representative of the strategies that we found to systematically underlie participants' query preferences. More precisely, there were no differences in accuracy (as determined by a given utility function) between the strategies employed by participants. This included across strategies that are conceptually opposite such as 'elimination' and 'frontrunner'. Although this might speak to the robustness of the utility functions, arguably this insensitivity or robustness might not be appropriate for a comprehensive descriptive framework of human information acquisition that is able to account for the seeker's own preferences and strategies.

S3. Experiment 2: Query selection Accuracy

We once again computed the percentage of correct choices within each condition according to each utility function (see Table S3a below).

Table S3a

Experiment 2: Percentage of correct participant query selections within each condition according to KL-D, IG, PG, PG_H and Impact.

Utility Function	Condition 1	Condition 2	Condition 3	Condition 4
KL-D	90%	21%	100%	98%
IG	90%	21%	100%	98%
PG	91%	21%	100%	100%
PG_H	89.7%	81%	84.7%	80.3%
Impact	91%	21%	100%	100%

A Chi-Square test of independence illustrated no significant difference in the percentage of correct/incorrect choices between the different utility functions within Condition 1, χ^2 (3) = 0.2, p = 0.98; Condition 2, χ^2 (3) = 0, p = 1; within Condition 3, χ^2 (3) = 0, p = 1; and within Condition 4, χ^2 (3) = 0.6, p = 0.88. As such, all utility functions were able to equally account for participants' query selections in Conditions 1, 2 and 3, and inaccurately predicted their choices in Condition 2. In this condition, all utility functions predicted 'primary item stolen' to be of greater informative value than the alternative query, thereby not reflecting participants' modal preference for the query 'burglary time' in this, and all other, conditions.

S4. Experiment 2: Strategies and Query Selection Accuracy

Given that strategies remained largely unvaried across conditions, when exploring their relationship to query selection accuracy (as defined by each utility function), we collapsed the conditions.

Out of the participants who utilised an 'identify culprit' strategy (N = 50), 80% chose the correct query according to KL-D/IG, and 84% according to PG, PG_H and Impact. Out of the participants who utilised a 'differentiation' strategy (n = 48), 81% chose the correct query according to KL-D/IG, PG and Impact, and 79% according to PG_H. Out of the participants who utilised a 'symmetry' strategy (n = 26), 77% chose the correct query according to KL-D/IG, 92% according to PG_H, and 81% according to PG and Impact. Out of those who utilised a 'zero-sum/risk aversion' strategy (n = 31), 83.9% chose the correct query according to PG_H, and 77.4% according to KL-D/IG, PG and Impact. Out of those

who utilised the 'highest percentage' strategy (n= 32), 81% chose the correct query according to KL-D/IG, 84% according to PG_H and 78% according to PG as well as Impact. Finally, out of those who utilised a 'identify culprit +zero-sum/risk aversion' strategy (n = 13), 100% of participants chose the correct query according to PG_H and 69% according to KL-D/IG, PG and Impact.

Utilising Chi-Square tests of independence we found no significant difference in the distribution of correct and incorrect query selections across strategies, when correctness was computed utilising KL-D/IG, χ^2 (6) = 3.7, p = 0.72, V = 0.12; PG, χ^2 (6) = 3.6, p = 0.73, V = 0.12; PG_H, χ^2 (6) = 5.7, p = 0.46, V = 0.15; and Impact, χ^2 (6) = 3.6, p = 0.73, V = 0.12.

S5. Experiment 3: Query Selection Accuracy

Table S5a

Experiment 3: Percentage of correct participant query selections within each condition and scenario according to KL-D, IG, PG, PG_H and Impact.

Utility Function	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
KL-D	91%	72%	24.3%	24.3%	31.8%	56%	76%	77%	98.5%	97%
IG	91%	72%	24.3%	24.3%	31.8%	56%	76%	77%	98.5%	97%
PG	91%	71.4%	83%	88%	62%	86%	76%	77%	80%	74%
PG_H	91%	71.4%	77.3	78.6	100%	100%	76.6%	77.2%	80.3%	71.4%
Impact	91%	72%	100%	100%	31.8%	56%	76%	77%	88%	90%

S6. Experiment 3: Strategies and Query Selection Accuracy

Within Condition 1, using Chi-Square tests of independence, a significant difference in the distribution of correct and incorrect query selections across strategies was found when correctness was computed utilising a KL-D/IG model, χ^2 (7) = 16.6, p = 0.02, V = 0.23. Further analysis revealed this was due to significantly more respondents using a frontrunner strategy (across scenarios) made correct query selections compared to those employing other strategies. No significant difference was found in the distribution of strategies across correct and incorrect queries when correctness was computed using Impact, χ^2 (7) = 8.32, p = 0.3, V = 0.16; PG, χ^2 (7) = 7.4, p = 0.39, V = 0.14; or PG_H, χ^2 (7) = 5.99, p = 0.54, V = 0.14. In these probabilistic environments it therefore appears that KL-D and IG are more capable of detecting changes in the responses dictated by participants' strategies as a result of task framing than the other utility functions.

Within Condition 2, a Chi-Square tests of independence found no significant difference in the distribution of correct and incorrect query selections across strategies when correctness was computed utilising any OED measure: KL-D/IG, $\chi^2(7) = 5.74$, $p = 0.57$, $V = 0.13$; Impact, $\chi^2(7) = 2.4$, $p = 0.93$, $V = 0.13$; PG, $\chi^2(7) = 5.8$, $p = 0.57$, $V = 0.13$; and PG_H, $\chi^2(7) = 10.1$, $p = 0.18$, $V = 0.17$. This suggests that these strategies, though demonstrated to systematically underlie participant's query selections and are to some extent sensitive to factors such as task framing, are not being consistently captured by any of the utility functions.

S6: Experiment 4: Query selection Accuracy

In order to evaluate whether any of the utility functions were able to account for participants' choices, we computed the percentage of correct choices at each decision stage according to each of these (see Table S6a below). Accuracy was calibrated with IB-OED models, meaning that if a participant chose an incorrect query at decision stage 1, they might still have chosen the correct query at decision stage 2, given the queries that were left. Decision stage 4 was not included in this analysis, and will not be included in any subsequent analyses, given that at this decisions stage, all participants selected the 'optimal' query (i.e., the remaining option). As can be seen from the below table none of the utility functions were able to fully account for participants' query selections at any of the decision stages.

Table S6a

Experiment 4: Participant accuracy at each decision stage according to each utility function

Utility Function	Decision Stage 1	Decision Stage 2	Decision Stage 3
Chance Level	25%	33%	50%
KL-D	53%	56.4%	52%
IG	53%	56.4%	52%
PG	54.7%	54.7%	55.6%
Impact	53%	56.4%	53%

Cochran-Q tests illustrated no significant difference in the proportion of correct query choices across the three decision stages, $\chi^2(2) = 0.44$, $p = 0.8$, when accuracy was measured utilising KL-D/IG. Similar results were found when accuracy was defined utilising Impact, $\chi^2(2) = 0.38$, $p = 0.83$, and PG, $\chi^2(2) = 0.02$, $p = 0.98$. As such, it appears that none of the utility functions are able to strongly predict participants' preferences at any decision stage, and that participants' accuracy in selecting the most

informative query, quantified by the different utility functions, did not significantly vary across decision stages. Although all models performed better than chance level at the first and second decision stage, by decision stage three - when participants and models only had two options left to choose from - none of the models performed better than chance level. We further test the predictive abilities of these models in the subsequent analyses.

S7. Experiment 4: Belief Updating and Query Selection Accuracy

We investigated the relationship between accuracy of belief updating and accuracy of query selections. To do so we first obtained a qualitative measure of participants' updating by ranking their posterior beliefs for each of the suspects at each decision stage in descending order (1 = highest posterior, 2 = medium posterior, 3 = lowest posterior) and subsequently explored the correlation between having a *correct* suspect rank order (coinciding with that obtained from B-OED models) at the decision stage preceding a query selection and the accuracy of that query selection (correct or incorrect).

Spearman-Rho correlation tests illustrated no significant correlation between a correct suspect rank order after the first stage and correct query selection at the subsequent decisions stage when correctness was measured utilising KL-D/IG, $r_s = 0.14$, $p = 0.14$; PG, $r_s = 0.13$, $p = 0.15$; and Impact, $r_s = 0.14$, $p = 0.14$. Similar results were found pertaining to obtaining a correct suspect rank order after the second decision stage and correct query selection at the third decision stage when correctness was measured utilising KL-D/IG, $r_s = 0.04$, $p = 0.7$; PG, $r_s = -0.04$, $p = 0.66$; and Impact, $r_s = 0.03$, $p = 0.74$.

S8. Experiment 4: Strategies and Query Selection Accuracy

When accuracy was measured utilised KL-D/IG, a general linear mixed effects model with 'strategy' and 'decision stage' as predictors of 'correct query', illustrated a main effect of strategy, F (6, 333) = 3.1, $p = 0.06$, no main effect of decision stage, F (2, 333) = 1.5, $p = 0.22$, and a significant interaction effect of decision stage and strategy, F (2, 333) = 7.7, $p < 0.0001$. Through post-hoc pairwise comparisons we found the only significant difference to be at decision stage 1, where an elimination strategy significantly predicted more inaccurate choices than a 'frontrunner' strategy, $p = 0.02$, and a 'differentiation' strategy, $p = 0.005$. A significant difference was also found at Decision Stage 3, where

a frontrunner focused strategy led to more inaccurate answers than an elimination strategy, $p < 0.001$. Similarly, a 'highest outcome' strategy was associated with inaccurate answers significantly more at decision stage 1 than decision stage 3, $p < 0.0001$. Virtually identical findings were found when accuracy was measured utilising PG and Impact, suggesting that although the utility functions are not able to aptly capture participants' query selections across the decision stages, they are somewhat sensitive to capturing the different employment of strategies throughout the task.

Given our findings that certain strategies underlie certain query selections, and this is dependent on decision stages, it is unsurprising that strategies are differentially related to accurate decisions across decision stages. For example, our finding that an elimination strategy is associated with more inaccurate decisions at decision stage 1 can be explained by our previous finding that the majority of participants employing an elimination strategy queried 'burglary time' which, at this decision stage, was not deemed to be the most informative query by all utility functions.