



# Knowing what to know: Implications of the choice of prior distribution on the behavior of adaptive design optimization

Sabina J. Sloman<sup>1</sup> · Daniel R. Cavagnaro<sup>2</sup> · Stephen B. Broomell<sup>3</sup>

Accepted: 12 March 2024 / Published online: 8 July 2024  
© The Author(s) 2024

## Abstract

Adaptive design optimization (ADO) is a state-of-the-art technique for experimental design (Cavagnaro et al., 2010). ADO dynamically identifies stimuli that, in expectation, yield the most information about a hypothetical construct of interest (e.g., parameters of a cognitive model). To calculate this expectation, ADO leverages the modeler's existing knowledge, specified in the form of a prior distribution. *Informative* priors align with the distribution of the focal construct in the participant population. This alignment is assumed by ADO's internal assessment of expected information gain. If the prior is instead *misinformative*, i.e., does not align with the participant population, ADO's estimates of expected information gain could be inaccurate. In many cases, the true distribution that characterizes the participant population is unknown, and experimenters rely on heuristics in their choice of prior and without an understanding of how this choice affects ADO's behavior.

Our work introduces a mathematical framework that facilitates investigation of the consequences of the choice of prior distribution on the efficiency of experiments designed using ADO. Through theoretical and empirical results, we show that, in the context of *prior misinformation*, measures of expected information gain are distinct from the correctness of the corresponding inference. Through a series of simulation experiments, we show that, in the case of parameter estimation, ADO nevertheless outperforms other design methods. Conversely, in the case of model selection, misinformative priors can lead inference to favor the wrong model, and rather than mitigating this pitfall, ADO exacerbates it.

## Introduction

Inferences made on the basis of behavioral experiments have the potential to influence both scientific consensus and personalized treatment recommendations. However, strong and accurate inferences can require a daunting number of observations, a requirement that can be prohibitive when resources, e.g., participant attention, are scarce. Thus, methods that maximize the information provided by each individual observation are extremely valuable.

SJS undertook this work primarily while affiliated with the Department of Social and Decision Sciences, Carnegie Mellon University.

✉ Sabina J. Sloman  
sabina.sloman@manchester.ac.uk

<sup>1</sup> Department of Computer Science, The University of Manchester, Oxford Rd, M13 9PL Manchester, UK

<sup>2</sup> College of Business and Economics, California State University, Fullerton, CA, USA

<sup>3</sup> Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

**Adaptive design optimization (ADO)** is a method that leverages observations from individual participants, on-the-fly, to identify the most powerful design in sequence (Cavagnaro et al., 2010).<sup>1</sup> At its core, ADO evaluates candidate stimuli with a *global utility function* that estimates, for each stimulus, the potential informativeness of possible responses to that stimulus. Because of its potential to automatically identify powerful designs, ADO has been used extensively for behavioral, psychometric, and psychiatric applications (Kwon et al., 2022). Such applications are facilitated by the combination of increased access to computational resources and the development of software packages that facilitate its implementation (Yang et al., 2020; Sloman, 2022).

ADO relies on the machinery of Bayesian inference, which requires that the user specify a prior distribution across models and parameter values that will generate their data, i.e., a distribution across possible values of the psychological characteristics underlying the observed stimulus–response relationship. When using optimal design methods like ADO,

<sup>1</sup> We use the convention that **terms in bold** refer to definitions, *terms in italics* refer to technical terms that will be defined later, and ‘terms in single quotations’ refer to vague or ill-defined concepts.

which rely on specified prior distributions in the design of the experiment itself, the choice of prior has dual consequences: Misinformative priors can bias inference and mislead the experimental design process. The prior distribution can have a substantial impact on ADO's behavior (Myung et al., 2013; Cavagnaro et al., 2016). Thus, choosing a prior distribution is an issue of enormous practical import, and requires that the experimenter balance multiple considerations, e.g., prior knowledge and analytical tractability (Myung et al., 2013).

The goal of the present work is to unpack the effects of these various considerations on the behavior of ADO. We consider a common paradigm in which the goal of the experiment is to measure some latent variable, representing a given psychological characteristic, at the participant level as precisely as possible. The assumption is that the behavior exhibited by a given participant can be perfectly captured by a single value of this latent variable, and that these values are drawn from a distribution characterizing the participant population.

In practice, experimenters usually specify a single prior that they use for a large number of experimental participants, their **specified prior**. If the specified prior matches the true distribution of relevant psychological characteristics in the participant population, ADO's criterion for evaluating stimuli can be interpreted as the amount of information the experimenter would receive, on average, across sufficiently many repetitions of the experiment. In this case, the design selected by ADO is optimal in the sense that it will lead the experimenter to correct inferences as quickly as possible, on average. If the specified prior does not match this population distribution, ADO's global utility function no longer admits this interpretation, and the designs selected by ADO may no longer lead the experimenter efficiently towards correct inferences.

Prior literature has devised ways to construct an informed specified prior by incorporating observations from similar past experiments (Tulsyan et al., 2012; Kim et al., 2014). However, this may be infeasible or impractical in many situations of interest, due to, e.g., resource limitations that restrict the number of total participants one can recruit, or a desire to endow all participants with the same prior knowledge for the sake of ethical considerations or the tractability of pooled analyses. In such situations, experimenters are forced to contend with some degree of uncertainty about the true population distribution, and run the risk of deviations between the prior they specify and the population distribution.

The goal of the present work is to study how deviations between the specified prior and the true population distribution affect the performance of ADO. We refer to the presence of such a deviation as **prior misinformation**. In the sections that follow, we introduce a novel conceptual and mathematical framework for investigating the effect of prior misinformation. We leverage this framework to identify both

(a) characteristics of specified priors that contribute to robust inference and (b) cases in which the threats of prior misinformation can only be mitigated by acquiring knowledge of the population distribution.

“**Preliminaries**” introduces the mechanics of ADO and its application to problems of inference about psychological characteristics, such as trait values and model structure. “**The prior's two lives**” presents the main conceptual tension addressed in our paper: Users of ADO implicitly rely on two distinct – and potentially opposing – interpretations of the specified prior. “**Extended notation**” gives a mathematical decomposition of the measure of information gain that reveals how prior information affects ADO's efficiency. “**Prior misinformation in the context of parameter estimation**” and “**Prior misinformation in the context of model selection**” interpret these results in the context of the problems of parameter estimation and model selection, respectively. These sections also present results from simulation experiments illustrating the effect of misinformation on the behavior of ADO in practice. “**Robust practices for ADO for model selection**” discusses and suggests practices users of ADO can adopt to enhance robustness to issues we will show can arise in the context of model selection. “**Discussion and limitations**” discusses limitations of the present work and avenues for future work, and “**Conclusion**” concludes.

## Preliminaries

### Notation

We use bolded, capital letters to refer to random variables, and lowercase, unbolded letters to refer to their corresponding realizations. The probability of a particular realization  $x$  of the random variable  $\mathbf{X}$  is  $p(x)$ , i.e.,  $\mathbf{X} : x \rightarrow p(x)$ .

### Cognitive models

Latent constructs, like those typically of interest in psychological research, are, by definition, unavailable for observation and thus difficult to measure. For many applications, experimenters specify cognitive models, which mathematically represent these constructs in such a way that facilitates their measurement. The scope of the present work is within-subjects estimation: estimating as precisely as possible the degree to which a given participant exhibits a psychological characteristic. We give example applications later in this section. First, we make more precise how cognitive models facilitate the measurement of latent psychological constructs.

We consider probabilistic cognitive models that associate stimuli, e.g., questions that could be asked in an experiment,

with probability distributions over possible responses.<sup>2</sup> We denote stimuli  $x$  and responses  $y$ , which are realizations of a random variable  $\mathbf{Y}|x$ . Models, denoted  $m$ , are families of functions indexed by a free parameter or parameters, denoted  $\theta$ . Models encapsulate substantive mechanistic accounts of the relevant psychological, cognitive, or perceptual processes. The parameters encapsulate psychological or behavioral traits that may vary between experimental participants, but which are consistent within a participant. Our framework assumes that there is some true model  $m^*$  and corresponding parameter value  $\theta^*$  that defines the true data-generating distribution for each stimulus  $x$ , given by  $\mathbf{Y}|x, \theta^*, m^*$ .

We consider separately the goals of parameter estimation and model selection. Parameter estimation is the problem of inferring the value of  $\theta^*$ , or measuring the degree to which a participant exhibits a particular trait (assuming a given model structure). For example, for educational testing, the examiner's goal is to identify the examinee's ability level (assuming a given item-response model). Model selection is the problem of inferring the identity of  $m^*$  from a set of candidate models  $M$ , i.e., determining which of several substantively different processes a participant exhibits. For example, a longstanding problem in psychophysics has been to distinguish among various functional forms for describing the relationship between physical dimensions of stimuli and the psychological experience they induce (Roberts, 1979). Both of these goals – parameter estimation and model selection – can be achieved using Bayesian inference, in which the experimenter places a prior distribution across models and parameter values ( $\mathbf{M}, \Theta$ ) and updates this prior according to observed data.

By specifying a prior distribution, the experimenter also implicitly specifies a **prior predictive distribution**  $\mathbf{Y}|x$ , for which each possible response to a stimulus has a corresponding marginal probability:

$$p(y|x) = \sum_{m \in M} p(m) \int_{\theta} p(y|x, \theta, m) p(\theta|m). \quad (1)$$

We can also compute the predictive distribution conditioned on a particular quantity, such as a parameter value or model.

## Adaptive design optimization

Different sets of stimuli have different degrees of power to identify the generating model and parameter value (Myung & Pitt, 2009; Cavagnaro et al., 2010; Young et al., 2012; Broomell et al., 2019). To address this, researchers have

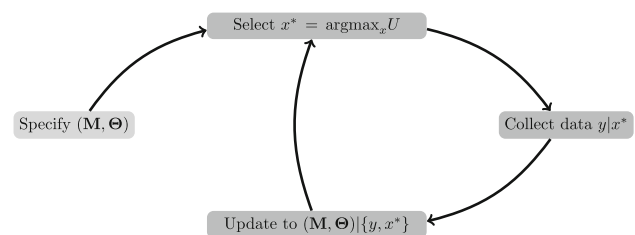
developed methods for the principled selection of stimuli to maximize the informativeness and efficiency of one's experiment (Myung & Pitt, 2009; Broomell & Bhatia, 2014). ADO is one such method (Cavagnaro et al., 2010). By basing its recommendations on the observations it has seen so far, ADO identifies experimental designs tailored to the response patterns of the current participant.

Experiments using ADO proceed across a sequence of mini-experiments, which we call trials. Each trial may consist of a single stimulus or a block of stimuli. ADO dynamically incorporates information throughout the experiment by using the posterior distribution from one trial as the prior distribution on the subsequent trial. This process is visualized in Fig. 1.

To identify the stimulus with the greatest information gain, users of ADO specify a **local utility function**  $u(x, y, \theta, m)$  which is a function of the candidate stimulus  $x$ , response  $y$ , and a possible model and parameter value  $\{m, \theta\}$  (together, a possible **state of the world**). The local utility function measures how much is learned from response  $y$  on stimulus  $x$  about the state of the world  $\{m, \theta\}$ . It can take a variety of forms, depending on the particular goals of the experimenter. The true state of the world and outcome of the experiment are unknown to the experimenter *a priori* – otherwise, there would be no need to run the experiment. Therefore, rather than maximizing  $u$ , ADO selects the stimulus that maximizes the expectation of  $u$  across possible models, parameter values and experimental outcomes according to the specified prior distribution. This yields the **global utility function**:

$$U(x) = \sum_{m \in M} p(m) \int_{\theta} \int_y u(x, y, \theta, m) p(y|x, \theta, m) p(\theta|m). \quad (2)$$

For our applications, we consider a specification of  $u$  such that Eq. 2 measures the amount of information the candidate stimulus  $x$  is expected to yield about some inferential quantity of interest. The amount of information one variable



**Fig. 1** Note. Flow chart of ADO experiment. The experimenter begins the experiment at the lightest grey node, by specifying a prior distribution over models and parameter values. On each trial, they select the stimulus that maximizes the global utility, observe responses to that stimulus, update the distribution over models and parameter values according to Bayes' rule, and then use the obtained posterior as the prior on the next trial

<sup>2</sup> In the remainder of this paper, the term 'cognitive model' can be read as 'probabilistic cognitive model.'

provides about another has been made mathematically precise in the field of information theory by the concept of mutual information (Cover & Thomas, 1991). Motivated by these information-theoretic principles, **global mutual information utility** is the mutual information ( $I$ ) between a focal quantity of interest, which we refer to as the **focus** and denote  $\phi$ , and responses to a stimulus (Bernardo, 1979). Then, the global mutual information utility of a stimulus is:<sup>3</sup>

$$U(x) = \int_{\phi} \int_y \log \left( \frac{p(\phi|y, x)}{p(\phi)} \right) p(y|x, \phi) p(\phi) \\ = I(\Phi; \mathbf{Y}|x). \quad (3)$$

In order for the global utility function to have the form in Eq. 3, the local utility function must take the form:

$$u(x, y, \theta, m) = \log \left( \frac{p(\phi|y, x)}{p(\phi)} \right) \quad (4)$$

which can be thought of as a measure of the information gained about the true value of  $\phi$  from  $y|x$ .

“**Parameter estimation**” and “**Model selection**” show how this specification is adapted to two of the most frequent applications of ADO: the problems of parameter estimation and of model selection. In the former case, the parameters  $\theta$  are the focus, and in the latter case, the model  $m$  is the focus.

Notice that Eq. 3 can be rewritten in terms of Kullback–Leibler divergence, an information-theoretic measure that captures the information gained in moving from one distribution to another. Specifically:

$$U(x) = \int_y \underbrace{D_{KL}(\Phi|y, x || \Phi)}_{\text{Focal divergence}} p(y|x) \quad (5)$$

where  $D_{KL}(\Phi|y, x || \Phi)$ , or what we will refer to as the **focal divergence**, is the Kullback–Leibler divergence from distribution  $\Phi|y, x$  to distribution  $\Phi$ . In other words, global mutual information utility captures, in an information-theoretic sense, how much an observed response to a particular stimulus  $x$  is expected to move the prior distribution assigned to the focus.

## Parameter estimation

Parameter estimation refers to the problem of maximizing the precision of one’s estimate of the parameters  $\theta$  given a particular model  $m$ . Applications of ADO to parameter estimation are useful if the experimenter is interested in

capturing individual variation, for the purpose of, e.g., generating personalized treatment recommendations on the basis of a behavioral assessment. In the educational testing setting mentioned above, the examiner’s goal is to identify each examinee’s ability level in order to make recommendations of areas of strength or potential improvement (Owen, 1969). In a medical application, Hou et al. (2016) used ADO to estimate participants’ degree of visual contrast sensitivity, a characteristic that can be used for diagnosis of eye disease and treatment recommendations.

In the context of ADO for parameter estimation,  $m$  is assumed known, and the focus of the utility function is the parameter  $\theta$ .

The global utility function is:

$$U(x) = \int_{\theta} \int_y \log \left( \frac{p(\theta|y, x)}{p(\theta)} \right) p(y|\theta) p(\theta). \quad (6)$$

## Focal predictive distributions

As mentioned in “**Cognitive models**”, we can compute the predictive distribution conditioned on any particular state of the world,  $\mathbf{Y}|x, \theta, m$  (Eq. 1).

In the context of parameter estimation, the value of  $m$  is known by assumption, so we can equivalently compute the predictive distribution conditioned on any particular value of  $\theta$ ,  $\mathbf{Y}|x, \theta$ . In this case, the set of predictive distributions characterized by possible parameter values are also the set of **focal predictive distributions**, or the predictive distributions associated with possible values of the focus.

We highlight two properties of the focal predictive distributions in the context of parameter estimation. First, since the true data-generating distribution is  $\mathbf{Y}|x, \theta$  for some value of  $\theta$ , the set of focal predictive distributions is in effect a set of possible data-generating distributions. The parameter estimation problem then (asymptotically) amounts to identifying which value of the focus has a corresponding predictive distribution that most resembles the distribution of observed data.

Second, because of this, the predictive distribution corresponding to a particular value of the focus does not depend on additional information like the current trial number or history of observations: While a particular value of  $\theta$  may become arbitrarily more or less likely, it will always elicit the same likelihood on a given stimulus–response pair.

## Model selection

Model selection refers to the problem of maximizing the precision of one’s estimate of the model  $m$ , assuming both  $m$  and  $\theta$  are unknown. The problem of model selection can be thought of as identifying the core psychological process governing a participant’s response distribution.

<sup>3</sup> If  $\Phi$  is a discrete random variable, as is the case in the problem of model selection (“**Model selection**”), the integrals in Eq. 3 are replaced by the analogous sums.



In the context of model selection, the focus of the utility function is the model  $m$ , which yields the global utility function:

$$\begin{aligned} U(x) &= \sum_{m \in M} p(m) \int_y \log \left( \frac{p(m|y, x)}{p(m)} \right) p(y|x, m) \\ &= \sum_{m \in M} p(m) \int_{\theta} \int_y \log \left( \frac{p(m|y, x)}{p(m)} \right) p(y|x, \theta, m) p(\theta|m). \end{aligned} \quad (7)$$

### Focal predictive distributions

In the case of model selection, the focal predictive distributions are the predictive distributions associated with possible values of the model  $m$ , which can be calculated as:

$$p(y|x, m) = \int_{\theta} p(y|x, \theta, m) p(\theta|m). \quad (8)$$

Experimenters faced with the model selection problem have two sources of uncertainty to contend with (the value of  $m$  and the value of  $\theta$ ), yet measure utility with respect to reduction in only one source of uncertainty. This is reflected in properties of Eq. 8: Unlike in the case of parameter estimation, here, the focus is not the only conditioning variable needed to completely specify a possible response distribution  $\mathbf{Y}|x, \theta, m$ ; full specification of the response distribution also requires knowledge of  $\theta$ .<sup>4</sup> In addition, unlike in the case of parameter estimation, the focal predictive distributions are a moving target: Because of their dependence on the parameter distributions, they shift as the parameter distributions are updated on the basis of observed data. These characteristics will become important in our discussion in “[Prior misinformation in the context of model selection](#)” of the impact of prior misinformation in the context of model selection.

### The prior’s two lives

In ADO, the specified prior plays two roles: It both facilitates estimation of the focus from data via Bayesian updating, and informs the design of the experiment that generates these data. These two roles, or ‘lives,’ of the prior map on to two traditions in Bayesian statistics: Bayesian inference and Bayesian decision theory. While the effect of the prior on the behavior of Bayesian inference has been well studied, specified priors that enjoy good theoretical guarantees in the context of Bayesian inference may not seem so appealing

when evaluated on the quality of a corresponding sequential decision-making policy. This section unpacks the reasons for this.

The goal of the present work is, in a sense, parallel to that of literature understanding the effect of priors on Bayesian inference: Our goal is to understand the effect of the choice of prior distribution on the quality of the corresponding sequential decision-making policy, and give guidance for users of ADO constrained to identify a single prior that lives both lives.

Sequential Bayesian inference is a core component of ADO: On each trial, the prior distribution is constructed as the posterior from the previous trial. In its first role, the prior can be seen as a launching pad for learning that will occur throughout the experiment. The prior is understood as an incomplete and ill-informed characterization of the distribution over possible states of the world, and is usually constructed on the basis of a variety of epistemic and pragmatic considerations. Considerations pertaining to – and guidance for constructing – the prior in the context of sequential Bayesian inference is the topic of a substantial body of existing literature (e.g., Lopes and Tobias (2011); Gelman et al. (2017)). *Uninformative* priors are often selected because of their pragmatic appeal in this role.

In its second role, the prior is used when calculating the global utility (Eq. 2) and thus informs the experimental design policy about the relative likelihoods of various outcomes. Bayesian decision theory refers to a prescriptive decision-making policy in which the costs and benefits of taking an action in different states of the world are averaged according to the probabilities of those states of the world (DeGroot, 2005; Berger, 2013). ADO’s policy of selecting the stimulus that maximizes the global utility is a special case of a Bayesian decision theoretic method. If the decision-making policy relies on a prior that mischaracterizes the relative likelihoods of candidate states of the world, the prescribed action is no longer defensible as the action with the highest expected benefit. Bayesian decision theoretic applications thus require a prior that is as informed as possible with available knowledge about the distribution of states of the world. Priors that ignore or mislead about the available knowledge cannot be easily justified from a decision-theoretic perspective, as they may bias the design selection toward stimuli that would not actually be the most informative across multiple experiments.

We assume that the relevant prior knowledge is the true distribution of relevant psychological characteristics in the participant population. Therefore, we will refer to the best decision-theoretic prior as the **population prior**. We do this for conceptual tractability; however, the analyses that follow require only that there is some defensible decision-theoretic

<sup>4</sup> For this reason, the problem of model selection is a special case of an embedded model problem (Foster, 2021), or inference in the presence of nuisance parameters (Paninski, 2005).

prior. Our results apply regardless of the basis on which that prior is constructed. In many cases, information in addition to or instead of a population distribution should inform the decision-theoretic prior. For example, in all but the first trial of an adaptive experiment, the decision-theoretic prior must condition on the observations seen in previous trials. In these cases, the decision-theoretic prior can be formed from the population distribution conditioned on the history of observations (our analyses incorporate this consideration, in a way that is stated more formally in “[Extended notation](#)”). More generally, our framework extends to any case in which other information, e.g., knowledge about relevant demographic characteristics or a participant’s past behavior, is available, as our results can be readily generalized by considering the ‘population’ as all participants with the same demographic or behavioral characteristics.

If the specified prior – the prior used in the context of the experiment – matches the population prior, the global utility (Eq. 3) is also the *expected focal divergence* – the degree of focal divergence one should expect if one were to run the experiment on a sufficiently large participant sample. On the other hand, if the specified prior is not well calibrated, the global utility values could be misleading about the expected focal divergence. “[Motivating example](#)” gives an example of this in the context of an item-response model, a common paradigm used for educational testing. Experiments identified by ADO may not have the power to precisely identify the true model or its parameters, leading to a situation where a characteristic indicative of a disease or needed intervention is not identified efficiently, or possibly at all.

## Types of priors

Priors are typically categorized as ‘informative’ or ‘uninformative.’ With an informative prior, a Bayesian analysis may reach a different conclusion than a conventional one because the prior injects information that is not in the data. For a single experiment aimed at identifying the model and parameter of an individual, the ideal informative prior would be a degenerate one that gives probability 1 to the true model and parameter. Such a prior is not feasible for the paradigm we consider here, where the same prior must be used for each participant drawn from a heterogeneous population. For this case, the best one could do would be to use a population prior. The logic of ADO implicitly assumes that the specified prior is the population prior. Therefore, we characterize the prior that coincides with the population prior as **informative**, and any prior that deviates from that population prior as **misinformative**.

Under our definition, priors that are usually referred to as ‘uninformative’ are typically misinformative when

considered in the context of decision-theoretic applications. ‘Uninformative’ priors are not supposed to inject information, but in the paradigm we consider here, they entail explicit assumptions about the population of participants in the study. We will here use **uninformative** in the context of parameter estimation to refer to a special class of misinformative priors that are agnostic about either the parameter value or the predictive distribution. Priors that are agnostic about the parameter value – are **uninformative in parameter space** – are disperse across the support of the parameter distribution. Priors that are agnostic about the data distribution – are **uninformative in data space** – have high density in regions of the parameter space that correspond to a wide variety of data distributions. These two properties do not necessarily, or even usually, coincide.

## Expected focal divergence

The primary innovation of our analysis is to decouple the two lives of the prior, and provide a framework within which one can reason separately about the process of sequential Bayesian inference and the distribution of observations upon which this inference is performed.<sup>5</sup> In this section, we more precisely define, motivate, and mathematically unpack the expected focal divergence, a concept that is central to the remainder of our analyses.

## Extended notation

In the remainder of our paper, it will be important to distinguish whether a random variable is distributed according to the population or specified distribution of the corresponding quantity. We will do this by subscripting variables that correspond to the population distribution with a 0, e.g., the population distribution of models and parameters becomes  $(\mathbf{M}, \Theta)_0$ , and the corresponding marginal distribution of observations becomes  $\mathbf{Y}_0|x$ . Analogously, we will subscript variables that correspond to the specified distribution with a 1, e.g., the specified distribution of models and parameters becomes  $(\mathbf{M}, \Theta)_1$ , and the corresponding marginal distribution of observations, i.e., the distribution of observations implied by the specified prior, becomes  $\mathbf{Y}_1|x$ . We will also use  $p_0$  and  $p_1$  analogously to refer to the probabilities of the implied random variables taking particular values under the true and specified distribution, respectively.

The notation for quantities used repeatedly is summarized in Table 1. While Table 1, and our discussion more generally,

<sup>5</sup> See Simchowitz et al. (2021) for a related analysis in the context of Bayesian decision-making algorithms more generally.

**Table 1** Extended notational system

Terminology	Variable	Realization	Evaluation	Known?
Candidate stimulus		$x$	Specified by experimenter	✓
Population prior	$(\mathbf{M}, \Theta)_0$		Property of the system under study	×
Specified prior	$(\mathbf{M}, \Theta)_1$	$(m, \theta)$	Specified by experimenter	✓
Population distribution of focus	$\Phi_0$		Subspace of $(\mathbf{M}, \Theta)_0$	×
Specified distribution of focus	$\Phi_1$	$\phi$	Subspace of $(\mathbf{M}, \Theta)_1$	✓
Response distribution	$\mathbf{Y}_0 x$		$y \rightarrow \sum_{m \in M} p_0(m) \int_{\theta} p(y x, \theta, m) p_0(\theta m)$	×
Prior predictive distribution	$\mathbf{Y}_1 x$	$y$	$y \rightarrow \sum_{m \in M} p_1(m) \int_{\theta} p(y x, \theta, m) p_1(\theta m)$	✓
Focal predictive distribution	$\mathbf{Y}_1 x, \phi$		$y \rightarrow p_1(y x, \phi)$	✓
Global utility		$U(x)$	$\int_y \int_{\phi} \log \left( \frac{p_1(\phi y, x)}{p_1(\phi)} \right) p_1(\phi y, x) p_1(y x)$	✓
Expected focal divergence		$U^1(x)$	$\int_y \int_{\phi} \log \left( \frac{p_1(\phi y, x)}{p_1(\phi)} \right) p_1(\phi y, x) p_0(y x)$	×

*Note.* Columns show, respectively, the terminology used for quantities repeatedly referred to, the corresponding random variable notation, notation used for realizations of the corresponding random variable, how the corresponding distribution is evaluated, and whether the corresponding distribution is available to the experimenter

refers to prior distributions, i.e., the distributions of random variables before conditioning on observations, all distributions should be interpreted to implicitly condition on the number of observations implied by context. For example, we write  $(\mathbf{M}, \Theta)_1$  to refer generally to the specified prior, regardless of how many experimental trials have elapsed. When considering the degree of prior misinformation on the second trial of an experiment, i.e., after an observation  $(x, y)$ , this can be read as  $(\mathbf{M}, \Theta)_1|\{y, x\}$  (recalling that the posterior from the first trial is the prior on the second trial). In the same way, the population posterior distribution is  $(\mathbf{M}, \Theta)_0|\{y, x\}$ , which can be interpreted as the appropriate decision-theoretic prior for the next trial given the history of observations.

### Definition of expected focal divergence

Equation 5 showed that global mutual information utility can be rewritten as an expectation of the focal divergence across the specified predictive distribution. In the context of the prior's two lives, the focal divergence can be thought of as the degree to which the prior fulfills its role of efficient Bayesian inference. Taking the expectation of the focal divergence across the specified predictive distribution then invokes the prior's decision-theoretic role: One uses the predictive distribution implied by the specified prior to calculate the relative likelihood of prospective observations.

In the case where the specified prior deviates from the population prior, i.e., the specified prior is misinformative, the global mutual information utility is not equivalent to the focal divergence an experimenter would achieve from a stimulus if they presented it to many members of the participant population. We refer to this latter quantity – the expectation of

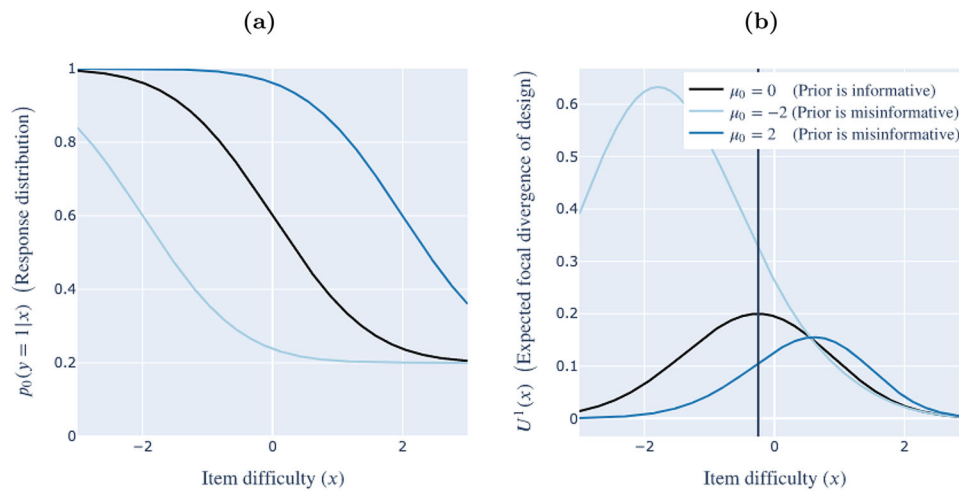
the focal divergence taken across the response distribution – as the **expected focal divergence**. The expected focal divergence associated with a stimulus  $x$ , denoted  $U^1(x)$ , is:

$$\begin{aligned}
 U^1(x) &= \int_y \int_{\phi} \log \left( \frac{p_1(\phi|y, x)}{p_1(\phi)} \right) p_1(\phi|y, x) p_0(y|x) \\
 &= \int_y D_{KL}(\Phi_1|\{y, x\} || \Phi_1) p_0(y|x), \quad (9)
 \end{aligned}$$

i.e., is the expected Kullback–Leibler divergence between posterior and prior under the response distribution  $\mathbf{Y}_0|x$ , or how much observations distributed according to the population distribution are expected to move the prior distribution.

### Motivating example

To illustrate our claim that misinformative priors can impact the effectiveness of ADO, we demonstrate how the population distribution can affect the expected focal divergence of a stimulus in the context of a simple item-response model. We consider an item-response model that uses a one-dimensional ‘proficiency’ trait  $\theta$  to predict the probability of a correct response to a multiple-alternative question with a given ‘item difficulty,’  $x$ . For a fixed value of  $x$ , higher values of  $\theta$ , i.e., greater proficiency, yields a higher probability of a correct response. For a fixed value of  $\theta$ , higher values of  $x$ , i.e., more difficult items, yield lower probabilities of a correct response, with the lowest possible probability being some value greater than zero that is consistent with random guessing. The goal of an experiment is to estimate the proficiency of each participant from their responses to items of various difficulty levels.



**Fig. 2** The effect of prior misinformation: Motivating example from item response theory. **(a)** Response distributions. **(b)** Expected focal divergence. *Note.* Effect of the population distribution on (a) response distribution  $p_0(y = 1|x)$ , and (b) the expected focal divergence of a

In prior work, Weiss and McBride (1983) found that priors that differed from the population distribution induced biases in inferences drawn from experiments designed using a version of ADO.<sup>6</sup> As our running example, we adopt the item-response model used in their simulation study:<sup>7</sup>

$$p(y = 1|x, \theta) = .2 + \frac{.8}{1 + e^{-2.72(\theta - x)}}. \quad (10)$$

The black curve in Fig. 2a shows, for each item difficulty  $x$  between -3 and 3, the predictive distribution associated with a prior  $\Theta_1 \sim \mathcal{N}(0, 1)$  (i.e., distributed according to a standard normal distribution). In the case this prior is informative, i.e., the population distribution is also  $\Theta_0 \sim \mathcal{N}(0, 1)$ , this curve also shows the empirical distribution of responses one should expect. The black curve in Fig. 2b shows the global utility corresponding to each candidate design under this prior. In the case this prior is informative, this curve also shows the expected focal divergence corresponding to each candidate design.

The blue curves show the distribution of observations and expected focal divergence values under two other possible populations. The light blue curves correspond to the population distribution  $\Theta_0 \sim \mathcal{N}(-2, 1)$ , and the dark blue curves correspond to the population distribution  $\Theta_0 \sim \mathcal{N}(2, 1)$ .

<sup>6</sup> Unlike us, Weiss and McBride (1983) did not provide analytical results, did not extend their analysis beyond item-response models, and did not examine the effect of general properties of prior distributions (e.g., dispersion).

<sup>7</sup> We set the item discrimination parameter to the middle of the range investigated by Weiss and McBride (1983), resulting in the constant 2.72 present in Eq. 10.

stimulus  $U^1(x)$ . Colors denote different true distributions  $\Theta_0$ . In all cases, the specified prior is  $\Theta_1 \sim \mathcal{N}(0, 1)$  (i.e., is a standard normal distribution). The vertical line indicates the stimulus, i.e., value of  $x$ , that would be selected by ADO under the specified prior

With reference to Fig. 2b, if the true population distribution is  $\Theta_0 \sim \mathcal{N}(2, 1)$ , the stimulus selected by ADO will yield much less focal divergence than ADO anticipates, on average. By contrast, if the population distribution is  $\Theta_0 \sim \mathcal{N}(-2, 1)$ , the stimulus selected by ADO will yield much more focal divergence than ADO anticipates, on average.

What accounts for this difference? Are there systematic properties of prior distributions that determine which will yield a greater or less expected focal divergence? The following section unpacks these questions.

## Decomposition of the expected focal divergence

The expected focal divergence  $U^1(x)$  decomposes into three terms, which provide insight into how prior misinformation may affect ADO's efficiency:

$$U^1(x) = \begin{aligned} & H(\mathbf{Y}_0|x) && \left. \begin{array}{l} \\ \\ \end{array} \right\} \text{Response variability} \\ & + D_{KL}(\mathbf{Y}_0|x \parallel \mathbf{Y}_1|x) && \left. \begin{array}{l} \\ \end{array} \right\} \text{Surprisal} \\ & + \int_y \int_\phi \log(p_1(y|x, \phi)) p_1(\phi|y, x) p_0(y|x). && \left. \begin{array}{l} \end{array} \right\} \text{Hindsight} \end{aligned} \quad (11)$$

Derivation is deferred to Appendix A.

**Response variability** is the entropy in responses to a given stimulus. Entropy is an information-theoretic measure of the uncertainty related to the possible outcomes of a random variable. If a random variable has only one possible outcome then it has no entropy, while a distribution with high entropy is very dispersed across its support. This captures the intuitive notion that questions are less informative when the experi-



menter already knows what the response will be. Response variability stems from a) uncertainty about the value of the focus, and b) uncertainty about the responses given a particular value of the focus. The source of the stochasticity will determine how this term affects inference, which we discuss more in “[Item response theory](#)”. Another important characteristic of response variability is that it is a function only of the response distribution, and so should not affect one’s choice of prior.

**Surprisal** is the Kullback–Leibler divergence between the specified prior predictive distribution and the response distribution. Higher surprisal contributes to higher expected focal divergence since the specified prior is forced to update in light of observed inconsistencies. Considered differently, high surprisal indicates that there is a lot to learn – i.e., the specified prior is in a sense more misinformed. Thus, despite its contribution to the expected focal divergence, one would generally prefer a specified prior that induces low surprisal.

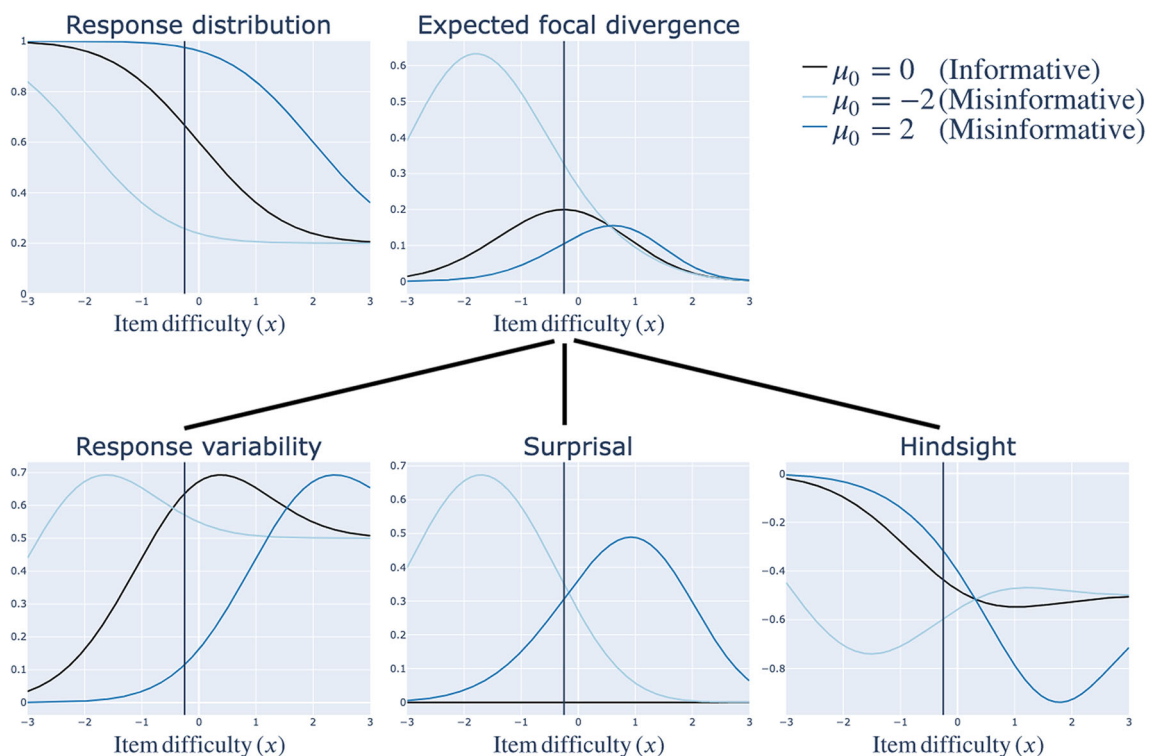
**Hindsight** is the expected posterior log likelihood of responses under the specified prior. Posterior likelihood is a function of both prior likelihood and the specified prior’s ability to ‘respond’ to observations. We discuss this property of ‘responsiveness’ more formally in “[Prior misinformation in the context of parameter estimation](#)”. Surprisal and hindsight will tend to be inversely related through the prior likelihood.

Our discussion of considerations in the specification of one’s prior, particularly in “[Prior misinformation in the context of parameter estimation](#)”, will focus on the effect of different specified priors on hindsight.

### Revisiting motivating example

Figure 3 shows the amount of response variability, surprisal and hindsight under each of the three population distributions shown in Fig. 2. This gives insight into the puzzle posed in “[Motivating example](#)”: Why does the zero-centered prior lead to a more powerful experiment when the population exhibits low values of the trait  $\theta$  than when the population exhibits high values of  $\theta$ ?

Figure 3 reveals that this is because of two (in this case, related) reasons: Both response variability and surprisal are higher in the low- $\theta$  population. Looking more carefully at the response distributions shown in the lefthand panel, the probability of an observation of  $y = 1$  is closest to .5 in the low- $\theta$  population. This makes sense: As discussed in “[Motivating example](#)”, low values of the trait lead to arbitrary responses – i.e., responses that are harder to predict. Thus, response variability is much higher. For the same reason, surprisal is also higher: The low- $\theta$  population surprises the specified prior by producing  $y = 0$  much more often than it anticipates. (The high- $\theta$  population also surprises the specified prior by



**Fig. 3** Reproduction of Fig. 2 along with the three components of the expected focal divergence curves in Fig. 2b: response variability, surprisal and hindsight (Eq. 11). *Note.* As in Fig. 2b, colors denote different true distributions  $\Theta_0$ . In all cases, the prior is  $\Theta_1 \sim \mathcal{N}(0, 1)$

producing  $y = 0$  less often than it anticipates, but the surprise is not as much as in the low- $\theta$  population.)

This section has shown that when the specified prior is misinformed, ADO's global utility may mislead about the expected focal divergence. The following sections explore the practical relevance of this misalignment. As stressed in “[Introduction](#)”, the motivation for our work is a situation where the population distribution is inaccessible to the experimenter. While our motivating example applied our framework to understanding the effect of variation in the population distribution, what is of more practical interest is what can be controlled by the experimenter: the prior they use, and whether they use ADO at all. “[Prior misinformation in the context of parameter estimation](#)” and “[Prior misinformation in the context of model selection](#)” address these questions in the context of parameter estimation and model selection, respectively.

## Prior misinformation in the context of parameter estimation

“[The prior's two lives](#)” and “[Expected focal divergence](#)” showed that under prior misinformation, ADO can be mistaken about the expected gain in information from a particular stimulus. In cases where it cannot reliably anticipate the expected focal divergence, does ADO still enjoy an advantage over other experimental design methods? In this section, we investigate this question in the context of the problem of parameter estimation. We show that even under prior misinformation, ADO facilitates identification of the correct parameter value faster than other sequential design methods. In many practical cases, using methods like ADO may be even more important when there is danger of prior misinformation, since this misinformation can be overcome comparably faster than under other experimental design methods.

As discussed in “[Decomposition of the expected focal divergence](#)”, when identifying properties of specified priors that are robust to prior misinformation, we are most interested in their effect on hindsight. With reference to Eq. 11, hindsight is composed of three terms:  $p_1(y|x, \phi)$ ,  $p_1(\phi|y, x)$  and  $p_0(y|x)$ . In the case of parameter estimation, these become  $p_1(y|x, \theta)$ ,  $p_1(\theta|y, x)$  and  $p_0(y|x)$ . Here, unlike in the case where the focus is the model, the focal predictive distributions do not depend on the specified prior, i.e.,  $p_1(y|x, \theta) = p_0(y|x, \theta)$ . Thus, of these three terms, only  $p_1(\theta|y, x) \propto p_0(y|x, \theta) p_1(\theta)$ , representing the specified posterior, depends on the specified prior. One way to achieve high hindsight given a misinformative prior is to specify a prior for which the likelihood dominates the posterior. As we discussed in “[Types of priors](#)”, this is the definitional property of priors that are uninformative in parameter space.

Indeed, empirical studies by Alcalá-Quintana and García-Pérez (2004) showed that in the context of the adaptive estimation of psychometric functions, uniform priors led to less bias than other commonly specified priors. These results lead us to expect that priors that are uninformative in parameter space will contribute to robustness in the face of prior misinformation.

## Empirical results

This section empirically tests the robustness of ADO to misinformation in two modeling paradigms: the item response paradigm introduced in “[Motivating example](#)”, and a paradigm used to measure a participant's capacity for memory retention. All experiments reported in this paper were run using the pyBAD package for ADO (Sloman, 2022).

### Item response theory

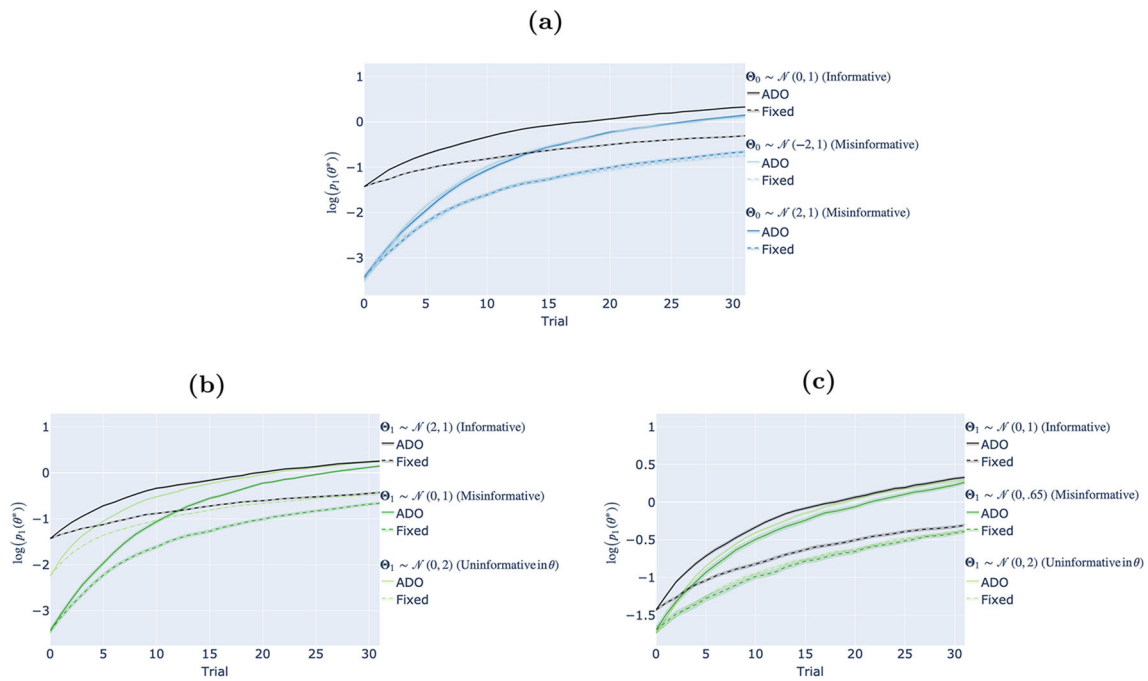
This section discusses simulation experiments to estimate the parameters of item-response models run under the modeling paradigm used as our motivating example.

**Experimental setup** We simulated experiments under two design methods: ADO and a fixed design method. Again drawing inspiration from Weiss and McBride (1983), who discretized the parameter space into 31 equally spaced levels ranging from -3 to 3, the fixed design was set *a priori* as all such 31 stimuli (presented in a random order). ADO was similarly constrained to select from amongst these 31 candidate stimuli. All experiments were run for 31 trials.

For the fixed design, this means that each stimulus would have been presented exactly once, while in the ADO experiments some of those candidate stimuli may be repeated or not presented at all. For each combination of design method, population distribution, and specified prior, we simulated a total of 1000 experiments. In each experiment, a new value of  $\theta^*$ , the parameter value governing the true distribution of responses, was sampled at random from the corresponding population distribution, and held fixed for that experiment. Data were generated according to Eq. 10. Both methods were initialized with the specified prior.

We here show the results of three sets of experiments:

1. Experiments that show the effect of changes in population distribution, with the specified prior held fixed, were run under the same conditions as shown in Figs. 2 and 3.
2. Experiments that show the effect of uncontrolled changes in specified prior fixed the population distribution to  $\Theta_0 \sim \mathcal{N}(2, 1)$  and varied the specified prior among an informative prior ( $\Theta_1 = \Theta_0 \sim \mathcal{N}(2, 1)$ ), a misinformative prior ( $\Theta_1 \sim \mathcal{N}(0, 1)$ ), and a more dispersed misinformative prior, i.e., a prior that is uninformative



**Fig. 4** Item response models: Empirical results. **(a)** Results for different populations with the same prior,  $\Theta_1 \sim \mathcal{N}(0, 1)$ . **(b)** Results as a function of uncontrolled changes in specified prior with population  $\Theta_0 \sim \mathcal{N}(2, 1)$ . **(c)** Results as a function of controlled changes in specified prior with population  $\Theta_0 \sim \mathcal{N}(0, 1)$ . Note. Lines track the posterior probability assigned to the true parameter value across trials, averaged over  $n = 1000$  simulation experiments. Shaded regions indicate standard errors. In Panel (a) shows results for different populations with the same prior,  $\Theta_1 \sim \mathcal{N}(0, 1)$ . The specified prior is fixed and

different colors correspond to different population distributions. Panel (b) shows results as a function of uncontrolled changes in specified prior with population  $\Theta_0 \sim \mathcal{N}(2, 1)$ . Panel (c) shows results as a function of controlled changes in specified prior with population  $\Theta_0 \sim \mathcal{N}(0, 1)$ . In each of Panels (b) and (c), the population distribution is fixed and different colors correspond to different specified priors. Black curves always denote the case where the specified prior is informative, i.e.,  $\Theta_1 = \Theta_0$ . Solid lines show the performance of ADO. Dashed lines show the performance of the fixed design

in parameter space ( $\Theta_1 \sim \mathcal{N}(0, 2)$ ). We refer to these manipulations as ‘uncontrolled’ changes because they do not control for the degree of prior misinformation: The uninformative prior assigns a higher prior log probability to  $\theta^*$ , and induces lower surprisal across part of the stimulus space. Thus, the misinformative prior is at an initial disadvantage but may learn faster because of the mismatch in surprisal.

- To isolate the effect of dispersion from prior misinformation, experiments that show the effect of controlled changes in specified prior fixed the population distribution to  $\Theta_0 \sim \mathcal{N}(0, 1)$  and varied the specified prior among an informative prior ( $\Theta_1 = \Theta_0 \sim \mathcal{N}(0, 1)$ ), a misinformative prior ( $\Theta_1 \sim \mathcal{N}(0, .65)$ ) and a more dispersed misinformative prior, i.e., a prior that is uninformative in parameter space ( $\Theta_1 \sim \mathcal{N}(0, 2)$ ). While these conditions are more artificial than those in our second set of experiments, they control for prior misinformation in the sense that the uninformative prior both tends to assign a lower prior log probability to  $\theta^*$ , and induces higher surprisal across the entire stimulus space.

**Results** Each panel of Fig. 4 shows results corresponding to one of the three sets of experiments described above. The x-axis of each panel indicates the trial number. The y-axis indicates the log posterior probability of the true parameter value.<sup>8,9</sup> In all cases, the black curve corresponds to the informative case, where the specified prior matches the population distribution.

First, comparing ADO (solid lines) to the fixed design (dashed lines), it is clear that ADO outperforms the fixed design in all three cases. In fact, ADO even under prior mis-

<sup>8</sup> The true parameter value was different in each simulated experiment, so, writing  $\theta_i^*$  for the true parameter value in experiment  $i$ , the average log posterior probability of the true parameter value is  $\frac{\sum_{i=1}^{1000} \log(p_1(\theta_i^*))}{1000}$ .

<sup>9</sup> When discussing our results, we measure the effectiveness of each design method by tracking  $\log(p_1(\theta^*))$  across trials. The log transformation reflects the structure of the global utility and expected focal divergence measures. Sometimes, qualitative trends of the non-logged probabilities differ from those shown in the figures in the main text. For completeness, we include corresponding plots of non-logged probabilities in Appendix B.

information ultimately results in stronger inference than the fixed design under an informative prior.

Taking a closer look at the first set of simulations in Fig. 4a, we find no discernible difference. Although Fig. 3 showed the low- $\theta$  population induced higher expected focal divergence, this difference does not translate into a difference in the rate of convergence on the correct parameter value. Recall from “Revisiting motivating example” that the higher expected focal divergence in the low- $\theta$  population was largely driven by higher response variability. If high response variability stems mainly from dispersion across values of the focus, this indicates that each value of the focus makes distinct predictions, facilitating identification of the correct value (Houlsby et al., 2011).

However, in the low- $\theta$  population, response variability stems mostly from higher guessing rates. More generally, as this example illustrates, high response variability that is inherent in the model, i.e., that does not disappear even when conditioning on a particular parameter value, inhibits identification of the correct parameter value.

Figure 4b and c show that the prior that is uninformative in parameter space generally converges more quickly on the correct parameter value, whether using ADO or the fixed design.<sup>10</sup> This is the case even when controlling for prior misinformation (Fig. 4c), since priors that are uninformative in parameter space are able to respond more effectively to unexpected observations.

## Memory retention

While the simplicity of the item-response paradigm allows careful control of our experimental conditions and facilitates interpretation, it potentially limits the generalizability of our findings. We now test whether the main finding – that ADO for parameter estimation outperforms other sequential design methods under prior misinformation – holds in a more complex modeling paradigm: estimating a participant’s capacity for memory retention.

Over a century of research on forgetting has shown that a person’s ability to remember information just learned drops quickly for a short time after learning and then levels off as more and more time elapses (Ebbinghaus, 1913; Laming, 1992). The simplicity of this data pattern has led to the introduction of a number of models to describe the rate at which information is retained in memory (Rubin & Wenzel, 1996).

One of these is the power-law model, which posits that the probability a participant will recall an item ( $y = 1$ )

$x$  seconds after presentation is (Wixted & Ebbesen, 1991):

$$p(y = 1) = a(x + 1)^{-b}. \quad (12)$$

The parameters of the model are  $a$  and  $b$ , where  $0 \leq a \leq 1$  encodes a baseline level of accuracy, and  $0 \leq b \leq 1$  encodes the forgetting rate.

**Experimental set-up** We again ran experiments under two design methods: ADO and a fixed design method. The design variable to be manipulated was the time delay between presentation of the target and the recall phase (i.e.,  $x$  in Eq. 12).

The fixed design method was a slight variation on a benchmark used by Cavagnaro et al. (2010), taken from previous literature (Rubin et al., 1999). In the fixed design method scheme, delays were {0, 1, 2, 4, 7, 12, 21, 35, 59, 99}. Each fixed-design experiment ran for 100 trials, allowing each of these ten delays to be repeated ten times. The order of stimuli was randomized separately for each experiment. ADO experiments also ran for 100 trials. In each ADO trial, the time delay could be any integer between 0 and 100 s.

We simulated experiments under two different population distributions, each combined with four types of specified priors. For the high  $b$  population, we set  $b_0 \sim \text{Beta}(2, 1)$ , i.e., the forgetting rate is high, on average, but negatively skewed. For the low  $b$  population, we set  $b_0 \sim \text{Beta}(1, 2)$ , i.e., the forgetting rate is low, on average, but positively skewed. For both populations, we set  $a$  to  $\text{Beta}(1, 1)$ , which is equivalent to a uniform distribution between 0 and 1. The four types of specified priors are as follows:

1. Informative priors matched the population distributions given above.
2. Priors that mistook the two populations: The specified prior for the high  $b$  population was  $a \sim \text{Beta}(1, 1)$ ,  $b \sim \text{Beta}(1, 2)$ , and the specified prior for the low  $b$  population was  $a \sim \text{Beta}(1, 1)$ ,  $b \sim \text{Beta}(2, 1)$ . In the context of these experiments, we refer to these as misinformative priors.
3. Priors that were uninformative in parameter space specified that  $a \sim \text{Beta}(1, 1)$ ,  $b \sim \text{Beta}(1, 1)$ .
4. Priors that were uninformative in data space resulted in maximally dispersed predictive distributions. The prior that achieves this is  $a \sim \text{Beta}(2, 1)$ ,  $b \sim \text{Beta}(1, 4)$  (Cavagnaro et al., 2010).<sup>11</sup>

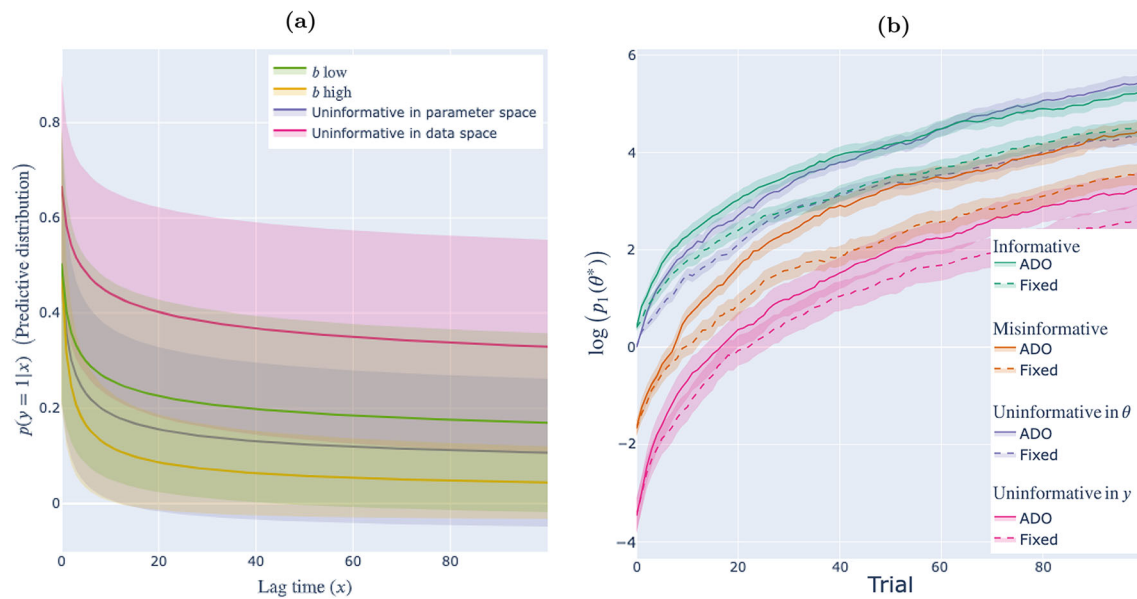
Figure 5a shows typical forgetting curves under each prior.

For each population and for each type of prior, we simulated 100 experiments, for a total of 2 design methods  $\times$  2 populations  $\times$  4 types of specified priors  $\times$  100 repetitions = 1600 experiments. In each experiment, a true parameter

<sup>10</sup> The difference appears small in the log space, but is illustrated more dramatically when probabilities are plotted on the linear scale, as shown in Fig. 8c.

<sup>11</sup> Note that it is only when the prior is uninformative in data space that the distribution over  $a$  is misspecified.





**Fig. 5** Memory retention models (parameter estimation): Empirical results. *Note.* Panel (a) shows the predictive distributions under each prior, with lines denoting mean predictions, and shaded regions indicating the standard deviation across the corresponding prior. Panel (b) shows performance across the course of the experiment. Lines denote the mean log posterior probability assigned to the generating parameter

$\theta^* = \{a^*, b^*\}$  was randomly drawn from the corresponding population distribution, the time delay on each trial was selected according to the design method, and data were generated according to Eq. 12.

**Results** Figure 5b shows how the correctness of inference evolves over the course of the experiment under each type of prior (results are pooled across the two populations). Values on the y-axes are the log probabilities assigned to the true, generating parameter value under each specified prior. This figure shows replication of our main result from “[Item response theory](#)”: ADO outperforms the benchmark for each population and every type of specified prior. Interestingly, unlike in the item-response paradigm, differences in performance at the end of the experiment are mostly accounted for by the type of prior: The fixed design under the informative prior generally does better than ADO under the misinformative or uninformative in data space priors (this is despite the fact that, unlike in the item-response example, ADO has access to a larger stimulus bank than the fixed design method).<sup>12</sup>

<sup>12</sup> If the experimenter is able to collect a sufficiently large amount of data, they will eventually see inference in all cases converge on the correct parameter value (or the set of parameter values that are indistinguishable from the true parameter value on the basis of responses to the available stimuli).

values, and shaded regions denote standard errors around those means (across  $n = 2$  populations  $\times$  100 repetitions = 200 simulated experiments). Different colors correspond to results under different specified priors. Solid lines show the performance of ADO. Dashed lines show the performance of the fixed design

In sum, in both simulation paradigms, ADO performed better than the fixed design method even under prior misinformation. In other words, we do not find that prior misinformation diminishes ADO’s relative advantage. In fact, our results suggest that using ADO when there is prior misinformation may help to overcome that misinformation more quickly than using other design methods.

### Choosing a prior distribution

Our results show that access to an informative prior offers a clear advantage: Experiments run under an informative prior lead to the quickest convergence on the true parameter value. However, both our mathematical and empirical results also show that specifying a prior that is uninformative in parameter space can be nearly as good: Experiments run under a prior that is uninformative in parameter space result in a final parameter estimate that is essentially as accurate as the estimate from experiments run under an informative prior. In other words, informative priors offer inference a ‘head start,’ while the efficiency of priors that are uninformative in parameter space enables inference to quickly catch up with inference under the informative priors.

Of course, when there is uncertainty about the population distribution, attempting to specify an informative prior exposes one to the downside risk that the specified prior actually slows convergence relative to what could be achieved

with an uninformative prior (cf., the orange line in Fig. 5b). Is this risk worth the potential head start, or are researchers better off specifying priors that are uninformative in parameter space as a rule? The answer to that question is that it depends.

The size of the informative prior’s upside advantage – the size of its head start – depends on how misinformed the uninformative in parameter space prior is. The true prior shown in Fig. 5 is quite similar to the corresponding uninformative in parameter space prior, and so the size of the head start was relatively small (and in practice, would likely be outweighed by the downside risk of misinformation). However, this may not be the case when prior knowledge more severely restricts the set of likely parameter values.

Another important consideration is the number of responses the experimenter can elicit from each participant: In our simulations, experiments were run for long enough that inference under the uninformative in parameter space priors could catch up to inference under the informative priors. In these cases, priors that are uninformative in parameter space offer the same upside advantage as informative priors without the downside risk. However, in practice, experimenters may have the resources to collect far fewer responses per participant – indeed, it is precisely such cases that motivate the use of ADO in the first place! In these cases, the head start provided by informative priors will result in a more accurate final inference.

## Prior misinformation in the context of model selection

“Prior misinformation in the context of parameter estimation” showed that in the context of parameter estimation, ADO usually leads to faster convergence on the true parameter value under prior misinformation than other sequential design methods. This section explores whether the same can be said in the context of model selection. It will turn out that, in the context of model selection, the effect of prior misinformation can be more damaging: It can lead one to favor the wrong model.

A common measure of the strength of evidence in favor of one model  $m_1$  over another model  $m_2$  is the Bayes factor, or relative likelihood of data  $y|x$  under  $m_1$  and  $m_2$ :

$$BF(m_1, m_2) = \frac{p_1(y|x, m_1)}{p_1(y|x, m_2)} = \frac{\int_{\theta} p(y|x, \theta) p_1(\theta|m_1)}{\int_{\theta} p(y|x, \theta) p_1(\theta|m_2)}. \quad (13)$$

Equation 13 reveals the sensitivity of model selections to prior misinformation: The apparent strength of evidence in favor of one model over the other is a function of the spec-

ified priors  $\Theta_1|m_1$  and  $\Theta_1|m_2$ . Under prior misinformation, the magnitude and even direction of the Bayes factor can be misleading – implying that it can lead to the erroneous selection of one model over the true, generating model (Vanpaemel, 2010; Lopes & Tobias, 2011).

This is an important concern in Bayesian inference, and addressing it through the choice of prior has been the subject of much literature (Vanpaemel, 2010; Lee et al., 2019). In this section, we show that this relates importantly to the consequences of the choice of prior in its decision-theoretic role.

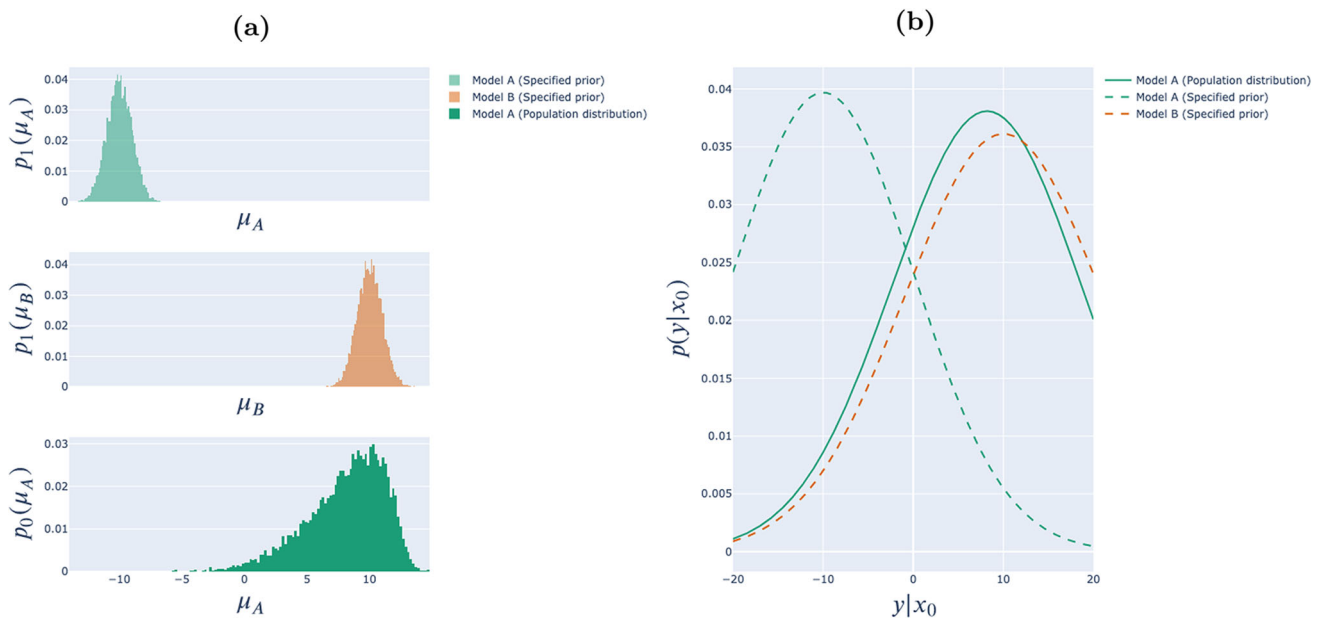
Recall Eq. 7, which gives the global mutual information utility in the context of model selection. Cavagnaro et al. (2010) showed that Eq. 7 can be rewritten as a function of the Bayes factors between all pairs of candidate models. This result implies that ADO results in the selection of stimuli that are expected to lead to extreme Bayes factors according to the specified prior. When the Bayes factors are misleading, this effect of ADO can exacerbate the amount of information encountered that leads one to the wrong model.

The results presented in the remainder of this section will show that in the case of model selection, like in the case of parameter estimation, ADO tends to accelerate convergence towards a particular model. However, under a deceptive prior, this might be the wrong model. In such cases, desirable behavior for an experimental design method would be to decelerate, rather than accelerate, convergence. We will show in “Empirical results” that in such cases other experimental design methods outperform ADO.

## Effect of prior misinformation through the lens of Bayesian inference

Before turning to our results on the effect of ADO, we first present a simple example illustrating the potential effect of prior misinformation in the context of Bayesian inference more generally. Consider the toy example shown in Fig. 6. The task is to distinguish between two models, Model A and Model B. Each model has a free parameter,  $\mu_A \sim \mu_A$  and  $\mu_B \sim \mu_B$ , respectively,<sup>13</sup> and makes predictions for a single stimulus  $x_0$ . Under Model A, responses to  $x_0$  are distributed as  $\mathbf{Y}|x_0, \mu_A \sim \mathcal{N}(\mu_A, 10)$ , while under Model B, responses to  $x_0$  are distributed as  $\mathbf{Y}|x_0, \mu_B \sim \mathcal{N}(\mu_B, 11)$ . Thus, the families of functions captured by the two models are distinguished by the inherent variance in responses. The experimenter is required in advance of the experiment to assign a prior distribution to  $(\mathbf{M}, \mu_A, \mu_B)$ , i.e., to both

<sup>13</sup> Here we simply bold the notation for the realization  $\mu_A$  ( $\mu_B$ ) to indicate its corresponding distribution,  $\mu_A$  ( $\mu_B$ ), to avoid confusion with the random variable that generically represents the distribution over models,  $\mathbf{M}$ .



**Fig. 6** Prior misinformation biases inference for model selection: Motivating example (see discussion in main text). *Note.* Panel (a) shows parameter distributions and Panel (b) shows focal predictive distributions

assess the relative likelihoods of Model A and Model B and to specify the distributions over  $\mu_A$  and  $\mu_B$ .

The top two panels of Fig. 6a show the prior parameter distributions the experimenter specifies for Models A and B. The corresponding focal predictive distributions for the two models are shown, respectively, as the green and orange dashed lines in Fig. 6b.

Now, consider a heterogeneous participant population in which everyone responds according to Model A (i.e.,  $\sigma = 10$ ), but with different values of  $\mu_A$  as represented in the bottom panel of Fig. 6a. The solid green line in Fig. 6b shows the distribution of responses from this population. The dispersion in this curve captures both the inherent variance in each participant's responses ( $\sigma = 10$ ), and variance due to the distribution of values of  $\mu_A$  across participants. Importantly, most participants will produce data that is more likely under Model B than under Model A, under their respective specified priors, yielding apparently strong evidence in favor of Model B.

The upshot is that the true state of the world, i.e., the true response distribution, may look very different from the focal predictive distribution corresponding to the generating model. In essence, the specified prior sets an expectation for what data from a given model will look like, but data from that model may look different in reality if the specified prior is far from the population distribution, and that can lead to wrong inference. In effect, unless the true state of the world happens to coincide exactly with the predictive distribution of  $m^*$ , each possible value of the focus is effectively mis-specified *a priori*. Notice that this doesn't matter in the case

of parameter estimation: In this case, the focal predictive distributions are unaffected by prior misinformation – as Eq. 1 shows, they are a function only of the model structure, which is (by assumption) known.

This example is albeit quite contrived to prove a point. However, such deceptive priors – priors that induce initial convergence towards the wrong model – can actually emerge in practice, as we show in “Empirical results”. In the remainder of this section, we explore – conceptually in “Effect of prior misinformation through the lens of Bayesian decision theory” and empirically in “Empirical results” – the degree to which this phenomenon persists in the context of ADO. The consistency of Bayesian inference guarantees that the experimenter in this example will eventually be able to recover  $m^*$ . However, when the amount of data collected is not large, relying on Bayesian decision-theoretic policies – i.e., choosing data on the basis of these misinformed inferences – has the potential to exacerbate the effect of misinformation.

### Effect of prior misinformation through the lens of Bayesian decision theory

In the toy example in “Effect of prior misinformation through the lens of Bayesian inference”, ADO would assign  $x_0$  a high global utility because it induces a large divergence between the predictions of Models A and B – even though these predictions are made on the basis of prior misinformation.

In general, when crafting a policy for selecting optimal designs, the goals of parameter estimation and model selection may come into conflict. A stimulus that ADO calculates

is optimal for discriminating between models may not be optimal for refining estimates of the distribution of parameter values. In other words, ADO for model selection faces a version of an explore – exploit dilemma: By acting on its prior beliefs about each model’s predictions, it may fail to explore parts of the sample space that could challenge these beliefs.

Thus, when the goals of model selection and parameter estimation are in conflict, ADO can actually exacerbate the problem. By aggressively ‘exploiting’ areas of the design space that appear to yield information about the models, ADO finds powerful evidence in favor of its prior beliefs. In contrast, by ‘exploring’ less apparently informative stimuli, other methods may have more of an opportunity to learn the correct parameter distributions before making strong conclusions about the generating model.

ADO’s aggressiveness is thus a double-edged sword: It converges quickly on conclusions based on what it believes about the predictions of the foci. However, in the case where prior beliefs do not reflect the population distribution, it does not seek opportunities to challenge these incorrect beliefs.

### Choosing a prior distribution

“Prior misinformation in the context of parameter estimation” showed that, in the case of parameter estimation, priors that are uninformative in parameter space can somewhat mitigate the damage of prior misinformation. One would hope that the issues that arise in model selection could be avoided by using similarly uninformative priors.

Unfortunately, this is not the case: As will be shown in the following section, priors that are uninformative in parameter space nevertheless associate models with particular response distributions, and are also prone to inducing biased inference. One could nevertheless hope that specifying such priors over the parameter distributions of candidate models might mitigate the problem by facilitating more rapid convergence on informative parameter distributions. Indeed, we find empirically that in one model selection context, recovery from biased inference is relatively fast under a uniform prior. However, it is difficult to disentangle the effect of the responsiveness of the uniform prior from its effect on the focal predictive distributions – in particular, how they diverge from the response distribution. We leave investigating whether specifying priors that are uninformative in parameter space mitigates biased inference in the context of model selection as an avenue for future work.

Is it possible to identify a prior that is instead ‘uninformative in model space’? In the case of parameter estimation, the important characteristic of an uninformative prior was that it was responsive: Areas of the parameter space quickly became represented in proportion to the relative likelihood they assigned to the history of observations. A prior that was uninformative in model space would facilitate the proportional

representation of models according to their relative conditional likelihood. But as emphasized in “Model selection”, the relative conditional likelihood of a model depends on the prior parameter distribution; indeed, the problem of not knowing the parameter distribution is in a sense the problem of not knowing the conditional likelihood distribution  $Y_0|x, m$ .

In summary, these results suggest the absence of concrete guidance for the case of model selection. The following section reinforces through simulation results that apparently uninformative priors can inadvertently induce biased inference.

### Empirical results

This section extends the memory retention paradigm introduced in “Memory retention” to model selection. The goal of these results will be to demonstrate that apparently uninformative priors can inadvertently bias inference, and that this bias is exacerbated by ADO.

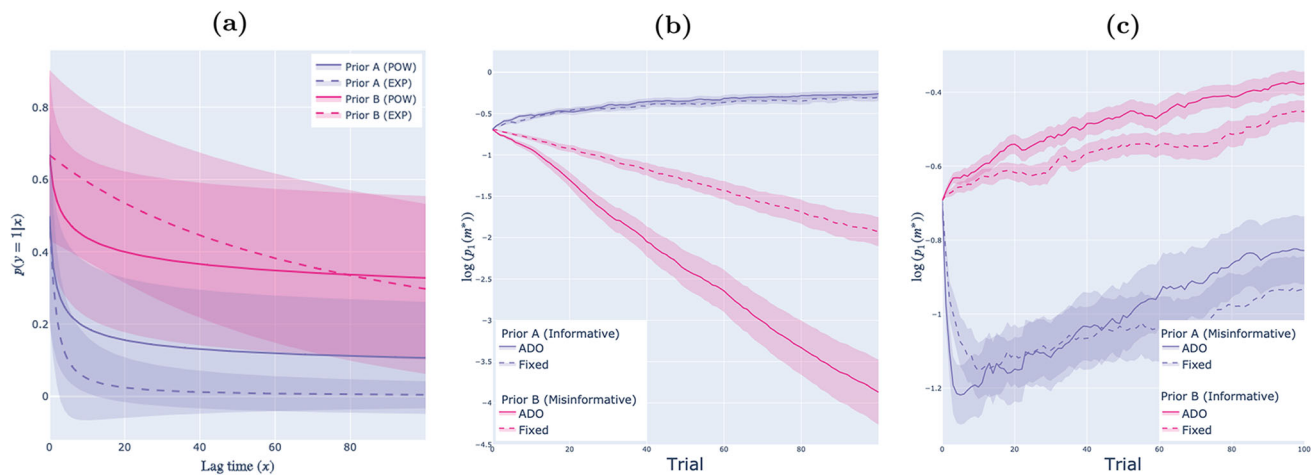
In these experiments, the goal is to distinguish the power-law model introduced in “Memory retention” (Eq. 12) from the exponential model of memory retention, which posits that the probability a participant will recall an item  $x$  seconds after presentation is:

$$p(y = 1) = ae^{-bx}. \quad (14)$$

**Experimental set-up** We considered two types of prior distributions, prior A and prior B. We also varied the population distribution such that each prior was informative in half of our experiments and misinformative in the other half.

1. **Prior A** assigns a 0.5 probability to each of the power-law and exponential models, with  $a \sim \text{Beta}(1, 1)$  and  $b \sim \text{Beta}(1, 1)$  under both models. According to prior A, all viable parameter values are equally likely. Therefore, prior A is uninformative in parameter space. When encountering the population characterized by this distribution, which we call population A, prior A is also the informative prior. When encountering any other population, prior A is misinformative.
2. **Prior B** also assigns probability 0.5 to each model, but  $a \sim \text{Beta}(2, 1)$  and  $b \sim \text{Beta}(1, 4)$  under the power-law model, and  $a \sim \text{Beta}(2, 1)$  and  $b \sim \text{Beta}(1, 80)$  under the exponential model. The predictive distribution under prior B exhibits an extremely diffuse set of behavioral patterns, as shown in Fig. 7a (see also Cavagnaro et al., 2010). Therefore, prior B is uninformative in data space. When encountering the population characterized by this distribution, which we refer to as population B, prior B is also the informative prior. When encountering





**Fig. 7** Memory retention models (model selection): Empirical results. *Note.* Panel (a) shows the predictive distribution of each model under each prior, with lines denoting mean predictions, and shaded regions indicating standard deviations across the corresponding prior. Panel (b) shows results where data were generated from population A. Panel (c) shows results where data were generated from population B. In each

of panels (b) and (c), lines denote the mean log posterior probability assigned to the generating model, and shaded regions denote standard errors around those means (across  $n = 2$  models  $\times$  100 repetitions = 200 simulated experiments). Different colors correspond to results under different specified priors. Solid lines show the performance of ADO. Dashed lines show the performance of the fixed design

any other population, such as population A, prior B is misinformative.

We simulated experiments with two different design methods: ADO and the fixed design described earlier. In each simulated experiment, data were generated from a given model, either the power-law or exponential model, with parameters randomly drawn from the corresponding population distribution (either A or B). We did not randomize the generating model, but rather simulated an equal number of experiments with each model (each population distribution assigns a probability of .5 to each of the power-law and exponential models, and so the prior over models was always correctly specified). In all, we simulated 1600 experiments: 100 for each combination of design method (ADO or fixed), population (A or B), prior (A or B) and model (power-law or exponential).

**Results** Figure 7b and c show how the correctness of inference evolves over the course of the experiment under each type of prior (results are pooled across the two generating models). Values on the y-axes are the log probabilities assigned to the generating model  $m^*$  under each generating prior. Like in the case of parameter estimation (Fig. 5b), when the specified priors are informative, inference converges steadily, and more quickly under ADO, towards the true model.

However, inference under the misinformative priors exhibits the dynamic explained in the previous subsections, favoring

the wrong model (at least initially) – and this is exacerbated by ADO. The reasons for this are precisely the reasons for the confusion illustrated in Fig. 6: As shown in Fig. 7a, in both cases the specified prior distributions are wildly off base about the expected behavior of the population characterized by each model.

Recovery from biased inference under the specified prior that is uninformative in parameter space (Fig. 7c) is quicker than recovery from biased inference under the specified prior that is uninformative in data space (Fig. 7b), potentially reflecting the capacity of the prior that is uninformative in parameter space to more quickly ‘respond’ to unexpected observations.<sup>14</sup> However, our setup here is not adequate to confirm this. Notice first that while the specified prior varies between the two panels of Fig. 7, so does the population distribution. More fundamentally, the specified prior changes the focal predictive distributions. Taken together, this implies that our setup does not (and perhaps cannot) control for qualitative differences in the divergence between the focal predictive distributions and the response distribution, which, as discussed in “Effect of prior misinformation through the lens of Bayesian inference”, is the source of the biased inferences.

<sup>14</sup> As in the case of parameter estimation, if the experimenter is able to collect a sufficiently large amount of data, they will see inference eventually converge on the correct state of the world and, by extension, on the correct model indicator. However, as Fig. 7b illustrates, the trajectory of inference can misguide the experimenter for at least 80 trials – far longer than the duration of many experiments in practice.

## Robust practices for ADO for model selection

The results from the previous section highlight the importance of taking steps to ensure one’s priors are informative – especially when used in conjunction with decision-theoretic methods like ADO, which amplify biases induced by prior misinformation. In the context of model selection, if an experimenter specifies a prior that faithfully captures their epistemic uncertainty, ADO will treat that uninformative prior as being a true representation of relative likelihoods in the world and select designs accordingly. Because the two roles of the prior here conflict, this can result in incorrect inferences.

While we framed our results in “[Prior misinformation in the context of model selection](#)” as applying to the problem of model selection – identification of model structure – note that these results apply to any situation in which knowing the value of the focus of interest does not completely identify the true data-generating distribution. In the case of model selection, this applies because one needs the value of both  $m$  and  $\theta$  to identify the data-generating distribution, yet evaluates performance based only on  $m$ . However, one could also apply ADO to, e.g., a parameter estimation problem for which some ‘nuisance parameters’ are not considered foci for inference (e.g., estimating only main effects in the presence of fixed or participant-level effects). In these cases, our results on model selection, not parameter estimation, would apply.

“[Prior misinformation in the context of model selection](#)” discussed the potential beneficial effect of specifying priors that are uninformative in parameter space in mitigating these biases. This section discusses additional methods to alleviate or anticipate this bias, some of which have been adopted by previous studies, and some of which provide promising avenues for future research.

### Additional trials to inform specified priors

One way to increase confidence in one’s specified priors is to devote a portion of one’s experimental resources to collecting observations from which to learn more informed parameter distributions. For example, when using ADO to distinguish between competing models of intertemporal choice, Cavagnaro et al. (2016) devoted three quarters of each experiment to parameter estimation, i.e., selecting stimuli to maximize the global utility function for parameter estimation, before using the inferred posteriors for each participant during the later model selection trials.

In a parameter estimation application, Kim et al. (2014) leveraged hierarchical modeling techniques to pool information across participants to construct informed distributions: Data from each sequential participant was used to refine the specified prior for the next participant. They showed that this

method led to better parameter estimates in the context of a psychophysical experiment.

While these methods offer promising solutions for many use cases, their application falls outside the scope considered by our work. As we stated in “[Introduction](#)”, we consider situations in which the experimenter wishes to use the same prior for every participant. This characterizes situations in which incorporating data from previous participants would be infeasible or unfair (e.g., educational testing), or when the experimenter cannot afford to spend scarce resources on additional parameter estimation trials. (Note that participants in Cavagnaro et al. (2016)’s study were required to complete 80 experimental trials. Conducting an experiment of this length would be at best difficult and at worst impossible in cases in which candidate stimuli correspond to potentially irritating or invasive tests such as a medical procedure.)

### Total entropy utility

Borth (1975) introduced the total entropy utility function in order to cope with the dual sources of uncertainty that characterize the model selection problem, i.e., uncertainty about both the model identifier and the parameter value. The total entropy utility function considers the entire state of the world as the focus of the utility function:

$$U(x) = \sum_{m \in M} p(m) \int_{\theta} \int_y \log \left( \frac{p(m, \theta | y, x)}{p(m, \theta)} \right) p(y | x, \theta, m) p(\theta | m). \quad (15)$$

We had hoped that running ADO using the total entropy utility function would, like Cavagnaro et al. (2016)’s method, lead to a balance between parameter estimation and model selection trials. We had further hoped that it would do so more efficiently than fixed or heuristic methods of achieving this balance.

To test this, we ran simulation experiments with exactly the same setup as those discussed in “[Empirical results](#)”, with the exception that when using ADO, the stimulus that maximized Eq. 15 (rather than Eq. 7) was selected. The results of these experiments, presented in Appendix C, did not show a consistent advantage of the total entropy utility function in leading to more robust selection of the correct model.

### Novel approaches to robust adaptive experiments

The previous two subsections discussed existing methods for coping with the effect of prior misinformation on model selection. However, these existing methods can be prohibitively costly (running additional trials to inform priors) or potentially ineffective (using the total entropy utility

function). An important direction for future research is the development of methods that increase the robustness of adaptive design methods to the pitfalls introduced in “Prior misinformation in the context of model selection”. To this end, in this section, we propose two steps experimenters can take in the design and implementation of adaptive experiments to increase their robustness. We leave further development and stress testing of these approaches as avenues for future research.

**1. Anticipating biases via prior sensitivity analyses** As mentioned in “The prior’s two lives”, the choice of prior distribution in the context of Bayesian inference is the topic of a substantial literature. One practice advocated in this literature (e.g., Lee et al. (2019)) is to perform prior sensitivity analyses, i.e., to perform data analysis under a variety of priors to ensure one’s inferences are robust to the specification of the prior.

We echo the importance of this practice. In the context of adaptive experiments, analogous prior sensitivity analyses are important to understand not only the direct effect of the prior on inference, but also the prior’s indirect effect through its effect on the data collected. For a given specified prior, experimenters should simulate sets of experiments where data is generated by parameter values distributed according to several different ‘participant’ populations. If these simulated experiments are reliably able to identify the true model, this will provide reassurance that actual experiments run under the specified prior will be able to recover the generating model, even if the true participant population differs slightly from the specified prior.

**2. Using a design policy that navigates the explore–exploit dilemma** Another approach is to respecify the utility function itself in a way that is more robust to such biases (Go & Isaac, 2022). The total entropy utility function (“Total entropy utility”) is one example of an alternative utility function designed for a similar purpose.

As we discussed in “Effect of prior misinformation through the lens of Bayesian decision theory”, in the context of model selection, ADO effectively faces an explore–exploit dilemma: Should it select a stimulus that ‘exploits’ what it thinks it knows about the predictions of the competing models, or a stimulus that has the potential to contradict these pre-existing beliefs? Designing decision-making policies that effectively navigate the explore–exploit dilemma has been the subject of literature spanning cognitive science (Hills et al., 2015) to machine learning (Schulz et al., 2018). Utility functions intended to navigate this dilemma in the context of model selection could draw from this literature.

One approach to sequential decision-making that navigates this dilemma in a principled way is known as upper confidence bound (UCB) sampling (Schulz et al., 2018):

Rather than sample where their expectation of the value of the local utility is highest, a UCB sampler would sample where an additive combination of this expectation and a measure of the variance around this expectation is the highest. UCB effectively constructs a confidence interval around the expectation, and samples at the upper bound of that confidence interval. During early trials, the variance measure usually dominates, inducing exploration. As the variance measure decreases, the expectation measure begins to dominate, and the sampler gradually turns to exploiting areas where the expectation of the utility is highest.

In Appendix D, we leverage our framework to suggest one way the global mutual information utility function could be modified to incorporate principles from UCB sampling.

## Discussion and limitations

While our work provides insight into the behavior of ADO under prior misinformation, it also has a number of limitations. Firstly, the scope of our theory is limited. In the case of model selection, our work shows that ADO can exacerbate inferential biases in the presence of prior misinformation, but does not provide additional guidance as to when or the extent to which one should anticipate these biases. While some work provides additional theoretical characterization of the conditions that lead to related biases in the context of ADO (Sloman et al., 2024) and other adaptive design schemes, e.g., optional stopping (Hendriksen et al., 2020; Heide & Grünwald, 2021), there remain many open theoretical questions.

Secondly, when ADO is used in practice, the correctness of the resultant inferences can be affected by multiple factors not captured by our mathematical and simulation-based framework. For example, parameter values may drift due to, e.g., changes in the environment. In the context of real-world applications of item-response theory, one of our motivating experimental paradigms, this can lead to biased proficiency estimates (Wells et al., 2002). However, we expect that the dynamics induced by such factors will usually interact with, rather than counteract, the dynamics revealed by our results.

One of the most important factors we have not considered is the potential for model misspecification, or misspecification of the functional form of the model  $p(y|x, \theta, m)$  itself. The case of model misspecification differs from our case, prior misinformation, in that the true state of the world is assigned a probability of zero: Since the functional form of the model does not update in light of observed data, one can never hope to counteract this form of misspecification. Under model misspecification, inference can stray further and further from the Bayes-optimal parameter estimate, resulting

in totally random designs leading to better estimates than ADO (Sloman et al., 2022). Recent work has proposed design policies that increase the robustness of ADO in such cases (Overstall & McGree, 2022; Catanach & Das, 2023). Alongside development of practices to increase ADO's robustness to prior misinformation, further development of methods to enhance ADO's robustness to model misspecification is an important avenue for future research.

## Conclusion

When performing Bayesian inference, there are many considerations experimenters must keep in mind. An important one is the specification of one's prior distribution. When using optimal design methods like ADO, which rely on specified prior distributions in the design of the experiment itself, this decision has dual consequences: Misinformative priors both bias inference, and mislead the experimental design process.

In this paper, we introduced a conceptual and mathematical framework for reasoning about the effect of prior misinformation on the efficiency of ADO. Our framework elucidated one general limitation of mutual information utility functions: While the implied expected focal divergence indicates the degree of posterior divergence, it does not in general indicate whether that divergence is in the right direction.

We applied our framework to two common use cases for ADO: the estimation of parameters that measure individually varying psychological characteristics, and the identification of model structure to inform the development of psychological theory. Through mathematical analysis and simulation experiments, we demonstrated counterintuitive pitfalls of using uninformative priors in the case of model selection. In the context of parameter estimation, our framework elucidated principles upon which users of ADO can base selection of their prior – namely, to favor priors that are uninformative in parameter space, rather than data space. In the context of model selection, we discussed and suggested several practices users of ADO can adopt to enhance the robustness of their design and analysis strategies to the biases we identified. Investigating these practices is a promising direction for future research.

## Open Practices Statement

All the simulation code used to generate the results reported in this paper is publicly available at <https://github.com/sabjoslo/prior-impact>.

## Appendix A: Derivation of Eq. 11

$$\begin{aligned}
 U^1(x) &= \int_y \int_\phi \log \left( \frac{p_1(\phi|y, x)}{p_1(\phi)} \right) p_1(\phi|y, x) p_0(y|x) \\
 &= \int_y \int_\phi \log \left( \frac{p_1(y|x, \phi)}{p_1(y|x)} \right) p_1(\phi|y, x) p_0(y|x) \\
 &= \int_y \int_\phi \log(p_1(y|x, \phi)) - \log(p_1(y|x)) p_1(\phi|y, x) p_0(y|x) \\
 &= \int_y \int_\phi \log(p_1(y|x, \phi)) p_1(\phi|y, x) p_0(y|x) \\
 &\quad - \int_y \log(p_1(y|x)) p_0(y|x) \\
 &= H(\mathbf{Y}_0|x \parallel \mathbf{Y}_1|x) \\
 &\quad + \int_y \int_\phi \log(p_1(y|x, \phi)) p_1(\phi|y, x) p_0(y|x) \\
 &= H(\mathbf{Y}_0|x) + D_{KL}(\mathbf{Y}_0|x \parallel \mathbf{Y}_1|x) \\
 &\quad + \int_y \int_\phi \log(p_1(y|x, \phi)) p_1(\phi|y, x) p_0(y|x) \quad (16)
 \end{aligned}$$

where  $H(\mathbf{X}_1 \parallel \mathbf{X}_2)$  denotes the cross entropy of the distribution that characterizes the random variable  $\mathbf{X}_2$ , relative to the distribution that characterizes the random variable  $\mathbf{X}_1$ .

## Appendix B: Experimental results using linear probability measures

Figures 8, 9 and 10 reproduce Figs. 4, 5b and 7b–c, respectively, with the values on the y-axis showing the average probability assigned to the true value of the focus, rather than the average log probability.

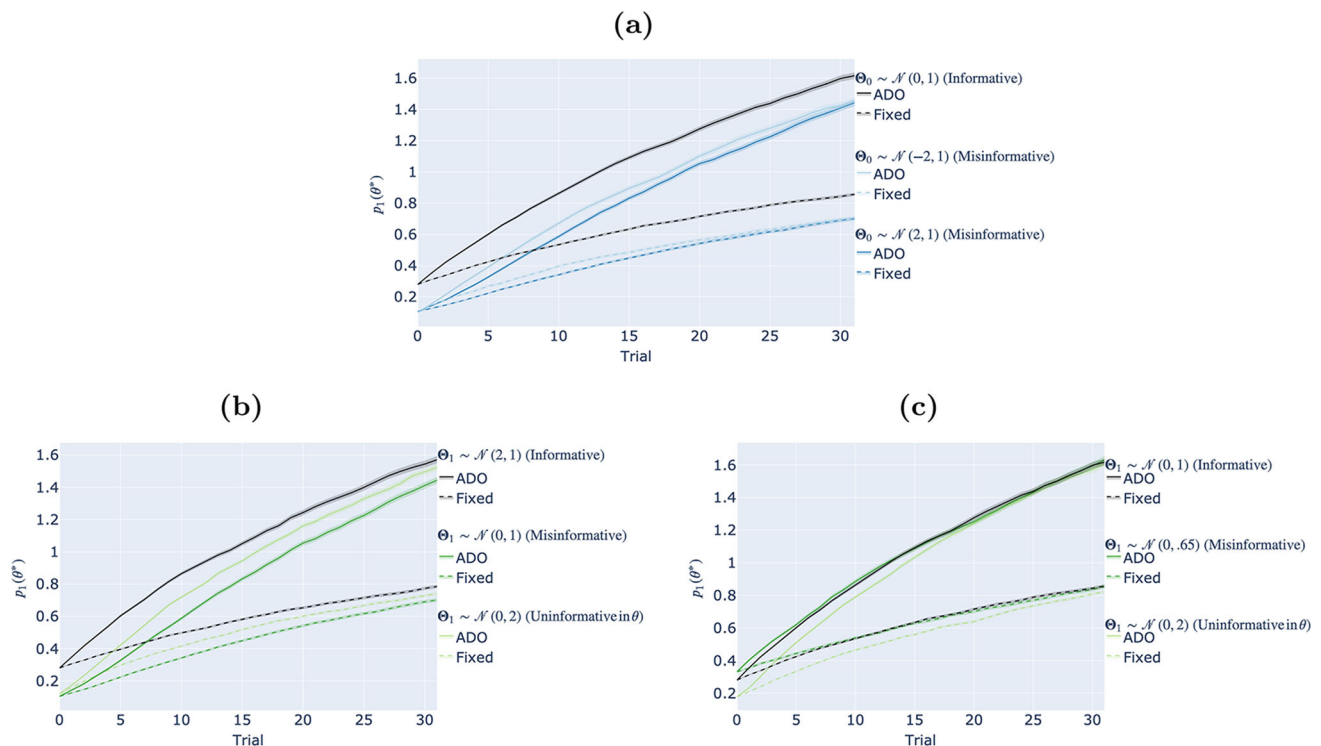
## Appendix C: Experimental results using the total entropy utility function

“Total entropy utility” introduced the total entropy utility function. Figure 11 reproduces the experiments shown in Fig. 7b–c, with the exception that the ADO experiments use the total entropy utility function. While it appears to make a difference in the experiments shown in Fig. 11a, it actually appears to exacerbate the problem in Fig. 11b. It therefore does not appear to be a consistent solution to the problem.

## Appendix D: Upper-confidence bound global utility

In “Novel approaches to robust adaptive experiments”, we framed ADO's failures in the case of model selection under the more general framework of an exploration–exploitation dilemma. Here, we leverage our framework to suggest one way the global mutual information utility function could be





**Fig. 8** Note. Panel (a) corresponds to Fig. 4a. It shows results for different populations with the same prior,  $\Theta_1 \sim \mathcal{N}(0, 1)$ . This highlights that focal divergence in the low- $\theta$  population is both higher and on average in the right direction (divergence in the wrong direction contributes to the overall trend more when probabilities are logged, since the log

operation exacerbates low probabilities). Panel (b) corresponds to Fig. 4b. It shows results as a function of uncontrolled changes in specified prior with population  $\Theta_0 \sim \mathcal{N}(2, 1)$ . Panel (c) corresponds to Fig. 4c. It shows results as a function of controlled changes in specified prior with population  $\Theta_0 \sim \mathcal{N}(0, 1)$

modified to incorporate principles from UCB sampling, an approach for navigating this dilemma discussed in “[Novel approaches to robust adaptive experiments](#)”.

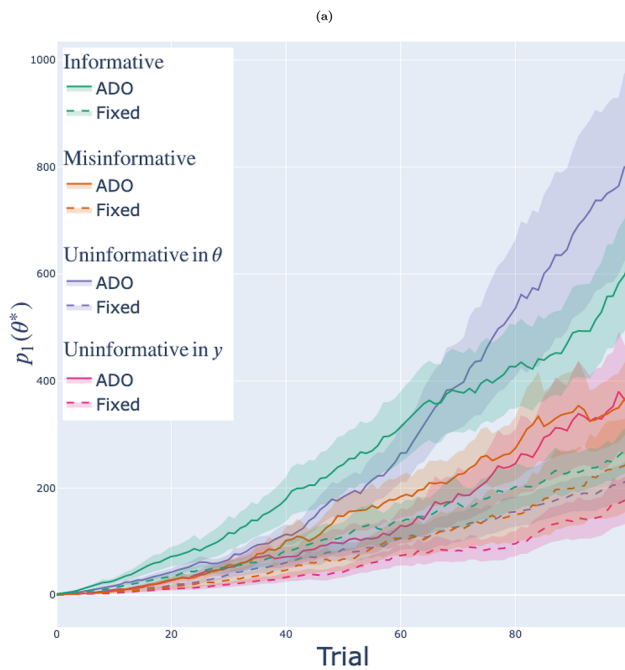
A direct application of UCB in ADO would involve incorporating a measure of dispersion of the local utility values around the global utility. However, this would not be sufficient to address our motivating problem: Recall that our goal for ‘exploration’ here is to challenge our pre-existing beliefs about the specified prior parameter distributions. First of all, notice that this naïve application of UCB targets uncertainty in the utility values, which is not what we care about. Secondly, in the same way that the global utility (the expectation of the local utility) is calculated on the basis of the specified prior (Eq. 2), the most natural way to calculate the analogous second moment would also be on the basis of the specified prior. Thus, rather than challenging our beliefs about the priors, this approach would actually incorporate additional reliance on them.

Nevertheless, we can leverage core principles of UCB – maximizing an additive combination of an exploitation and exploration measure that dynamically adjusts over time – to construct a decision-making policy that targets the dual goals of model selection and parameter estimation. As discussed, existing measures of global mutual information

utility effectively exploit specified prior knowledge. To construct a UCB policy, we can directly use this as a measure of exploitation. As a measure of exploration, we seek a quantity that both reflects the degree to which we will learn about the parameter estimates, and shrinks as these estimates become more precise.

With reference to our decomposition of the expected focal divergence (Eq. 11), notice that the response variability and surprisal terms are shared by both the expected focal divergence corresponding to model selection and to parameter estimation. If a decision-making policy for model selection selects stimuli that induce high response variability and/or surprisal, this will facilitate not only the explicit goal of model selection, but also the implicit goal of parameter estimation. Thus, together, response variability and surprisal achieve our first criterion for an appropriate measure of exploration: They reflect the degree to which the experimenter can be expected to learn about the parameter values.<sup>15</sup>

<sup>15</sup> Although recall from “[Decomposition of the expected focal divergence](#)” the caveat that the effect of response variability on inference will depend on the source of the variability, i.e., whether it stems from uncertainty about the parameter value, or uncertainty about responses even conditioned on a particular parameter value.



**Fig. 9** Corresponds to Fig. 5b

Combined, these terms will also tend to achieve the second criterion: Surprisal, by definition, will shrink as the parameter estimates converge.

Therefore, one could consider the combination of response variability and surprisal as an exploration measure. Equation 17

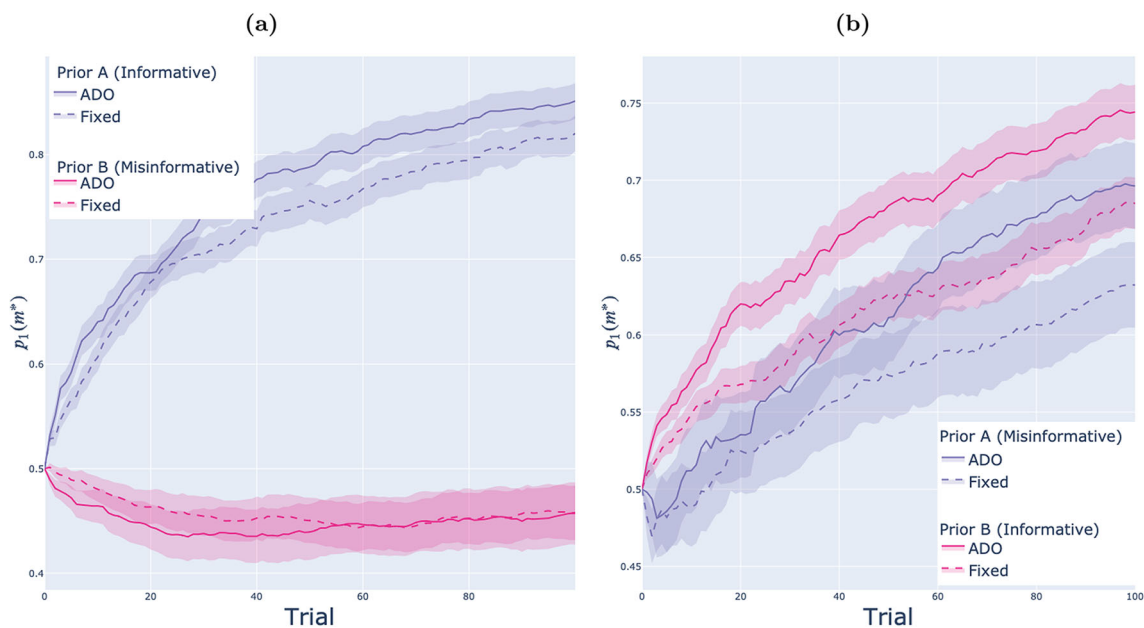
gives the corresponding expected focal divergence function:

$$\begin{aligned} U_{UCB}^1(x) &= \underbrace{U^1(x)}_{\text{Exploitation term}} + \underbrace{H(\mathbf{Y}_0|x) + D_{KL}(\mathbf{Y}_0|x \parallel \mathbf{Y}_1|x)}_{\text{Exploration term}} \\ &= U^1(x) + H(\mathbf{Y}_0|x \parallel \mathbf{Y}_1|x) \\ &= \int_y \int_\phi \left( \log \left( \frac{p_1(y|x, \phi)}{p_1(y|x)} \right) - \log(p_1(y|x)) \right) p_1(\phi) p_0(y|x) \\ &= \int_y \int_\phi \log \left( \frac{p_1(y|x, \phi)}{p_1(y|x)^2} \right) p_1(\phi) p_0(y|x) \end{aligned} \quad (17)$$

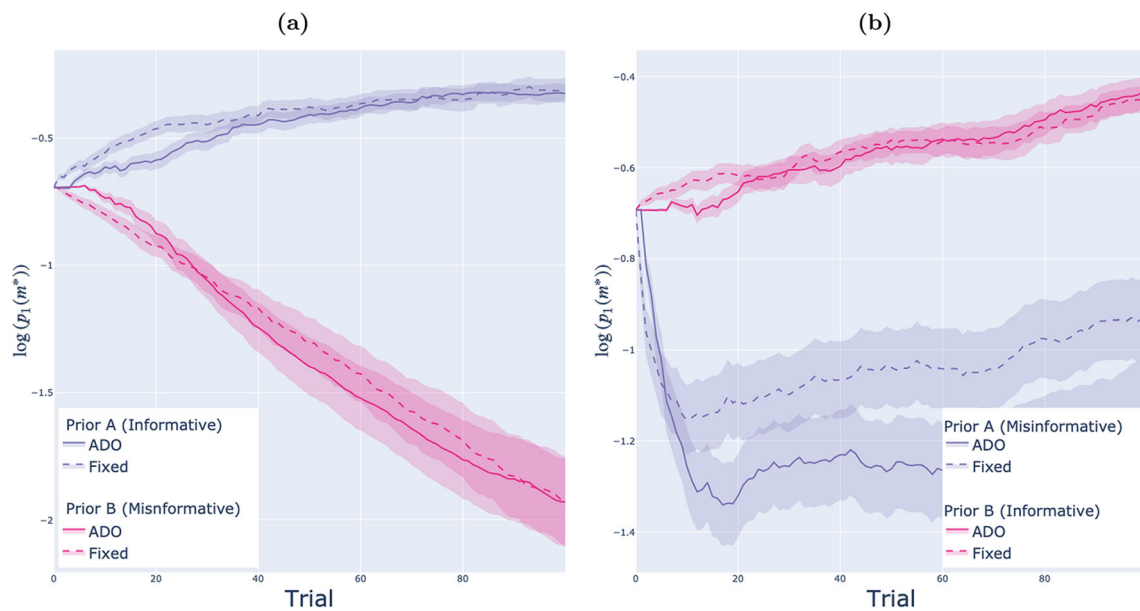
where  $H(\mathbf{Y}_0|x \parallel \mathbf{Y}_1|x)$  denotes the cross entropy of the predictive distribution relative to the response distribution.

Of course, in practice we are not maximizing the expected focal divergence (the expectation of the focal divergence under the population prior), but rather the global utility (the expectation of the focal divergence under the specified prior). Equation 18 gives the global utility function implied by Eq. 17, i.e., what one would actually maximize in practice:

$$\begin{aligned} U_{UCB}(x) &= \int_\phi \int_y \log \left( \frac{p_1(\phi|y, x)}{p_1(\phi) p_1(y|x)} \right) p_1(y|x, \phi) p_1(\phi) \\ &= \int_\phi \int_y \log \left( \frac{p_1(y|x, \phi)}{p_1(y|x)^2} \right) p_1(y|x, \phi) p_1(\phi) \\ &= I(\Phi_1; \mathbf{Y}_1|x) + H(\mathbf{Y}_1|x). \end{aligned} \quad (18)$$



**Fig. 10** Note. Panel (a) corresponds to Fig. 7b. It shows results in population A. Panel (b) corresponds to results in population B. Here, not taking the logs and thus not penalizing for extremely small values helps ADO, which tends to result in more extreme posterior probabilities than the fixed design



**Fig. 11** Memory retention models (total entropy): Empirical results. Absolute performance across the course of the experiment. *Note.* Panel (a) shows results where data were generated from population A. Panel (b) shows results where data were generated from population B. In each of panels (a) and (b), lines denote the mean log posterior probability

Equation 18 is an additive combination of the mutual information between  $\Phi_1$  and  $\mathbf{Y}_1|x$ , i.e., our original measure of global utility, and the entropy of  $\mathbf{Y}_1|x$ , a criterion used for an alternative sampling scheme known as uncertainty sampling (Lee et al., 2021).

Both Eqs. 17 and 18 are written using the more generic notation of  $\phi$ , to emphasize their potential application in any case the value of the focus of interest does not completely identify the true data-generating distribution. For the problem of model selection, Eq. 18 would more specifically become:

$$\begin{aligned}
 U_{UCB}(x) &= I(\mathbf{M}_1; \mathbf{Y}_1|x) + H(\mathbf{Y}_1|x) \\
 &= \sum_{m \in M} p_1(m) \int_{\theta} \int_{y} \log \left( \frac{p_1(y|x, m)}{p_1(y|x)^2} \right) p(y|x, \theta, m) p_1(\theta|m).
 \end{aligned}
 \quad (19)$$

In other words, a relatively straightforward combination of two common sequential experimental design strategies – one that targets mutual information, and one that targets uncertainty – can be theoretically motivated to achieve the dual goals of model selection and parameter estimation in the presence of prior misinformation.

**Acknowledgements** Thanks to Danny “Muscles” Oppenheimer for comments and feedback. SJS was supported by a Tata Consultancy Services (TCS) Fellowship at Carnegie Mellon University while contributing to this work. DRC was supported by National Science Foundation grant SES # 20-49896. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 (Towns

et al., 2014) (specifically, the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center), and the Computational Shared Facility at The University of Manchester.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alcalá-Quintana, R., & García-Pérez, M. A. (2004). The role of parametric assumptions in adaptive bayesian estimation. *Psychological Methods*, 9 (2).
- Berger, J. O. (2013). *Statistical decision theory and bayesian analysis*. New York, NY: Springer Science & Business Media.
- Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, 7 (3).
- Borth, D. M. (1975). A total entropy criterion for the dual problem of model discrimination and parameter estimation. *Journal of the Royal Statistical Society: Statistical Methodology*, 37.
- Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision*, 1 (4).

- Broomell, S. B., Sloman, S. J., Blaha, L. M., & Chelen, J. (2019). Interpreting model comparison requires understanding model-stimulus relationships. *Computational Brain & Behavior*, 2.
- Catanach, T. A., & Das, N. (2023). Metrics for bayesian optimal experiment design under model misspecification. (Accessed via [arXiv:2304.07949](https://arxiv.org/abs/2304.07949))
- Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, 52.
- Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, 22 (4).
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.
- DeGroot, M. H. (2005). *Optimal statistical decisions*. Hoboken, NJ: John Wiley & Sons.
- Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University. (Translated by Ruger, H. A. and Bussenius, C. E.)
- Foster, A. E. (2021). Variational, monte carlo and policy-based approaches to bayesian experimental design. Unpublished doctoral dissertation, University of Oxford.
- Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19.
- Go, J., & Isaac, T. (2022). Robust expected information gain for optimal bayesian experimental design using ambiguity sets. In: *Proceedings of the thirty-eighth conference on uncertainty in artificial intelligence* (uai 2022).
- Heide, R. de, & Grünwald, P. D. (2021). Why optional stopping can be a problem for bayesians. *Psychonomic Bulletin and Review*, 28.
- Hendriksen, A., Heide, R. de, & Grünwald, P. (2020). Optional stopping with bayes factors: a categorization and extension of folklore results, with an application to invariant situations. (Accessed via [arXiv:1807.09077](https://arxiv.org/abs/1807.09077))
- Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & the Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19 (1).
- Hou, F., Lesmes, L. A., Kim, W., Gu, H., Pitt, M. A., Myung, J. I., et al. (2016). Evaluating the performance of the quick csf method in detecting contrast sensitivity function changes. *Journal of Vision*, 16.
- Houlsby, N., Huszar, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. (Accessed via [arXiv:1112.5745](https://arxiv.org/abs/1112.5745))
- Kim, W., Pitt, M. A., Lu, Z.-L., Steyvers, M., & Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Computation*, 26.
- Kwon, M., Lee, S. H., & Ahn, W.-Y. (2022). Adaptive design optimization as a promising tool for reliable and efficient computational fingerprinting. (Accessed via [arXiv:https://psyarxiv.com/8emcu/](https://psyarxiv.com/8emcu/))
- Laming, D. (1992). Analysis of short-term retention: Models for brown-peterson experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18 (6).
- Lee, M. D., Criss, A., Devezar, B., Donkin, C., Etz, A., Leite, F. P., et al. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, 2.
- Lee, S. H., Kim, D., Opfer, J. E., Pitt, M. A., & Myung, J. I. (2021). A number-line task with a bayesian active learning algorithm provides insights into the development of non-symbolic number estimation. *Psychonomic Bulletin & Review*, 29.
- Lopes, H. F., & Tobias, J. L. (2011). Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in bayesian analysis. *The Annual Review of Economics*, 3.
- Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, 57.
- Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116 (3).
- Overstall, A., & McGree, J. (2022). Bayesian decision-theoretic design of experiments under an alternative model. *Bayesian Analysis*, 17 (4).
- Owen, R. J. (1969). *A bayesian approach to tailored testing* (Research Bulletin No. RB-69-92). Princeton, New Jersey: Educational Testing Service.
- Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17.
- Roberts, F. S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences*. Reading, MA: Addison-Wesley.
- Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5).
- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103.
- Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85.
- Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D., Lykouris, T., Dudík, M., et al. (2021). Bayesian decision-making under misspecified priors with applications to meta-learning. In *Advances in neural information processing systems* 35 (neurips 2021).
- Sloman, S. J. (2022). Towards robust bayesian adaptive design methods for the study of human behavior. Unpublished doctoral dissertation, Carnegie Mellon University.
- Sloman, S. J., Bharti, A., Martinelli, J., & Kaski, S. (2024). Bayesian active learning in the presence of nuisance parameters. (Accessed via [arXiv:2310.14968](https://arxiv.org/abs/2310.14968))
- Sloman, S. J., Oppenheimer, D. M., Broomell, S. B., & Shalizi, C. R. (2022). Characterizing the robustness of bayesian adaptive experimental designs to active learning bias. (Accessed via [arXiv:2205.13698](https://arxiv.org/abs/2205.13698))
- Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., et al. (2014). Xsede: Accelerating scientific discovery. *Computing in Science & Engineering*, 16 (5).
- Tulshyan, A., Forbes, J. F., & Huang, B. (2012). Designing priors for robust bayesian optimal experimental design. *Journal of Process Control*, 22.
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apology for the bayes factor. *Journal of Mathematical Psychology*, 54.
- Weiss, D. J., & McBride, J. R. (1983). *Bias and information of bayesian adaptive testing* (Research Report No. 83-2). Minneapolis, MN: Computerized Adaptive Testing Laboratory, Department of Psychology, University of Minnesota.
- Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26 (1).
- Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2.
- Yang, J., Pitt, M. A., Ahn, W.-Y., & Myung, J. I. (2020). Adopy: a python package for adaptive design optimization. *Behavior Research Methods*, 53.
- Young, M. E., Cole, J. J., & Sutherland, S. C. (2012). Rich stimulus sampling for between-subjects designs improves model selection. *Behavior Research Methods*, 44.