Probabilistic models for human judgments about uncertainty in intuitive inference tasks

Philipp Schustek

TESI DOCTORAL UPF / 2018

Thesis supervisor

Prof. Rubén Moreno-Bote
Department of Information and Communication Technologies

Universitat
Pompeu Fabra
Barcelona

Acknowledgment

I am grateful for having been able to dedicate four years of my life to educate myself in the scientific study of a subject that is of great interest and significance to myself - and hopefully to others. This is to acknowledge those deserving special attention for their support without which I personally would have found it difficult to reach this point.

The biggest thank you belongs to those who have been there for me ever since. I was lucky to grow up in a family which has fostered curiosity, valued education and which has always provided a cushion of support to count on that any exploratory endeavor needs. Thanks to Mum, Dad and my sister Celine.

I would also like to thank my supervisor, Dr. Rubén Moreno-Bote, for bolstering my studies and for the freedom to venture into unchartered territory. A big thank you to all the 'lab' members over the years for dealing with me on a daily basis. In particular to Ramón who helped me to settle down in Barcelona and for fervent and controversial discussions. Likewise to Iñigo for his advice, openness and awesome companionship. To Alex especially for his support and encouragement to overcome some of the bigger obstacles along the way.

The Center for Brain and Cognition at the UPF stands out for its warm welcome. Thanks in particular to Dr. Salvador Soto-Farraco for his helping hand to carry out experiments. I also owe a lot to the friendly support of Dr. Jordi Navarra and Irune Fernandez at the Fundació San Joan de Déu where my adventure began.

The persons who have accompanied me along the way and prevented excessive dwelling in the academic ecosystem will be part of my fondest memories: Pierrick, Fran, Mireia and many others who I didn't mean to leave out. Thanks to Rosnel and Martina for their friendship and for not just sharing a flat but a home. I feel greatly indebted to those who I could turn to, even though we were sometimes physically far apart. Thank you for being there: Johannes, Jessica, Mayra, Peter and Christoph.

Abstract

Updating beliefs to maintain coherence with observational evidence is a cornerstone of rationality. This entails the compliance with probabilistic principles which acknowledge that real-world observations are consistent with several possible interpretations. This work presents two novel experimental paradigms and computational analyses of how human participants quantify uncertainty in perceptual inference tasks. Their behavioral responses feature non-trivial patterns of probabilistic inference such as reliability-based belief updating over hierarchical state representations of the environment. Despite characteristic generalization biases, behavior cannot be explained well by alternative heuristic accounts. These results suggest that uncertainty is an integral part of our inferences and that we indeed have the potential to resort to rational inference mechanisms that adhere to probabilistic principles. Furthermore, they appear consistent with ubiquitous representations of uncertainty posited by framework theories such as Bayesian hierarchical modeling and predictive coding.

Keywords: Uncertainty Representation, Rationality, Bayesian Statistics, Probabilistic Inference, Cognitive Modeling, Behavioral Analysis, Confidence, Decision Making, Generalization, Cognitive Biases.

Zusammenfassung

Ein Grundstein rationalen Denkens ist die Anpassung vorherrschender Überzeugungen infolge empirischer Evidenz. Dies beinhaltet die Berücksichtigung von wahrscheinlichkeitstheoretischen Prinzipien, die anerkennen, dass jede Beobachtung, abgesehen von Idealisierungen, stets mit mehreren Interpretationen konsistent ist. Diese Arbeit präsentiert zwei neuartige experimentelle Versuche, um zu analysieren, wie Menschen Unsicherheit bezüglich ihrer wahrnehmungsbasierten Überzeugungen abschätzen. Verhaltensdaten weisen komplexe Muster probabilistischen logischen Schlussfolgerns auf, wie zum Beispiel zuverlässigkeitsgewichtete Anpassungen hierarchischer Zustandsrepräsentationen. Trotz charakteristischer Abweichungen beim Verallgemeinern, kann das Verhalten nicht auf alternative heuristische Erklärungen zurückgeführt werden. Die Ergebnisse legen nahe, dass die interne Darstellung von Unsicherheit ein wesentlicher Bestandteil unserer wahrnehmungsbasierten Schlüsse ist, und dass wir durchaus das Potential besitzen uns rationaler Inferenzmechanismen, die wahrscheinlichkeitstheoretische Prinzipien befolgen, zu bedienen. Darüber hinaus scheinen sie darauf hinzudeuten, dass interne Unsicherheitsrepräsentationen allgegenwärtig sind, was von Rahmentheorien wie *Bayesian hierarchical modeling* und *predictive coding* vorausgesetzt wird.

Stichwörter: Representation von Unsicherheit, Rationalität, Bayessche Statistik, Probabilistische Inferenz, Modellierung kognitiver Prozesse, Verhaltensanalyse, Konfidenz, Entscheidungsfindung, Verallgemeinerung, Kognitive Verzerrungen.

Resumen

Un pilar fundamental de la racionalidad es actualizar las creencias con la finalidad de mantener la coherencia con la evidencia observacional. Esto implica cumplir con principios probabilísticos, los cuales reconocen que las observaciones del mundo real son consistentes con varias interpretaciones posibles. Este estudio presenta dos novedosas pruebas experimentales, así como análisis computacionales, de cómo participantes humanos cuantifican la incertidumbre en tareas de inferencia perceptiva. Sus respuestas conductuales muestran patrones no triviales de inferencia probabilística, tales como la actualización de creencias basadas en la confiabilidad sobre las representaciones jerárquicas del estado del entorno. A pesar de los sesgos característicos de generalización, el comportamiento no puede ser correctamente explicado con descripciones heurísticas alternativas. Estos resultados sugieren que la incertidumbre es una parte integral de nuestras inferencias y que efectivamente tenemos el potencial para recurrir a mecanismos de inferencia racional, los cuales adhieren a principios probabilísticos. Además, dichos resultados son compatibles con la idea de que representaciones de incertidumbre internas son ubicuas, lo cual presuponen teorías generales como *Bayesian hierarchical modeling* y *predictive coding*.

Palabras claves: Representación de incertidumbre, racionalidad, estadística bayesiana, inferencia probabilística, modelación cognitiva, análisis de comportamiento, confidencia, toma de decisiones, generalización, sesgos cognitivos.

Preface

How can we be sure of our knowledge? This question is of tremendous relevance, for instance, to determine if we can trust a medical diagnosis before undergoing risky surgery. From an epistemic standpoint, it is related to the old philosophical question what truth is. Our success as a species seems very coincidental if we did not possess a mechanism that allows us to distill truthful statements from false ones.

However, the applicability of certain formalisms, such as Aristotelian syllogistic logic, is too limited to serve as a general rule - largely due to their incapacities to handle uncertainty. Rational reasoning crucially relies on respecting belief uncertainty for which a theoretically normative theory is available with probability theory. Nevertheless, it is generally claimed to be inadequate to describe how we actually reason because of frequently observed biases - deviations from the norm. It is puzzling that our inference skills, as e.g. evidenced by visual scene understanding, are very powerful and reach far beyond the immediate observations. But yet we systematically fail on the simplest inferences in explicit reasoning tasks. Ironically, scene understanding has turned out to be very difficult to replicate in machines, while our reasoning errors typically do not pose great challenges.

Statements regarding the subjectively assessed certainty of knowledge are expected to provide important insights into the power and flaws of human inferences. We hypothesize that humans have access to probabilistic computations for judgments about uncertainty and experimentally further explore their extent and limitations. Crucially, testing must respect the subjective 'boundary conditions' such as prior knowledge. As an analogy, numerical weather prediction requires both physical principles (fluid dynamics) but also initial and boundary conditions (today's weather) to be combined. We must disentangle whether errors arise from disobeying principles or from inappropriate boundary conditions. This work contributes to an emerging research field at the intersection of machine learning and cognitive science. It follows a constructive approach to understanding the mind and attempts to describe cognition and behavior in computational terms. After all, the basic perceptual problems are shared by both biological and artificial agents.

Chapter 1 introduces the fundamental problem of perception along with important concepts. The following Chapter 2 discusses the implications for experimental testing of rational inferences. Supplementary information is given in 'info boxes' which may be omitted without compromising further understanding. The subsequent two Chapters 3 and 4 present two comprehensive experimental and computational studies which can be read independently. Finally, Chapter 5 sums up their contributions and puts them into context with the literature.

Contents

Chapter 1

Probabilistic principles for rational inference

1.1 Perception under uncertainty

Any agent must determine its behavior based on uncertain information about the state of its environment. For our purposes, an agent is an entity that can interact with and sense aspects of the environment in which it is embedded [1]. Natural agents, such as humans, must acquire knowledge of their surroundings to avoid potentially harmful states. Instead, they must seek out valuable resources to maintain physiological homeostasis and to procreate. The better the internal map to navigate the environment, the better an agent can adapt its behavior accordingly.

This is complicated by the fact that our sense organs are adapted to only pick up certain signals, the stimuli, of the physical world. We do not have access to the environment as it is, we only have access to our sensations. The brain is enclosed in our skulls and only provided with access to high-dimensional, raw signals of the sense organs [2]. These impinging sensory signals are important to the organism only insofar as they bear information about the environment. In order to access this information, the neural code needs to be deciphered to reveal the environmental state they originated from. This interpretative process is commonly referred to as perception: "Perception is the organization, identification, and interpretation of sensory information in order to represent and understand the environment" [3].

The central idea is that perception leads to a representation of the environment. In its most elementary form, a representation is a surrogate for the thing itself [4]. The percept is the "internal re-creation" of the distal object from the neural activity called the proximal stimulus [5]. Internal operations on the surrogate "substitute for operations on the real thing" [4] - much like a simulation. Therefore, a correspondence must be established between the real environment

and the objects referred to. Perception may be considered a canonical information processing operation that extends to increasingly abstract concepts, but which are nevertheless firmly based on sensations. Apart from the semantics, the accuracy of the representation is of great importance as mental surrogates are inevitably imperfect.

1.1.1 Uncertainty: An inescapable problem

The dissociation between internal representations and real-world objects poses an inference problem. Because our senses provide only limited access, perception is uncertain, especially in realistic or complex environments. Uncertainty refers to the problem that, given a set of observations, different conclusions may be drawn about the environment.

It underlies many common ambiguous visual percepts (Fig. 1.1A-B) such as the Rubin vase [6]. However, uncertainty is no less prevalent for inferences in more abstract domains such as natural language (Fig. 1.1C-D). In the example, it is impossible to resolve the uncertainty about the meaning of these sentences without making additional assumptions about their context. In other words, the

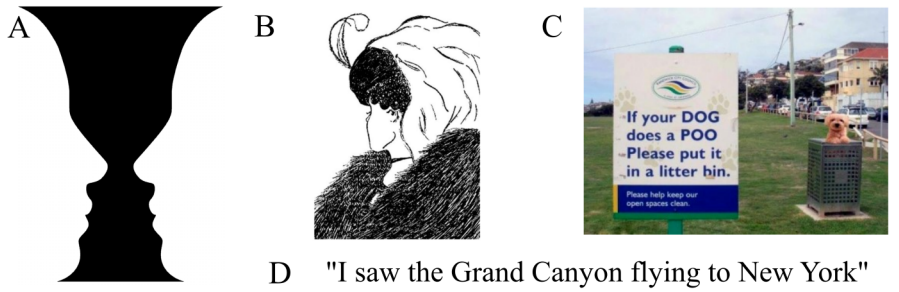

Figure 1.1: Ambiguous percepts (A) Face-vase illusion. (B) The picture can be interpreted as either an old or a young woman [7, 8]. (C-D) Sentences with several possible meanings [9, 10].

observations do not sufficiently constrain the set of hypotheses to arrive at a certain conclusion. Correspondingly, this so-called "poverty of stimulus" [11] has been considered a fundamental problem to provide a unique description of the world. Instead, suitable assumptions are indispensable to reach truthful conclusions about our environment.

Such inferences under uncertainty are very common in every-day life, e.g. when determining the trustworthiness of an unknown person based on her actions. A typically strategy is to resolve uncertainty by gathering more information. In fact, this may be the reason why people find watching crime stories interesting.

However, the inference we just engaged in (assuming it is a fact that many people like crime stories) is actually consistent with many other explanations. Not only in this example do our conclusions critically depend on assumptions how humans behave and for what they strive. Overall, the uncertainty of our inferences is fundamental as it stems from two unsurmountable origins.

Origin of Uncertainty

"*Little Susie was told by her parents to never open the basement door. Why?*"[1]. In popular games called situation puzzles, people are given some puzzling facts calling for an explanation. The goal is to explain the situation in the fewest steps possible by asking yes-no-questions to the host who has made up the narrative behind the scenario. This is an instance of active inference in which information should be sampled such that it most effectively reduces uncertainty akin to deliberate hypothesis testing.

Gathering information lessens uncertainty about possible interpretations. Non-observability [1] on the other hand introduces uncertainty. We can only sample a small subset of the state-space of the environment, e.g. because our sensors are only sensitive to particular physical stimuli. For objects more than a few meters away, stereo-vision is virtually ineffective. We only have access to 2D projections on our retinas to determine what the 3D world looks like (Fig. 1.2A). Our percepts largely correspond to the outside world because the brain effectively exploits previously acquired background knowledge [12] to draw the most likely conclusion, e.g. about the shape of an object. As finite agents, our percepts are governed by states that are beyond our access. For instance, a central characteristic of a stock's price evolution is its lack of patterns. Consequently, it is difficult to predict which out of several possible states it will adopt next. This randomness, or stochasticity, is mostly due to non-observability.

If we had access to the intentions of all potential investors, we could account for a great amount of this variation that otherwise appears to occur without obvious origin (assuming behavior can be fore-casted). In a completely deterministic world, one could in principle explain away all variation by recursively uncovering its causal factors. This is similar to exploring which branches of a maze lead to the exit or to a dead-end (Fig. 1.2B).

However, from physics we known that events on a microscopic quantum-level are nondeterministic, i.e. there is irreducible randomness. The photon interaction with a receptor in the retina is governed by laws that only allow for probabilistic statements, not for statements about single events. The molecular components of the nervous system are susceptible to individual physical quanta

[1] http://puzzlewocky.com/brain-teasers/situation-puzzles/, 05.01.2018

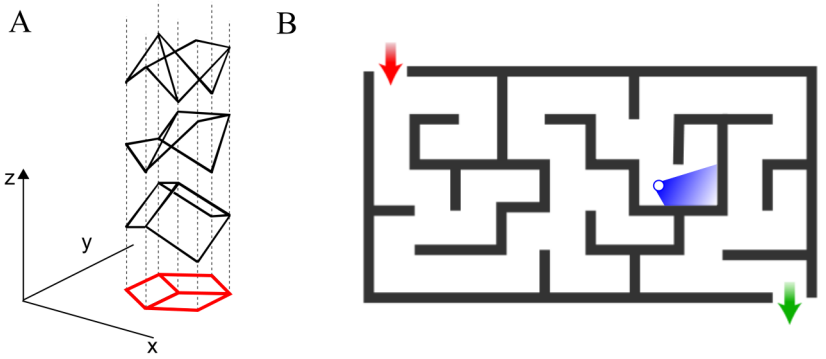

Figure 1.2: Non-observability introduces uncertainty (A) Several 3D objects (black) yield the same 2D-projection (red). Adapted from [12]. (B) Finding its way through an unknown maze, the access to the environment of the agent (blue) is limited to its line of sight (blue cone). It is *a priori* unclear which decision leads to a dead-end and which to the exit (green). Adapted from [13]

.

[14]. The brain itself is highly nonlinear so that intrinsic noise may be amplified and might be translated into behavioral variability.

Altogether, nondeterminism and non-observability are the reasons why our observations are subject to changes for which we cannot account [1]. These stochastic processes introduce uncertainty into our inferences because we face the 'problem of induction'. The question how one can generalize from "instances of which we have had no experience resemble those of which we have had experience"(D. Hume, e.g. [15]). Uncertainty is an inescapable problem for every agent, apart from artificially limited environments to which idealized descriptions, such as propositional logic, apply.

1.1.2 Perception as optimization

The idea of perception as an inference process that occurs unconsciously and prerationally in the brain dates back to Helmholtz [16]. Critically, the uncertainty involved in the process converts perception into an optimization problem because the best matching hypothesis to explain the observations needs to be found. To attain convergence of the internal world representation with the actual state of the environment, one needs an optimization objective. Probability theory offers precisely this, a way of scoring competing hypotheses in the light of ambiguous evidence (see also [17, 18]).

While we motivated the need to reduce uncertainty of the internal representation in rather intuitive terms, a more rigorous argumentation with respect to

thermodynamics and information theory has been proposed. A natural agent must seek out ways to (locally) resist the tendency of the thermodynamical forces of the environment to destroy its structural integrity by disordering [19]. In order to maintain homeostasis, its attained internal sensory and physiological states must be limited to some few states. Mathematically speaking, this means that the probability distribution over sensory states must have low entropy. This in turn can be achieved by minimizing a information theoretic equivalent of free-energy which allows to avoid surprising events in the environment. Minimizing free-energy is tightly related to posterior inference in Bayesian statistics which is discussed next.

1.2 Probability theory: A calculus for uncertainty

Above, we argued that uncertainty is inherent in the inference problem with which the environment confronts an agent. Due to the fact that we do not have access to all of the causes of our sensations, our observations are only a randomly sampled subset of a population containing all possible outcomes. As a consequence, finite and especially small samples feature fluctuations, i.e. their statistics deviate from the corresponding measure on the population.

An agent may choose to ignore uncertainty and somehow limit its internal representation to just one interpretation. A better, because more adaptive, option is to explicitly handle it by constructing appropriate representations. Probability theory is commonly regarded as the formalism of choice when uncertainty needs to be represented [1, 20].

Here we must limit the discussion to relevant aspects for the following studies and refer to the standard literature for mathematical foundations or details. Readers who prefer some more background information on the basic assumptions that establish the probabilistic formalism and its relation to reasoning are kindly referred to consult the appendix (Sec. A.1).

1.2.1 Probabilistic inference

Because of sampling, there is no deterministic correspondence in the form of a mapping between an observation and any feature of the population. Here, we illustrate the practical application of probabilistic inference with an example of an urn problem that forms the basis of an experimental task which was carried out with human participants (Chapter 3, Fig. 3.1).

Suppose that we blindly pick four items from a large urn (population) which contains only red and blue items. The goal is to determine if there were more blue or red items inside the urn before drawing. Assuming that the sample $\boldsymbol{d}$ happens to consist of three blue and one red item (Fig. 1.3, top), we cannot say for

sure what value the population proportion μ must take because the observations $\boldsymbol{d}$ are sampled from the random variable $D \sim \text{Bin}(N = 4, \mu)$. Except for the

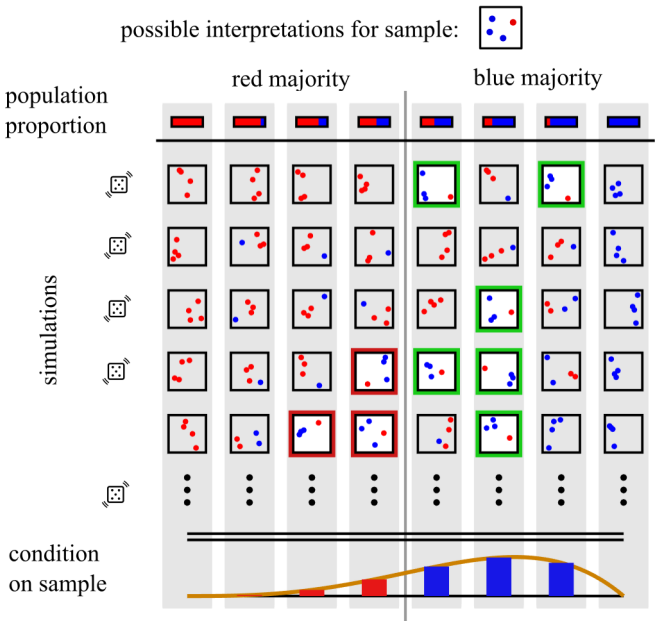

Figure 1.3: Probabilistic inference determines the hypothesis (population proportion) that is most likely to result in the observed data/sample (top). Conditioning on the observed sample (colored frames) yields the posterior distribution over possible population proportions. The simulations that generated the sample from a red majority (red frames) provide misleading evidence as the sample suggests a blue majority.

extremes of only red or blue items, we cannot exclude any value of μ with certainty as each urn proportion is capable of producing the sample (Fig. 1.3, simulated observations). However, some are more likely to generate the actually observed sample $\boldsymbol{d}$ than others.

Probabilistic inference is performed by deriving the distribution over μ when we condition on the actually observed instances $\boldsymbol{d}$. In the example (Fig. 1.3), the number of simulations that generated three blue and one red item is tallied up for each possible population proportion (columns). The normalized histogram (bottom) approximates the posterior distribution as specified by Bayes theorem (see Eq. A.1) for an infinite number of simulations (bottom, orange curve).

$$p(\mu|\boldsymbol{d}) \propto \text{Bin}(\boldsymbol{d}|N = 4, \mu)p(\mu) \tag{1.1}$$

The distribution $p(\mu)$ is termed the prior distribution (or just prior) as it reflects the state of knowledge about μ prior to observing $\boldsymbol{d}$. Ideally, it should correspond to

the base rates, i.e. the actual frequency $p(\mu)$ with which such urn proportions μ occur. While the prior distribution can be any valid probability distribution, it was chosen to be uniform for simplicity here.

Overall, observing $\boldsymbol{d}$ has changed the belief state from $p(\mu)$ to $p(\mu|\boldsymbol{d})$. Nevertheless, uncertainty about the value of μ remains which is reflected by the fact that $p(\mu|\boldsymbol{d})$ is a distributional estimate over all possible values that μ might take. Altogether, the environment's sampling process introduces uncertainty into our inferences which should result in the construction of distributional estimates over possible environmental states. The estimated posterior distribution $p(\mu|\boldsymbol{d})$ is the optimal, i.e. most accurate, belief about the population proportion μ. Any agent attempting to perform truthful inferences should ideally comply to such belief updating (see *Relation to normativity*).

Relation to normativity

Justifying beliefs in terms of their empirical evidence is at the heart of science and of rationality considerations. Our goal is to make truthful inferences to establish beliefs that achieve close correspondence to the state of the environment. In the example above, the posterior yields a justified belief if the prior distribution is sensibly chosen.

Formal arguments have been put forward that establish probabilistic inference as a normative theory of belief updating [21], claiming that probability theory is indeed the only sensible way to reason under uncertainty [20, 22]. Accordingly, the framework of Bayesian inference can be considered a normative theory for belief updating in the light of empirical evidence.

1.3 Probabilistic agents

Before, we have outlined the problem of establishing truthful beliefs about the environment that every agent faces. We followed the arguments that suggest probabilistic inference as a solution to that problem. The framework of probabilistic inference is formal and does not concretely specify neither the probabilistic models and prior distributions, nor how the necessary computations are carried out on physical hardware such as a nervous system.

However, it can be used to formulate computational models of subjective beliefs. These internal 'boundary conditions' can be tested and inferred themselves using similar inference methods as for cognitive modeling itself (see *Computational approaches to cognition*). Proponents of the Bayesian brain hypothesis [23–25] believe that this framework provides a principled approach to understanding cognition and to guide further inquiries. In essence, probabilistic inference is clai-

med not to be just a normative language but also a descriptive one for internal models of cognition [26].

The main implication for a probabilistic agent concerns the nature of its knowledge representation. The "probabilistic approach is first and foremost about representing knowledge as probability distributions" [25]. This entails that information about possible world states must be instantaneously available as "trial-to-trial neural representation of uncertainty" [27]. Consequently, the central task of a probabilistic agent is density estimation of a (probability) distribution over unobserved world states. Beyond that, belief updating from observational evidence must follow the probabilistic rules (appendix A.1).

Computational approaches to cognition

The present conceptualization of perception under uncertainty is typical for the approach of cognitive science. Cognition is a broadly used term to refer to mental processes which typically emphasizes the information processing viewpoint. It rests on the tenet that "thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures." (Paul Thagard, [28]).

If agents are well adapted to the challenge that a task poses, their solutions should be similar to formal computational approaches to the problem. In such cases, knowledge of the agents' goals and a formal model of the environment should suffice to explain and predict behavior. Such arguments underlie the rational analysis methodology [29–31] which emphasizes the structure of the environment as it is expected to provide tighter constraints on a theory of cognition than the specific structure of the brain [32]. Consequently, such accounts reside on a high level of abstraction encompassing the computational and the algorithmic levels of description according to Marr's and Poggio's classification [33, 34].

If the problem drives the solution, formal cognitive models are believed to provide a suitable starting point which can then be refined to capture deviations, e.g. due to resource constraints [35]. Critically, this is by no means a claim that a computational level analysis provides an adequate description for all kinds of tasks or phenomena. Important constraints of cognitive processing and behavior are expected to be found on algorithmic or implementational levels which more concretely describe an algorithm's realization in physical hardware. Indeed, bridging levels of analysis is a central topic of ongoing research and among the ultimate goals of cognitive neuroscience [36].

1.3.1 Generative models

Inherent in the probabilistic approach to inference is a model of the observations. It is defined in terms of hidden variables H that are used to construct a probability distribution over observable variables D (Fig. 1.4A). These hidden or latent

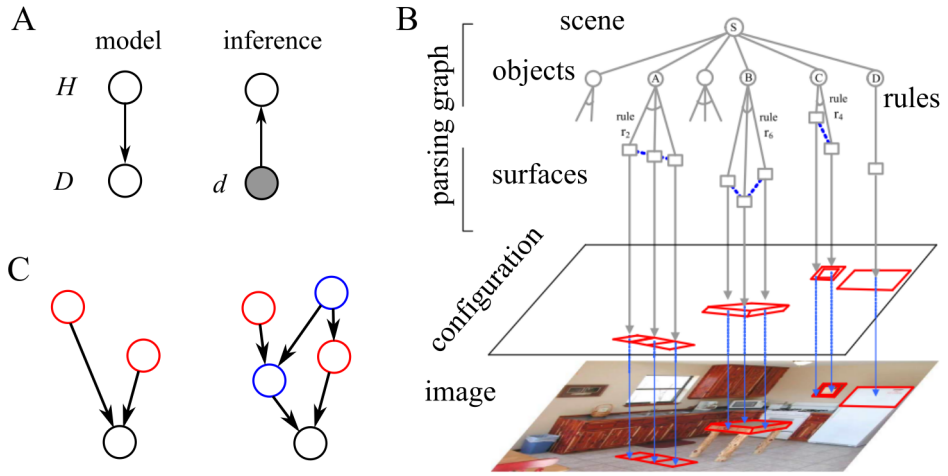

Figure 1.4: Generative models (A) A model can generate data D for a given hypothesis H. Inference determines the most likely hypothesis depending on actually observed instantiations $\boldsymbol{d}$ of D. (B) Schematic of compositional construction rules for objects from rectangular primitives embedded in visual images which can be used to parse a scene. Adapted from [37]. (C) A hypothesis (red variables) may be expanded in terms of further variables (blue) for a more flexible representation.

variables may be interpreted as possible world states, as the urn proportion in the example above (Sec. 1.2.1). As a consequence, inference follows an "analysis by synthesis" approach [38] that, simply put, inverts the constructive process (Fig. 1.4A). In the example of the urn problem, a generative model was evaluated for some candidate world states to quantify their respective consistency with the sample that was actually observed (Fig. 1.3, posterior).

Here, we use H to generically refer to such hidden variables or hypotheses. It may stand for the space over model parameters as in traditional inferential statistics, e.g. the parameters defining a mixture distribution of Gaussians. However, also models over complex spaces D such as images can be defined (Fig. 1.4B). The model specification may e.g. include a certain inventory of objects of a visual scene similar to the settings of a computer graphics engine to render images [37].

In inferential statistics, a generative model $p_M(D, H) = p(D|H, M)p(H|M)$ refers to a joint probability distribution over both observable D and hidden variable(s) H. It specifies which variations in sample space

could be explained away by knowing the value of H. The model M defines the distribution over the random observations D in terms of H and potentially further parameters. Ultimately, the distinction between what constitutes a hidden random variable and what a parameter is artificial and depends on assumptions about the context and the scope of modeling.

Our observations result from a multitude of interrelated causal, physical processes. Hypotheses may be defined in terms of further intermediate states or latent variables (Fig. 1.4B,C). In probabilistic graphical models the probabilistic chain rule (appendix A.1) allows to expand the latent structure to richer representations.

To predict the observations well, the structure of the hypotheses H ought to correspond to the causal structure in the environment. Several organization properties of the physical world such as symmetry, locality, compositionality, and polynomial log-probability [39] have been suggested to facilitate representation learning. Mirroring the outside dependence structure internally with a web of interdependent variables gives rise to a mental "small-scale" model of the world [40]. Strictly speaking, it is only a generative model of the observations D with some correspondence to the actual processes in the environment. Over its individual perceptual history, the agent thus learns an approximate world-representation, including uncertainty, by refining the distribution over the latent variables H.

1.3.2 Learning as belief updating

The Bayesian inference equation (Eq. A.1) can be understood as an update rule for a statistical model of the observations. A 'good' internal representation allows the agent to truthfully generalize from a sample $\boldsymbol{d}$ to the actual, unknown population $p_E(D)$ of the environment E. The adequacy of an internal representation can be objectively assessed, e.g. by predictive accuracy. Mathematically, generalization from a sample $\boldsymbol{d}$ is expressed as the predictive distribution over unseen observations D which is a posterior-weighted sum (or integral) of the probabilistic model.

$$p(D|\boldsymbol{d}) = \sum_H p(D|H)p(H|\boldsymbol{d}) \tag{1.2}$$

The internal model of the environment $p(D|H)p(H)$ needs to adapt to the environment $p_E(D)$ in order to make good predictions and to avoid surprising events (Sec. 1.1.2). In principle, an agent can always compute how probable a given set of observations $p(\boldsymbol{d}|H)$ is under its current hypothesis H of the environment. Critically, Bayesian model evidence integrates out competing *a priori* values of the hidden variables H and thereby automatically penalizes complex models with widely dispersed prior distributions $p(H)$ [41, 42] that overly adapt

to the data $\boldsymbol{d}$,

$$p(\boldsymbol{d}) = \sum_H p(\boldsymbol{d}|H)p(H) \ . \tag{1.3}$$

Maximizing the probability of the observed data, the model evidence - or marginal likelihood - provides an optimization objective to more accurately represent the environment (Sec. 1.1.2). The agent is free to tune the prior $p(H)$ and the model structure H itself, i.e. the distribution over the latent variables of the statistical model which is typically governed by further hyper-parameters.

The free energy formulation provides deep analogies between the minimization of (environmental) surprise (free energy) and the maximization of a lower bound on evidence [19, 43]. Intuitively, adaptive changes to better represent the environment should be limited to those that increase model evidence. Posterior (Bayesian) inference amounts to a suppression of free energy [19]. Hence, the free-energy principle may be seen as a justification of the Bayesian framework linking information processing in an representational agent with first principles from thermodynamics.

So far, we have mainly discussed what we ought to do as perceptual agents who have to act in only partially observed environments. Probability theory provides a measure of evidence that allows to find those hypotheses which offer the best account of the problem. However, for the difficult real-world inference problems that humans face, often not even the problem structure itself is clear. We may always test a particular hypothesis but hypothesis generation is outside of the formalism discussed above. Hence, the basic practical problems of a probabilistic agent are model selection (Sec. 1.4) and computability (Sec. 1.5) which are discussed next. We will see that the situational context and tractability naturally impose restrictions for applications embedded in realistic environments.

1.4 Inductive biases and inference models

The fundamental problem of induction (Sec. 1.1.1) reappears in the Bayesian inference formulation as the choice of the model. The specification of a particular probability distribution out of all possible distributions over some space of outcomes is high-dimensional. We need to impose constraints on possible distributions by assuming a certain parameterization or latent structure such as the objects that might be present in a visual scene. These constraints then impose biases for inference onto what might be identified. If the only represented categories are cats and dogs, even a picture of a car is forced to be categorized as either of the two. Correspondingly, a model can be seen as a compact representation of all possible data [44].

One can explore whether the assumptions expressed by a particular model are also adopted by humans when they perform an experimental version of the problem [45, 46]. Accordingly, these assumptions have been termed inductive biases [32], as they affect how an agent generalizes from a sample to the population. Hence, the model itself constitutes a form of prior knowledge by imposing inductive biases [32].

1.4.1 Embedding in the causal context of task

A task-specific knowledge representation should be constrained by the context. The hierarchical structuring of nature [39] allows to extract knowledge that is applicable in a wider range of contexts. There is considerable experimental evidence that humans organize and interpret their input in a hierarchical manner [47–49]. Moreover, also behavioral responses feature hierarchical patterns [50] and e.g. follow optimal action hierarchies [51].

A concrete formalism are probabilistic graphical models (PGM, e.g. [52]) which describe the construction of probability distributions in terms of a (potentially large) web of interrelated latent random variables whose dependence structure can be represented with a graph. Moreover, this allows to incorporate knowledge of unidirectional, causal dependencies among variables [53–57].

In the language of PGMs, higher-level, ancestral nodes define and thus constrain the prior distribution for lower-level variables (see Fig. 1.4B). Importantly, hierarchical dependence structures allow the agent to learn constraints from observational data itself as learning takes places on all levels (see task in Chapter 3). Thus, learning increasingly abstract representations on multiple levels of a hierarchy partly resolves the conundrum of what priors an agent should choose for a particular task [58]. Hence, in principle at least, the model can be inferred as well, e.g. by learning more abstract graph-production rules [59]. However, such a hierarchical extension just defers the problem one level up. If the correct 'rule' is not included, inference will suffer from biased estimates. For some applications such 'abstracting away' might suffice, but it is generally unclear what 'rules' one should follow to generalize well.

1.4.2 Structural uncertainty and instance-based approaches

Beyond uncertainty regarding the designated latent variables, there can be structural uncertainty [60]. More specifically, this refers to uncertainty about what the latent variables are, their dependencies and the parameters defining a model. This is very common in real-world environments and a substantial problem for the generative approach to inference which requires choosing a model.

We illustrate that with a simple example of clustering in some arbitrary feature space (Fig. 1.5). The agent is assumed to know that the features are normally distributed for each cluster. It however does not know any parameter of the Gaussian mixture distribution, nor how many clusters there are in the first place (Fig. 1.5A). It is an example of unsupervised learning in which only the unlabeled observations are given. Specifically, there is no information about the latent structure such as the cluster identity from which a data point originated.

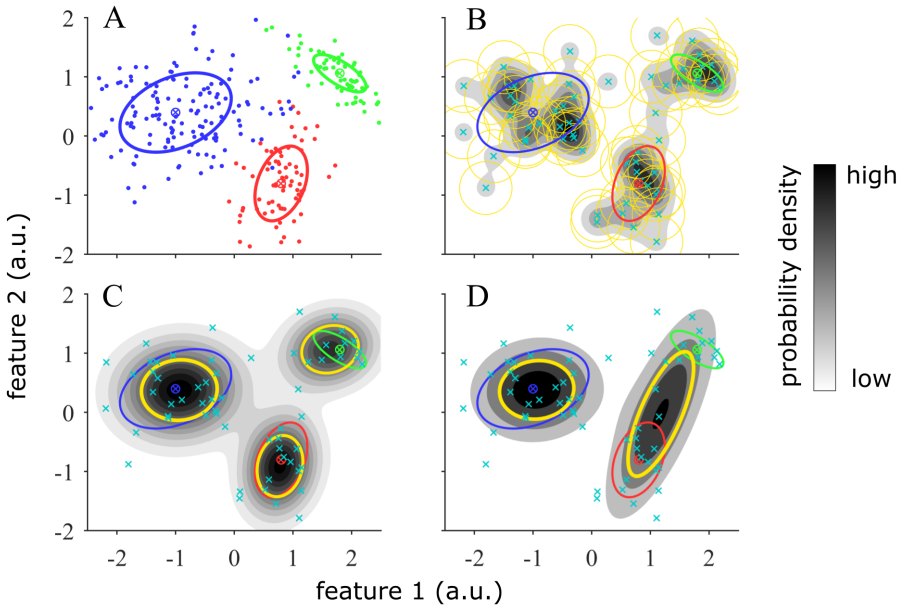

Figure 1.5: Estimating density in terms of an unknown number of clusters (A) The 'unknown' generative structure consists of three 2D-Gaussian distributions (blue, red, green) whose means and covariances are indicated. Values are sampled according to a mixture distribution over the Gaussian and indicated in their respective colors. (B) Kernel density estimation (KDE) sums up the contribution (color code) of Gaussian basis distributions (yellow) placed on each observed data point (cyan crosses). (C) Bayesian nonparametric clustering infers the appropriate number of basis distributions (yellow) for the data. (D) As opposed to nonparametric approaches, an imposed limitation to two categories results in biased estimates.

The agent may approach this inference problem in different ways depending on how freely its internal model may adapt to the observational data. Purely instance-based inferential methods take an extreme stance in that they impose only minimal constraints on the distribution to be inferred. They effectively assume that the sample is highly representative of the population.

A simple example is nonparametric kernel density estimation (KDE, Fig.

1.5B). The estimated density is constructed by placing generic basis distributions (Gaussians of certain width, yellow circle) on each data point (cyan) and by then summing up their normalized contributions. The inferred distribution is explicitly conditional on the data. As opposed to parametric approaches, the data does not just enter through the updated latent variables H (see e.g. Eq. 1.2). As a consequence, the model complexity grows with the amount of data which is retained as building blocks of the model. This is the defining property of a more general class of inferential methods termed nonparametric [44, 61].

The strengths of instance-based methods are complementary to their weaknesses. In terms of the bias-variance classification of generalization errors, they fall at the low-bias end of the spectrum (see also [41]). As they do not impose strong structural constraints, such approaches may be used as a starting point in environments of unknown structure. This is essential to acquire new concepts such as entirely new categories or clusters of features. People have been shown to imagine and flexibly generate new category exemplars and categories in experiments [62]. In our example, the density inferred by KDE will be more concentrated over the generative clusters if enough data is provided.

However, purely instance-based, nonparametric methods come with major liabilities. First, their low estimation bias comes at the expense of a high susceptibility to fluctuations in the sample. Especially in high-dimensional spaces, the assignment of probability (density) is typically sparse and local. This is reminiscent of the over-fitting problem inherent in the maximum likelihood procedure to model fitting. Without any restrictions, the model that strictly maximizes the likelihood $p(D|H)$ is discontinuous and distributes density only to the data points which become singular. This extreme case highlights the importance of including constraining prior (structural) knowledge to avoid that the model fits the noise.

As a consequence, instance-based inferential methods are detached from contextual or prior knowledge. Strictly speaking, KDE only implicitly captures the three clusters. It actually estimates one cluster for each data point (Fig. 1.5B). In terms of causal modeling, each data point is treated as its own cause, as if the observed instances would consistently reproduce in the long run. Such a proliferation of latent variables results in inefficient representations. They do not harness the structure inherent in the data and poorly transfer to different contexts.

1.4.3 Efficient representations

Part of the inefficiency evidently arises from the need to memorize all training data. Another problem is the large overlap of the basis distributions which makes the representation very redundant (see Fig. 1.5B). To alleviate this, the efficient coding hypothesis [63] posits an additional constraint known as redundancy re-

duction. Nevertheless, even more crucial than a mere "compressed representation of sensory experience" [64] is a representation whose factors are driven to be statistically independent. Such disentangled representations are considered to be important for generalization and have been suggested to enable zero-shot inference [65]. Studies have claimed that the internal representation of human visuo-motor errors are sparse, i.e. they are made up of a low number of basis distributions [66]. Additionally, they were found to be close to orthogonal, i.e. non-overlapping, and to have compact support which is achieved by tiling the space with simple adjacent basis distributions.

In the statistical realm, Bayesian nonparametric methods (BNP, see e.g. [61]) have been suggested to alleviate both problems of purely nonparametric methods. First, they allow for an incorporation of prior knowledge. And second, they implement "an automatic Occam's razor embodied in Bayesian inference [that] trades off model complexity and fit to ensure that new structure (in this case, a new class of variables) is introduced only when the data truly require it" (Tenenbaum et al., [42]).

In the clustering example, the BNP-model correctly infers the dominance of the three underlying clusters (Fig. 1.5C) without any explicit restrictions. Such a strategy effectively is a hybrid between instance-based nonparametric approaches and structured parametric approaches. This framework is used to explain human inductive biases for predicting future events (Chapter 4). More generally, enforcing sparsity and orthogonality of the latent structure is believed to provide a link towards theory-based Bayesian models of inductive learning and reasoning [67, 68].

1.4.4 Generalization biases

Ultimately, it is not generally clear what the right level of flexibility is. The basic challenge is: "balancing constraint and flexibility, or the need to restrict hypotheses available for generalization at any moment with the capacity to expand one's hypothesis spaces, to learn new ways that the world could work." (Tenenbaum et al., [42]). We must learn representations that are not fixed and that may grow with the number of observations [69]. However, a purely instance-based generalization scheme too readily posits new categories and thus detaches from prior knowledge such as in the extreme case of KDE (Fig. 1.5B). Nonparametric models generally follow a strict bottom-up approach as the resulting representation is (almost) solely determined by the data.

On the other hand, if we make the *a priori* assumption in our example that there are just two clusters, the estimated density will be biased no matter how much data we collect (Fig. 1.5D). Such expectation-biased generalization enfor-

ces a structure that is very rigid. The data points are only assigned to either of the two clusters which is inappropriate in this case. More generally, we must communicate whether the hypothesized model is still adequate and revise beliefs about the context (the model) if necessary.

Consequently, a feature of hierarchical generative models is a bidirectional information flow [58] between higher-level, contextual variables and lower, more input-related variables (see study in Chapter 3). In such a scheme, the balance of top-down compared to bottom-up influences is critically determined by uncertainty. To select a proper context, uncertainty must be ubiquitously available across all levels of the hierarchy (see e.g. [43]).

After all however, the choice of a model is not a purely representational question. It also depends on the cost of performing inference on that knowledge representation.

1.5 Performing probabilistic inference

In the previous section, we mainly discussed the generative model $p(D|H)p(H)$ that allows to generate data D conditional on a hypothesis H. For inference we must compute the posterior distribution $p(H|D)$ over the entire hypothesis space which often poses a big practical problem.

1.5.1 The computability problem

Rich latent structures with many hidden variables H are necessary to map onto the generating processes in the environment. If their domain is unbounded and they are assumed to vary independently, finding the best matching hypothesis $p(D|H)$ requires searching the Cartesian product space $H = X_1 \times ... \times X_n$ of all latent variables $X_1, ..., X_n$, that grows exponentially with the number of variables. For interesting real-world problems this is prohibitively expensive.

This "curse of dimensionality" [41, 70] can be mitigated by exploiting the fact that the laws of physics constrain generative variables to a subspace of lower dimensionality. The Helmholtz machine [71] e.g. limits this combinatorial explosion by hierarchical self-supervised learning which maximizes a lower bound on the probability of the observations. It has been shown that inference in Bayesian belief networks (also PGM) belongs to the nondeterministic polynomial-time (NP)-hard complexity class [72]. There is evidence that human cognition is similarly affected by computational complexity measures [73].

Furthermore, perceptual inference is a highly non-convex optimization problem that requires a rather global optimization scheme to avoid local optima. In the brain, neural variability has been suggested to enable this exploration of diffe-

rent possible interpretations [74]. More formally speaking however, even powerful approaches, such as genetic algorithms, are heuristic solutions (see [75]) as they cannot guarantee global optimization. Interestingly though, certain evolutionary dynamics resemble Bayesian inference describing "populations of beliefs" [76].

1.5.2 Facilitating assumptions and approximations

One way to attain tractable posterior inference is a restriction to simpler representations. Within the appropriateness of these representations, often even exact posterior inference is possible. If we restrict a probability distribution to have a certain shape which can be described by a function and its parameterization, the parameters efficiently encode the whole distribution. Operations on distributions can then be substituted by operations on their parameters. If a conjugate prior is chosen, the posterior has the same functional form (parameterization) as the prior and inference may be formulated as iterative updates of the parameters (see e.g. [41]). Similarly, an analytic expression for normalization may be found that otherwise requires a numerical integration over the whole space of hypotheses which is often prohibitively expensive.

An easier, factorized joint probability distribution may be obtained if conditional independences between latent variables can be assumed [41]. Inference can then be performed by local computations (marginalization) on the factors and by passing messages between factor nodes (see also [77]). All together, such approaches can greatly facilitate the computation of the posterior and have led to tractable inference methods even for large belief networks [41, 52, 78, 79]. However, often tractable computations come at the cost of concessions to representational accuracy. Too many times, a certain distribution is chosen for mathematical ease rather than being warranted by the generative processes of the data.

Alternatively, to render inference tractable, one may forgo exact inference and compute the posterior distribution only approximately. For instance, a distribution can be approximated by a small number of samples to reduce computational cost. This comes with an asymptotic guarantee because the sampling approximation approaches exact posterior inference as more samples are taken. Applications for which computationally efficient sampling schemes can be found are very powerful [68].

A more radical solution is to deliberately limit inference to a belief subspace. Such categorical commitments are similar to decision making among several mutually exclusive ways to act and have been reported for humans [80].

Links to algorithmic and implementational accounts

Important constraints for a theory of cognition (see e.g. [25, 57] for overviews) may be found on levels that describe how abstract information processing is carried out on physical hardware.

The sampling hypothesis (see [81]) suggests that the brain is a sampler whose dynamically changing states over time correspond to a probability distribution over these states [82]. There is experimental evidence, such as behavioral variability, for which the sampling hypothesis provides an elegant explanation [83, 84]. The results range from perceptual bi-stability [85] over variability in the decision processes [86] to causal inference [87]. There appear to be deep connections between posterior modes and attractor basins in dynamical models [88]. Further studies attempted to map sampling-based probabilistic representations onto neural models in cortex [89] and interpreted neural variability as probabilistic inference through Markov chain Monte Carlo sampling [90].

A different implementation for probabilistic machinery in the brain are probabilistic population codes (PPC) [27, 91, 92]. They suggest that neurons of different tuning properties represent a basis function decomposition of a distribution over a stimulus variable. Contrary to sampling, a complete distribution is maintained by PPCs at all times which may make them more demanding in terms of memory. For such representations, message passing akin to inference in PGMs has been suggested [77]. Overall, there are further ways to represent uncertainty, and the brain may use more than one encoding type. Accordingly, there is evidence that "distinct neural encoding (including summary statistic-type representations) of uncertainty occurs in distinct neural systems" [93].

Despite the wealth of environmental features that are represented by the brain, neural processing is suggested to follow a small set of information processing operations (e.g. common to all input modalities) [94]. Canonical microcircuits (see e.g. [95, 96]) have been suggested to support population-level integrative operations such as divisive normalization [94]. At the implementational or physical level, for instance, the neuromodulators acetylcholine and norepinephrine are believed to signal expected and unexpected uncertainty respectively [97].

1.6 The rationality of inferences

So far, we have mainly discussed the problem of making truthful inferences that closely correspond to the environment. However, we as humans have not evolved with the purpose of making the best possible inferences of the world. If at all, we

are maximizers of evolutionary fitness with a wide array of competing proximal goals.

1.6.1 Valuation and objectives

Natural selection across ancestral generations has equipped us with basal motivations that are an evolutionary proxy for beneficial states to attain in order to maximize evolutionary fitness (see [98] for a comprehensive review). The brain features dedicated neural structures, the reward system, that provide signals to virtually all parts of the brain through neurotransmitters such as dopamine [99]. Evolutionarily important primary rewards trigger learning and define the proximal reward functions of everyday behavior (e.g. money or candies). It is even possible that curiosity or learning itself is intrinsically rewarding.

Beyond that, if an agent possesses an accurate representation of the environment, it is able to make better inferences, predictions and select better actions. Hence, truthful representations are of at least instrumental value as they allow to obtain other rewards more efficiently. They may thus increase the agent's evolutionary fitness in comparison to one with a poor knowledge of its surroundings. Nevertheless, truthful inferences are just one among many competing goals. To arbitrate between several actions which lead to differently valued goals, decision theory suggests the following.

1.6.2 Combining values and beliefs in decision theory

A decision maker should always choose the option of highest expected utility (see also [100, 101]).

$$E\left[U(A)\right] = \sum_{O} p\left(O|A\right) U(O) \tag{1.4}$$

A decision or action A leads to outcome O with probability $p(O|A)$ and has utility $U(O)$ (subjective value). The expected utility $E[U]$ depends on action A and is the sum over the product of all possible outcomes (or states) and utility. This is assured by the von-Neumann-Morgenstern utility theorem if basic assumptions about an agent's preferences are satisfied [102]. As a consequence of the uncertainty about the state of the environment, decision outcomes are uncertain as well. This rule is practically difficult because values need to be assessed and the outcome probability depends on the expected (inferred) state of the environment.

1.6.3 Cost, Effort and Motivation

The expected utility hypothesis of decision theory is rather a desideratum than a descriptive rule for human cognition [103, 104]. The theoretical separation bet-

ween probabilities and utilities is lifted for effort considerations of state inference [105]. There is evidence that mental effort is similarly important for human cognition as computability costs are for machines (Sec. 1.5.1). Accordingly, human participants display a strong incentive to avoid mental engagement [106]. Their responses are systematically biased by a cost to act [107]. Likewise, the passage of time for evidence accumulation has been associated with an increasing internal cost that shapes the timing of a decision [108]. These examples indicate that cognition comes at a substantial internal cost. Importantly, also the two-systems framework of Kahneman [109] mainly separates cognitive mechanisms along an axis of mental costs.

1.6.4 Computational and bounded rationality

As the goals of a natural agent are typically not perfectly aligned with any given task at hand, there is no strict incentive to optimize. Additionally, to cut through complexity, humans are suggested to rather "satisfice" (H. Simon, [110]), i.e. search until an acceptable solution has been obtained.

This notion has reemerged as computational rationality [111] for human cognition and artificial computing systems. An explicit incorporation of the cost of computation into the objective function makes an agent also deliberate about the most suitable way of solving a problem. If there is e.g. no time-pressure, it may be worth investing more time into finding a close-to-optimal solution, whereas under time pressure one better goes with the default option instantaneously.

While simple in principle, it requires a quantification of the internal costs to make better (meta-) decisions of resource allocation. A trade-off must be found between the accuracy of an approximation and the costs to compute it. Such considerations have been shown to introduce e.g. reward-modulation of sensory receptive fields [112] - i.e. they interfere with the representation of beliefs. Experimental evidence has been provided that the brain disposes over several behavioral control systems and ways of arbitrating between them [113]. Cognitive costs may be tracked based on the degree to which control mechanisms are employed [114]. This arbitration constitutes a form if meta-optimization with potentially rational cost-benefit trade-offs [115].

1.6.5 Rationality and optimality

All the considerations above must be taken into account when the rationality of behavior is experimentally assessed. A mere comparison to some devised task optimal strategy is generally insufficient [116, 117].

Optimality is used to refer to an ideal solution to a specific task or agreed

upon, objectively known problem. Strictly speaking, task-optimality is an impossible problem even for a machine for virtually all problems of real-world complexity [118].

As opposed to optimality, rationality acknowledges the perspective of the agent. Most scholars such as Kant identified two dimensions behind the concept of rationality. This largely accords with the differentiation between theoretical and practical rationality [119, 120] which we build upon here.

Theoretical rationality refers to internal belief consistency. Beliefs must be grounded in empirical support. Logical consistency has a formal, objective dimension such as the use of Bayes theorem to update beliefs.

Practical rationality Optimization of the use of cognitive resources to achieve goals with respect to their relative importance (meta-optimization).

Table 1.1 attempts to illustrate these concepts with some examples for rational and irrational inference behavior.

Table 1.1: Examples for different notions of rationality

	Theoretical rationality	**Practical rationality**
Rational	Bayesian inference over all known candidate structures	Heuristic inferences to quickly finish a dull and repetitive task
Irrational	Neglecting prior knowledge; Inconsistent beliefs; Local or subspace optimization	Heated debate due to emotional reactions limiting inferences to deeply entrenched stereotypes

Equating theoretical rationality with task-optimality grossly overlooks competing internal goals and the costs of cognition [121]. It furthermore neglects that the agent might solve the problem under (slightly) different assumptions which may be justified given its perceptual history [117, 122]. Hence, irrationality does not automatically follow from suboptimal task performance.

In this work, we are predominantly interested in testing for theoretical rationality which translates to representing knowledge as distributions and Bayesian belief updating. However, for the participants in our experiments, the practical notion of rationality is equally important as they face mentally effortful tasks and certainly possess competing goals such as a desire to finish early. The next Chapter addresses how our rational inference abilities have been interpreted and how they can be measured.

Optimal vs. probabilistic

The terms 'optimal' and 'probabilistic' should not be conflated [123]. Giving optimal responses in some task does not imply probabilistic processing

and vice versa. Probabilistic inference based on wrong assumptions leads to suboptimal results. Optimal responses in turn can under some conditions be learned associatively without resorting to internal probabilistic representations. "The probabilistic approach, however, is not about optimality per se, [...] the probabilistic approach is first and foremost about representing knowledge as probability distributions [...]." (Pouget et al., [25])

Chapter 2

Probing the rationality of human inferences

Dealing with uncertainty and adhering to probabilistic principles of belief updating are key criteria for rational inferences. The experimental assessment how humans internally quantify uncertainty is expected to reveal characteristics of the underlying representations and inference processes. We will also revisit how our rational abilities have been interpreted based on experimental evidence.

2.1 Measuring uncertainty processing

The defining characteristic of probabilistic computations is a "trial-to-trial neural representation of uncertainty" [27]. Knowledge of latent variables must be represented in the brain as distributions instead of as point estimates [124]. Ideally, we would like to elicit a behavioral statement that cannot be made on a trial-by-trial basis from point estimates only. The experimenter may induce variations regarding the certainty with which the stimulus supports a conclusion, such as a decision, across experimental trials. The responses of the participants, such as decision confidence, should then follow (covary with) these induced "trial-by-trial" variations of uncertainty. Participants must be prompted to reveal and outwardly communicate their internal estimates. If those estimates take the format of distributions, a suitable (scalar) "summary statistic" must be elicited [125]. As an example, the computation of Bayesian decision confidence is discussed next.

2.1.1 Bayesian confidence hypothesis

Uncertainty refers to the whole distribution over the set of all hypotheses H under consideration. Decision confidence in contrast may be regarded a summary statis-

tic that reports the fraction of probability mass/density assigned to the subset Q of these hypotheses that corresponds to the choice made [125–127].

For instance, in the urn problem introduced earlier (Sec. 1.2.1), we might like to compute the confidence that the majority of items in the urn is blue, irrespective of the actual proportion. To derive decision confidence, uncertainty is quantified through the integration over all hypotheses $H \in Q$ that correspond to the decision made (see also Fig. 1.3, blue part of histogram at the bottom).

$$c(Q) = p(Q|\boldsymbol{d}) = \int_{H \in Q} p(H|\boldsymbol{d}) \mathrm{d}H \tag{2.1}$$

This Bayesian or posterior-based approach construes confidence as expected decision accuracy as it acknowledges that the sample may also be consistent with $H \notin Q$ (Fig. 1.3, red part of histogram). Such a quantification of uncertainty makes use of the normalization property of probability theory and will be used in both experiments.

If the agent's model $p(D|H)p(H)$ matches the generative model of the data, confidence predicts the probability that a decision turns out to be correct. In this case, a decision maker is said to give calibrated confidence reports. The Bayesian confidence hypothesis has received empirical support also due to its amenability to animal research (see also [128–130]). Specifically, we will use it to model confidence reports in the empirical prior task (Chapter 3).

Other experimental measures

Probing representations of uncertainty is experimentally difficult due to the subjectivity of uncertainty representations (e.g. the many degrees of freedom). Decision making is most commonly used [131] because the decision boundary in hypothesis space may be clearly communicated. Nevertheless, the participant still has to map his internal estimate (summary statistic) onto the response variable provided by the experimenter (e.g. saccade endpoint [132]). By converting it to discrete ratings, natural language or numerical scores certain nonlinear distortions may be introduced obscuring the actual estimate [133].

Other approaches use indirect reports, e.g. by introducing a gamble or wager [134]. However, effects such as loss aversion on post-decision wagering must be taken into account [135]. Few methods such as Matching Probability (no-loss gambling) [136] circumvent such pitfalls at the expense of a more complicated incentive structure of the task which must be understood

by the participant.

Some experiments indirectly evidence uncertainty estimates through their effects on other behavioral responses [86]. For instance, participants were found to counteract a lateral distortion based on uncertain evidence when making reaching movements [137].

Not all experiments elicit trial-by-trial estimates of uncertainty. Some exclusively rely on recorded decisions (e.g. [138]) and estimate per-trial confidence from the fraction of correct trials. Hence, only trial-averaged statistics are shown to conform to those of a probabilistic agent. Other studies like this work directly probe momentary representations of uncertainty [139–141].

2.1.2 Comparison to a model-free approach

So far, we have mainly discussed the benefits of a probabilistic generative model for inference. Here we point to important consequences if we drop that assumption. We compare a probabilistic agent (A1) to a hypothetical contestant (A2) that merely learns appropriate responses to stimuli. This is not intended to be exhaustive as there are many other learning strategies available. Nevertheless, it attempts to highlight that judgments about uncertainty are very difficult for non-probabilistic, model-free agents even in the simplest tasks especially if supervising feedback is withheld.

We build on the urn problem introduced before (Sec. 1.2.1) where the agent was asked to decide whether the majority of items inside the urn is either blue or red. Beyond that, the goal here is to provide accurate estimates of the chances that a decision will turn out to be correct. The task consists of independent instantiations of this problem across trials whereby both the unobserved urn proportion and the sample size vary unpredictably (for details see appendix A.2).

The following provides a brief comparison between critical aspects of both learners (see Table 2.1 for an overview). The probabilistic agent (A1) is assumed to known the appropriate probabilistic model to generate observations. It performs posterior inference over the latent urn proportion and reports decision confidence as expected accuracy (Eq. 2.1). However, if we assume that it uses an inappropriate uniform prior distribution that does not correspond to the actual base rates of the latent urn proportions, its estimates are slightly suboptimal throughout the task (Fig. 2.1, green line).

The agent A2 on the other hand does not possess a generative model. Whereas agent A1 builds a representation of the environment, agent A2 merely learns to make appropriate actions for different stimuli. Such mapping-based approaches are also referred to as model-free (e.g. [142, 143]) because they cannot generate observations, e.g. simulate hypothetical samples for a given urn proportion.

Table 2.1: Comparison of a probabilistic model-based and a model-free agent.

Agent	Probabilistic (A1)	Model-free (A2)
Model	Generative	None
Inference	Posterior over urn proportion	Heuristic: Sample proportion
Estimate	Distribution	Point estimate
Uncertainty	Distribution contains full information	No explicit representation
Confidence	Expected accuracy	Learned function of input
Objective	Model evidence	Error across trials if supervised
Task experience	Not necessary	Needs feedback
Origin of procedure	Hierarchical context	Unclear
Task generality	Environment-specific, not task-specific	Specific to learned task objective
Advantage	Generalization, contextual extension	Simplicity
Disadvantage	Biased inference for wrong assumptions	Task specific, handcrafted

Instead, agent A2 may infer the urn proportion heuristically by computing the proportion of blue samples which is actually accurate in the limit of infinite sample size. Both agents report the same decisions which always follow the sample majority.

While the estimate of agent A1 is a distribution over all possible urn proportions supported by the data, the estimate of agent A2 commits to one singular interpretation. As opposed to the distributional estimate of agent A1, the scalar state estimate of agent A2 does not contain explicit information regarding the uncertainty of the urn proportion. Therefore, agent A2 has no principled objective for deriving decision confidence on a single-trial basis. However, it may use the feedback about the correctness of its actual choices to learn appropriate confidence estimates over trials (Fig. 2.1, A2).

Consequently, and given enough flexibility, it may approximate the mapping that the inference procedure of agent A1 defines. In fact, the behavior derived from the posterior distribution is nothing but a particular, albeit complicated, mapping from sensory inputs to actions. To suppress that possibility in experiments and to evidence the reliance on an internal probabilistic model, supervising feedback about correct behavior must be withheld [124]. Remarkably, even for this simplistic task, many trials are required to reach comparable levels of accuracy.

A data-efficient and thus fast way to learn is batch processing (Fig. 2.1,

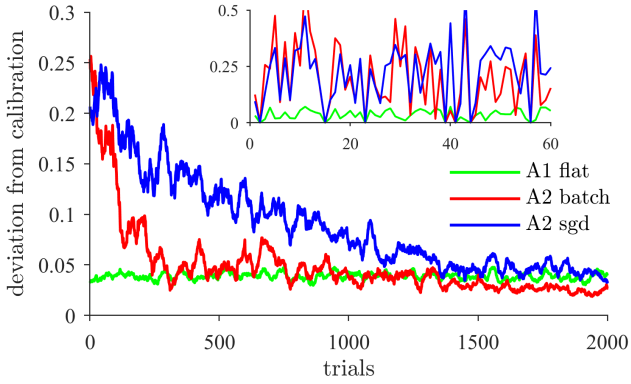

Figure 2.1: Model-free agents must learn from experience The absolute difference between decision confidence judgments and the actual probability of making correct decisions for different learners (smoothed). Agent 1 (A1, green) is a probabilistic model-based agent correctly describing outputs as a function of the unobserved proportion of items inside the urn. Its flat prior distribution is mismatched so that there is small mis-calibration. The model-free learner (A2) can only learn from experience over trials. Batch learning (red) requires memorizing all statistics and outcomes across trials but converges faster as on-line learning (blue, stochastic gradient decent (sgd)). The inset zooms in on the first trials (unsmoothed).

red). Critically, this requires the agent A2 to memorize all decision outcomes for each experienced condition. A more realistic assumption is some form of on-line learning, i.e. small parameter updates which only depend on the outcome of the preceding trial. This however leads to highly fluctuating confidence estimates and slower convergence than batch learning (blue). In contrast, the probabilistic agent A1 does not rely on any feedback from prior experience with the task and can make principled confidence judgments from the first trial on (Fig. 2.1, green).

For both agents, the origin of the inference procedure is not readily clear. However, probabilistic generative models are ideally suited to be embedded in larger hierarchical or contextual structure. On the other hand, the problem of selecting a suitable heuristic appears to be more severe. As opposed to probability theory, heuristics lack a theory of evidence so that the agent does not possess a within-trial optimization objective to select the 'best fitting' heuristic to explain the data. Due to these problems, integrative models rather than collections of heuristics have been proposed to describe human inferences [144].

Moreover, a generative model is a task-independent representation of the environment. Therefore, it leads to task generality because the representation is still useful when the objective changes (e.g. reporting the probability that the urn contains at least ten percent of red items). It also generalizes more readily to

seemingly minor changes in the generative process such as to smaller or larger sample sizes. The mapping-based agent A2 would in both cases have to revise its confidence mapping by learning from experience again.

The main advantage of a heuristic approach is its simplicity which is the main motivation that heuristic accounts have been considered. The downside is their task specificity and that it is not clear what a 'good' heuristic is. On the other hand, a disadvantage of probabilistic models are biased estimates if their assumption do not (fully) match the generative process as shown here. However, such models may resolve contextual uncertainty and e.g. learn about the prior. We will address a related problem of contextual inference in Chapter 3 in which the capability to infer appropriate prior knowledge based on its reliability is examined.

2.2 Interpretations of suboptimal human performance

The power of today's computational methods enables rigorous testing of behavior against benchmarks which are often termed optimal conditional on some assumptions about the task. One may categorize the observed behavioral deviations from task optimality as follows:

(1) Fundamental insufficiency to reason theoretically rationally (Sec. 1.6.5) Violation of formal rules such as Bayesian inference

(2) Systematic biases through the use of approximations (Sec. 1.5)

(3) Structural problem mismatch (Sec. 1.4) Subjective assumptions by the participant that do not match the generative structure of the task

(4) Response noise Residual variations due to (input-independent) influences beyond the experimenter's control

The forth point is neither avoidable nor surprising given the influence of many extraneous factors on experimental responses, such as motor noise. Upon more careful inspection, the noise carries information about its origin [145] which is however not further examined in this work. Next, we will revisit the scientific literature that has cast doubt on the notion that humans adhere to rational mental processing (interpretation 1).

2.2.1 Heuristics and biases

In the wake of the heuristics and biases program pioneered by Kahneman and Tversky [109, 146], many studies have reported cognitive biases, flawed reasoning and decision making errors in humans.

Many participants were reported to obey the so-called representativeness heuristic. Loosely speaking, it refers to the idea that humans generalize according to similarity, e.g. that a person belongs to a group because she matches the stereotypical description of the group, irrespective of other factors that are important for judgments of probabilities. This entails an insensitivity to base rates and to sample size [147]. Correspondingly, the 'law of small numbers' [148] states that an observed sample - no matter how small - is representative of the population in all its aspects [149]. The notion of representativeness is conceptually somewhat vague and commonly illustrated through a set of examples [150]. The general idea of a tendency to equate the population with the sample is reminiscent of instance-based generalization bias (Sec. 1.4.4) and is extensively discussed in both studies.

We will address other systematic errors, such as a confirmation bias which is characterized by an interpretation of ambiguous evidence such that preexisting beliefs are selectively supported [151, 152], for instance by neglecting contradictory evidence, which may lead to belief perseverance [153]. We have already discussed such expectation-biased generalization in the clustering example before (Sec. 1.4.4). There, a restriction to only two clusters neglects evidence for the more appropriate choice of three clusters and thus leads to a confirmation bias even though inference is otherwise completely rational.

Furthermore, humans are frequently reported not to treat probabilities linearly but to "apply" a systematic nonlinear weighting function that over-weighs small probabilities and vice versa (probability distortion, see e.g. [154]).

Overall, research has compiled a huge list with myriads of biases that might actually stem from common mechanisms. Much of it may be considered phenomenological because of its unsatisfactory insight into the origins [153]. Common to many accounts is the conclusion that the rational faculties of humans are poorly developed or rarely used in practice.

If the rational abilities of humans are poorly developed, some other mechanisms must be able to explain the undeniably big cognitive achievements of humans. We were suggested to instead use an "adaptive toolbox" of heuristics [155]. The defining characteristic of heuristics is their cognitive or computational simplicity and their reliance on some way of effort reduction by ignoring part of the information. In other words, they are justified by 'satisficing' an immediate goal, while they do not and cannot guarantee optimal results.

Yet, there is no clear-cut notion what constitutes a heuristic. For instance, the coarser a computational approximation (Sec. 1.5), the more heuristic an inference may appear. Furthermore, justifying the extent of such an approximation may be regarded a heuristic choice unless it results from a clear optimization rationale that trades off its cost against its benefits (Sec. 1.6.4). Critically, in perception under uncertainty, heuristics are not straightforwardly linked to a measure of evi-

dence that allows to arbitrate which heuristic performs 'better' inference on just one instance of the problem (see also Sec. 2.1.2).

2.2.2 A different perspective by cognitive neuroscience

The availability of modern computers to conduct experiments and to perform model-based analysis has scrutinized these interpretations. As a result, describing inferences in probabilistic terms, underpinned by behavioral and computational studies, has gained considerable interest in the scientific community and suggests that the human brain's capability to perform probabilistic inference may have been under-appreciated.

Experimental evidence in support of internal probabilistic computations mainly rests on behavioral studies. Many have shown that human behavior responds to uncertainty in ways that is consistent with probabilistic processing [156]. A corollary of probabilistic belief updating is uncertainty-weighted integration which relies more strongly on information sources of low uncertainty, or conversely high reliability. As an example, psychophysical experiments of cue combination (e.g. [157]) claimed a reliability-based weighting of conflicting cues from different modalities that is consistent with distributional estimates provided by each modality [138].

The visuo-motor domain is particularly amenable to experimental testing. Studies reported that the planning of reaching movements depends on representations of motor-error [158, 159]. Participants were also found to counteract a movement perturbation based on the reliability of a visual cue [137].

Beyond the visuo-motor domain, human participants are reported to adhere to principles of probabilistic inference in more abstract cognitive tasks [160, 161]. Interestingly, such claims have even been made for pre-verbal infants [162] suggesting some independence from language and cultural learning. Humans were also able to give sensible confidence judgments in a complicated hierarchical learning task [139]. Furthermore, they demonstrated the ability to learn patterns of abstract (unsignaled) hierarchical visual concepts without explicit awareness [163].

Beyond that, there is evidence from neuroscience that sensory uncertainty represented in visual cortex can be decoded and used to predict behavior [164]. Activity in certain brain areas, such as the anterior cingulate cortex (ACC), shows a specificity to uncertainty, e.g. to the estimated volatility of the environment [160].

Additionally, experimental paradigms have been developed to test uncertainty representation in behaving animals [129]. Their results suggest that monkeys possess an internal notion of confidence which is used to guide behavior [128, 130, 165]. Altogether, these and similar findings show that nervous systems may

adhere to principles of rational inference.

2.2.3 Reinterpreting biases

In the light of these seemingly contradictory findings, what are the appropriate interpretations for human insufficiencies to comply to task optimal behavior presented at the outset of this section 2.2? As opposed to denying rationality (interpretation 1), a closer examination of the actual constraints that the agent faces more strongly emphasizes an explanation due to the use of approximations (2) and internal constraints to adequately represent the problem (3). For instance, approximate Bayesian belief updating (e.g. [166]) and efficient coding may account for systematic deviations [167]. Similarly, sampling approaches to posterior inference have been shown to naturally generate a variety of systematic probabilistic reasoning errors [82].

Critically, the assumption that the task structure is transparent to the participant is actually a strong and unrealistic one [116]. Performing inference with a mismatched model can lead to severely biased inferences [168]. In more extreme cases of structural uncertainty, participants may resort to model-free behavior [60].

Strikingly, many of the tasks for which violations of rational reasoning were reported feature a commonality. They tend to be description-based and strongly rely on working memory and natural language. These developmentally relatively recent faculties may constitute a severe impediment to interact with inferential systems that are capable of rational inference [133].

2.2.4 Design of experiments

Following these ideas, and to evidence the actual potential of human inferences, the two studies introduced in this work attempted to avoid these interrogation problems. The task design was guided by the 'principle of intuitiveness'. In short, outward reporting of internal estimates should be facilitated and reduce the involvement of both working memory and natural language. Correspondingly, the tasks avoid any explicit deliberation that would allow our participants to easily explain their behavior.

The first study (Chapter 3) investigates the human ability to infer and use contextual information through the construction of empirical priors. Special attention is given to generalization biases due to imbalances between bottom-up and top-down influences which are more specifically investigated in the subsequent chapter. The second study (Chapter 4) is an examination of the inductive biases that underlie the generalization of continuous random variables.

Chapter 3

Empirical priors for confidence judgments

3.1 Abstract

As a consequence of the uncertainty inherent in perception, our observations must be supplemented with suitable contextual assumptions to make better inferences. This is challenging as for real-world problems neither the context is certain. While this theoretically requires uncertainty representations on both the task and the contextual level, little is known how humans solve such problems. We present a novel hierarchical cue integration task in which human participants may learn a contextual prior belief from a series of ambiguous cues across trials. This contextual prior provides additional constraints for inference of a latent variable on the trial level. There, participants freely express their decision confidence which is found to closely correspond to actual decision accuracy. Behavior exhibits several nontrivial patterns of probabilistic inference such as sample size effects. Despite the high degree of sophistication, commonly reported reasoning fallacies are not generally present and neither do participants appear to rely on simple heuristics. Instead, information integration can be captured with reliability-based message passing between latent variables across hierarchical state representations. This is evidence for ubiquitous representations of uncertainty similar to a probabilistic agent.

The content of this study is currently prepared for publication, P. Schustek and R. Moreno-Bote.

3.2 Introduction

Updating beliefs to maintain coherence with observational evidence is a cornerstone of rationality [120, 169]. Real world outcomes are imbued with uncertainty such as the preference to vote for a certain party of a randomly polled voter. To infer the unknown voting preferences of the whole population from a finite number of polls, one faces uncertainty in that the results are consistent with many possible interpretations. Probabilistic inference acknowledges uncertainty by representing degrees of belief with distributional estimates [25] as opposed to point estimates [124], such as committing to only one possible election outcome.

The scarcity and insufficiency of the data to constrain the conclusions may be mitigated if powerful assumptions about the task's context can be made. For instance, prior knowledge that voters in a certain context (e.g. their profession) tend to have similar preferences. Indeed, the power of human inferences is believed to crucially rest on selecting appropriate contextual knowledge to supplement sparse stimulus data [42].

This is complicated by the fact that the contextual structure generating our observations is not fully certain. The central question is where reliable information about the context itself originates from. Theoretically, Bayesian inference can be extended hierarchically so that upper-level, ancestral latent variables constrain lower-level, more task-related variables [59]. Hence, contextual information itself can be inferred across multiple levels of a hierarchy [43]. Hierarchical Bayesian inference selects contextual constraints depending on the degree of empirical support received in related situations. Such hierarchical dependencies among latent variables are characterized by bottom-up and top-down information flow that conveys empirical and contextual information respectively [58]. Previous studies evidenced that humans internally represent uncertainty and adapt behavior accordingly [25, 124, 138].

However, the scope of explicit uncertainty representations is unclear. (1) A fully probabilistic agent maintains a probability distribution over its entire latent structure which ideally mirrors the generative process in the outside world. Such a complete representation allows to derive statements about uncertainty such as decision confidence [125–127]. (2) Humans may instead apply crude approximations to simplify the problem. One example are categorical commitments to an interpretation once sufficient evidence has been accrued for it (see e.g. [80]). If contradictory information is subsequently neglected, this could lead to a confirmation bias [151] of existing beliefs. (3) Alternatively, we may use rather task-specific heuristics, such as learned cues to uncertainty [17], which are detached from representations.

In the latter case, rational probabilistic principles are rejected to describe

internal information processing [155]. This interpretation is favored by a wealth of studies reporting reasoning biases [109] such as an insensitivity to sample size for confidence judgments [147, 149, 150].

This, in turn, is challenged by tasks in more perceptual domains that typically found humans to adhere to principles of rational belief updating. However, evidence is limited to rather simple tasks in which the unobserved problem structure was low-dimensional and is often assumed to be completely transparent to the participant.

This study developed a novel paradigm that allows to probe the acquisition and subsequent use of contextual information. To our knowledge, only very few studies (e.g. [47, 139]) have experimentally addressed the maintenance of explicit uncertainty representations across a hierarchy of latent variables which are posited by framework theories such as predictive coding [58, 170, 171].

We asked whether participants acquire and use contextual information in a graded manner that respects contextual uncertainty? To which degree are their inference patterns specific to a probabilistic agent? And furthermore, what systematic biases are displayed? Despite mostly expected deviations, participants performed remarkably well and displayed reliability-based belief updating over hierarchical latent structures. These findings invite challenging the popularized notion that we are not suitably endowed with the necessary preconditions for rational mental processing.

3.3 Results

Many formal random processes are embedded in real world problems. The example of election forecasting above is essentially an instance of an urn problem. In this study, we use a hierarchical extension of a basic urn problem to construct an inference task with hierarchical latent structure and map it onto a more accessible metaphor phrased in commonly understandable terms to avoid resorting to formal mathematical descriptions. As this structure is latent, we must assure that the participant attributes the observations to the right processes and variables.

We attempted to carefully align the actual structure of the problem with the one that the participant assumes. To achieve this, participants were incrementally exposed to the full complexity of the problem (see Methods 3.5.3-3.5.4). Specifically, they first completed a basic version of the task on a separate appointment (Experiment 1) that we will address first. Afterwards, we will discuss the hierarchically extended learning task (Experiment 2).

3.3.1 Experiment 1: Empirical support of decision confidence

The basic urn problem is phrased as follows (details in appendix C.2): The sampled dots on the screen represent passengers randomly exiting an airplane that transported only two kinds of passengers (Fig. 3.1a). One must decide whether the flight

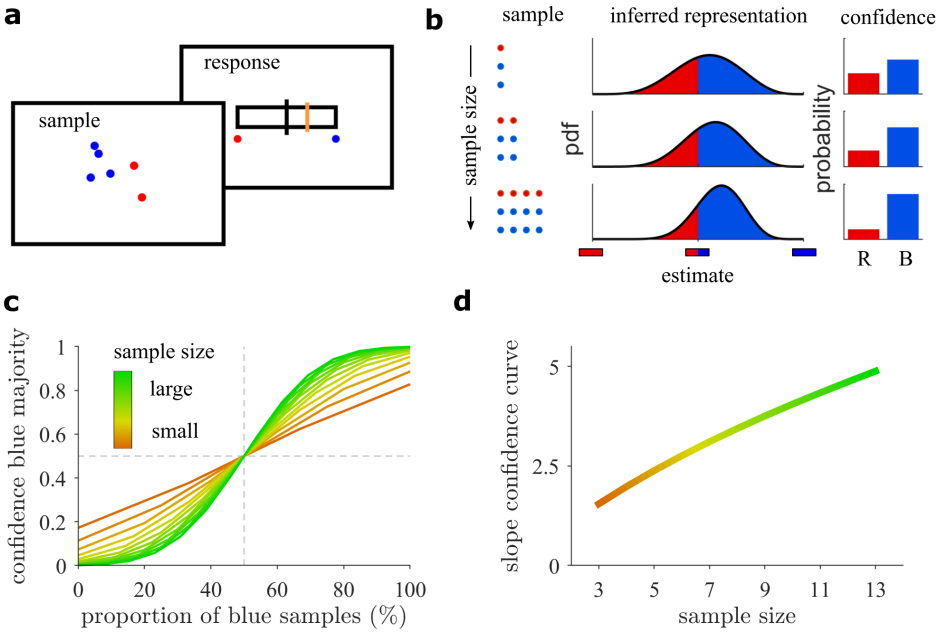

Figure 3.1: Posterior-based confidence features sample size effects
(a) Task: The colored dots (sample) represent two kinds of passengers (blue and red) that disembarked an airplane. The participants are subsequently asked to report the confidence in their decision that the airplane carried more blue or red passengers by horizontally moving the cursor line (orange). **(b)** Sample size increases posterior-based confidence in a (blue) trial majority (right) suggested by the (blue) sample majority. Confidence (right) is computed as expected accuracy from the area under the curve for the inferred proportion (middle). **(c)** Posterior-based confidence in a blue airplane majority increases with the proportion (%) of blue samples. A higher sample size (color coded) increases decision confidence for a given sample proportion and leads to a higher slope at the category boundary (50 % red-blue). **(d)** Consequently, the slope parameter of fitted sigmoidal functions increases with sample size.

carried more blue or red passengers based on the small sample on the screen. Our participants freely determined when to proceed to the response screen where they were instructed to report their confidence in the correctness of their decision. The more confident they are that the airplane majority is blue, the farther right from the middle they should position the response cursor and vice versa (Fig. 3.1a, vertical yellow line).

The only feature that distinguished the sampled passengers was the dot color that we chose to be either blue or red. Because the positions of the dots are communicated not to be informative, the sample is completely summarized by the sufficient statistics, i.e. the numbers of blue and red passengers N_B, N_R. We will mostly use the equivalent formulation $D = q = N_B/N, N = N_B + N_R$ expressed through the presented sample proportion q and the sample size N. As the presented sample proportion on the screen can be assumed to virtually equal the subjectively perceived proportion, we simply refer to it as 'sample proportion' (appendix C.3).

Each trial consists of an independent instantiation of this problem. The sample size and the latent airplane proportion are independent draws from constant distributions across trials (Methods 3.5.3). After each trial, the participant receives feedback about the correctness of his decision but no supervising feedback regarding his confidence estimate. On pauses every five trials, only trial-averaged feedback based on the absolute deviation from actual performance was provided to motivate task engagement and to determine a bonus payment at the end of the entire experiment (Methods 3.5.2). The airplane metaphor was chosen for convenience, but the mathematical problem could surely be mapped onto a different metaphor instead.

Inherent in the probabilistic approach to inference is a generative model that allows to simulate outcomes. The agent knows that the observations are random draws from a population (airplane) of blue and red passengers whose proportion is unknown. This population proportion is the latent variable that needs to be inferred. To achieve this, the probabilistic agent calculates the posterior distribution over possible population proportions (Methods 3.5.6, Eq. 3.4) and thus acknowledges that several latent proportions are consistent with the sample.

However, some are more likely to generate the observed sample than others. In the example (Fig. 3.1b, middle column), posterior probability is predominantly assigned to population proportions corresponding to a blue majority. Thus, the agent believes that a blue population majority is more likely and hence the rational response is to report a blue majority, even though a red one still receives substantial empirical support. The proportion of posterior probability assigned to a blue majority (the choice) thus corresponds to the expected decision correctness and equals the proportion of the blue area under the posterior distribution (Fig. 3.1b). Correspondingly, a probabilistic agent may naturally report confidence as expect decision accuracy (Methods, Eq. 3.5). The better its internal model matches the generative process in the environment, the more correct its subjective estimates of decision accuracy will be. Importantly, given the choice of a model, no previous task experience is required.

A central feature of posterior-based confidence is sample size dependence. For example, if we consider three different samples with the same sample pro-

portion but of increasing sample size (Fig. 3.1b, rows), the posterior distribution becomes more concentrated over population majorities suggested by the sample. Consequently, confidence in a blue majority, i.e. expected decision correctness, increases with sample size. In contrast, for any approach that neglects sample size, any systematic confidence report would be constant for all samples. Such behavior was reported for humans [109, 149] but we challenge its generality.

In the experiment, sample size unpredictably varies across trials. For a probabilistic agent, confidence in a blue majority is a function of both the proportion of blue samples and the sample size (Fig. 3.1c). Sample size acts as a magnifier for the population majority which is suggested by the sample proportion. It thus provides a proxy to experimentally control the uncertainty of evidence and to evidence intricate patterns of probabilistic inference.

More concretely, sample size effects can be evidenced by the slope of the confidence curves when we condition on sample size (Fig. 3.1c). We fitted sigmoidal functions to the reported confidence in a blue majority whose slope parameter can vary separately for each sample size (Methods 3.5.7). For a probabilistic agent, this slope pattern features a steady increase with sample size (Fig. 3.1d).

3.3.2 Confidence judgments are predictive of their performance

First, we sought to establish a correspondence between the confidence estimates of our participants and their actual trial-by-trial decision correctness as estimated by expected accuracy of the optimal inference model. The raw experimental response was linearly scaled to an interval between zero (red) and one (blue) and hence it can be interpreted as a confidence estimate of a blue trial (airplane) majority. As a relative quantity, it can be straightforwardly converted to the belief in a red trial majority (see Methods 3.5.2). We found that human confidence judgments are highly predictive of their trial-by-trial decision accuracy despite systematic deviations from giving calibrated responses (linear correlation, $\rho = 0.81$, $p = 1.27 \cdot 10^{-45}$, details in appendix C.1).

3.3.3 Participants adjust confidence to sample size

A hallmark of probabilistic inference is a dependence on sample size (Fig. 3.1c-d). As a function of the sample proportion, experimental confidence shows sample-size-conditional curves (Fig. 3.2a, solid lines). We found that human confidence judgments increase with larger sample sizes as evidenced by the increasing slope of these sample-size-conditional curves (Fig. 3.2b, linear correlation, pooled across participants, $\rho = 0.30$, $p = 7.05 \cdot 10^{-6}$). This even holds individually for 19 out of 24 participants (linear correlation, shuffling test, threshold $p = 0.05$). Remarkably,

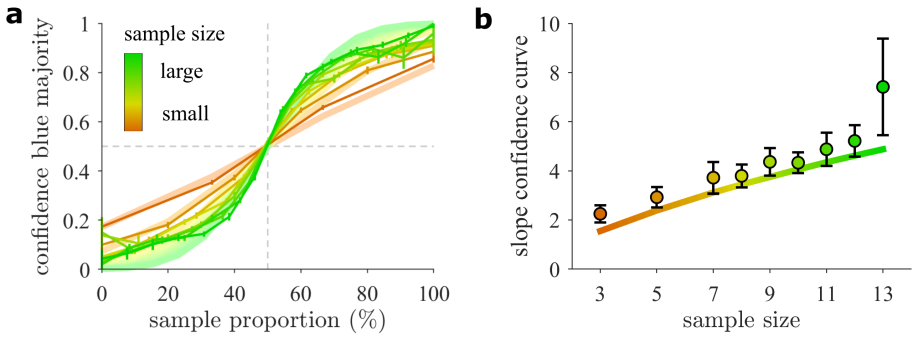

Figure 3.2: Human confidence estimates vary with sample size as described by probabilistic inference
(a) Confidence in a blue majority is a monotonic function of the proportion of blue samples (solid lines). For a given sample proportion, sample size (color coded) increases the respective decision confidence (optimal model in light colors). **(b)** The slope of the confidence curve in (a) increases with sample size. Participants feature a quantitatively similar increase as the optimal model (solid line). Error bars indicate SEM across participants.

we found a high quantitative match of the average responses across participants (data points) to the optimal model's responses (solid lines) calculated on the same trials (Fig. 3.2b, linear correlation of the averages, $\rho = 0.90$, shuffling test $p < 0.001$). Hence, participants show quantitatively similar sample size effects as a probabilistic agent.

Consistently, sample size was found to be crucial to predict confidence judgments as determined by the comparison of the optimal inference model to two different heuristic estimators. The ratio model (ratio) reports confidence as a function of the sample proportion q alone. This could be the result of a simpler approach in which the population estimate is a point estimate corresponding to the sample proportion. The ensuing implicit assumption that the sample is representative of the population is actually accurate in the limit of large sample sizes. The difference model (diff) on the other hand estimates confidence as a function of the difference of blue and red samples $N_B - N_R$. As the ratio heuristic, this statistic is informative of decision correctness but additionally even covaries with sample size. The output of the optimal model and the heuristic estimators are fed into a sigmoidal function, called response mapping (Methods 3.5.6), to map the estimates onto the unit interval or to account for distorted reports.

The comparison between the optimal model (opt) and the ratio model (ratio) shows that the latter is clearly rejected because of its incapacity to take sample size into account (details in appendix C.3, Fig. C.2). Even though the confidence estimates of the difference model (diff) are sensitive to the sample size, they typically do not correspond to the notion of uncertainty that our participants report. We can

thus dissociate the experimental reports from these simple but covariant heuristics and conclude that the response patterns of our participants typically follow a probabilistic inference approach.

3.3.4 Experiment 2: Learning inferential constraints from prior data

In order to test the ability to handle contextual uncertainty, we extended the basic task (Experiment 1) hierarchically by forming blocks of always five consecutive trials which are governed by a single binary contextual variable. The generative

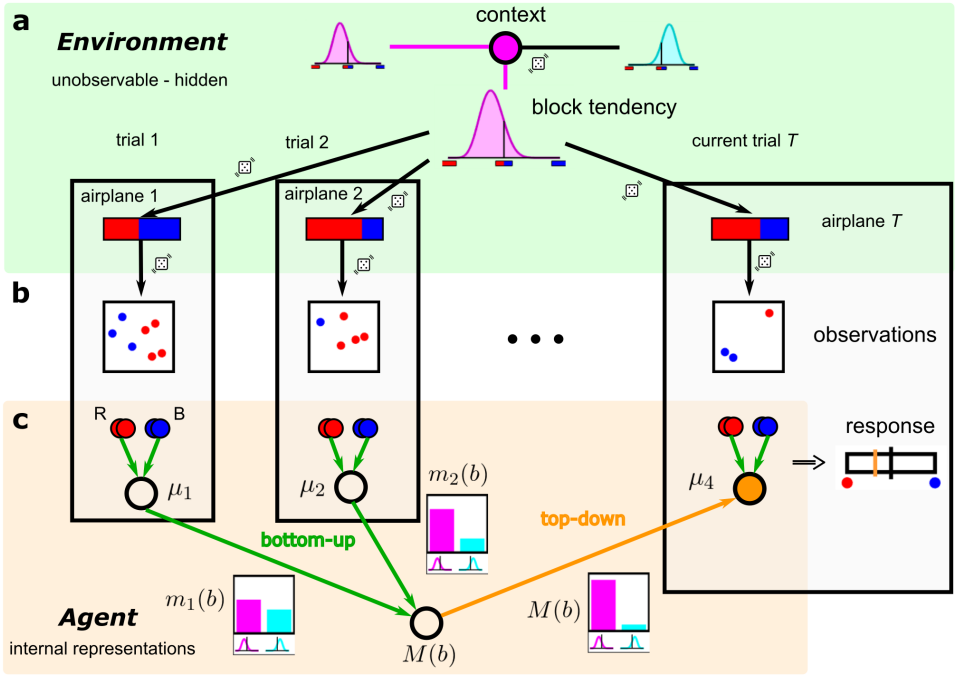

Figure 3.3: Schematic of the hierarchical structure for learning empirical priors Participants are told that across trials $(1, 2, ..., T)$ within a block (schematic) they will see passengers from different airplanes arriving to the same airport. They are still asked to report their confidence about the airplane majority on the current trial T. **(a)** Within a block of trials, the hidden airplane majorities are drawn from a common skewed distribution (cyan or magenta), the block tendency, which is selected by a contextual variable b drawn once for each block. In the example, the context favors airplanes of red majorities. **(b)** Sample generation given the airplane majority is the same as for the basic task. **(c)** The internal representation of the agent (orange background) mirrors the dependence structure in the environment (green background). Probabilistic inference is performed by message passing between the nodes which internally represent the hidden variables. Previous trials $(t < T)$ provide evidence about the block tendency through messages $m_t(b)$. They are probabilistically integrated into a belief about the block tendency $M(b)$ which provides top-down constraints on the inference of a new airplane's majority (orange node).

structure of the observations which were shown to the participant is illustrated in Fig. 3.3a (green shading). In the example, the contextual variable b is randomly chosen to favor red airplane (trial) majorities in the block by selecting a skewed Beta-distribution (details in Methods 3.5.5). On each trial, the airplane majority is independently drawn from this distribution (magenta shading). Given the airplane majority, the procedure to select the presented sample is identical to Experiment 1 (Fig. 3.3b). Importantly, also the task remains the same as inference is still asked to be made about the airplane majority of the current trial T (Fig. 3.3c, orange variable).

The participant is informed of the block tendency by extending the task metaphor. The block is supposed to correspond to several airplanes that arrive consecutively at an airport of a particular city. The city is known to host an event that tends to attract either more red or more blue passengers and thus corresponds to the binary contextual variable b. The contextual variable itself is unknown, so that joint inference must be performed over the trial-level variables and the contextual block-level variable which introduces a dependence on all previous trials ($t < T$) within a block.

Such prior knowledge provides additional information beyond the momentary sample by confining the space of hypotheses that may explain the data. For instance, knowledge of being in an environment in which there is a prevailing tendency to observe 'red' airplane majorities should raise the corresponding confidence, even for very ambiguous or even contradictory samples. The probabilistic agent inverts the generative structure to perform inference (Fig. 3.3c, orange shading) by passing messages $m_t(b)$ to update latent variables on the graph (Methods 3.5.6, Eq. 3.6-3.10). In the example, trials encountered prior to trial T provide information about the contextual variable b through the messages $(m_1, ..., m_{T-1})$ which result from a marginalization operation over the trial-level latent variables $(\mu_1, ..., \mu_{T-1})$ of previous trials. These messages are integrated and define the belief $M_T(b)$ over the binary contextual variable b prior to trial T. This 'prior' is passed downwards to the trial-level to constrain inference about the current airplane's majority (Fig. 3.3c, orange) and to generate the response (Methods, Eq. 3.10). At the beginning of each block, the context is unknown. It may only be acquired through inference across trials within a block as there is no feedback (Methods 3.5.4). In the whole task, the inferred block tendency $M(b)$ is never asked to be reported. It is only indirectly revealed through the top-down effect on the responses.

3.3.5 Features of the probabilistic inference model

We use the optimal inference model that is assumed to know the generative model of the task to illustrate experimentally testable probabilistic inference patterns (Fig. 3.4). Human participants cannot be expected to fully comply with these particular task assumptions. In the following, we highlight the variation patterns that we consider specific to probabilistic belief updating but which do not presume knowledge of the exact Beta-Binomial mixture model used in the task (see Methods 3.5.6).

Integrating information from previous trials leads to different confidence reports conditional on the real (hidden) block tendency (Fig. 3.4a). The inferred contextual belief $M(b)$ correlates with the actual contextual variable and thus leads to an increased decision confidence (on average) for a blue trial majority in a context that favors blue majorities and vice versa. We must control for evidence from the momentary sample to show actual prior effects. This is achieved by plotting confidence as a function of the sample proportion as the generative sample size distribution is by construction independent of all other quantities. Ideally, the vertical separation of the context-conditional curves is maintained over the whole range and features point symmetry around the point of indifference (Fig. 3.4a, center).

The uncertainty of the inferred contextual belief $M(b)$ gradually affects confidence reports (Fig. 3.4b). The higher the prior belief in a blue context, the larger should be the top-down modulation of the response towards a blue trial majority and vice versa. To increase statistical power, we plot the (block-) aligned confidence. This refers to the confidence that matches the actual context in the environment which can always be obtained from the normalization property. Likewise, the aligned quantities of the sample proportion and the inferred contextual prior belief are statistically independent due to the symmetry of the skewed generative distributions defining the context (Fig. 3.3a). Ideally, aligned confidence is a monotonically and gradually increasing function whose slope indicates the modulatory strength of the prior belief on trial-level inference.

Sample reliability governs the integration of momentary evidence with the prior belief (Fig. 3.4c). Our task requires an integration of bottom-up signals from the momentary sample $D_T = (q, N)$ with top-down contextual signals $M(b)$, e.g. to resolve conflicts when the sample suggests a blue airplane majority while the inferred context suggests a red majority. For a given sample proportion, larger samples reduce uncertainty about possible trial (airplane) majorities more strongly and consequently should be relied upon more. Ideally, the modulation of aligned confidence with the momentary aligned sample proportion is stronger for larger sample sizes leading to the crossover of the two conditional curves at the point where sample evidence is indifferent.

Samples of high reliability influence bottom-up belief revision more strongly (Fig. 3.4d). When learning the context, one should rely more strongly on those trials that allow to draw strong conclusions, as opposed to trials whose sample is likely to be a result of fluctuations. Thus, the modulation of aligned confidence with the sample proportion of the previous trial should be stronger when the previous sample was large. This is an effect that is only indirectly revealed through the influence of the bottom-up messages $m_t(b)$ on the belief about the inferred context $M(b)$.

Each trial provides the same information about the context compared to all others on average across blocks. The test must be separated with respect to the number of previous trials as normalization effects reduce the influence of each trial as more trials are added (Fig. 3.4e, Methods 3.5.6). We used a model that allows to selectively adjust the influence that each trial's sample proportion q has on the judgments in later trials (Methods 3.5.7). Ideally, for an infinite amount of data, all trial weights for the same number of previous trials should be positive and equal.

Information from all previous trials within a block should be accumulated and thus aligned confidence should increase on average as more trials are observed (Fig. 3.4f). Theoretically, hierarchical integration does not require memorization of the samples after bottom-up belief revision. Merely the belief corresponding to the contextual variable $M(b)$ must be kept in memory. Ideally, aligned confidence for the optimal model is a monotonically (and sub-linearly) increasing function of the number of previously observed trials within a block.

Importantly, all these patterns refer to covariation, they do not claim absolute confidence values. Extraneous factors such as subjective assumptions about the strength of the block tendency naturally lead to deviations such as vertical offsets of most curves.

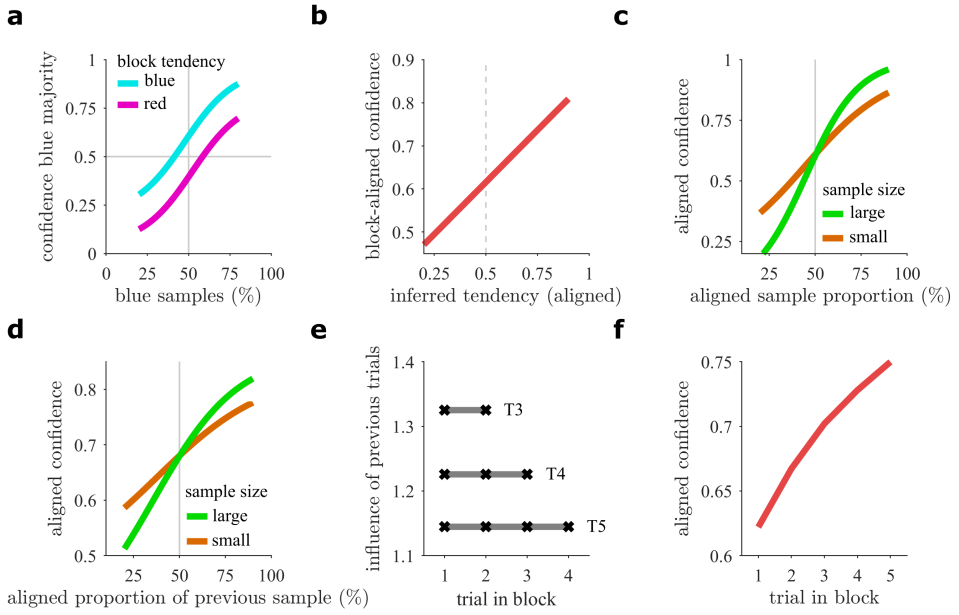

Figure 3.4: Behavioral patterns of probabilistic inference in the hierarchical inference task

(a) For a given sample proportion, confidence in a blue airplane (trial) majority should be larger in a block that favors blue majorities (cyan) than in a block favoring red majorities (magenta). **(b)** The belief in the presence of a certain block tendency should gradually increase the confidence in the corresponding trial majority. Thus, responses can be pooled with respect to the real block tendency. We refer to it as 'aligned confidence' and use the same concept for other relative quantities below. **(c)** Confidence in the aligned airplane majority increases with the aligned sample proportion. This modulation is stronger for larger sample sizes (green) compared to smaller ones (orange) while it has no effect for an indifferent sample (50 % sample proportion). **(d)** Likewise, the aligned confidence increases with the aligned sample proportion of the preceding trial and is modulated by its respective sample size. **(e)** The influence of all previous trials should be equal on average (e.g. trials $1-2$ on trial 3, T3). However, it decreases with the number of previous trials due to normalization. **(f)** Aligned confidence increases across trials within a block because of evidence accumulation regarding the block tendency. All patterns are derived from the optimal model (Methods 3.5.6).

3.3.6 Prior observations constrain future inferences

We found that participants, on average, manage to correctly learn the real block tendency (Fig. 3.5a, pattern: Fig. 3.4a). As a function of the proportion of blue

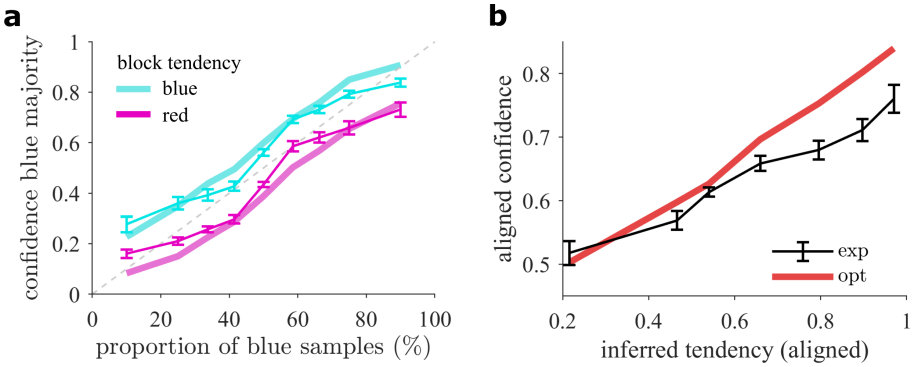

Figure 3.5: Learned belief about the block tendency affects confidence reports
(a) Confidence in blue majority is higher when the block tendency favors blue majorities than when it favors red majorities. Experimental results (data points) are shown along with optimal behavior (solid lines) indicating an integration of sample information with a learned belief about the block tendency. **(b)** Behavior (black) increases with the optimally inferred belief about the block tendency and is a close correlate of the optimal response (red). This suggests that participants internally track a graded belief based on previously available evidence. Error bars indicate SEM across participants.

samples, their responses show a higher confidence in blue majorities when we condition on a real blue context than vice versa (details of plotting in Methods 3.5.7). As a comparison we additionally plotted the response of the optimal model for the same data in all following figures.

The actual inferred context is subjective to the participants and only manifests itself through its influence on the responses. However, we can ask to what extent these observed influences correspond to those of the optimal model. Participants (Fig. 3.5b, black) were found to track a close correlate of the optimally inferred block tendency $M(b)$ (red) (linear correlation of binned values in Fig. 3.5b, pooled across participants, $\rho = 0.77$, $p = 5.12 \cdot 10^{-33}$). The relationship appears to be monotonic and close to linear which suggests a gradual integration of subjective prior information (pattern: Fig. 3.4b).

An obvious deviation from the optimal model is the reduced sensitivity to prior information apparent by the smaller range. Correspondingly, we found smaller linear slopes compared to the optimal model of individually fitted linear functions (one-sided signed rank test across participants, $p = 0.0043$). This discrepancy mainly stems from under-confidence for high evidence for the aligned block tendency. As in the basic task, the confidence judgments of our participants are found

to correspond closely to the actual decision correctness as derived from the optimal model (appendix C.1).

3.3.7 Uncertainty governs hierarchical information integration

In the hierarchical dependence structure of latent variables, disparate information about the context from different trials must be fused. Using the same analysis as for the basic task (Methods 3.5.7), we found that the size of the momentary sample also increases decision confidence (Fig. 3.6a). As before, we evidence an increase of the slope of the confidence response curves with sample size (linear correlation of slope with sample size, pooled across participants, $\rho = 0.49$, $p = 8.67 \cdot 10^{-14}$). This measure averages over fluctuating values of the prior belief and again demonstrates a good overall match with the optimal sample size dependence (linear correlation, pooled across participants, $\rho = 0.52$, $p = 1.22 \cdot 10^{-15}$). Crucially, this pattern cannot be reproduced by any heuristic estimate ignorant of sample size and indicates that behavior preserves this feature in the hierarchically extended task of higher complexity.

Beyond the finding above that participants learn the block tendency (Fig. 3.5a), they should use it selectively and rely more strongly on the sample compared to the prior when sample evidence is reliable (Fig. 3.6b, pattern: Fig. 3.4c). Indeed, the modulation with the aligned sample proportion is stronger for larger sample sizes and leads to the crossover of the two conditional curves (signed difference of conditional slopes from linear regression, signed rank test across participants, $p = 1.44 \cdot 10^{-5}$). This pattern is expected from a probabilistic agent that constantly adjusts the relative strengths of bottom-up and top-down influences to update hierarchical state representations.

As for the momentary sample, the influence of the previous trial depends on its reliability. Behavior is more strongly modulated if the previous sample size was large (Fig. 3.6c, pattern: Fig. 3.4d) (linear regression, signed difference of conditional slopes, sign rank test across participants, $p = 0.002$). This pattern is weak and superseded by noise, yet we can even determine a significantly larger slope for nine out of 24 individual participants (shuffling test of high/low sample size conditions, threshold $p = 0.05$). Interestingly, participants do not generally appear to discard evidence that contradicts the already established belief about the block tendency. On average across blocks, a presented percentage $< 50\ \%$ (Fig. 3.6c) contradicts the established belief. If a confirmatory bias were present, there should be no modulation with contradictory evidence because it is simply discarded. However, we found no general confirmation bias, as the slopes of fitted linear functions are significantly larger than zero in this range (signed rank test of slopes across participants, $p = 0.0051$).

A major challenge is to integrate evidence (messages) from different trials to determine the best estimate of the block tendency. Even several trials, each providing weak evidence, may jointly allow to draw strong conclusions about the block tendency. The optimal estimate of the block tendency is a complicated nonlinear function of the individual evidence distributions involving pointwise multiplication and subsequent renormalization (Methods 3.5.6). To assure that the behavioral pattern caused by the learned prior belief of the block tendency is unlikely to be produced by a comparably simple, heuristic integration mechanism, we attempted to reject three alternative prior accumulation schemes that differently estimate $M(b)$ (overview in appendix, Tab. C.1).

The averaging model (avg) computes an average of the presented percentages of previous trials in a block and thus neglects sample size (Methods 3.5.6). The tally model (tly) in contrast, tallies the total number of blue samples vs. the number of all points observed so far within a block (Methods 3.5.6). This is similar to pooling the samples of all trials, as if they were drawn from a common population. The tally model can be seen as an estimate its proportion through the (pooled) sample proportion. As larger samples contribute more points, this tally estimate is sensitive to sample size. In addition, and similar to Experiment 1, we test a difference model (diff) that relies on a running average of the differences between the number of blue and red samples in previous trials to compute the belief of the block tendency.

Even though all three heuristic approaches are close correlates of the optimal prior belief $M(b)$, all are determined to be insufficient as a model of behavior. On the group level, the optimal model is significantly more likely to predict the data of a randomly chosen participant compared to any other model (Fig. 3.6d). However, large confidence intervals suggest that few participants may be better described by rather heuristic approaches. For fitting, we attempted to impose the fewest constraints possible on the implementation of the integration of the empirical prior M with the sample $D = (q, N)$. Thus, to reduce the noise in the estimation process, we modeled this stage with a flexible function $(D, M) \rightarrow C$ onto the confidence report C which is adaptive to the idiosyncrasies of behavior (Methods 3.5.6).

Overall, we found strong support for sample size effects suggesting an important role for uncertainty to guide the information flow for inference of interdependent latent variables.

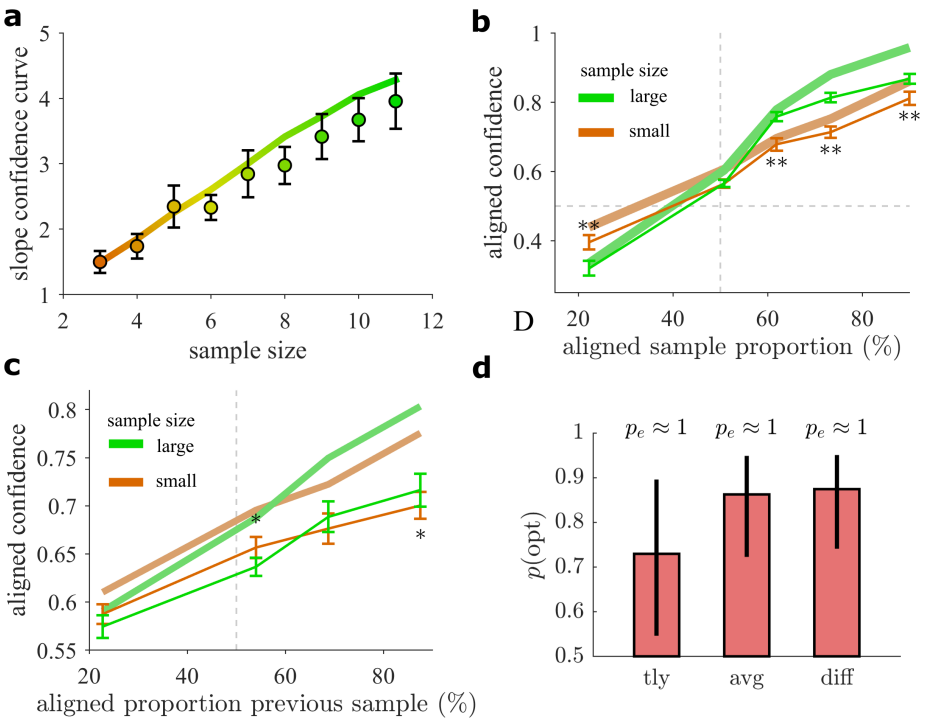

Figure 3.6: Sample size effects evidence reliability-based information integration
(a) Confidence reports increase with sample size (data points) and tightly follow the optimal pattern (solid line). As in Fig. 3.2b, the slope of the confidence curve is shown. **(b)** The modulation of aligned confidence with the aligned sample proportion of the current trial is larger when the sample size is high (green) than when it is low (orange). Significant signed differences of a bin-wise one-sided signed rank test are indicated, $* : 0.01 < p \leq 0.05$, $** : p \leq 0.01$. **(c)** The modulation of aligned confidence with the aligned sample proportion of the previous trial is larger when the sample size of the previous trial is high (green) than when it is low (orange), similar to the previous panel. Error bars indicate SEM across participants in (a-c). **(d)** Binomial probability of the optimal model to account for the data of a randomly chosen participant (error bars are 95 %-CI, see Methods 3.5.6). Pairwise comparisons to the models (tly, avg, diff) show that probabilistic information integration yields better predictions on the group level. Additionally, the exceedance probability p_e is used to quantify how much more likely the optimal model is.

3.3.8 Incremental prior learning is consistent with hierarchical evidence accumulation across trials

Behavior is remarkably consistent with hierarchical integration in which evidence between trials is mediated solely via the context-level variable. A central prediction of the probabilistic model is that all previous trials should have equal influence on behavior on average across blocks (pattern: Fig. 3.4e). We determined their influence from a regression analysis on the confidence report (see Methods 3.5.7) and found a rather balanced influence of all previous trials (Fig. 3.7a, black, pattern: Fig. 3.4e). Accordingly, no significant trend could be evidenced through

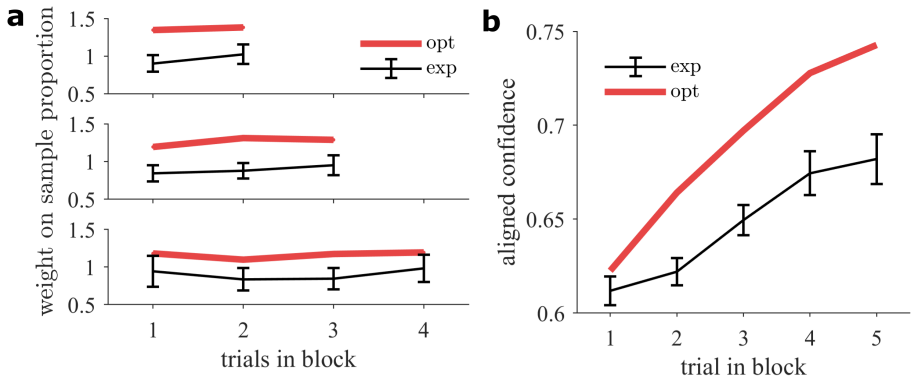

Figure 3.7: Behavior reflects hierarchical evidence integration across trials
(a) On average across blocks, all previous trials provide the same information about the block tendency irrespective of their distance to the current trial. From top to bottom, trials number 3-5 of each block are predicted from the indicated previous trials (sample proportion). Participants show a balanced weighting despite smaller weights compared to optimal inference in the hierarchical task (red). **(b)** Participants accumulate evidence about the block tendency in a gradual fashion. Aligned confidence increases over trials within a block despite a smaller effect compared to the optimal model (red). Error bars indicate SEM across participants.

another linear regression analysis of the trial index against aligned confidence (regression on means across participants, p-values $(0.47, 0.51, 0.87)$ for trials with $(2, 3, 4)$ previous trials respectively). We remark that consistent with the above findings (Fig. 3.6c), there is no general confirmation bias which is characterized by selective evidence integration once a belief has been established. If it were present, later trials should be disregarded more often which would result in a lower influence here. In addition, this rather balanced weighting is also inconsistent with some sort of leaky prior integration scheme in which evidence presented long ago is fading from memory. Hierarchical integration offers a better explanation instead as it does not require explicit memorization of previous samples after they

are integrated into the context-level variable.

Consistent with an accumulation of evidence for the block tendency, aligned confidence increases over the trials within a block (Fig. 3.7b, black, pattern: Fig. 3.4f). A linear regression analysis of the trial index against aligned confidence clearly shows the expected increase (regression on means across participants, $p = 3.43 \cdot 10^{-9}$). Further evidence for accumulation of evidence was provided above by the correlation with the optimally inferred belief about the block tendency (Fig. 3.5b). As this correlation is high, participants are expected to also accumulate evidence. Participants generally refrain from making extreme responses in the hierarchical task both for high sample evidence (e.g. Fig. 3.6b) as well as for high contextual evidence (Fig. 3.5b). This is expected to more severely affect later trials in a block which, on average, should allow for stronger trial-level inferences. Such deviations, as e.g. the markedly smaller increase in Fig. 3.7b, are more specifically addressed next.

3.3.9 Limitations of the probabilistic inference model to account for behavior

So far, we have found that behavior well matches characteristic variations that are expected from probabilistic inference (Figs. 3.5-3.7, patterns: Fig. 3.4). However, despite finding close correlations, the direct output of the task-optimal probabilistic model is not sufficient to provide a close fit to behavior because there are further perturbations. Compared to the optimal model, most patterns feature a substantial vertical offset and participants appear to rely less on evidence from previous trials (see e.g. Fig. 3.5b).

Participants are not expected to possess matching a priori knowledge of the block tendency to our modeling choice of a mixture of two skewed Beta distributions (Methods 3.5.5). One possible reason for such a deviation could be that participants assume that the asymmetry introduced by the block tendency is weaker than the generative one which is supposed to be known be the optimal model. To explore this, we fitted a model that can account for this fact by allowing for a differently skewed Beta-distribution implementing this block tendency. In addition, it accounts for some nonlinear distortions on the response. According to that, participants appear to subjectively assume a somewhat weaker block tendency as evidenced by the expectation value of the skewed Beta-distribution (optimal 0.61, quartiles across participants $(0.54, 0.56, 0.60)$, one-sided signed rank test, $p = 0.0011$). We achieved a better fit although systematic deviations from the fitted model remain for the probabilistic inference patterns (see appendix C.4). Not surprisingly, this suggests that participants likely rely on a different parameterization of the block tendency.

Next, we tested whether there is further evidence for partial ignorance of prior information which is not restricted to this modeling assumption. The interpretation of the deviations, such as the vertical offset in Fig. 3.6c, is not straightforward as every response is the result of tracking prior information about the block tendency $M(b)$ and the evidence provided by the momentary sample D. Consequently, deviations may result from a lower dependence on previous trials and/or from a lower dependence on sample evidence. Nevertheless, there is evidence that participants are less affected by prior evidence overall. Aligned confidence is less strongly modulated with the presented percentage of the previous trial compared to the optimal model (see Fig. 3.6c) (individual slopes from linear regression, one-sided signed rank test, $p = 0.0036$). Consistently, the estimated weights on the sample proportion of previous trials (Fig. 3.7a) are typically lower than the respective weights of the model (red) (pooling weights across participants and trial index, signed rank test, p-values for trials with $(2, 3, 4)$ previous trials $p = (1.50, 0.05, 0.90) \cdot 10^{-5}$). This is consistent with an accumulation of contextual evidence in Fig. 3.7b that is weaker than the optimal model).

However, participants might generally respond weaker, that is also to evidence from the sample. There is evidence that participants also depend somewhat less on the sample than the optimal model. This is quantified by the weights of the fitted sigmoidal function which are found to be slightly smaller than the corresponding weights of the model (Fig. 3.6a, one-sided signed rank test for smaller slope, $p = 4.62 \cdot 10^{-5}$, pooled across participants and sample sizes). We also tested whether the aligned confidence is less strongly modulated with the sample proportion (see Fig. 3.6b, linear regression, one-sided signed rank test for smaller slope, $p = 0.13$). The result is only suggestive (but non-significant) of such a trend and consistent with a tendency to restrict confidence reports in particular for large sample fractions (Fig. 3.6b, sample fractions over $60\ \%$).

In summary, participants depend weaker than optimally on previous trials and presumably, but not decisively, weaker on momentary evidence. Nevertheless, the variation with the momentary sample is of a larger magnitude (scales in Fig. 3.6b-c) so that a considerable part of the systematic deviations might arise from integration with momentary evidence and is not solely due to prior belief tracking.

3.3.10 Dominance of bottom-up influences

A selectively weaker dependence on previous trials might hint at a characteristic imbalance between bottom-up and top-down influences on hierarchical integration. Here, we report two relative measures of the influence of (1) the sample D over (2) a subjective prior belief $M(b)$. The first measure uses conditional variance while the second relies on the frequency that judgments oppose sample evidence.

If participants did barely depend on inferred prior information, all variation in their behavior would be exclusively determined by the sample and noise. We controlled for variations in the sufficient statistics of the sample and estimated residual variance which is attributed to (1) tracking prior information of the block tendency and (2) response noise (Methods 3.5.7). Compared to the total variance computed on the same responses, we estimated a ratio of about one third (median across participants 0.34 $(0.24, 0.53)$, 95 %-CI) for the residual variance fraction. While the corresponding fraction for the optimal model yields a similar value 0.36 $(0.39, 0.42)$, a direct comparison is somewhat biased by response noise which only enters in the estimate of the participants. If we however add realistic levels of response noise (appendix C.3) to the model, its prior related fraction is larger than for the participants (median 0.4786 $(0.45, 0.504)$, one-sided signed rank test $p = 0.038$). This suggests that participants respond less strongly to previous trials relative to the variations induced by the current trial.

Further insight is provided by trials in which an optimal agent would e.g. estimate a red majority despite more blue samples because of a high prior belief in a red tendency. We found that most participants likely make these evidence-opposing choices (see Methods 3.5.7, one-sided signed rank test with respect to non-hierarchical ratio model with realistic response noise, $p = 0.008$). There is however a tendency to stay on the side of the category boundary that is suggested by the momentary evidence, as they make significantly fewer opposing choices than the optimal model (one-sided signed rank test, $p = 0.008$). Remarkably, evidence-opposing choices are virtually absent in the basic task suggesting that there is almost no sensory-motor noise leading to misjudgments of the sample majority (appendix C.3).

Altogether, we found evidence that participants are less strongly driven by previous trials in comparison to the effect that the current trial has. On the group-level, bottom-up influences tend to dominate judgments even though more robust test are needed to corroborate these findings. Special attention should be payed to the conspicuously large variations across participants which may render simple group-level measures unrepresentative.

3.4 Discussion

This study presented a challenging hierarchical integration task which requires human participants to respect uncertainty about all jointly inferred variables in order to make truthful confidence judgments. Participants appear to impose contextual constraints on their inferences to the extent that contextual evidence is reliable. Correspondingly, behavior exhibits several nontrivial patterns of probabilistic pro-

cessing such as sample size effects. Moreover, the inference procedures involve complex and nonlinear operations such as normalization and marginalization [172] which could not be reduced to explanations with simple heuristics [155].

Beyond that, there is a close correspondence between the probabilistic, posterior-based conception of decision confidence [125–127] and the uncertainty estimates of our participants. Remarkably, they were completely free to report their subjective estimates of their decision correctness on a finite and quasi-continuous interval spanned by the two possible choices. No clues, nor supervising feedback was provided to guide their confidence estimates which means that they must possess surprisingly accurate internal trial-by-trial representations of uncertainty [123, 124]. Overall, their behavioral patterns closely match the ones of a probabilistic agent [24, 25] which possesses ubiquitous representations of uncertainty through distributions over all latent variables.

Our basic choice and confidence judgment task (Experiment 1) is comparable to several commonly performed studies [140], although not all systematically assessed sample size effects [161]. In one study [132], decision time is a proxy for sample size and was similarly reported to influence confidence judgments. Importantly, our task reaches beyond commonly employed basic visuo-motor tasks [138, 140, 156] and probes uncertainty representations for more abstract, higher-level concepts. Apart from the insight into probably domain-general mechanisms, our computational modeling approach largely sidesteps the idiosyncrasies of the perceptual stage.

Overly simplistic and un-naturalistic tasks have been criticized for limiting brain processing to a domain where its power is hidden [77]. The hierarchical inference task drastically increases complexity as joint inference of several latent variables at different levels of a hierarchy needs to be performed which capture the contextual embedding that is typical for real-world inferences [42]. Our participants could successfully adjust to perform inference in the hierarchical task without guiding feedback and repeated exposure. Such rapid domain adaptation [32, 59, 173] might be enabled by internal representations that harness the compositionality [68, 174] of our hierarchical problem which is a modular extension of the basic task.

Such powerful generalizations are hard to conceive without relying on an internal model of the observations. This is in line with previous studies (e.g. [175, 176]) which conclude that human inferences are model-based or use internal simulations [177]. Veridical judgments about the uncertainty of our inferences require a representation of the possible worlds [178] that are consistent with our data – even of those which are not most strongly supported. Correspondingly, the probabilistic approach to inference always faces the problem of model selection.

Estimating uncertainty about latent variables is a particularly difficult pro-

blem for heuristic bottom-up approaches which do not acknowledge the distributional (probabilistic) format that our estimates should take [25, 124], e.g. by committing to one interpretation. In our task for instance, learning calibrated confidence reports would require repeated exposure to the same sample together with supervising feedback about the actual latent variable (airplane majority). Even for very simple problems, the scarcity of such data makes this frequentist approach to uncertainty estimation practically difficult and thus un-ecological.

To select a model, participants must first infer the problem structure itself from the instructions. We constructed a task metaphor relating to airplanes to convey the mathematical assumptions in an intuitive manner. We believe that the task metaphor is exchangeable as long as it manages to communicate or trigger the underlying assumptions equally well.

An appealing feature of hierarchical Bayesian models is their ability to infer suitable constraints from data. Lower, task-level variables are constrained by higher-level contextual variables. The latter may be acquired empirically from related situations as joint inference is performed simultaneously at all levels. This results in a bidirectional information flow across levels [58] to select the best matching contextual constraints. Consistent with the use of such top-down constraints, a recent imaging study [179] has provided evidence for the activation of so called stimulus templates.

More generally, such a hierarchical scheme is believed to underlie visual processing [180]. The behavior of our participants fits well into this framework. Information integration can be well captured with reliability-dependent message passing between latent variables at different levels of a hierarchy. Such ubiquitous representations of uncertainty are a crucial ingredient for framework theories such as predictive coding [170, 171]. A virtue of explicit, probabilistic representations is that uncertainty estimates can naturally emerge from the knowledge representation itself, without requiring a meta-representation [181]. If reliability information is separable from other aspects of the estimate [165], and if it follows a hierarchical organization [47], may be elucidated with further (imaging) studies.

When comparing behavior against normative approaches, the interpretation of deviations should respect the internal constraints of the participant as much as possible [116, 117]. The goal of the participant is not necessarily veridical inference but the maximization of some subjective cost-benefit measure [111]. Furthermore, sometimes different assumptions about the problem may be internally justifiable but outwardly appear as irrational to solve the problem [168]. Even though our instructions were evidently successful, participants likely committed to slightly mismatching structural assumptions such as the parameterization of the block tendency. This may explain the typically high variability across participants [182] and the failure of some few participants to engage in the task.

If the structure of the problem is uncertain, one might intuitively rely more strongly on the sample. Structural uncertainty has been evidenced to affect performance [140, 175, 183] and may even lead to model-free behavior in severe cases [60]. We found evidence that top-down information is relied upon less strongly relative to information from the specific instances of the sample. However, its origin, or whether it is due to structural uncertainty, is not conclusive. Importantly, this bias is not dichotomous but rather graded which might result from approximate computations. For instance, a sampling framework [81] may produce biases such as a base rate neglect [82]. Behavior features other systematic perturbations, such as a systematic probability distortion [154], whose origin is unclear and beyond the scope of this study. However, not all deviations must be due to inference as e.g. the movement-related control problem may obscure the actual estimates of the participant.

Strikingly, many commonly reported biases [109, 146] are not generally observed in this study, e.g. sample size insensitivity [147, 149]. The inferences in our tasks are mathematically almost identical (e.g. [147]) or of even higher difficulty in the hierarchical task. Hence, the brain may in principle carry out these computations. We even observed that sample size was taken into account on a quantitatively accurate level. Furthermore, we did not find a general confirmatory bias or base rate insensitivity even though participants appear to gravitate towards such behaviors.

A recent proposal suggested that the ability to interrogate rational inferential systems through developmentally recent systems which involve natural language and working memory is limited [133]. Consequently, to minimize their involvement, the task was designed to be intuitive. We attempted to make its structure maximally transparent by incremental familiarization and by repetitive exposure. Such structural alignment under precisely controlled conditions is probably crucial to interpret and possibly account for many behavioral biases [182].

Ultimately, what experimental conditions impede or enable more rational and/or veridical inferences must be addressed by specifically designed studies which e.g. control for task instructions or cognitive effort [106]. Such studies may also investigate the relationship between the ability to make rational inferences and the degree of conscious awareness underlying the common understanding of reasoning.

After all, our results suggest that uncertainty is an integral part of our inferences and that we indeed have the potential to resort to rational inference mechanisms that adhere to probabilistic principles. The extent and why we fail to use them may crucially depend on context and how the problem is communicated.

3.5 Methods

3.5.1 Participants

All participants were required to complete three sessions on separate appointments on different days within three consecutive weeks. The sessions were targeted to take about 35 minutes (Session 1) and 45 minutes (Sessions 2,3). In total 25 participants (15 female, 10 male) were recruited mainly among students from the Pompeu Fabra University in Barcelona. The median age was 25 (minimum 20, maximum 43). We accepted all healthy adults with normal or corrected to normal vision. We obtained written confirmation of informed consent to the conditions and the payment modalities of the task. Irrespective of their performance, they were paid 5 € for session 1 and 7 € for sessions 2 and 3.

Additionally, they had the chance to obtain a bonus payment which was determined by the mean of their final score after removing the worst trials ($2.3\ \%$). The score $S = 1 - |y - y_{opt}|$ of a response y was computed with respect to the optimal response y_{opt}. The payment was determined by comparison to an array of five thresholds that were set according to the $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ cumulative quantiles of the empirical score distribution across prior participants. A higher score S corresponds to a better performance so that participants were payed an additional bonus of $\{1, 2, 3, 4, 5\}$ € if their final score was higher or equal to the quantile thresholds. This is a relative way of rewarding their efforts to optimize their responses.

Written task instructions explained that we would score their responses with respect to the chances that their decision turns out to be correct and that bonus payments would be based on that score. Additionally, they were informed that their score was to be compared to the other participants and that the experimenter could monitor their behavior on-line via a second screen from outside.

3.5.2 Stimuli & Responses

The task was presented with Matlab Psychtoolbox 3.0.12. Immediately after trial onset, our participants were shown the sample consisting of red and blue solid circles arranged on a two-dimensional grid about the screen center. The goal was to make the sufficient statistics easily perceptible while making the display appear otherwise completely random. Adequate grid spacing was introduced to prevent the circles from overlapping. Furthermore, red and blue samples never appear intermingled (details in appendix C.2).

The display is static until the participant makes a response by clicking the USB-mouse which clears the display of the sample. After a short delay of 300 ms, the program shows a centered horizontally elongated response bar of random hori-

zontal extent with a vertical line marking its center. In addition, the response cursor (Fig. 3.1a, short vertical line) is shown at a random initial horizontal position along the response bar. Participants can precisely adjust the horizontal position of the response cursor by moving the mouse horizontally and confirm the input with a click. The movement range of the response cursor was bounded to the horizontal extent of the response bar. Their raw response is linearly mapped onto an interval between $[0, 1]$ and interpreted as the confidence in a blue trial majority y. The corresponding quantity for the confidence in a red majority is $1 - y$.

The program then either proceeded to the next trial or to a feedback and/or pause screen. Participants may receive a short time-out which is signaled by a horizontal 'progress' bar which linearly diminishes over time indicating the fraction of the waiting time left. During time-out, there is nothing a participant can do to proceed but wait. Apart from that, the participants are free to proceed at their own pace without restrictions.

Every five trials, a pause screen is shown which provides information about how many out of all trials have already been completed. To motivate engagement in the task, we gave motivational feedback as an average over the trials since the last pause (blocks for hierarchical task) of the score $\langle S \rangle$. Additionally, they also received a time-out of some few seconds proportional to $1 - \langle S \rangle$.

3.5.3 Experiment 1: Procedure & Instructions

First, participants read detailed written instructions of the task. We introduced the task metaphor that relates to judging the (hidden) majority of passengers on a flight and used it to explain the mathematical assumptions in more intuitive terms (see appendix C.2).

Additionally, our participants were given 30 trials to familiarize with the handling of the task through a short interactive session. The subsequent experimental session (session 1) consisted of 280 trials with pauses together with feedback after every 5 trials. The sample sizes where independent and identically distributed (i.i.d.) samples from $\{3, 5, 7, ..., 13\}$ while the hidden airplanes proportions were i.i.d. samples from a Beta$(4, 4)$-distribution. After confirming the response, participants received extra feedback about the correctness of each decision. Partly, this was done to emphasize the dissociation between sample and population majority. In addition, a two second time-out was presented for false decisions.

3.5.4 Experiment 2: Procedure & Instructions

Experiment 2 comprises the sessions 2 and 3 and was carried out with the same 25 participants as in Experiment 1 (session 1). Later, we excluded two of them because one did not complete the experiment and one showed too little compliance with the hierarchical task (appendix C.4). Despite the hierarchical extension across blocks of five trials, the handling of the task and the presentation of the sample is virtually the same. The changes to the latent structure should lead to a different interpretation of the information which we attempted to convey by an extension of the task metaphor (appendix C.2).

As for Experiment 1 and prior to starting session 2, participants completed two very short training sessions. First, they were given 20 trials (4 blocks) with a strong block tendency (sample sizes $\{8, ..., 11\}$, block asymmetry $\text{Beta}(15, 7)$). Then another 30 trials under slightly harder conditions (sample sizes $\{3, ..., 11\}$, block asymmetry $\text{Beta}(15, 7)$). Importantly, this only permits them to understand the structure of the reasoning task. However, they cannot deduce how they have to make their judgments because we do not give informative, supervising feedback to learn from.

Afterwards, our participants completed 270 trials of the experimental session 2 with an even more difficult setting of the parameters (sample sizes $\{3, ..., 11\}$, block asymmetry $\text{Beta}(14, 9)$). On the third session, on a different appointment, the participants just continued the instructed task of session 2 for 300 trials with identical settings to obtain more data.

3.5.5 Generative model for the stimuli of the prior learning task

First and once for every block, the binary variable b governing the prevalence for either red or blue trial majorities is drawn from a Bernoulli distribution $b \sim \text{Bern}(0.5)$ in which b stands for a blue block tendency. For simplicity, we use the same variable names for the generative process (Fig. 3.3a) as for the optimal agent (Fig. 3.3c), although in general, an agent's representation is not necessarily the same as the generative process in the environment. For every trial, the latent airplane proportion μ is drawn from one of two Beta distributions depending on b. More formally, this can be written as a mixture distribution:

$$p(\mu|\nu_1, \nu_2, b) = b \cdot \text{Beta}(\mu|\nu_1, \nu_2) + (1 - b) \cdot \text{Beta}(\mu|\nu_2, \nu_1) \tag{3.1}$$

The Beta distribution is parameterized by two parameters ($\nu_1 = 14, \nu_2 = 9$). They are chosen such that the resulting distribution over the trial majority μ is skewed. By convention, $\text{Beta}(\mu_t|\nu_1, \nu_2)$ is positively skewed ($\nu_1 \geq \nu_2$) and models a blue block tendency. The greater the expectation $\nu_1/(\nu_1 + \nu_2) \approx 0.609$ the more ex-

treme this effect.

Sampling then proceeds as for Experiment 1. First, an i.i.d. sample is drawn from a uniform categorical distribution $\mathrm{Cat}(N|1/n, ..., 1/n)$ over all n sample sizes $N \in \{3, \ldots, 11\}$. Then, the sufficient statistics of the sample are determined by a draw from a Binomial distribution $N_B \sim \mathrm{Bin}(N, \mu)$. Hence, the sampling distribution for one trial is:

$$p(N_B, N, \mu|\nu_1, \nu_2, b) \propto \mathrm{Bin}(N_B|N, \mu) \cdot \mathrm{Cat}\left(N|1/n, ..., 1/n\right) \cdot p(\mu|\nu_1, \nu_2, b) \quad (3.2)$$

The geometric placement on the screen is not considered to be part of the generative model as we make the assumption that only the sufficient statistics matter. The expression in Eq. 3.2 defines the probability distribution for the sufficient statistics of the observations of trial t to which we refer more concisely as $p(q_t, N_t, \mu_t|b, \nu_1, \nu_2)$, thus equivalently expressing it in terms of the sample proportion $q = N_B/(N_B + N_R)$ and the sample size $N = N_B + N_R$. We drop the conditioning on the parameters of the categorical distribution over sample sizes to keep the notation uncluttered. Using this expression, the entire sampling distribution over all variables of all trials within a block is:

$$p(q_1, ..., q_5, N_1, ..., N_5, \mu_1, ..., \mu_5, b|\nu_1, \nu_2) = p(b) \prod_{t=1}^{5} p(q_t, N_t, \mu_t|b, \nu_1, \nu_2) \quad (3.3)$$

Given the block tendency b, the per-trial quantities, such as μ_t, are independent. The parameters, e.g. (ν_1, ν_2), do not result from sampling but serve to define other distributions.

3.5.6 Computational models

Inference using the probabilistic generative model of the basic task

Due to the choice of a conjugate distribution $p(\mu)$ for the Binomial probabilistic model $N_B \sim \mathrm{Bin}(N, \mu)$ above, posterior inference yields a Beta-distribution over the latent airplane proportion μ.

$$\mathrm{Beta}(\mu|N_B + r_B, N_R + r_R) \propto \mathrm{Bin}(N_B|N, \mu) \cdot \mathrm{Beta}(\mu|r_B, r_R) \quad (3.4)$$

Specifically, to give calibrated responses, the prior distribution used for inference must correspond to the actual base rates specified by $\mathrm{Beta}(\mu|r_B = 4, r_R = 4)$. The confidence in e.g. a blue trial majority $c(B)$ is expressed as the belief that choosing a blue majority is correct by integrating over the corresponding subspace

of inferred blue majorities.

$$c(B) = 1 - c(R) = p(\mu > 0.5 | N_B, N_R) = \int_{0.5}^{1} \mathrm{Beta}(\mu | N_B + r_B, N_R + r_R) \quad (3.5)$$

Inference using the probabilistic generative model of the hierarchical task

The optimal inference model inverts the generative structure of the task. It maintains a probability distribution over the observations of all in-block trials and their respective latent variables $(\mu_1, ..., \mu_T)$ up to the current trial T. The parameters (ν_1, ν_2) are part of the generative structure and assumed to be known. Consequently, inference amounts to an updating of the distribution over the latent variables through a calculation of the posterior distribution conditional on the observations. We identify distributions by their respective arguments and e.g. write $p(D|\mu)$ for the distribution over the sufficient statistics of the sample. We use the abbreviation $D = (q, N)$ for the observations, omit parameters and index according to in-block trials t.

$$p(\mu_1, ..., \mu_T, b | D_1, ..., D_T) \propto p(b) \prod_{t=1}^{T} p(D_t | \mu_t) p(\mu_t | b) \quad (3.6)$$

The current trial is labeled T and we would like to compute the probability of a blue latent trial majority, namely that μ_T is larger than 0.5. For this purpose, all nuisance variables that are not of interest (previous trials) must be integrated out.

$$\begin{aligned} p(\mu_T \geq 0.5 | D_1, ..., D_T) &= \frac{1}{\psi} \sum_{b=\{0,1\}} \int_{0.5}^{1} p(D_T | \mu_T) p(\mu_T | b) \, \mathrm{d}\mu_T \\ &\cdot \quad p(b) \prod_{t=1}^{T-1} \int_{0}^{1} p(D_t | \mu_t) p(\mu_t | b) \, \mathrm{d}\mu_t \end{aligned} \quad (3.7)$$

Because of conditional independence given the block tendency b, the high-dimensional distribution factorizes so that only one-dimensional integrals over the latent variables of previous trials must be performed. Examining the graph structure, we see that they may be considered messages $m_t(b)$ which are passed upwards to update the block-level variable b.

$$m_t(b) = \frac{1}{\psi_{m_t}} \int_{0}^{1} p(D_t | \mu_t) p(\mu_t | b) \, \mathrm{d}\mu_t \quad (3.8)$$

For proper normalization ψ_{m_t}, they are themselves probability distributions which convey bottom-up evidence for the block tendency variable $b = \{0, 1\}$ based on the observations D_t.

These bottom-up messages from different trials are integrated to update the belief $M_T(b)$ about the block tendency b prior to trial t through point-wise multiplication and proper renormalization ψ_M.

$$M_T(b) = \frac{1}{\psi_M} p(b) \prod_{t=1}^{T-1} m_t(b) \tag{3.9}$$

As more evidence is gathered (trials), more factors can be absorbed into the belief about b without having to memorize data from all previous trials as it is efficiently encoded in $M_T(b)$. Subsequently, this knowledge serves as top-down constraint on future inferences on the trial level. Consequently, to derive the probability of a blue trial majority on the next trial, the integration of momentary evidence (Eq. 3.7) can be expressed as

$$c_T(B) = \frac{1}{\psi} \sum_{b=\{0,1\}} M_T(b) \int_{0.5}^{1} p(D_T|\mu_T) p(\mu_T|b) \, \mathrm{d}\mu_T \tag{3.10}$$

Proper normalization can be obtained analytically (appendix C.4).

Hierarchical heuristic average percentage model (avg)

To derive the belief in a blue block tendency, this model computes the average of the presented fractions of blue samples $q = N_B/(N_B + N_R)$ in the trials t prior to the current trial T.

$$M_T^q(b = 1) = \frac{1}{T-1} \sum_{t=1}^{T-1} q_t \tag{3.11}$$

It neglects sample size and corresponds to the implicit assumption that each trial's population is well captured by a point estimate, i.e. by its respective sample proportion. To integrate information from each trial, equal weight is given to each trial ignoring the fact the some trials provide more information than others due to different sample sizes. As for the other models below, indifference is assumed on the first trial $M_{T=1}^q(b = 1) = 0.5$.

Hierarchical heuristic tally model (tly)

Similarly, this model computes a tally of all blue samples observed prior to the current trial T versus the number of all samples observed in a block so far.

$$M_T^{\pm}(b=1) = \frac{\sum_{t=1}^{T-1} N_{B_t}}{\sum_{t=1}^{T-1} N_{B_t} + N_{R_t}} \tag{3.12}$$

This corresponds to pooling the samples of all trials, as if they were drawn from a common population of unknown population proportion whose ML estimator is $M_T^{\pm}$.

Hierarchical heuristic difference model (diff)

The heuristic difference model considers the difference between the number of blue and red samples $d_t = N_{B_t} - N_{R_t}$ as informative to establish a belief about the block tendency. Across trials, it is accumulated by computing:

$$M_T^{d}(b=1) = \frac{1}{1 + \exp\left[-\omega \cdot \sum_{t=1}^{T-1} d_t/(T-1)\right]} \tag{3.13}$$

The logistic sigmoidal function ensures that the result always takes a value between zero and one and that it can be interpreted as a proper belief. The parameter ω adjusts the sensitivity to the sample-difference statistics d_t and can be determined by a fit to behavioral data.

Response mapping captures distorted reports of internal confidence estimates

Behavior is influenced by additional factors and subjective assumptions of the participant which are difficult to model explicitly. Instead, we implicitly model those which can be captured by a nonlinear transformation of the confidence estimate through the effects they exert on the response. By allowing for additional freedom through a mapping, we can capture that participants may not report their internal estimate in an unperturbed way, e.g. due to motor control constraints, without having to model its origin.

All models compute a confidence estimate $c \in [0, 1]$ and may be supplemented with this mapping. First, we standardize the output $c' = 2(c - 0.5)$ which then enters the argument of a logistic sigmoid function through the polynomial $Z = \omega_0 + \omega_1 c' + \omega_2 c'^3$.

$$\hat{y} = \frac{1}{1 + \exp(-Z)} \tag{3.14}$$

As we assume symmetry, only odd powers of c' are used. In other words, the

perturbed confidence estimate $\hat{y}$ should lead to the same decision confidence irrespective of the sign (red or blue) of the unperturbed (standardized) confidence estimate c'.

This function is flexible and able to approximate a wide range of distorted reports including the identity mapping and various forms of probability distortion. It only accounts jointly for all effects which affect the final judgment. Other systematic deviations during confidence estimation which are conditional on a subset of the input space can only be partially accounted for, e.g. deviations for extreme values of the sample proportion.

Flexible sigmoidal mapping (zmap)

This is a flexible extension of the response mapping described before. It is used to construct an approximation of low estimation bias to the sample integration stage in the hierarchical task. More concretely, we must integrate any given prior belief M, not necessarily derived from a probabilistic model, with the momentary sample $D = (q, N)$ and map it onto the final response $(q, N, M) \to \hat{y}$. As a mere function approximator, it is agnostic to the mechanisms that participants may use to combine information. Correspondingly, its parameters ω must be determined by a fit to the experimental data. Here, this process is approximated by a polynomial function Z of the input (q, N, M) that is fed into a logistic sigmoid as in Eq. 3.14.

$$\begin{aligned} Z &= \omega_1 + \omega_2 q' + \omega_3 q' N + \omega_4 M + \omega_5 q'^3 + \omega_6 q'^3 N \\ &+ \omega_7 N M' + \omega_8 M'^3 + \omega_9 N M'^3 \end{aligned} \tag{3.15}$$

The argument Z contains only odd powers of q and M because we assume symmetry and no preference for estimating either red/blue majorities. Correspondingly, both quantities are standardized beforehand by the function $f(x) = 2(x - 0.5)$. As they are also independent from one another, no corresponding product terms are included.

Preliminary testing revealed that the inclusion of nonlinear terms is important to capture finer-grained patterns of behavior. The sample size N is introduced into some terms to model its magnifying effect on the signed quantities (q, M). We performed a weight normalization by the SD of each polynomial (for the input data) which was absorbed into the indicated weights ω. The particular choice of the terms in Eq. 3.15 balances flexibility with model complexity (and optimization for finite behavioral data). We manually tested different parameterizations but did not find crucial differences for reasonable choices of the mapping. However, we remark that behavior certainly features more subtle variations that cannot be captured well but only approximated by this functional choice.

The response distribution

The probability of obtaining the behavioral response y_t on trial t conditional on the data d_t and the model parameters is assumed to be a Gaussian distribution truncated to the interval from zero to one $N_{[0,1]}(y_t|\hat{y}_t, \theta)$. The mean parameter of the normal distribution is set to the model prediction $\hat{y}_t$. The latter is denoted by $\hat{y}$ to distinguish it from the response y of the participant which is formally represented by a draw from the response distribution to account for task-intrinsic behavioral variability beyond the variations captured by the model. The standard deviation (SD) parameter θ of the Gaussian is assumed to be constant and robustly estimated from the data (appendix C.2).

As our data might be contaminated by other processes such as lapses, we take precaution against far outlying responses. The response likelihood is calculated for all R responses $\boldsymbol{y}$ as:

$$p(\boldsymbol{y}|D_1, ..., D_R) = \prod_{t=1}^{R} (1 - \epsilon) N_{[0,1]}(y_t|\hat{y}_t, \theta) + \epsilon \tag{3.16}$$

Additionally, to prevent isolated points from being assigned virtually zero probability, we generally add a small probability of $\epsilon = 1.34 \times 10^{-4}$ to all. This corresponds to the probability of a point at four standard deviations from the standard normal distribution. For non-outlying points this alteration is considered negligible.

Estimating model evidence

The evidence that each participant's data lends to each model is derived from predictive performance in terms of the cross-validation log likelihood (CVLL). For training, we maximized the logarithm of the response likelihood (Eq. 3.16). To maximize the chances of finding the global maximum even for non-convex problems or shallow gradients, every training run first uses a genetic algorithm and then refines its estimate with gradient based search (MATLAB ga, fmincon). The CVLL for each participant and model is summarized by the median of the logarithm of the response likelihood (Eq. 3.16) on the test set across all cross validation (CV) folds (appendix C.2).

Differences in model evidence, Δ, are reported on a log-scale in decibans (also decihartleys, abbreviated dHart) that may be used to interpret the significance of the results of individual participants. According to standard conventions, we consider a value of $5 > \Delta$ barely worth mentioning, $10 > \Delta \geq 5$ substantial, $15 > \Delta \geq 10$ strong, $20 > \Delta \geq 15$ very strong and $\Delta \geq 20$ decisive.

Group level model comparison

Instead of making the assumption that all participants can be described by the same model, we use a hierarchical Bayesian model selection method (BMS) [184] that assigns probabilities to the models themselves. This way, we assume that participants may be described by different models. That is a more suitable approach for group heterogeneity and outliers which are certainly present in the data. The algorithm operates on the CVLL for each participant ($p = 1, ..., P$) and each model ($m = 1, ..., M$) under consideration and estimates a Dirichlet distribution $\mathrm{Dir}(r|\alpha_1, ..., \alpha_M)$ that acts as a prior for the multinomial model switches u_{pm}. The latter are represented individually for each subject by a draw from a multinomial distribution $u_{pm} \sim \mathrm{Mult}(1, \boldsymbol{r})$ whose parameters are $r_m = \alpha_m/(\alpha_1 + ... + \alpha_M)$. We use the CVLL and assume an uninformative Dirichlet prior $\boldsymbol{\alpha}_0 = \mathbf{1}$ on the model probabilities. Later, for model comparison, exceedance probabilities, $p_e = \int_{0.5}^{1} \mathrm{Beta}(\alpha_i, \sum_{j \neq i} \alpha_j)$, are calculated corresponding to the belief that a given model i is more likely to describe the data than all other models under consideration. High exceedance probabilities indicate large differences on the group level. We consider values of $p_e \geq 0.95$ significant (marked with $*$) and values of $p_e \geq 0.99$ very significant (marked with $**$). A comparison of the parameters of the models reported in the main text can be found in the appendix C.2.

3.5.7 Other analyses

Regression for sample size dependence

Separate regression analyses conditional on sample size N are used to determine the slope of the psychometric curves of the confidence judgments in a blue trial majority over the proportion of blue samples q (Figs. 3.2, 3.1 and 3.6). For a given sample size N, we use a logistic sigmoid with a weight ω_N to relate the standardized sample proportion $q'_N = 2(q_N - 0.5)$ to the modeled response $\hat{y}$.

$$\hat{y} = \frac{1}{1 + \exp(-\omega_N \cdot q'_N)} \tag{3.17}$$

We note that with this parameterization symmetric and unbiased judgments are assumed. Conditioning reduces the number of data points available for fitting. To avoid numerical singularities due to finite data (sigmoid collapses to step function), we use the likelihood function (Eq. 3.16) but with the truncated Gaussian replaced by a Gaussian. This effectively leads to weighted regression assigning less probability density to responses close to the extremes (e.g. a response of 1 is assigned $1/2$ of the density due to spill-over of the Gaussian into $[1, \infty)$). In this heuristic scheme, outlying responses are given less importance which translates into higher

stability of the weight estimate.

Regression for previous trial weights

To estimate the weight on the sample proportion of previously presented in-block trials on the current confidence estimate, we perform a regression analysis (see Figs. 3.4e and 3.7a). Probabilistic integration of evidence for the block tendency M (Eq. 3.9) results in a nonlinear increase of aligned confidence with the number of previously observed trials which saturates due to normalization. Hence, as the relative contribution of each trial decreases as more trials are observed, we perform the regression analysis separately for different numbers of predictors $(2, ..., T-1)$ (previous trials).

$$\hat{y} = \frac{1}{1 + \exp\left[-\sum_{t=1}^{T-1} \omega_t \cdot q'_t\right]} \tag{3.18}$$

As before, we use a logistic sigmoid with a weight ω_t to relate the standardized sample proportion $q'_t = 2(q_t - 0.5)$ of each previous trial t to the modeled response $\hat{y}$. Again, this conditioning reduces the number of data points available for fitting ($570/5 = 114$ trials) from which up to four weights have to be determined. To avoid numerical singularities due to finite data, we use the likelihood function (Eq. 3.16) but with the truncated Gaussian replaced by a Gaussian (see above).

Residual variance when conditioning on the sample

If we control for the sufficient statistics of the sample (q, N), the residual variation may be attributed to variations due to the prior belief and non-input related response noise. We searched for all trials with the same sufficient statistics of the momentary sample. If there were ten or more trials for a particular sufficient statistic, we computed their squared deviation from the mean. Subsequently, we pooled all squared deviations calculated this way and took the mean individually for each participant. This was used as an estimate of the residual variance conditioned on the sample. As a reference, the total variance on the same trials was computed. The residual variance was then expressed as the ratio of the sample-conditioned variance with respect to the total variance. We added realistic levels of response noise ($\mathrm{SD} = 0.1$, appendix C.3) to the optimal model for a less biased comparison of the prior related variance fraction with experimental data.

Evidence opposing choices

The sample proportion is converted to a frame of reference in which it corresponds to evidence for the real (latent) block tendency, called aligned sample proportion

and denoted by $\tilde{q}$. In this frame, evidence from the sample opposes the tracked prior belief (on average) when $\tilde{q} < 0.5$. If we record a response that reports the other category $\tilde{y} > 0.5$ for such a trial, we call this an evidence opposing choice (confidence judgment). This can be attributed to an influence of an opposing prior belief or task-intrinsic response noise (independent of input). To avoid biased estimates because of the latter, the analysis is conditional on trials $\tilde{q} < 0.5$ that provide opposing evidence on average.

Crucially, in Experiment 1, we found that noise basically does not lead to evidence opposing choices (appendix C.3). Nevertheless, we make a conservative estimate by comparison to a model whose evidence opposing choices just result from noisy responses in the absence of any prior belief tracking. This reference model $\hat{y} = \tilde{q} + \epsilon$ just reports the aligned sample proportion $\tilde{q}$ plus independent noise ϵ drawn from a truncated Gaussian distribution of standard deviation $\mathrm{SD} = 0.1$.

Binning for visualization and analyses

To impose minimal constraints on data for visualization (see Figs. 3.5-3.7), we plotted the responses by grouping them into approximately equally filled bins across participants. The number of bins was manually chosen to achieve an appropriate trade-off between resolution and noise of the estimated bins values. Importantly, this only affects visualization. Unless stated otherwise, the underlying un-grouped data is used for testing. The conditional curves in Fig. 3.6b and c were determined by the cumulative quantiles Q of the sample size distribution (many $\geq Q(0.6)$, few $< Q(0.4)$) and (many $> Q(0.5)$, few $\leq Q(0.5)$) respectively.

Chapter 4

Inductive biases for inference

4.1 Abstract

While previous studies have shown that human behavior adjusts in response to uncertainty, it is still not well understood how uncertainty is estimated and represented. As probability distributions are high dimensional objects, only constrained families of distributions with a low number of parameters can be specified from finite data. However, it is unknown what the structural assumptions are that the brain uses to estimate them. We introduce a novel paradigm that requires human participants of either sex to explicitly estimate the dispersion of a distribution over future observations. Judgments are based on a very small sample from of a centered, normally distributed random variable that was suggested by the framing of the task. This probability density estimation task could optimally be solved by inferring the dispersion parameter of a normal distribution. We find that although behavior closely tracks uncertainty on a trial-by-trial basis and resists an explanation with simple heuristics, it is hardly consistent with parametric inference of a normal distribution. Despite the transparency of the simple generating process, participants estimate a distribution biased towards the observed instances while still strongly generalizing beyond the sample. The inferred internal distributions can be well approximated by a nonparametric mixture of spatially extended basis distributions. Thus, our results suggest that fluctuations have an excessive effect on human uncertainty judgments because of representations that can adapt overly flexibly to the sample. This might be of greater utility in more general conditions in structurally uncertain environments.

The content of this study is prepared for publication: "Instance-based generalization for human judgments about uncertainty", P. Schustek and R. Moreno-Bote, under review.

4.2 Author Summary

Are three heavy tropical storms this year compelling evidence for climate change? A suspicious clustering of events may reflect a real change of the environment or might be due to random fluctuations because our world is uncertain. To generalize well we should build a probability distribution over our observations defined in terms of latent causes. If data is scarce we are forced to make strong assumptions about the shape of the distribution ideally incorporating our prior knowledge. In our task, human behavior is consistent with probabilistic inference but reveals a tendency to generalize based on observed instances enhancing the effect of random patterns on behavioral judgments. This decreased context-sensitivity corresponds to a dominance of bottom-up sensory information. Maintaining a balance with expectation-driven top-down information is crucial for proper generalization. Our work provides evidence for the necessity to include graded instance-based generalization into the mathematical formulation of cognitive models. The investigation of the determinants and neural substrates of this inferential bias is expected to give insights into the richness but also fallibility of human inferences.

4.3 Introduction

Determining from limited data when observations reflect a consistently appearing pattern or when they are merely the result of randomness is important to faithfully represent the environment (see e.g. [185]). Suppose you want to assess the skill of a dart player to thrown darts at the bullseye (center) of the board. For a single bad throw, it is hard to discern whether it was due to bad luck or to the general inability of the player. For several throws however, the dispersion of the darts around the center should more closely reflect the skill of the player.

To represent uncertainty of our knowledge in this and more general situations, normative considerations suggest that an agent should explicitly represent knowledge as probability distributions instead of as point estimates [25, 124]. Several studies have shown that under certain conditions humans behave as if the uncertainty about a task-relevant variable was available to them as a distribution over its possible values [137, 186].

For instance, judging the skill of the dart player corresponds to estimating the spread of the distribution around the observed values. This requires constraining structural assumptions about the "shape" of the underlying probability distribution (e.g. a parameterized function such as a Laplacian or Gaussian) and it is generally unknown what assumptions are used by humans when dealing with uncertainty. Ideally, previous knowledge about the data generation process such as an expectation for the darts to cluster around the center corresponding to the goal (of

the example) is incorporated. As opposed to visuo-motor uncertainty [66], there is little evidence for the shape of inferred trial-by-trial perceptual representations in the small sample limit. In several previous studies such as cue combination [138], distributional estimates are taken to be normally distributed. While this may be justifiable under certain conditions, we challenge the general validity of this assumption.

Here we asked what kind of internal structural assumptions humans employ to generalize from sparse observations. Human participants are asked to quantify uncertainty about future events by estimating the dispersion of a normally distributed random variable. Although the instructions and the framing of the task suggested a simple, centered, unimodal, bell-shaped distribution, human behavior was not consistent with structural assumptions based on a close to normal probability distribution. Instead, human behavior was better explained by instance-based generalization whereby observed samples were used to build an internal representation of the underlying probability distribution, not necessarily unimodal or symmetric. The resulting internal representation is a mixture of several components and hence less sparse than necessary. Participants demonstrated faithful trial-by-trial estimates of uncertainty while the opportunity to learn a suitable response mapping from feedback was suppressed [124]. All alternative heuristic explanations proved insufficient to explain complex and consistently made estimates. Hence, our results support the notion that approximate probabilistic processing underlies behavior.

4.4 Results

We asked human participants to estimate the dispersion of future events from a small sample by indicating a range in which they predicted 65% of all future events to fall. The task instructions alluded to judging the ability of a dart player to hit the target based only on the outcome of previous attempts (Fig. 4.1).

Ideally, this task could be accomplished by inferring the dispersion of the generative distribution which in accordance to the task was chosen to be Gaussian. More specifically, participants were asked to judge the unknown accuracy of a "dart player" to hit the center of the board (Fig. 4.1A). On a given trial, of a total of 320 trials, the participants are shown four points representing the "darts" thrown by one unobserved player of unknown accuracy to hit the center of the board. Based on the four observed "darts", participants must predict where future darts from the same player might strike the board. Specifically, participants were asked to capture $65\ \%$ of all future imaginary darts by adjusting the width of the rectangular frame of size $2y$ symmetrically about the center (y is the ho-

rizontal, one-sided distance of the lateral borders of the rectangle to the center). Only the horizontal dispersion of the dots is relevant to estimate the accuracy of the dart player, while vertical displacements are added just to improve visibility of the samples. The choice of $65\ \%$ is convenient as it does not depend on an accurate estimate of the distribution's tail and conveniently allows to examine a limiting case of instanced-based generalization. Participants were informed that they would see a new player of unknown accuracy to hit the center in every trial, that there would be just as many amateur as expert level players and that the order of appearance is unpredictable.

Based on the observed samples, a probabilistic agent would infer a predictive probability distribution over the position of the next sample to accurately estimate the size of the frame that would capture 65% of the imaginary darts thrown by the same dart player. Inference requires the specification of a generative model of the observational data. However, the actual generative model in the environment (controlled by the experimenter) and the model the agent uses for inference is generally different. Nevertheless, in order that inference is optimal, the agent's probabilistic model needs to match the generative process (in the environment). Exploiting knowledge that a normal distribution $d_n \sim N(\mu = 0, \sigma)$ centered at zero is responsible for the $N = 4$ observations $\boldsymbol{d} = (d_1, \ldots, d_N)$, estimation of the predictive density $p(x|\boldsymbol{d})$ over an unseen event x amounts to inference of the only unknown quantity σ parameterizing the standard deviation of the zero-mean Gaussian distribution. Maximizing the likelihood function $p(\boldsymbol{d}|\sigma)$ with respect to σ yields $\sigma_{ML} = \sqrt{1/N \sum_{n=1}^{N} d_n^2}$. This corresponds to the expression for the standard deviation with a known mean of zero. The predictive distribution may be directly based on maximum likelihood estimation (MLE) $p(x|\sigma = \sigma_{ML}(\boldsymbol{d}))$ which is illustrated in Figure 4.1B. However, given the observations it is not possible to determine σ with certainty. The maximum likelihood estimator σ_{ML} and the number of observations N can only be regarded as sufficient statistics for σ.

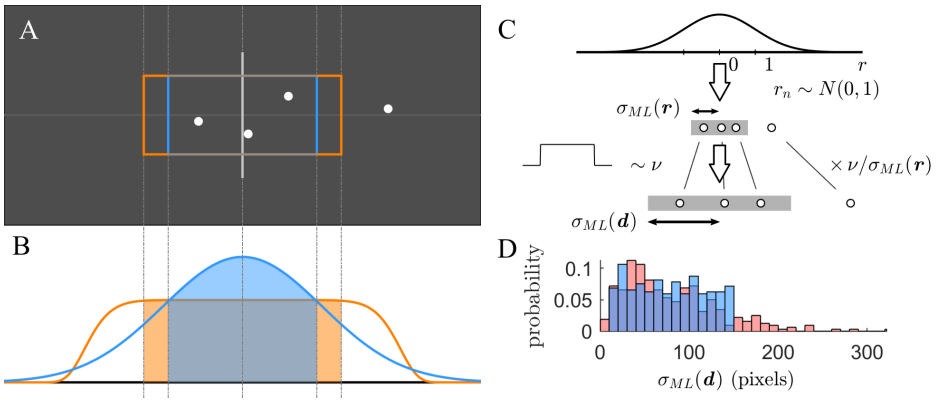

Figure 4.1: Human participants perform a task consisting in estimating the dispersion of future events based on a few observations (A) Schematic of one trial of the task. Participants were asked to judge the unknown accuracy of a "dart player" to hit the center of the board (gray rectangle). Based on the four observed "darts" (white dots), participants must predict where future darts might strike the board. Specifically, participants were asked to capture $65\ \%$ of all future imaginary darts by adjusting the width of the rectangular frame (colored frames, see below). Only the horizontal dispersion of the dots is relevant to estimate the accuracy of the dart player, while vertical displacements are added just to improve visibility of the samples. **(B)** Based on the observed samples, the participants might infer a predictive probability distribution over the position of the next sample. Two hypothetical predictive distributions are shown, representing different structural assumptions about how the samples might have been generated, corresponding to maximum likelihood estimation based on a Gaussian distribution (blue) or a generalized normal distribution with shape parameter $p = 10$ (orange) (see Methods 4.6.3). Based on the predictive probability distribution, the participant can set the frame's width so that it matches the target percentage of $65\ \%$ (colored frames in panel A). Note that for the assumption of a generalized normal distribution, the posterior is more sensitive to data points far from the center and hence a larger frame is chosen. **(C)** The horizontal positions of the points with respect to the center were generated as follows. First, all samples $\boldsymbol{r} = (r_1, \ldots, r_4)$ were generated independently from a standard normal distribution. Second, the samples were scaled by the factor $\nu/\sigma_{ML}(\boldsymbol{r})$, where $\sigma_{ML}(\boldsymbol{r}) = \sqrt{1/N \sum r_n^2}$ is the maximum likelihood estimator (MLE) for a normal distribution centered at zero and ν is drawn from a uniform probability distribution over the range of $[10, 140]$ pixels. The scaled samples $\boldsymbol{d} = \nu/\sigma_{ML}(\boldsymbol{r}) \cdot \boldsymbol{r}$ feature a MLE given by $\sigma_{ML}(\boldsymbol{d}) = \sqrt{1/N \sum d_n^2} = \nu$. This method allows choosing any desired distribution of $\sigma_{ML}(\boldsymbol{d})$ by setting ν correspondingly. **(D)** Histogram of $\sigma_{ML}(\boldsymbol{d})$ across 320 trials (blue). For comparison, the red histogram indicates the results for a sample scaling $\boldsymbol{d} = \nu \cdot \boldsymbol{r}$ without normalizing by $\sigma_{ML}(\boldsymbol{r})$. Both samples have a comparable mean, but the red distribution features few but extremely outlying values, which are avoided by our scaling method

In a Bayesian treatment, the posterior distribution $p(\sigma|\boldsymbol{d})$ requires the specification of the distribution of prior knowledge $p(\sigma)$.

$$p(\sigma|\boldsymbol{d}) \propto \prod_{n=1}^{N} N(d_n|0,\sigma) \cdot p(\sigma) \tag{4.1}$$

The prior is part of the agent's subjective knowledge. However, to be optimal it must equal the actual distribution over σ in the environment, i.e. the base rate at which the hidden variable σ occurs. To then predict the probability of the next event at position x given $\boldsymbol{d}$, σ has to be marginalized out. The predictive distribution results from the probabilistic model $N(x|0,\sigma)$ weighted by the posterior over σ.

$$p(x|\boldsymbol{d}) = \int_0^{\infty} N(x|0,\sigma) \cdot p(\sigma|\boldsymbol{d}) \, \mathrm{d}\sigma \tag{4.2}$$

More generally, the predictive distribution $p(x|\boldsymbol{d})$ corresponds to the belief about future events after observing data $\boldsymbol{d}$.

Now, we turn to the problem of how the agent might set the frame in a principled way based on the estimated predictive probability distribution. For a given setting of the rectangular frame, z, one can determine the fraction of future events within that interval, the capture probability c, by calculating the integral

$$c(z) = \int_{-z}^{z} p(x|\boldsymbol{d}) \, \mathrm{d}x \tag{4.3}$$

The belief in Eq. 4.3 is subjective but it yields a clear objective to determine the response y (half-frame size) by optimization so that $c(y)$ matches the target probability of $65\ \%$ (Fig. 4.1B).

$$c(y) \stackrel{!}{=} 0.65 \tag{4.4}$$

For our purposes, we are mainly interested in inference strategies regarding the probabilistic model given task instructions and input $\boldsymbol{d}$. For data generation, we dispense with the definition of an explicit latent σ variable for the normal distribution as we are interested in the subjective assumptions underlying inference. We used a sampling scheme which reduces response noise and keeps outlying conditions to a minimum translating into improved discriminatory power for model comparison (see Methods 4.6.1). This was achieved by renormalization of the raw samples (Fig. 4.1C). We directly sampled $\sigma_{ML}(\boldsymbol{d})$ from a uniform distribution over the desired range of dispersions. Defining an explicit latent σ-variable over a

finite range would have led to a long-tailed $\sigma_{ML}(\boldsymbol{d})$ distribution with undesirable properties (Fig. 4.1D) which is avoided by our approach. Inference of a normal distribution whose width is assumed to vary parametrically across trials is devised as a reference model (benchmark) for comparison with behavior. It follows the inference strategy of Equations 4.1-4.2 and assumes a uniform prior over the range of $[0, 140]$ pixels corresponding to the task instructions. The Bayesian benchmark model was chosen as reference for motivational feedback and bonus payments to incentivize engagement in the task (see Methods 4.6.2).

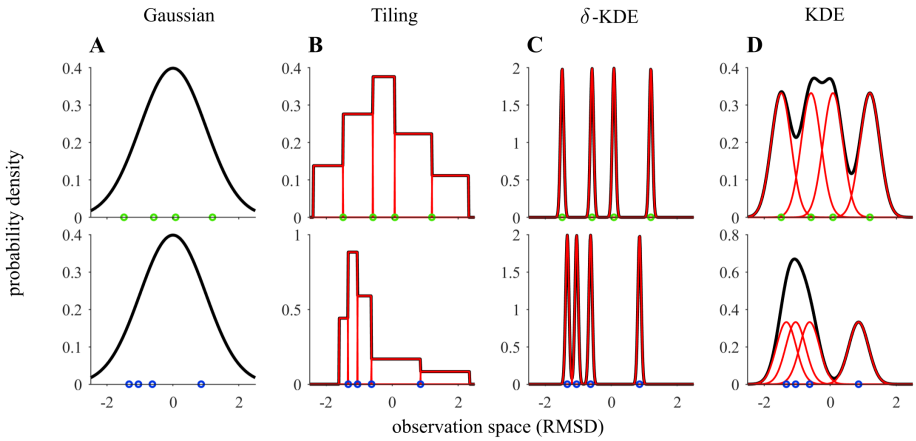

Figure 4.2: Generalization beyond the observed sample is governed by the parametric assumptions of the distribution Each row shows examples of probability densities (black lines) for a different sample (green and blue dots, four observations) in units of its root mean squared deviation (RMSD). **(A)** A zero centered unimodal Gaussian distribution is used to account for the whole sample. All point positions $\boldsymbol{d}$ enter via the estimated standard deviation parameter, $\sigma_{ML}(\boldsymbol{d})$ (RMSD), determined by probabilistic inference. Whereas for instanced-based generalization the sample points effectively enter as parameters themselves. **(B-D)** Different additive basis distributions (red) to cover the observation space can be used. The tiling model covers the space with adjacent non-overlapping uniform basis distributions resulting in a compressed distribution around spatially proximal points **(B)**. Additionally, models can be constructed from simpler components by centering a Gaussian kernel on each observation (see Methods 4.6.3). In the limit of vanishing kernel widths **(C)** there is no generalization beyond the sample while for larger widths **(D)** a smoothed density over the whole domain is obtain due to overlapping basis distributions.

The goal is to determine which inductive biases participants employ for generalization and whether that conforms to the structural assumptions suggested by the framing of the task. More specifically, we attempted to distinguish between inference of a centered, unimodal, bell-shaped distribution (Fig. 4.2A) and variants of instance-based generalization (Fig. 4.2B-D), such as kernel density estimation (KDE), which make only very few assumptions about the distribution to be infer-

red. We furthermore investigated whether participants might derive their behavior from an internal representation of a probability distribution. Alternatively, any measure that correlates with the dispersion to be estimated might serve to inform behavior. These heuristics are primarily chosen to facilitate processing and not to achieve a more accurate representation of the environment. Our task allows explicit testing of some heuristic short-cuts to the task.

4.4.1 Faithful tracking of trial-by-trial uncertainty

First, we tested whether participants demonstrate the ability to faithfully estimate the dispersion of the centered normal distribution assumed to be responsible for the observations. The MLE of the Gaussian, σ_{ML} (Fig. 4.3A, red), is the sufficient statistic to inform the optimal response (green).

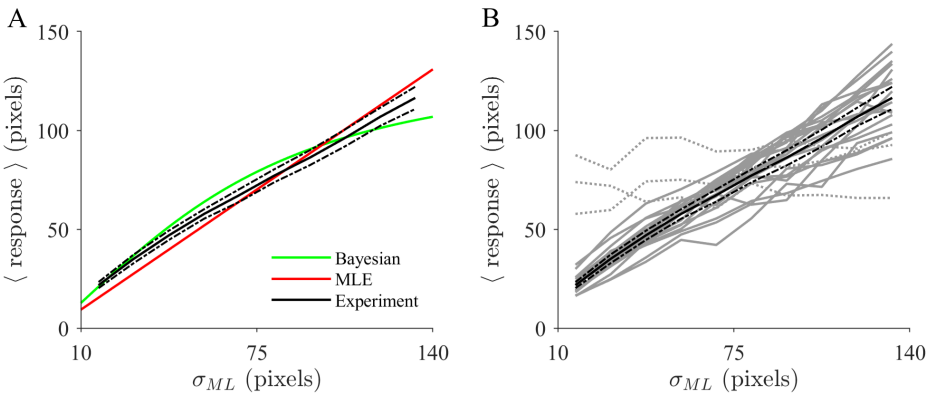

Figure 4.3: Human behavior closely tracks trial-by-trial uncertainty of future events **(A)** Mean response across participants plotted as a function of the MLE of the sample, $\sigma_{ML}(\boldsymbol{d})$, in ten equally spaced bins (black, error bars 95 % CI). Basing behavior on a Gaussian estimated by ML (red, $N(x|0, \sigma_{ML}(\boldsymbol{d}))$ results in responses proportional to the estimate. The prior distribution that is assumed by the devised Bayesian benchmark model (green) biases responses towards intermediate values (see Methods). **(B)** Individual response curves of all 23 participants tested (gray lines). Three participants displaying poor compliance with the instructed task (dotted) were excluded from further analysis. Average across participants is superimposed (black).

The averaged mean response across participants (black) is closely related to it in an almost linear relationship (see Methods 4.6.3). Assuming that participants use the Gaussian distribution for inference (Methods 4.6.3, normal-model) yields good predictive performance and accounts for a substantial amount of the variance (regression, cross-validated median $R^2 = 0.80$, 95 %-CI, $(0.73, 0.82)$, across participants). Uncertainty tracking is also apparent on an individual participant level (Fig. 4.3B) (cross-validated median R^2 ranging from 0.47 to 0.93). On average,

the responses appear to be systematically biased towards intermediate values with respect to the ML approach (Fig. 4.3A, red) resembling the effect of a prior distribution (green) incorporating knowledge about the range of dispersions across trials. This bias from proportionality is quantified by the loss of predictive performance of a model restricted to proportional outputs (Methods 4.6.3, Eq. 4.7). It is strongly inferior to a linear mapping (Methods, Eq. 4.6) even on an individual level (cross validation log likelihood (CVLL) difference $\Delta \geq 20$ for 12 participants, $\Delta \geq 10$ for 17 participants).

4.4.2 Evidence for an internal trial-by-trial objective

Next, behavior is examined with respect to the objective participants were instructed to obey. Namely, if their estimates are quantitatively accurate and correspond to the $65\ \%$ target percentage. For independent trials, participants must infer the dispersion anew on each trial. Inferring a probability distribution over future events allows behavior to be derived from a principled trial-by-trial objective regarding the target percentage (see Fig. 4.1B and Methods, Eqs. 4.3-4.4). By construction, our task objective demands a quantification of the relative frequency of all future events and was intended to require participants to approximate distributional estimates.

To examine how well participants performed with respect to the devised optimal inference strategy, we calculated the capture percentage by evaluating (Eq. 4.3) with respect to the optimally inferred probability distribution (Eq. 4.1-4.2). The distribution of the per participant median capture percentage across all trials is clustered close to the target of $65\ \%$ (Figure 4.4). In this measure opposing deviations cancel, so that it evidences an overall compliance to the target percentage across all trials. The median across participants is close to the target percentage, which indicates that participants quantify uncertainty in a quantitatively similar manner as the probabilistic benchmark model. The median of the absolute deviation per response is $6.54\ \%$ ($95\ \%$-CI, $(5.83, 7.28)\ \%$) with respect to the external objective of the task. However, it is possible that behavior has been produced from an internal objective (see Eq. 4.4) in which the percentage is matched much more closely to $65\ \%$. There are at least two contributions that inflate the deviation from the external measure (Fig. 4.4A). First, there is intrinsic response noise which would even occur for fixed stimuli on the screen, e.g. through motor-related variability. Second, there are deviations due to mismatched inference with respect to our benchmark model (see e.g. [168]). The latter are deterministic and the result of e.g. different prior knowledge from the one assumed by our benchmark model. Altogether, the median absolute deviation (Fig. 4.4A) is a conservative upper bound estimate for an internal trial-by-trial objective of the capture percentage so

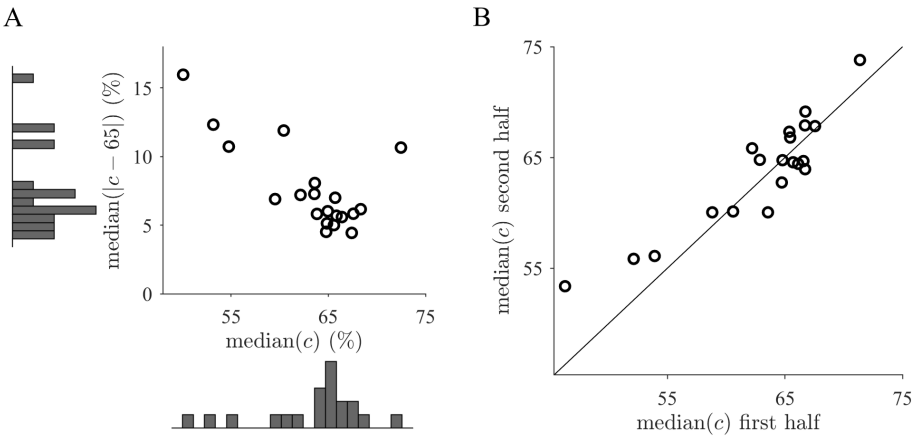

Figure 4.4: Behavior is consistent with participants possessing a subjective but well calibrated trial-by-trial internal objective that remains stable over the experiment. **(A)** Across trials participants tend to comply well to the objective despite per trial deviations due to systematic biases and response noise, as the capture percentage c is typically around the target value $65\ \%$ (vertical axis) and the median deviance is relatively small (horizontal axis). Histograms correspond to marginal distributions. **(B)** Participants display stable behavior throughout the experiment, as they do not appear to adjust their responses closer to the task objective over time. Median capture percentages c are calculated separately for the first and second halves of the experimental session.

that the quantitative match with the target percentage can be considered high.

If participants did not possess an internal trial-by-trial objective, they could instead associate stimuli with suitable responses by a learning a behavioral function. Next, we tested whether behavior is consistent with this alternative approach. We checked for temporal transients adapting to the externally provided objective via feedback and across-trial dependencies. Remarkably, the median capture percentage appears not to adjust closer to the target percentage as indicated by similar values calculated separately for the first and second half of the experimental session for each participant (Fig. 4.4B). The absolute difference of the median capture deviation is small and not significantly different from zero (right-tailed Wilcoxon signed rank test, $p = 0.48$) despite the fact that the trial-averaged feedback about the capture percentage in the experimental session may have allowed to derive some global adjustments. Accordingly, too high a capture percentage on average should subsequently lead to the choice of smaller response frames. Hence, a decrease of the feedback error would be expected over time. We also confirmed that the previously presented feedback about the capture percentage did not influence behavior (regression, exceedance probability $p_{exc} = 2.04 \cdot 10^{-4}$ compared to baseline model, see Methods 4.6.3). Similarly, no considerable de-

pendencies across trials were found (Methods 4.6.3).

Overall, participants typically predict the dispersion of future darts in a quantitatively accurate manner. They appear to have relied on an internal trial-by-trial objective regarding the target percentage as they largely conform to trial independence, feature stable processing across time and virtually ignore feedback. This is consistent with internal probabilistic processing.

4.4.3 Systematic deviations from inference of a Gaussian

Thus far, behavior appears to be close to the optimal inference strategy defined by the benchmark model, but we have also observed deviations (Figure 4.3, 4.4A). If

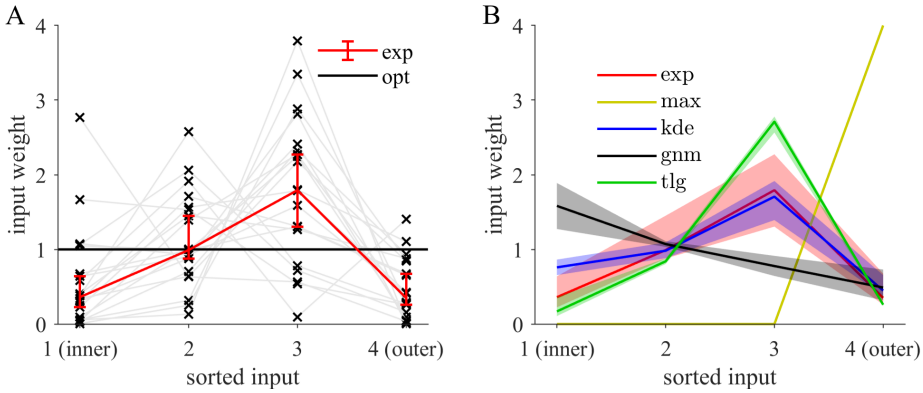

Figure 4.5: The weighting pattern of the observed samples deviates from inference of a close-to-normal distribution and matches kernel density estimation (KDE) Evaluation of the normalized weights ω_n of the weighting-model $\hat{S}(\boldsymbol{d}) = \sqrt{1/N \sum_n \omega_n d_n^2}$ as a generalization of the MLE of a zero-centered Gaussian. The points are indexed according to their distance from the center. **(A)** Input weight that each participant (gray lines) assigns as a function of the weight index. If participants followed optimal MLE based on a Gaussian centered at zero, all input weights should be equal (black line). Fitting of the weighting model (see Methods 4.6.3) shows a systematic deviation of the median across participants (red, error bars 95 %-CI). Participants tend to overweigh the third most extreme value compared to the others. **(B)** Among all models tested, only KDE (blue) qualitatively matches the characteristics of the experimental weighting pattern (red, same as panel A). The other models fail to capture the behavioral weighting pattern (fits of the weighting model to the other indicated models' output).
Model abbreviations: kde - kernel density estimation, tlg - tiling, gnm - generalized normal, max - maximum

behavior follows from inference of a normal distribution, it can only depend on the sample via the sufficient statistic, $\sigma_{ML}(\boldsymbol{d}) = \sqrt{1/N \sum d_n^2}$. This means that the squared position of each point should contribute equally to the final estimate. We tested this with a weighting model that generalizes σ_{ML} by assigning a tunable

weight ω_n to each input depending on its excentricity, $\sqrt{1/N \sum \omega_n d_n^2}$. Excentricity refers to the distance from the center irrespective of the side where the sample occurs.

Experimentally, the weights of the individual points tend to take unequal values (Fig. 4.5A). Participants put more emphasis on the third most excentric point and down-weigh the first and the fourth point. We also tested whether other models of behavior are able to reproduce this pattern (Fig. 4.5B).

In the following, models will be compared by both the (i) weighting pattern (Fig. 4.5B) as well as their (ii) overall ability to predict behavior (Fig. 4.6). Consistent with the weighting pattern observed in our data, the normal model (nm) is far from providing the best predictions of behavior. This can be seen from the pairwise model comparison matrix (Fig. 4.6). There, the binomial probability that the model indexing the row (vs. the model indexing the column) is more likely to account for the data of a randomly chosen participant is depicted as color code. Additionally, entries with high exceedance probabilities are considered significant (Methods 4.6.3) and marked with asterisks. For instance, the comparison between the weighting model in row (wgt) to the normal model in column (nm) shows that the latter is clearly rejected ($p_{exc} > 0.999$). Beyond the group level, the normal model can be decisively ruled out individually for many participants despite the fact that generally different participants are best described by different models.

We tested whether generalizations of the Gaussian can account for the systematic deviations that were observed before. The generalized normal model (gnm) allows for more freedom in the representation of the inferred density through a shape parameter governing its kurtosis (see Fig. 4.1B) by generalizing the square in the exponential function to other powers than two leading to an unequal weighting pattern of the samples (Fig. 4.5B). This model predicts significantly better than the Normal-model (Fig. 4.6, $p_{exc} > 0.999$) by making use of the additional shape parameter to represent heavier tailed distributions (quartiles across participants $Q = (0.79, 1.24, 1.68)$). Heavier tailed distributions discount outlying and enhance the influence of inlying points on judgments (Fig. 4.5B, black line). The experimental pattern (red) is not matched well suggesting that it does not reflect how participants behave. In addition, the weighting model still outperforms the generalized normal model (Fig. 4.6).

4.4.4 Simple heuristics are poor predictors

Before, we determined that responses are on average relatively close to the target but that the finer-grained behavioral patterns are inconsistent with inference of a close-to-Gaussian distribution. That raises the question whether simpler, heuristic strategies might offer a better account of behavior which might also unequally

weigh sample information.

We first tested the established heuristic models that use perceptually simple statistics and only a subset of the available information. The maximum model (max) only depends on the most excentric point which leads to a weighting pattern (Fig. 4.5B, yellow) which is highly inconsistent with the experimental one (red). The participants' weighting is more balanced and typically features weights smaller than four (normalization to number of sample points). The range model (rng) is based on the sample's range and predicts worse than the maximum model (Fig. 4.6). On the group level, both are clearly refuted by all other models.

Another heuristic strategy is attending to just one point when sorting them according to their excentricity. In particular, the third most excentric point is important as it closely corresponds to the target percentage of $65\ \%$ on the sample and is the response in the limiting case of pure instance-based generalization (see δ-KDE model, Methods 4.6.3). Participants typically take all point positions into account. The four (unnormalized) weights are significantly different from zero for many individual participants (weighting model, 10000-fold permutation test, $(14, 20, 20, 19)$ out of all 20 participants feature a p-value < 0.05 for the weights $(w_1, ..., w_4)$ respectively). Furthermore, for each individual, at most one weight is insignificant showing that it is not an effect of grouping. Consistent with integration of the whole sample, the maximum of the normalized weights is considerably lower than four (Fig. 4.5A).

Altogether, this is evidence that among all participants only few tend to exploit heuristics. The clear majority however resorted to some more sophisticated weighting inconsistent with the simple heuristics tested.

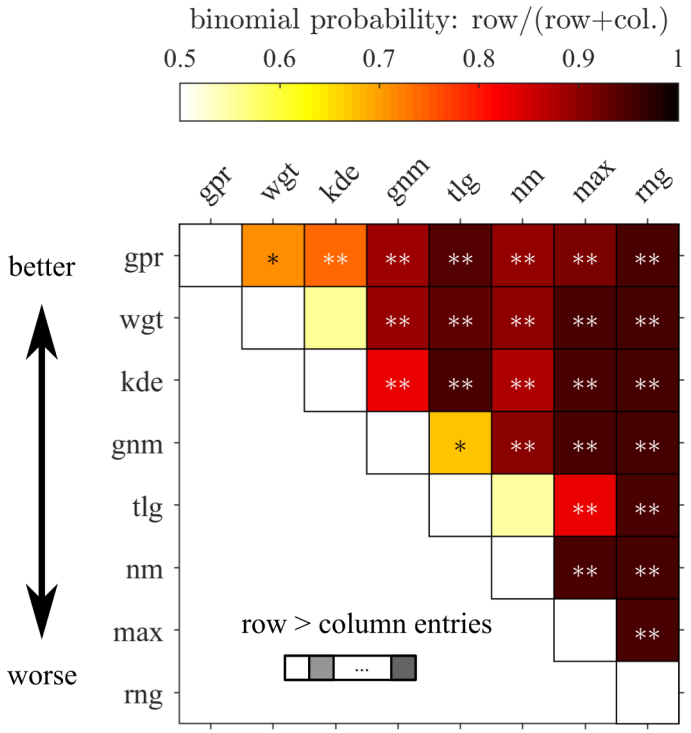

Figure 4.6: Pairwise model comparison evidences an inclination to resort to instance-based generalization, indicating that fluctuations have a profound effect on the inferred representations Summarized results of a hierarchical Bayesian model comparison procedure that estimates probability distributions over models. Pairwise comparisons (each square) are performed to evidence relative differences in prediction for models with different features. The color code over each square shows estimates of the parameter of the binomial distribution governing the probability by which the model indexed by the row is more likely than the one indexed by the column. This corresponds to the expectation value that a given model is considered responsible for generating the data of a randomly chosen participant. Superimposed are large differences of the exceedance probability ($* \mathrel{\widehat{=}} (0.99 > p_{exc} \geq 0.95)$; $** \mathrel{\widehat{=}} p_{exc} \geq 0.99$) which quantifies the belief that the row model is more likely to have generated the data of a randomly chosen participant compared to the column model. Model abbreviations: gpr=Gaussian process regression, wgt=weighting, kde=kernel density estimation, gnm=generalized normal, tlg=tiling, nm=normal, max=maximum, rng=range

4.4.5 Behavior relies on instance-based generalization

So far, participants appear to violate the assumptions of a close to Gaussian distribution centered at zero that was suggested by the task instructions and the dart metaphor. Alternatively, the probability distribution to be inferred may be directly constructed from the observed instances by imposing only minimal structural constraints on the data. That corresponds to the assumption that the sample is representative of the unknown population to be estimated.

Our tiling model (tlg) implements such an approach with spatially confined basis distributions. It places a uniform distribution in between observations and hence the resulting density is increased around clusters and reduced elsewhere (Methods 4.6.3). It adapts to the fluctuations which are present in the sample. Consequently, the target capture percentage of $65\ \%$ is by construction very close to the third most excentric point. As a result, this model emphasizes the third most excentric point (Fig. 4.5B, green) and thus captures an important characteristic of behavior (red).

The kernel density estimation (KDE) model uses Gaussian basis functions to implement instance-based generalization. It centers a Gaussian distribution on each data point and thus assigns density to its vicinity depending on the standard deviation parameter. The experimental weighting pattern (black) is closely captured by KDE (Fig. 4.5B, blue). It is very successful at predicting behavior and superior to both the normal and the generalized normal model considered before (Fig. 4.6). The small and insignificant difference of the model probability (Fig 6, wgt vs. kde) indicates that KDE predicts on a similar level as the weighting model even though the latter has more adaptable parameters and thus may be considered more flexible. The weighting model does not explicitly construct a probability density but can be viewed as a functional approximation that can capture similar dependencies of behavior on the sample.

In summary, participants do not sufficiently exploit the structural constraints suggested by the task but instead give more freedom to the specific instances of the observations to determine their responses. The tendency to assume that even small samples are representative of the population could be well captured by nonparametric kernel density estimation.

4.4.6 Inferred representations feature overlapping and redundant kernels

Probability distributions over perceptual variables should be embedded in the context of more general knowledge of the task's context. From a causal inference perspective, they should be attributed to the causal variables already known to exist.

Treating all observations as (new causal variables) if they originate from their own cause makes purely nonparametric methods seem of limited applicability in wider contexts. In this sense, KDE itself may be considered a heuristic approach as it largely ignores prior (structural) knowledge. Examining the inferred representations, we argue here that there is reason to belief that behavior is not purely nonparametric but can rather be conceived of as an instance-based modulation (bias) to causal inference.

If we infer very narrow kernel functions for our participants that would indicate that there is very little generalization from the sample. For close to orthogonal kernel functions with virtually no overlap (e.g. delta-distributions) the output reduces to a mere counting of observations. First, we tested how strong this instance-based bias is on the level of raw responses by comparing them to the predictions of δ-KDE (Fig. 4.7A). Both axes are normalized to the MLE, σ_{ML}, of the sample (i.e.

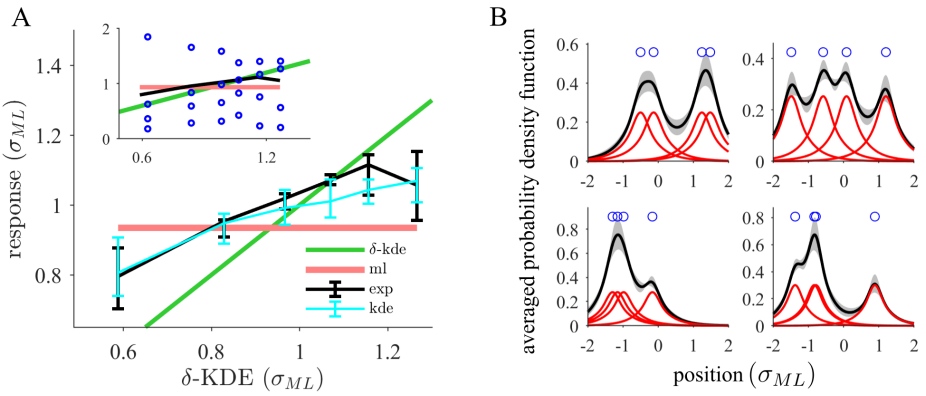

Figure 4.7: Strong generalization consistent with the possibility of integrating prior knowledge about the task structure (A) Responses (black) show higher consistency with inference of a single Gaussian than with approaches generalizing only weakly beyond the sample such as δ-KDE (limit of vanishing kernel widths; third most excentric sample point). The plot shows aggregated (median across participants, 95 %-CI) bin medians of the responses (normalized by σ_{ML}) and the fitted KDE model (cyan) as a function of the δ-KDE output (approximately equally filled bins). By construction, inference of a Gaussian results in a horizontal line (red) while δ-KDE (green) yields a linear function of slope one. The experimental curves are less steep indicating a rather moderate instance-based modulation compared to a Gaussian model. The inset is a zoomed out version additionally showing the relationship of the responses to the distribution of sample points (median of absolute value within each bin). **(B)** The KDE model infers internal distributions that are smoothed and spatially extended around the sample points. The mean probability density function across participants (black, 95 % CI) is shown for four different samples (blue circles). The inferred density is smooth featuring fewer modes than the number of basis distributions (red curves). This is a consequence of the large fitted Gaussian kernel widths which lead to substantial overlap of the basis distributions.

the draws from the standard normal distribution, see Methods 4.6.1). All responses are plotted as a function of the δ-KDE output. Thus, by construction, predictions of δ-KDE (green) itself follow the unity line while predictions of inference using a Gaussian likelihood function follow a constant line of slope zero (red). Values of the optimal benchmark model (not shown) would fluctuate because of varying prior beliefs that average to a constant independent of the sample given the MLE. The slope of a linear function fitted to the experimental responses is far from one as expected from δ-KDE (Fig. 4.7A, regression, median slope across participants $0.27\ (0.24, 0.38)$). As opposed to the δ-KDE model, the KDE model (cyan) can predict the behavioral pattern (black) well because its kernel width parameter takes large values (Fig. 4.7B, red) (median across participants $0.40\ (0.35, 0.59)$, in units of σ_{ML}). Participants capture a varying number of points with the response frame (Fig. 4.7A, inset) which is only possible if the constructed density is a non-local function of the specific sample configuration on the screen. This slope pattern is not entirely inconsistent with inference of a Gaussian likelihood function as responses actually vary around its value as a function of the sample configuration. On the contrary, the normal model reaches high predictive performance in absolute values as shown before. However, additional to the responses derived from Gaussian inference, there are subtle instance-based variations which can be captured by the KDE model. At the level of responses, behavior may be understood as inference of a normal distribution that is modulated by KDE.

Interestingly, KDE predicts behavior significantly better than the tiling model (Fig. 4.6). The main difference is that the tiling model relies on spatially confined basis functions while Gaussian kernels are spatially extended. The weighting pattern shows that the tiling model (Fig. 4.5B, green) overweighs the third most excentric point even more than behavior (red). The tiling model too closely resembles the purely instance-based approach of δ-KDE while behavior is not so strongly influenced by the third most excentric point. Of all models tested, KDE (blue) best captures the weighting pattern (Fig. 4.5B) because the large kernel width exhibits a non-local effect so that the positions of all points influence judgments leading to a more balanced pattern.

A large kernel width makes spatially extended kernels overlap (Fig. 4.7B, red). Accordingly, we typically find fewer than four modes in the inferred densities of the participants (median of the per participant mean across trials 2.0375 $95\ \%$-CI $(1.50, 2.27)$). Thus, increasing the kernel width may be understood as a reduction of the effective number of components in the mixture distribution as measured by the number of modes (Pearson correlation coefficient, $\rho = -0.95$, $p = 6.01 \cdot 10^{-11}$. Our data requires the KDE model to perform close to a regime where it must approximate inference of some smooth distribution which is closer to unimodal. Despite being the best approximation explored, it is nevertheless

possible that the inference method used by our participants is structurally more constrained than KDE and uses some prior knowledge of the task structure.

From a representational point of view, the large overlap of the basis distributions (Fig. 4.7B, red) is a rather redundant and thus inefficient way of representing the whole distribution. For a large degree of overlap, several kernel functions could be well represented by a single kernel function whose free parameters are tuned to accommodate all their contributions. Bayesian nonparametric mixture models [61] can effectively reduce the number of redundant mixture components and minimize shared responsibility to account for the data points. The number of components can adapt to the position and number of data points in the sample. It gives less freedom to the data than KDE but implements soft and gradual constraints towards sparsity. A preference for sparser or denser representations can be specified by a prior. Likewise, prior knowledge such as a zero-centered population may be included in this way. We suggest this as a connection to theoretical principles.

We found that participants show different preferences for instance-based generalization. The average number of modes of the inferred densities (according to the KDE-model) almost covers the full range of possible values (minimum 1.01, median 2.04, maximum 3.63, across participants). Even with wide kernels, KDE is limited in its ability to represent unimodal near-Gaussian distributions. Correspondingly, the difference in predictive performance (CVLL) between the KDE and the normal model is larger for smaller kernel widths (linear correlation coefficient, $\rho = -0.54$, $p = 0.0075$). Consistent with previous results, the slope in Figure 4.7A decreases with the kernel width (Pearson correlation coefficient, $\rho = -0.66$, $p = 7.30 \cdot 10^{-4}$). The determinants of the participants' preferences are unclear from this experiment. We remark however, that participants who infer more redundant densities tend to respond faster (Spearman correlation coefficient, $\rho = 0.35$, $p = 0.064$) although the result does not reach significance.

In summary, using KDE we found very wide overlapping kernels leading to densities which could be more sparsely represented. This hints at a more sophisticated inference approach than pure instance-based generalization. It may be considered a modulation of causal inference by a kernel based approach. We suggest a connection to nonparametric Bayesian methods in statistics that allow to incorporate prior knowledge and sparsity constraints.

4.4.7 Explanation close to ceiling level

There are many possible ways in which this task might be approached by our participants. Thus, we attempt to estimate an upper bound of the predictable structure in the data regardless of how the task was solved by the participant. Gaussian process regression (GPR) is used to find a low-bias functional approximation between

input $\boldsymbol{d}$ and behavior y. Hence, if a model reaches similar predictive levels this is an indication that it captures the most relevant computational operations. GPR is indeed found to be the best model (Fig. 4.6) on the group level. However, the differences to the KDE-model are not disconcertingly large (median CVLL difference across participants, 13.3 dHart, $95\,\%$-CI, $(-3.1, 23.5)$ dHart). Overall, KDE can predict on a comparable level as GPR. This is remarkable as for interpretable models, all factors need to be specified explicitly. For instance, even motor related variations with $\boldsymbol{d}$ would have to be incorporated. Moreover, as probability densities are high dimensional and subjective, the achieved match is not trivial. However, we conclude that behavior is somewhat predictable beyond what is captured by the KDE-model but that it manages to capture the most important computations reflected by behavior.

Whether the unexplained variations may be captured by more sophisticated approaches to density modeling is left for future investigations. Nevertheless, to give a more concrete demonstration of these ideas, additional experimental tests were conducted with a smaller sample of participants on a variation of this main experiment. Its primary purpose is to further substantiate the claims regarding probabilistic processing and to flesh out modeling with Bayesian nonparametric methods. The task design was slightly varied in that the task objective demands a more flexible use of the sensory representation of uncertainty. The reader is referred to the appendix B for further details. The additional experiment fully underscores the principal claims made by the main experiment. Moreover, it provides a proof of concept that Bayesian nonparametric mixture models are indeed suitable to describe the internal representations.

4.5 Conclusions

This study attempted to elucidate how sensory representations of uncertainty are constructed from sparse data. We have described a new experimental task that allows to measure quantitative judgments of uncertainty in response to a noisy stimulus with high precision. We find that (1) participants give faithful judgments about uncertainty on a trial-by-trial basis which are irreducible to simple heuristics. (2) Their behavior is not in agreement with the structural assumptions of a Gaussian suggested by the framing of the task. Instead, according to their behavior, participants are biased to judge the sample as representative of the population and that random fluctuations in the sample will reproduce in the long run. A connection to nonparametric Bayesian models is suggested to model this inclination towards instance-based generalization. (3) Furthermore, behavior is consistent with the idea that participants internally represent the variable of interest probabilistically

as a normalized distribution over its possible values.

The idea that perception constitutes some form of probabilistic inference process was suggested long ago [16]. It has a particular appeal for deriving subjective estimates of uncertainty as it emerges naturally from the knowledge representation itself, i.e. from the posterior distribution, without requiring a meta-representation [108, 126, 181].

Experimentally, one must elicit the read-out of a suitable summary statistic of the sensory representation. In previous work, participants were typically asked to report their confidence in that the latent variable to be inferred lies beyond some fixed decision boundary [129]. Instead, we allowed participants to freely estimate the dispersion of the inferred density. There is virtually no demand on working memory and participants do not need to resort to language to perform the task. Both aspects are believed to be critical for promoting rational behavior [133]. In addition to being intuitive, this task requires an ability to deal with uncertainty to construct an internal trial-by-trial objective regarding the target percentage. Critically, this task was designed to minimize sensory and motor noise to obtain a sensitive probe of behavioral variations of dispersion estimates. As opposed to prior work, e.g. using the random dot motion stimulus [132, 187], here mainly the task-relevant stimulus dimensions (dot positions) drive behavior. This study more specifically investigates the process of density estimation that is embedded in other (hierarchical) tasks. Previously, several studies tested how multiple inferred sensory representations are combined. The reliability based weighting of conflicting cues from different modalities suggests that distributional estimates are provided by each modality [138]. Another study also supplied evidence by means of a dot cloud [137] but assumed normally distributed noise. Many previous studies made the strong assumption that participants know the generative process of the task. Very often it is chosen to be a normal distribution [132, 161]. It may be a reasonably good proxy to model cognitive processes for simple, nonlinear and low-dimensional stimulus tasks with abundant evidence. However, we challenge the adequacy for inference in complex environments or sparse observations. These assumptions evade the deeper question of choosing a suitable model that the agent faces. In hierarchical models and depending on context, the upper levels provide constraints as to what the important causal factors are. We framed the task by alluding to a commonly known random process of throwing darts conforming to prior structural assumptions of a centered, unimodal and bell-shaped distribution that is close to Gaussian.

Nevertheless, we find that most participants fall short of these assumptions but rather give systematically biased estimates. Because of the low number of samples, our task allows testing what inductive biases [32] participants exhibit. They appear to give more freedom to the model's structure to adapt to the sample. Thus,

their judgments seem to assume that fluctuations in the sample are representative of the population [150]. However, we found evidence that their inferences are somewhat more constrained as purely kernel-based estimates leading to potentially sparser representations. We propose to view this in the framework of Bayesian nonparametric mixture models [61, 188] which may infer the appropriate complexity for each sample based on a prior expressing a preference for the sparsity of the final estimate (the number of components). In this context, the bias towards instanced-based generalization can be considered a prior that favors more complex solutions. This is reminiscent of findings in the literature where human abilities to learn functions are described by a hybrid of nonparametric and parametric approaches [189].

We can only speculate about the reasons behind this inductive bias. First, it might be due to considering the cost of computing [111] in an attempt to simplify judgments. However, we found a tendency towards more complex representations whereas sparser representations are typically believed to be more economical. E.g., decomposing high-dimensional objects such as continuous probability density functions of human visuo-motor errors into simple non-overlapping (uniform) basis distributions was suggested to be a solution to complexity by obtaining a sparser representation [66]. Instead, we speculate that the bias towards instanced-based generalization might be related to structural uncertainty about the causes of their observations. Structural uncertainty has been shown to lead to model-free learning [60]. Similarly, a sensitivity to small alterations in the task setting has been found to affect the optimality of behavior [140]. Furthermore, we might be equipped with a more fundamental bias to perceive causes behind patterns even for little evidence [109].

By construction, our task objective only applies to a normalized distribution over future outcomes regardless of its functional shape. Various studies have claimed that internal processing is probabilistic or at least demonstrated a "lower bound for the sophistication of confidence evaluation" [17]. Typical approaches derive an optimal solution to the task and show that behavior is reasonably close to it. However, strong claims require preconditions [190] such as testing alternative models [191] for non-trivial optimal processing. We do not claim optimal processing but emphasize systematic deviations that nevertheless might originate from internal probabilistic computations. Often as in our case, a clearly suboptimal strategy yields near-optimal results.

In fact, instead of a trial-by-trial objective for the target percentage derived from a density estimate, a learned stimulus responses mapping might be used. Our task design minimized the possibility to optimize a reward measure through trial-and-error over trials by omitting informative feedback. Consequently, the chances of acquiring a stimulus-response mapping are minimized. Furthermore,

simple heuristic approximations [155] to behavior have been ruled-out explicitly. Additionally, we found that the implementation of instance-based generalization by KDE is within reasonable bounds of an estimate of the predictable structure in behavior [190] suggesting that we have captured the important computations.

Ultimately, the degree to which claims to probabilistic processing seem substantiated depends on the propensity to belief that the task could alternatively be solved by a well-tuned mapping or heuristic estimator acquired prior to the experiment. This task is rather artificial and humans are seldom prompted to state or give error intervals in terms of percentages. Accordingly, the situations to learn from are sparse. Uncertainty about (latent) variables is rarely made explicit (especially in numerical terms) but rather implicitly used by the agent to integrate and update beliefs. Generally, there is little information about the frequency with which events happen in our world across instances of the same situation. Even though learning calibrated mappings from specific situations is in principle possible, it is highly uneconomical and thus regarded unlikely. Likewise, it seems unrealistic that evolutionary training across generations has provided us with well-tuned heuristics for specific situations such as this task. After all, we deem it more plausible to assume that most participants estimated some (approximate) probabilistic distribution to derive their judgments.

In conclusion, our results suggest that human judgments about uncertainty are guided by an internal probabilistic objective. However, there is a tendency to identify fluctuations in the sample as representative for judgments about the population. This may be captured by a representation endowed with a preference to adapt overly flexibly to the observed instances.

4.6 Materials & Methods

4.6.1 Sampling scheme to generate observations

On each of the 320 trials, the horizontal positions of the points with respect to the center were generated as follows (Fig. 4.1C). First, always $N = 4$ sample values $r = (r_1, \ldots, r_4)$ are independently drawn from a standard normal distribution $r_n \sim N(0, 1)$. Second, the samples were scaled by the factor $\nu/\sigma_{ML}(\boldsymbol{r})$, where $\sigma_{ML}(\boldsymbol{r}) = \sqrt{1/N \sum r_n^2}$ is the maximum likelihood estimator (MLE) for a normal distribution centered at zero of the samples $\boldsymbol{r}$ and ν is drawn from a uniform probability distribution over the range of $[10, 140]$ pixels. The scaled sample $\boldsymbol{d} = \nu/\sigma_{ML}(\boldsymbol{r}) \cdot \boldsymbol{r}$ always has a MLE given by $\sigma_{ML}(\boldsymbol{d}) = \sqrt{1/N \sum d_n^2} = \nu$. This method allows choosing any desired value of $\sigma_{ML}(\boldsymbol{d})$ by setting ν correspondingly. Setting $\sigma_{ML}(\boldsymbol{d})$ directly, which is the main determinant for inference, has the advantage that observations $\boldsymbol{d}$ and the MLE $\sigma_{ML}(\boldsymbol{d})$ take less extreme values which translates into increased numerical stability for model comparison. Defining an explicit latent σ-variable over a finite range instead would have led to a long-tailed $\sigma_{ML}(\boldsymbol{d})$ distribution with undesirable properties (s. Figure 4.1D). The ability to tell apart models with similar predictions is enhanced if response noise and outlying conditions are kept at a minimum.

However, because of this way of generating the dots, the optimal inference model with respect to the actual generative model in the environment is not readily defined. Nevertheless, participants do not know these alterations how the dots were generated. The best they can do is to follow the instructions and their prior knowledge suggested by the dart metaphor to explain the data. We do not define the optimal model with respect to the generative model in the environment. Instead, we define it as an optimal inference strategy based on a normal distribution whose width varies parametrically across trials. It follows the inference strategy of Eq. 4.1-4.2 and assumes a uniform prior over the range of $[0, 140]$ pixels. As this prior arguably matches the task instructions it was chosen as the basis for our Bayesian benchmark model and the feedback in the experiment.

4.6.2 Participants & Experimental Procedure

In total 23 participants (15 female, 8 male) were recruited mainly among students from the Pompeu Fabra University in Barcelona. We accepted all healthy adults with normal or corrected to normal vision. We obtained written confirmation of informed consent to the conditions and the payment modalities of the task. The training and the experimental session were carried out on a single appointment that nominally lasted 75 min. First, participants read detailed written instructions of the task. In a brief training session, they were given 40 trials to familiarize with the

handling of the task through a short interactive session with feedback after every trial. The feedback consisted of the actual percentage c_t (using Equations 4.1-4.3) they would have captured in trial t according their response y_t and our benchmark model. In addition, they were given a deviation score (mean squared error (MSE)) from the target percentage $\delta_t = (c_t - 0.65)^2 \cdot 1000$.

In principle, a participant could learn how a pair consisting of observations $\boldsymbol{d}$ together with his response y, (d, y), relates to the capture probability p from experience in the 40 training trials. For a given learned mapping $(d, y) \rightarrow p$ he would have to adjust y such that $p = 0.65$. We regard this as unlikely for the following reasons. First, 40 trials do not provide a lot of data to learn from. Second, the mapping is nonlinear and its domain is high-dimensional which makes it hard to learn and susceptible to the specific instantiations of $\boldsymbol{d}$ across trials – as well as the choice of y. (d, y) and p are never simultaneously visible on the screen. And finally, batch learning requires memorizing all presented pairs which seems infeasible for participants. While on-line learning is possible, it typically suffers from slower convergence rates.

Participants could ask any questions to the experimenter prior to the experiment. The subsequent experimental session consisted of 320 trials with pauses together with feedback after every 5 trials. In the experiment, the feedback consisted of 5-trial averages of the quantities c_t and δ_t above that were computed since the last pause. Participants were supposed to minimize the deviation score and were payed more compensation when having a smaller deviation score to incentivize optimization. This supposedly promoted high motivation to prevent participants from resorting to computationally cheaper heuristic shortcuts. The task circumvents risk aversion since there is practically nothing that the participant can do to prevent losses other than stating the response as accurately as possible.

The bonus payment was determined by the mean of their final deviation score after removing the eight worst trials. The payment was determined by comparison to an array of five thresholds that were set according to the $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ cumulative quantiles of the empirical deviation score distribution across prior participants. A lower score corresponds to a better performance so that participants were payed an additional bonus of $\{5, 4, 3, 2, 1\}$ € if their final deviation score was less or equal to the quantile thresholds. This is a relative way of rewarding their efforts to optimize their responses. Irrespective of their performance they were paid 10 € and hence on average received 11,50 € per session. The experiment was carried out with 23 participants. Later we excluded three of them because their behavior had very little dependence on the stimulus.

The task was presented with Matlab Psychtoolbox 3.0.12. Participants made input with an USB-mouse that allowed them to precisely adjust the width of the response frame and confirm it with a click. Immediately after trial onset, they were

presented with the dots and could start to expand/shrink the frame from a random initial width by moving the mouse up/down-wards. The points remained visible throughout the entire time until the participant confirmed his response with a click. The program then either proceeded to the next trial or to the feedback/pause screen that indicates the averages over the five last trials of the percentage the participant would have captured as well as the numerical deviation score. In addition, information about how many of all trials have already been completed was presented. The participant could proceed at his own pace.

4.6.3 Computational Models

We attempted to examine whether the behavior of our participants can be described by inference of probability distributions. More specifically, we attempted to infer whether their internal structural assumptions correspond to unimodal near-Gaussian distributions (Fig. 4.2A) or might be better described by instance-based (nonparametric) approaches (Fig. 4.2B-D) such as kernel density estimation. In addition, we checked whether selected heuristics can also account for the behavioral data.

Response mapping accounts for nuisance factors

Behavior is influenced by various factors and subjective assumptions of the participant which are difficult to model explicitly. Among these are subjective prior knowledge and probability distortion. Even for a probabilistic agent there exists some mathematical freedom as to what prior distribution over the latent variables to use. We did not explicitly include prior knowledge into our models but instead endow the model with flexibility to approximately account for such effects.

We make use of the fact that ultimately, behavior such as the one derived from a probabilistic inference model just amounts to a specific mapping $\boldsymbol{d} \to \hat{y}$ from inputs onto the response $\hat{y}$. Generally, for probabilistic models, the mapping $\boldsymbol{d} \to \hat{y}$ can be written in two steps. (i) Computing the sufficient statistic $\hat{S}$ which is then (ii) mapped onto the response, $\boldsymbol{d} \to \hat{S} \to \hat{y}$, such as $\hat{S} = \sigma_{ML}(\boldsymbol{d})$ for the Gaussian. We use $\hat{S}$ to refer to any dispersion estimate and call $\hat{S} \to \hat{y}$ the response mapping. For nonprobabilistic estimators, it just allows for additional tuning of the dispersion estimate. The introduction of the response mapping permits the construction of computationally simple models that may accommodate subjective knowledge of latent variables, like σ, in the second step.

This is illustrated in Figure 4.3A for the theoretical response curves (red, green). For maximum likelihood estimation (MLE), the response (red) is nothing but a linear mapping of the sufficient statistic $\sigma_{ML}(\boldsymbol{d})$ onto its output $\hat{y}$. The Bay-

esian benchmark model (green) also takes the sample size $N = 4$ and a uniform prior distribution over σ into account. Compared to MLE, its main effect is a bias of the responses towards intermediate values. The effect of a different prior on σ would merely manifest as a somewhat different mapping onto the response because $\sigma_{ML}(\boldsymbol{d})$ and N are sufficient statistics for σ. In other words, the model will produce the same results even when input $\boldsymbol{d}$ changes as long as the sufficient statistics remain the same. They compactly sum up all the information that is to be known about the hidden variables of a probabilistic model from the sample $\boldsymbol{d}$. Hence, distributions such as the posterior $p(\sigma|\boldsymbol{d})$ or the prior $p(\sigma)$ do not have to be explicitly represented in our model. Instead they are implicitly considered through the effects they exert on the response by allowing for additional freedom through a mapping. Apart from that, the mapping $\sigma_{ML}(\boldsymbol{d}) \rightarrow \hat{y}$ also depends on the target percentage that the model is required to capture. A larger target percentage leads to a larger dependence on σ_{ML} and would e.g. manifest as a larger slope of the ML response (Fig. 4.3A, red). The model may however account for the fact that participants suffer from probability distortion such that their internal target probability does not exactly match the one of a probabilistic agent (Eq. 4.4).

The response mapping from the dispersion estimate to the response, $\hat{S}(\boldsymbol{d}) \rightarrow \hat{y}$, is chosen to be the same for all models and is intended to be flexible enough to jointly account for all these implicit effects. Empirically we found that a quadratic polynomial is only minimally better than a linear mapping (using the weighting-model, Sec. 4.6.3). The improvements on the group level are significant (increased median cross-validation log likelihood (CVLL) across participants, Wilcoxon signed rank test, $p = 0.0027$) but small in absolute terms (median CVLL difference 3.66 dHart, 95 %-CI $(0.34, 7.15)$ dHart, Sec. 4.6.3). For this weak nonlinearity and to obtain a sparse model formulation, we consider a polynomial of first order to be a sufficiently good approximation to represent the response mapping.

$$\hat{y} = \beta_0 + \beta_1 \hat{S}(\boldsymbol{d}) \tag{4.5}$$

The models that we consider differ only in how they compute the dispersion measure $\hat{S}$. They may introduce additional parameters which are detailed below. We start by describing approximative models that do not make use of distributions first. In addition, we will explicitly consider heuristic models. In general, heuristics are not linked to optimal responses in a principled way but might nevertheless yield satisfactory results. Every estimator that correlates with σ_{ML} contains some useful information about the dispersion and may thus be used. As heuristics are frequently associated with less effortful processing we consider simple and visually salient quantities that may be readily assessed by the participants. As another approximate model, we test a weighting model that emphasizes certain stimulus

features. We will then describe probabilistic models that derive responses from different distributional estimates and conclude with a predictive model intended to serve as an estimator of the upper bound on predictability given our data.

Maximum model

This model uses the distance of the point that is farthest away from the center, that is, $\hat{S} = \max(|\boldsymbol{d}|)$. This function can be considered a simple heuristic approach because it reduces the input information to be processed, but as this distance strongly correlates with σ_{ML} it is expected to be predictive of behavior.

Range model

This model uses a dispersion estimate based on the difference between the leftmost and rightmost point $\hat{S} = \max(\boldsymbol{d}) - \min(\boldsymbol{d})$. Again, this quantity is correlated with σ_{ML}.

Weighting model

The maximum likelihood estimator σ_{ML} can be generalized in that it assigns different weights to individual points when calculating the root mean square deviation. The observations $\boldsymbol{d}$ are indexed according to their excentricity, i.e. their absolute deviation from zero such that $|d_n| \geq |d_m|$ for $n > m$.

$$\hat{y}(\boldsymbol{d}) = \beta_0 + \hat{S}(\boldsymbol{d}) = \beta_0 + \sqrt{\frac{1}{N}\sum_{n=1}^{N} \omega_n d_n^2}, \quad \omega_n \geq 0 \tag{4.6}$$

The parameter β_1 of the response mapping $\hat{y} = \beta_0 + \beta_1 \hat{S}$ (Eq. 4.5) is factored into the ω_n and set to one to avoid under-constrained solutions for regression. We may enforce the summation constraint, $\sum_n \omega_n = N$, on the weights after fitting to interpret the weights as relative contributions with respect to the case of $\omega_n = 1$, which corresponds to inference of a Gaussian. This can be done by factoring out a term $\sqrt{N/\sum_n \omega_n}$ which can be formally assigned to β_1. We consider the equal weighting of the square of each point's position $\sigma_{ML} = \sqrt{1/N \sum_{n=1}^{N} d_n^2}$ a non-trivial pattern of inference of a normal distribution. Within this model, we also test the heuristic of considering just one out of all $n = 1, ..., N$ points, $\hat{S}(\boldsymbol{d}) = |d_n|$. In this case, just one of the four weights should be four while the others would become zero due to the summation constraint. The task is constructed such that the position of the third most excentric point closely corresponds to the target

percentage. Yet, we found that this heuristic is evidently exploited by just one participant (normalized $\omega_3' = 0.95$, d_3 almost explains full variance, $R^2 = 0.96$).

Because of the generality and the computational ease with which optimization can be performed for this model, we used it to test variants of the response mapping Eq. 4.5. We tested whether participants behave in accordance to a prior belief about the range of dispersions across trials. A pure ML approach ignores prior knowledge and leads to responses proportional to the dispersion estimate $\hat{S}(\boldsymbol{d})$ (Fig. 4.3A, red). If that was sufficient to predict behavior, a model whose output is restricted to be proportional to the dispersion estimate (omitting constant term in Eq. 4.5) should perform equally well.

$$\hat{y}(\boldsymbol{d}) = \hat{S}(\boldsymbol{d}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} \omega_n d_n^2} \tag{4.7}$$

Likewise, a model which additionally features a quadratic term $\hat{y} = \beta_0 + \hat{S} + \beta_2 \hat{S}$ is used to test for the nonlinearity of the response mapping. The weighting model is chosen for these tests as it can flexibly account for other systematic biases in behavior that are not related to prior knowledge.

Normal model

Making inference using a normal distribution is equivalent to the mapping $\boldsymbol{d} \rightarrow \hat{S} \rightarrow \hat{y}$ in which $\hat{S} = \sigma_{ML}(\boldsymbol{d})$ is the sufficient statistic and the MLE of the Gaussian. To match the responses of our benchmark model the response mapping $\hat{S} \rightarrow \hat{y}$ must equal the green curve in Fig. 4.3A. The chosen response mapping for regression Eq. 4.5 can only provide a linear approximation to this curve but was chosen based on considerations regarding model sparsity and the empirical evidence to be sufficient to capture behavior.

Generalized normal model

The dart metaphor and the task instructions suggest that the distribution of darts follows some symmetric and bell-shaped curve centered at zero. As a perfect match between the true distribution and the one that was assumed by our participants is not expected, we consider a generalized normal distribution which has an additional shape parameter $p > 0$ so that it can represent a larger family of distributions.

$$p(x|\mu, \alpha, p) = \frac{p}{2\alpha\Gamma(1/p)} \exp\left[-\left(|x-\mu|/\alpha\right)^p\right] \tag{4.8}$$

It effectively generalizes the exponent of the normal distribution for which it takes a value of $p = 2$. For small p the distribution is more peaked whereas it approximates a plateau like distribution for larger values (Figure 4.1B). We assume that the exponent parameter p is constant across trials and treat it as an additional fitting parameter. For a known mean of zero, $\mu = 0$, the maximum likelihood estimator for α is $\hat{S} = \left(p/N \sum_{n=1}^{N} |d_n|^p\right)^{1/p}$ which we identify with the dispersion estimate $\hat{S}$. In the limit of $p \to \infty$ it corresponds to the heuristic MaxAbs-model above. We also tested a generalized normal model which infers μ on a trial-by-trial basis for a given exponent p to test whether dropping the assumption of a centered distribution can better explain behavior. In this case, Eq. 4.4 is explicitly solved, and its result is assigned to $\hat{S}$. As it was found to be worse than the centered normalized distribution on the group-level (exceedance probability $p_{exc} > 0.999$) we chose to only report results using a centered distribution.

Gaussian kernel density estimation model

If one imposes only minimal structural constraints, more freedom is given to the data to determine the inferred density. One may assume that even small samples represent the population well and that future observations will cluster around the already observed instances. One way to do so is to estimate $p(x|\boldsymbol{d})$ over future events x based on a kernel method. It generalizes observed data points d_n by assigning probability density proportional to a kernel function $k(x, d_n)$ to their vicinity and thus constitutes a data smoothing problem (Fig. 4.2D). For the whole observational data $\boldsymbol{d}$, kernel density estimation centers a kernel on each observation and sums up their contributions to determine $p(x|\boldsymbol{d})$ as:

$$p(x|\boldsymbol{d}) = \mathrm{NP}(x|\eta, d_1, ..., d_n) = \frac{1}{N} \sum_{n=1}^{N} k(x|d_n, \eta) \tag{4.9}$$

It is a nonparametric method because it does not assume a certain parameterized family of probability distributions for $p(x)$ apart from the kernels. The kernel function k typically decays with the distance between x and d_n. Here we assume that it has the shape of a normal distribution $k(x|d_n, \eta) = N(x|d_n, \eta)$. The kernel width $\eta = \eta(\boldsymbol{d})$ is in principle a free parameter, but it needs to be sensibly chosen with respect to the dispersion of the data. Manual testing revealed that $\eta = a \cdot (d_3 + d_4)/2$, with fitted proportionality parameter a, is a reasonably good approximation to the unknown $\eta(\boldsymbol{d})$ function. Thus, potentially even better performance might be achievable than the one reported here. The model's dispersion estimate, $\hat{S}$, regarding the 65 % capture probability is determined by inserting the inferred distribution (4.9) into Eq. 4.2 and then solving Eq. 4.4.

In the limit of vanishing kernel widths, $\eta \to 0$ (δ-distributions), the response for the target percentage of $p_t = 65\ \%$ converges to the third most excentric point. We refer to this approach as δ-KDE (Fig. 4.2C). In this limiting case, one would merely capture the target fraction p_t of observed points on the screen, thus replacing an estimation of the target fraction p_t of the population with a corresponding estimation of p_t on the sample.

Tiling model

To capture a certain percentage of points of the sample one must have some sort of quantile function that outputs the region containing the desired percentage. Explicit density models such as KDE entail a quantile function. A simple alternative way is to construct some normalized histogram. We attempt to do so with the constraint that an observation point only exhibits a local effect on the constructed density (Fig. 4.2B). Specifically, the contribution to the overall density of one data point only depends on its own position and on the position of its adjacent points. More formally, this can be achieved by tiling the space between observations into rectangular, adjacent but non-overlapping basis functions. We adhere to the additional constraint that the N ordered points correspond to the $(0.5/N, 1.5/N, ..., (N-0.5)/N)$ cumulative quantiles. Hence, each basis function spanned between points has to be normalized by N. To assign the remaining probability $0.5/N$ below the lowest point d_1, we use a uniform distribution $\mathrm{U}(d_1 - d_2, d_1)$ whose support equals the distance to its only adjacent point d_2 (and likewise for the largest point). Representations of probability densities based on orthogonal basis functions are suggested as a solution to tractably represent complex densities [66].

Gaussian Process Regression

Gaussian Process Regression (GPR) [192] is used to estimate the upper bound on the predictability of our participants' behavior. It does not lend itself readily to an interpretation of how participants solve the problem on a given trial. It is however very flexible and successful in prediction by exploiting consistency between input $\boldsymbol{d}$ and output y across pairs of trials (i, j). We used GPR since it is a virtually bias free estimator of the distribution $p(y|\boldsymbol{d})$ which is assumed to be normally distributed with a constant intrinsic noise parameter σ_I. We chose a Gaussian kernel function

$$k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \theta \cdot \exp\left[-\frac{1}{2}\sum_n \left(d_{in} - d_{jn}\right)^2 / \sigma_n^2\right] \tag{4.10}$$

that defines a scalar measure of similarity and the entries of the covariance matrix of the GP as $C_{ij} = C(\boldsymbol{d}_i, \boldsymbol{d}_j) = k(\boldsymbol{d}_i, \boldsymbol{d}_j) + \sigma_I^2 \delta_{ij}$. Input pairs $(\boldsymbol{d}_i, \boldsymbol{d}_j)$ that are considered similar in this sense should result in comparable responses (y_i, y_j) if the process $p(y|\boldsymbol{d})$ is consistent. Prediction is more strongly influenced by those trials' responses y for which $(\boldsymbol{d}_i, \boldsymbol{d}_j)$ are similar. To make predictions for a new input $\boldsymbol{d}_\nu$, we evaluate the mean of the predictive distribution $\hat{y}(\boldsymbol{d}_\nu) = \boldsymbol{k}^T C^{-1} \boldsymbol{y}$. Here $\boldsymbol{k}$ has the entries $k(\boldsymbol{d}_i, \boldsymbol{d}_\nu)$ with i indexing all trials in the training data. Likewise, C and $\boldsymbol{y}$ are constructed from all the training data used to derive predictions. For each trial $\boldsymbol{d}_t = (d_{t1}, ..., d_{tN})$ symmetry is exploited by sorting the points in ascending order of excentricity. To set the hyperparameters of the GP, $(\theta, \sigma_1, ..., \sigma_N, \sigma_I)$, its generalization error is minimized. To do so, the mean of the test sets of Eq. 4.13 of a 5-fold cross validation (CV) procedure is calculated. This procedure is part of training the GPR. We also attempted to predict behavior using a simple 1-hidden-layer feedforward neural network. Despite being a successful predictor, its performance was inferior to the GPR which is why we chose to only report the latter.

Baseline model

The baseline model is chosen to provide a simple lower bound estimate for predictability that is independent of the trial-by-trial variations of the stimulus. This model calculates the mean of the responses of all its input $\boldsymbol{y}_{\text{in}}$ (training set). It thus makes the same prediction on every trial t.

$$\hat{S}_t = \langle \boldsymbol{y}_{\text{in}} \rangle \tag{4.11}$$

Inter-trial and feedback dependence

We investigated the influence of other quantities on behavior that participants might have (erroneously) utilized to guide their responses. To test for a dependence on the preceding trial, the estimator $\hat{S}$ is chosen to be the previously stated response.

$$\hat{S}_t = y_{t-1} \tag{4.12}$$

There is a significant effect with respect to baseline (exceedance probability, $p_{exc} > 0.99$) yet the effect on behavior is virtually negligible as the overall predictive performance is very low (median cross-validation log likelihood across participants -318 dHart, $95\,\%$-CI $(-356, -300)$ dHart, with respect to the best model for each participant). The influence of the previously presented feedback about the capture percentage is similarly tested but its effect is found to be even weaker.

Table 4.1: Overview of model parameters

Model	Abbreviation	Fitting parameters
Maximum	max	β_0 β_1
Range	rng	β_0 β_1
Weighting	wgt	β_0 ω_1 ω_2 ω_3 ω_4
Normal	nm	β_0 β_1
Generalized normal	gnm	β_0 β_1 p
Kernel density estimation	kde	β_0 β_1 a
Tiling	tlg	β_0 β_1
GPR	gpr	Nonparametric; hyperparameters: $(\theta, \sigma_1, ..., , \sigma_N, , \sigma_I)$

Overview of model parameters

The models used have a different number of parameters depending on the dispersion estimate $\hat{S}$. The ones reported later are summarized in Table 4.1.

The response distribution

The probability of obtaining the response y_t on trial t conditional on the data $\boldsymbol{d}_t$ and the model parameters is assumed to be a mixture distribution of two contributions. The first and dominant term is a normal distribution centered on the model prediction $\hat{y}_t$ modeling task intrinsic noise around the estimates. Upon preliminary inspection of the data we found considerable heteroscedasticity with higher unexplainable response variability for larger sample dispersions.

To account for this feature of the response data, we assume that the standard deviation (SD), θ, of the distribution over response y_t, $N(y_t|\hat{y}_t, \theta(\hat{y}_t))$, is a function of the model output $\hat{y}_t$. The model output is denoted by $\hat{y}$ to distinguish it from the response y of the participant which is formally represented by a draw from the response distribution to account for behavioral variability. Instead of assuming a parametric relationship and the need to include further parameters, we make a parameter free estimate by assuming a discretized function, as follows. We divide the whole model output $\hat{\boldsymbol{y}}$ into Q equally filled quantiles $q \in \{1, ..., Q\}$ by assigning trial t to quantile q_t. For every quantile q the SD is estimated separately by calculating $\theta_q = (\sum_j (y_j - \hat{y}_j)^2 / J)^{1/2}$ ($j = 1, ..., J$ indexes trials belonging to quantile q). Hence, whenever there is heteroscedasticity, the true function $\theta(\hat{y})$ is approximated by the estimated bin values. For homoscedasticity all θ_q are the same and collapsing bins would make no difference. The resolution of the function is higher when many quantile divisions are used provided the θ_q can still be estimated faithfully. We consider $Q = 5$ a suitable choice for our problem. As our

data might be contaminated by processes other than dispersion estimation, such as lapses, we take precaution against far outlying responses. We calculate a trimmed standard deviation, i.e. before calculating θ_q we remove values below or above two interquartile ranges from the lower or upper quartiles respectively. However, this applies to θ_q estimation only. No points are removed from calculating the response likelihood

$$p(\boldsymbol{y}|\boldsymbol{d}_1, ..., \boldsymbol{d}_T) = \prod_{t=1}^{T}(1-\epsilon)N(y_t|\hat{y}_t, \theta_{q_t}) + \epsilon \; . \tag{4.13}$$

Additionally, to prevent isolated points from being assigned virtually zero probability we generally add a small probability of $\epsilon = 1.34 \cdot 10^{-4}$ to all. This corresponds to the probability of a point at four standard deviations from the standard normal distribution. For non-outlying points this alteration is considered negligible.

Estimating model evidence

The evidence that each participant's data lends to each model is derived as its predictive performance in terms of the cross-validation log likelihood (CVLL). For training, we maximized the logarithm of the response likelihood (Eq. 4.13). To maximize the chances of finding the global maximum even for non-convex problems or shallow gradients, every training run first uses a genetic algorithm and then refines its estimate with gradient based search (MATLAB ga, fmincon). The CVLL for each participant and model is summarized by the mean of the logarithm of the response likelihood (Eq. 4.13) on the test set across all cross validation (CV) folds.

As cross validation (CV) is a computationally expensive method, we use a random 5-fold split of data into training and test sets such that each training point is used four times for training and once for testing. However, to make splits more representative of the sample, we use a stratified version of CV by ensuring that the mean target variable is approximately equal in all folds. This is done by assigning data points to one of the 8-quantiles of the distribution of the target variable. We constructed strata that contain one value from each quantile. Subsequently, we sampled strata to create the 5-fold CV splits. To improve the reliability of per participant estimates of the model evidence (CVLL), we repeated this procedure with different random splits and aggregated the output so that in total 10 CV splits are performed for each participant and model.

Differences in model evidence, Δ, are reported on a log-scale in decibans (also decihartleys, abbreviated dHart) that may be used to interpret the significance of the results of individual participants. According to standard conventions, we consider a value of $5 > \Delta$ barely worth mentioning, $10 > \Delta \geq 5$ substantial,

$15 > \Delta \geq 10$ strong, $20 > \Delta \geq 15$ very strong and $\Delta \geq 20$ decisive.

Group level comparison

Instead of making the assumption that all participants can be described by the same model we use a hierarchical Bayesian model selection method (BMS) [184] that assigns probabilities to the models themselves. This way, we assume that participants may be described by different models. That is a more suitable approach for group heterogeneity and outliers which are certainly present in the data. The algorithm operates on the CVLL for each participant ($p = \{1, ..., P\}$) and each model ($m = \{1, ..., M\}$) under consideration and estimates a Dirichlet distribution $\mathrm{Dir}(r|\alpha_1, ..., \alpha_M)$ that acts as a prior for the multinomial model switches u_{pm}. The latter are represented individually for each participant by a draw from a multinomial distribution $u_{pm} \sim \mathrm{Mult}(1, \boldsymbol{r})$ whose parameters are $r_m = \alpha_m/(\alpha_1 + ... + \alpha_M)$. We use the CVLL and assume an uninformative Dirichlet prior $\boldsymbol{\alpha}_0 = \mathbf{1}$ on the model probabilities. Later, for model comparison, exceedance probabilities, $p_{exc} = \int_{0.5}^{1} \mathrm{Beta}(\alpha_i, \sum_{j \neq i} \alpha_j)$, are calculated corresponding to the belief that a given model is more likely to have generated the data than any other model under consideration. High exceedance probabilities indicate large differences on the group level. We consider values of $p_{exc} \geq 0.95$ significant (marked with $*$) and values of $p_{exc} \geq 0.99$ very significant (marked with $**$).

Chapter 5

General discussion

5.1 Summary of contributions

The introductory chapters argued that task-optimality is an insufficient criterion to assess the rationality of human inferences. This is due to subjective assumptions leading to problem mismatch and due to competing goals in combination with cost-sensitive cognition. As a consequence, tests for theoretical rationality must more specifically assess Bayesian belief updating while respecting the individual internal boundary conditions [116]. I sought to estimate the human potential to resort to rational inference mechanisms. For this purpose, I conceived and developed two novel experimental paradigms on my own. These tasks bridge low-level perceptual and higher-level cognitive domains in an attempt to expand experimental evidence beyond commonly tested basic visuo-motor tasks (Sec. 2.2.2).

The experimental responses show non-trivial patterns specific to internal probabilistic processing while the inference procedures involve complex and non-linear operations such as normalization and marginalization. Behavior is found to be stable across trials without relying on supervising feedback, suggesting that actions are guided by internal objectives derived from internal representations of uncertainty. While, the results were tested against many other conceivable approaches, simple heuristics are typically insufficient to account for behavior. Consequently, mechanisms that at least approximate probabilistic inference are suggested to be available for similar higher cognitive tasks.

Beyond that, behavior is highly consistent with a jointly learned representation at several levels of a hierarchy in which upper contextual levels constrain inferences of lower-level latent variables. Moreover, information integration can be well captured with reliability-dependent message passing between latent variables of a generative model suggesting that representations of uncertainty are ubiquitous.

Group-level results or well-performing individuals demonstrate that both tasks are cognitively feasible. However, the failure of some individuals to at least remotely perform the task suggests that proper problem alignment is crucial. Problem mismatch was partly made explicit by model-based analyses but also during briefing and debriefing of our participants. Systematic misconceptions and behavioral biases are very heterogeneous across participants and supposedly stem from extraneous factors.

The most notable inferential bias is a tendency of the momentary sample to dominate judgments against previously available information. The second experiment allowed to explicitly link this behavioral observation to an internal representation which tends to be overly dominated by the momentarily observed instances. Bayesian nonparametric approaches to density estimation were used for modeling and are suggested as a connection to further theoretical developments (see model selection, Sec. 1.4).

Crucial ideas and questions that this work addressed are further discussed in the following sections while the task-specific discussions can be found in Secs. 3.4 and 4.5.

5.2 Model-based probabilistic inference

Probabilistic inference is inherently model-based. The functioning of many everyday mental abilities such as counterfactual thinking, imagination, dreaming, prediction and planning is hard to conceive without reliance on a model that may go beyond the data that has ever been observed. There is scientific evidence that human decision making is model-based and not consistent with model-free learning [175]. Model-based inference may occur largely unconsciously and is suggested to underlie complex tasks such as physical scene understanding [177].

There are increasingly many accounts that attempt to interpret neural processing as an inference process of the causes behind their bodily influences [193]. Besides evidence from cognitive science [42], convergent evidence in support of the framework of Bayesian hierarchical inference [74] has indeed led to the development of quantitative and testable models of implementational (neural) aspects of brain function [96]. For instance, a recent neuro-imagining study has reported the preactivation of stimulus templates by expectations that is similar to actual stimuli [179]. Such top-down information flow is characteristic for model-based, hierarchical inference and for brains that are driven to adjust to best predict stimuli [43, 58].

The results of the present work support the notion that humans rely on probabilistic models of their sensations. Judgments about uncertainty, such as estima-

ting the probability of a correct decision (Bayesian decision confidence), require a representation of the possible worlds that are consistent with the data - even of those which are not most strongly supported. We found that many participants naturally chose to express "their" confidence as a quantity tightly related to Bayesian decision confidence.

This is a very difficult task for a non-probabilistic, model-free agent as argued before (Sec. 2.1.2). In the absence of a probabilistic world representation, estimating the frequency of occurrence of possible worlds requires re-experiencing a situation conditional on the observations. Even for few, stable and repetitive conditions with correctly supervising feedback, a model-free agent needs many repetitions to reach comparable levels of behavioral accuracy (Sec. 2.1.2, Fig. 2.1). From birth, we are exposed to a plethora of extremely complex situations in different contexts, whose latent structure is not signaled so that we must learn in an unsupervised manner. To make things even more difficult, many situations, under the same environmental distribution, are just experienced once. In the light of these arguments, the frequentist approach to estimating uncertainty appears very unecological.

Another argument in favor of perceptual models concerns the transfer of knowledge [194–196]. Models are a task-independent representation of the environment. But of course they may be used to construct task specific objectives. The extension of the second experiment exemplifies this because the read-out of the sensory representation has to adapt to the momentary target capture percentage. Such representations may be efficiently re-used for different tasks. On the other hand, directly learned stimulus-response mappings typically exhibit a high degree of task specificity and are thus of limited use when behavior must transfer to new environments or objectives (see also Sec. 2.1.2). Our participants transfered rapidly and successfully to more complex environmental distributions in both tasks. These powerful generalization abilities suggest that they make use of internal models that are inferred based on the task instructions.

The results of the empirical prior study are highly consistent with message passing between latent variables across a hierarchy and suggest that uncertainty information is ubiquitously available for reliability-based integration [197]. A feature of probabilistic representations is that uncertainty estimates can naturally emerge from the knowledge representation itself, without requiring a meta-representation [181, 198]. The degree to which the biological brain and human cognition rely on a separate re-representation (or meta-representation) is unclear (see also [199]). Some animal studies have claimed an anatomical locus for a meta-cognitive reports that is distinct from the processes required for perceptual decisions [165]. Other studies using causal neuro-physiological tests reported that reasoning on the object- and the meta-level could not be dissociated [130, 200] so

that overall evidence is inconclusive.

5.3 Model selection problem

The question what internal model humans select and how they generate new hypotheses is experimentally difficult to test and hence poorly understood. Many times, participants are simply assumed to know the problem structure which is tied to the faulty practice of equating task-optimality with theoretical rationality (Sec. 1.6.5).

Humans were speculated to monitor uncertainties about the world's causal structure [197]. Experimental evidence suggest that human inferences possess the ability to select among several structures (models) [183, 201], especially if the set of candidate structures is clear [202]. The empirical prior learning task confirms this notion (Chapter 3). Even more, our participants learned and rationally used a representation of the respective uncertainty of different structures. A similar information processing scheme was suggested to underlie human vision: "recurrent feedforward/feedback loops in the cortex serve to integrate top-down contextual priors and bottom-up observations so as to implement concurrent probabilistic inference along the visual hierarchy" [180]. The idea that humans internally maintain generative models, akin to Bayesian hierarchical inference in which higher-level variables constrain lower-level states, may extend to more abstract concepts beyond vision. Apart from behavioral data in our task, there is evidence, e.g. from a brain imaging study of a hierarchical planning [203], in support of this notion.

Every task is specified by a number of implicit and explicit assumptions. First, participants must infer the structure of the problem itself from the task description. In this process, description-based methods may fail to entirely communicate the problem, e.g. the base rate information as the specification of distributions is high-dimensional and difficult to convey in words. Unfortunately, this process is hard to test in a rigorous and controlled manner. While we only informally addressed this by making short, Q&A sessions before and after the experiment, it would be interesting to explore this more rigorously with a larger number of participants.

Systematic reasoning errors are suggested to arise to a large extent because inference based on a somewhat mismatched model of the environment is made which can lead to severely biased estimates [168]. Even though we attempted to clearly convey the dependence structure among the latent variables in the empirical prior task, the problem is not completely transparent to the participants, as we could e.g. not communicate the strength or magnitude of the block tendency to them. To reduce uncertainty, they may make subjective assumptions or even infer it across trials by an upward extension of the hierarchical latent structure (see Sec. 3.3.4, Fig. 3.3). In such an attempt to infer the problem, participants may

supplement the instructions with automatic but inappropriate assumptions. In this context, another study has explicitly attributed task suboptimality to structural learning [201]. Different subjective assumptions may largely explain the substantial inter-individual differences found in behavior. As the exposure to learning situations over the entire life is individually specific, different 'empirical priors' might emerge even if we were entirely rational agents (see Sec. 5.6 below).

After all, the assumption of a transparent problem in inference tasks is difficult to satisfy. The chances are high that participants attempt to solve a slightly 'mismatched problem'. Structural problem alignment is absolutely crucial and is claimed to increase compliance with Bayesian norms [204]. Particularly, the success to evidence intricate patterns of probabilistic inference in this work is believed to stem from careful instructions of incremental complexity (empirical prior task, Secs. 3.5.3-3.5.4).

Increased structural uncertainty such as a "small unusual twist or additional element of complexity" can adversely affect performance [140]. Similarly, we could ask what would happen if participants were not instructed and instead just exposed to the block tendency in the empirical prior task. We did not test this specific case but consider it an interesting extension that may further clarify the role of structural uncertainty.

More generally, when the problem structure behind the observations is highly uncertain, a nonparametric, instance-based approach to inference is sensible. This might actually be one explanation for instance-biased generalizations that were observed in the second experiment even though participants typically stated that the task was clear. Considering also the simplicity of the sampling process, this explanation seems less likely.

In the absence of structural knowledge, another study reported that participants resort to model-free behavior [60]. We however did not give supervising feedback to our participants which precludes a model-free approach. This, on the other hand, might explain the failure of few participants to even remotely engage in the tasks.

5.4 Biases through approximations

We sought to increase the motivation of the participants to engage in the task by using sparse motivational feedback and by posing economic incentives through bonus payments. This presumably leads to a high priority of performing rational, optimized inference (Sec. 1.5). It was intended to preclude that participants resort to computationally cheap short-cuts or heuristics which may introduce behavioral biases (Sec. 2.2.1).

Human inferences have been interpreted in terms of a sampling approach to inference when drawing samples is assumed to be costly [84]. Many cognitive biases, such as a base rate neglect, are reproducible by a sampling framework [82] or they are suggested to originate from noisy internal processing [205, 206]. A substantial part of the overall behavioral variability might actually be introduced by the inference process itself [87, 145]. For inference through finite sampling, for instance, behavioral response variability should covary with the width of the posterior distribution [86]. In our tasks, response variability introduced at the inferential stage cannot easily be disentangled from other noise sources, e.g. introduced through motor control. Thus, unfortunately, we cannot make claims about the use of sampling-based approximations here which is a limitation owing to compromises in the design.

In the second task, we found a tendency for instance-based generalizations. Such nonparametric or instance-based approaches to inference may also be considered an approximation because they detach the problem from the context. If they can be assumed to be cognitively cheaper, there is a practically rational incentive to use them even at the expense of poorer inference. The idea of having both an instance-based encoding scheme and a structured one, which allows for powerful generalizations, underlies recent advances in artificial intelligence [207]. These complementary learning systems parallel the functional roles that have been attributed to the hippocampus and the neocortex respectively in biological brains [208]. However, the degree to which they rely on spatially separated cognitive systems in the brain is not clear [209]. Generally, it may be beneficial to forgo precision of the world representation whenever it is not strictly important for the momentary task objective. Flexible schemes, such as utility-weighted sampling [210], may allow for a gradual incorporation of approximations into just one learning system. There might actually be a similar interaction between instance-based and structured learning systems that provide the hybrid functionality of Bayesian nonparametric models.

There is a remarkable parallel between the findings of both experimental tasks here in that behavior is a distorted, nonlinear function of optimal probability estimates. This could merely be caused by concurrent or extraneous processes such as motor control. On the other hand, it could reflect systematic biases of the inferred representations. The second task (Chap. 4) suggested that an instance-based representation leads to distorted reports in that objectively small target percentages are overstated while large ones are understated. In the first task (Chap. 3), we observed a similar form of probability distortion. Decision confidence was too low for objectively hard trials and too high for easy ones. At the same time, an increased reliance on the momentarily sample was found through the use of under-constrained internal representations. It is not clear if this is coincidental, or

if it may indicate that participants also construct internal representations which are overly influenced by the very instances of the data they were exposed to. Particularly, because there is evidence for a tendency of the sample to dominate judgments on the level of their behavioral responses. Correspondingly, to elaborate on this, modeling efforts would e.g. have to focus on hierarchical extensions of Bayesian nonparametric methods.

Overall, these ideas are speculative and their experimental testing probably requires adaptations in the task design, e.g. to control the extent that participants have an incentive to resort to approximations. Nonetheless, this is a vast area of active research with potentially beneficial interchange between artificial intelligence and cognitive neuroscience [207].

5.5 Generalization biases

The behaviorally found bias of the sample to dominate judgments corresponds to a form of dominance of bottom-up influences in a hierarchical model. It was linked to internal instance-based representations which give too much freedom to the model's structure to adapt to the sample. The origin of instance-biased generalization is suspected to be due to structural uncertainty (Sec. 5.3) or the use of effort-reducing approximations (Sec. 5.4).

However, there might be a third explanation which rests on a more fundamental, ecological argument. We might be equipped with a fundamental bias to perceive causes behind our observations even for little evidence (e.g. [109]). A virtue of (partially) instance-based methods such as Bayesian nonparametric models is that they may readily incorporate structural changes such as new causal factors. This might be adaptive in a structurally uncertain and changing world compared to parametric models of fixed structure.

On the other hand, completely instance-based approaches are inefficient because they ignore contextual information. We found behavioral evidence that inferences are more constrained than purely instance-based estimates. This is reminiscent of recent accounts of function learning [189] claiming that humans follow hybrid approaches between instance-based (nonparametric, similarity based) and structured (rule-based) methods. The second study here is limited in that it provides only weak evidence that internal distributional representations follow such a hybrid approach.

To generalize well in realistic and thus uncertain environments, an agent must disentangle noise from systematic patterns. Bottom-up sensory stimuli need to be integrated with top-down prior expectations about the context of the task. To strike a balance between both, the perceptual representation needs to be equip-

ped with the right degree of flexibility to adapt to the observations. Theoretically, uncertainty provides a key link between top-down and bottom-up influences as partly contradictory information must be combined into the posterior (see e.g. [43]). Experiments have confirmed the importance and the use of uncertainty representations for multi-stage [211] and hierarchical decision making [187]. Excessive instance-based generalization corresponds to a bottom-up surplus giving more emphasis to the sample at the expense of prior (structural) knowledge. In extreme cases, insufficient integration might lead to incoherent beliefs. Interestingly, attempts have been made to explain the positive symptoms of schizophrenia in terms of disturbed belief updating in a hierarchical Bayesian framework [212]. Internal representations of schizophrenic patients are commonly considered too fragmented which bears certain resemblance with the consequences expected from excess bottom-up dominance [213, 214].

This work suggests a link between bottom-up dominance of information flow and instance-based computational representations. At the same time, it is unclear why an instance-based generalization scheme tended to dominate in both experiments. This is not expected to be general as expectation-biased generalization through top-down dominance undoubtedly exists in human inferences. It might for example be a consequence of the rather dull, low-value and repetitive nature that is common to many laboratory tasks.

5.6 Are we rational agents?

The necessity to handle uncertainty has initiated the probabilistic turn to describe sound reasoning [31] - a paradigmatic shift away from the formalisms of traditional logic and towards probability. Theoretically rational inferences are not about a task result, they are about a method of inference.

We evidenced patterns of variation that are highly specific to probabilistic (Bayesian) belief updating which suggests that human participants have the potential to maintain coherent reliability-based beliefs over complex latent structures. These results support the thesis formulated in the beginning that humans have access to internal mechanisms of rational inferences.

Irrational responses, on the other hand, can result from an agent not making full use of its available resources (see also [119]). In this sense, rationality is not dichotomous but a continuum. The observed instance-based generalization approach is somewhat theoretically irrational unless there are justifiable reasons to believe that e.g. the environment undergoes drastic changes and that the previous context is not applicable anymore. Context-detached reasoning may lead to incoherent beliefs, e.g. by not taking all available prior knowledge into account.

Hence, "local rationality" would actually be a more descriptive term in such cases.

Practical rationality concerns the degree to which cognitive resources are allocated to achieve an internal set of goals, i.e. whether meta-reasoning is rational. As opposed to theoretical rationality, no claims regarding practical rationality can be made here because the internal motivations of our participants are opaque. In the light of "resource-rational analysis" [35], an habitual response may be more rational than a deliberate one if the avoidance of fatigue outweighs the expected losses from fast and frugal performance. A top-down reduction from a principled approach, such as an approximation, may be practically rational but outwardly appear like a theoretically irrational heuristic. Generally, however, I am skeptical whether meta-reasoning can be understood in an overarching rationality framework as it would require quantifications of the costs of possible approximations (see [215]).

It is possible that many of the task-irrationalities found by the heuristics and biases program can be explained away by structural problem mismatch (see e.g. [216–219]). The high-level tasks of economic decision making, which arguably cover an important domain of real life, are suspected to be particularly susceptible to misunderstandings and wrong implicit assumptions. Our results support the notion [133] that many biases may rather be due to a deficiency of reporting than a fundamental inability to reason probabilistically (Sec. 2.2.3).

Instead of targeting explicit deliberation, both tasks were designed to be intuitive. As an example, I avoided the use of number and discrete scales to facilitate undistorted, intuition-based reporting and I attempted to preclude explicit mental deliberation such as arithmetic. Our participants typically could not give an explanation about how they performed the task as evidenced by debriefing. For instance, the quantitative match of the magnifying effect of sample size on decision confidence is mathematically too complicated to be carried out explicitly. We take this as an indicator of intuitive reporting even though the concept certainly needs elaboration. Additionally, the sequential, repetitive exposure across trials is hypothesized to lead to a natural familiarization with the task (see also [220]). Similarly, we do not believe, but cannot provide evidence against, that explicit knowledge about mathematics and statistics has any influence apart from understanding the problem structure in our experiments.

Generally, the measures that we apply to test behavior should be robust to 'rational biases' [121], i.e. to extraneous factors and idiosyncrasies such as prior distributions that are unequal to the actual base rates [220]. Correcting for those confounds, the "ability to make decisions seem rather good, although not perfect, in both sensory-motor and cognitive domains" [182]. Nevertheless, abstract domains might still be more susceptible to problem mismatch so that care must be applied when interpreting systematic deviations from normative task behavior [80,

221, 222].

Critics may object that one might, in principle, always construe an objective whose optimum corresponds to the explanandum, i.e. rationalize all behavior or inferences. "Bayesianist" models have been criticized for having excess freedom of fit and for lacking falsifiability (e.g. [191, 223]). However, the present findings also clearly demonstrated the limitations that formal Bayesian models have. We could clearly reject probabilistic (Bayesian) models by implementing a wide array of (non-Bayesian) models that incorporate different assumptions. As always, Bayesian models are only appropriate if their assumptions can be justified. Nonetheless, the fact that the assumptions have to be made explicit is believed to be a virtue of the Bayesian framework (see also [224]).

After all, we seem to possess the potential for rational inference but we often fail to use it appropriately. This is remarkably similar to findings regarding behavioral control deficits of patients suffering from obsessive-compulsive disorder who "develop an accurate, internal model of the environment but fail to use it to guide behavior" [225].

5.7 Beyond this work

This work pioneered two experimental paradigms from the ground up. They are expected to hold potential for further contributions to the scientific community. Extensions could focus on scrutinizing the origin of the sample dominance by e.g. varying structural uncertainty in the instructions. Ways of controlling motivation or effort, e.g. through adaptations similar to demand selection tasks [106], may reveal whether cognitive approximations are responsible for (generalization) biases. Some physiological or brain related measures might provide important further clues. Pupil dilation for instance is reported to be linked to uncertainty induced arousal [226, 227], and might serve as a proxy for cognitive effort. Electroencephalographic correlates of subjective decision confidence [228] and correlates of hierarchical probabilistic inference [47] have been found. Similarly, it might be fruitful to explore traces of bottom-up dominance or the interaction of complementary learning systems with brain imagining techniques.

Interindividual differences have not been properly explored here because of the relatively small number of participants tested. For instance, the data is suggestive that the extent of sample dominance leads to shorter response times which might be a proxy for task engagement. Held against common statistical standards, this relationship slightly misses conventional significance measures, but it might well bear out to be robust for larger samples. Similar measures might provide useful clues of the origins of behavioral biases.

Across-participant variation and idiosyncrasies are indeed very common and should be taken into account [157, 166, 182, 229]. Studies have suggested a relation between behavioral biases and more general cognitive abilities [230], e.g. the extent of executive control to engage in cognitively demanding tasks [231]. In addition, truthful inferences are the key enabler to reach whatever goals in any environment. General inferential ability is thus the basis for intelligent behavior which "measures an agent's ability to achieve goals in a wide range of environments" [232, 233]. It is tempting to ask if the presented tasks could be elaborated to provide insight into more general cognitive abilities or disabilities such as those of clinical populations suffering from schizophrenia [212].

The field of research to which this work contributed may be seen as a (partial) resurrection of the notion of humans as "intuitive statisticians" [234]. Science shapes our world by enabling technology and by shaping our perception through the terms it coins. The very terms that we use to convince and influence one another. Opinions about human rationality tend to oscillate between its affirmation and its polar opposite. Without doubt, our rational capabilities are easily disengaged or taken over by more automatic control mechanisms which are part of our evolutionary heritage and which were maybe rational under ancient conditions. One may consider this a maladaption to an ever more complex world in which conflict resolution has to increasingly rely on justified arguments.

Currently, it is almost a truism that our thinking and action is deeply flawed. This is a frequently used argument and rationalization for paternalistic political measures. For instance, the idea that we almost need to be protected from ourselves and "nudged" to make better choices has led to the Nobel Memorial Prize in Economic Sciences in 2017 (e.g. [235]). However, we should be critical towards readily accepting such a strong societal top-down bias and thus ultimately an aggregation of power. It calls for a proper justification 'who' is ascribed this authority, its checks and balances, and that it is not driven by self interest. On a societal level, the current implementations and accepted procedures to aggregate beliefs (see also [236–238]) and to deliberate about desirable outcomes need amendments.

In this respect, I would like to point to the underexplored potential of artificial intelligence and machine learning. Can we manage to put artificial intelligence into our service as a more objective, interest-free "nudge" towards rationality and hence more truthful decisions? Could it help us improve our inferences, identify inappropriate assumptions or simply remind us to always remain a little bit doubtful about overly strong conclusions? Would it be desirable and acceptable, if we all had access to such tools with the ease of use of a spell checker? We all may thus dedicate more time to truly achieving our goals. As

a beneficial side effect, successful goal alignment would depend on 'cognition-computation interaction' in which I see a vital role especially for one: Humans.

"Not to be absolutely certain is, I think, one of the essential things in rationality."

Bertrand Russell

Appendix A

Supplements to the introduction

A.1 Probabilistic formalism

Probability is a measure that is assigned to an event. An event can be any subspace of the set of possible outcomes. This allows both discrete and continuous outcome spaces to be treated within an axiomatized formulation known as measure theory (e.g. [239]). As a formal theory, probabilities can be defined over (almost) arbitrary sets or objects making it an extremely versatile tool. The basic Kolmogorov axioms are:

1. The probability $P(E)$ of event E is a positive number: $P(E) \geq 0$
2. Assumption of unit measure. At least one of all outcomes O will occur: $P(O) = 1$
3. Assumption of σ-additivity. The probability of the union of countably many disjoint sets E_j is: $P\left(\cup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} P(E_j)$

In addition, these laws of probability may be derived (with the restriction of finite additivity of the third axiom) by making basic common sense assumptions about coherent reasoning (Cox's Theorem [21]). This has led some scholars to consider probability theory a natural extension of propositional logic to uncertainty [20].

For practical purposes, a probability distribution assigns probabilities (or density) to possible outcomes $D \in O$ and is typically some parameterized function $p(D|\text{parameters})$ over the space of outcomes O. Direct consequences of the basic laws of probability are the sum and the product rule for the joint distribution $p(X, Y)$ over two random variables X and Y.

1. sum rule: $p(X) = \sum_Y p(X, Y)$
2. product rule (chain rule): $p(X, Y) = p(X|Y)p(Y)$

Here, we assume they are discrete and that $\sum_Y$ sums over the space of outcomes of Y. For continuous random variables, summations must be replaced by integrations to obtain the respective expressions. The sum rule describes the marginalization operation to obtain the marginal distribution over X unconditional to Y. The product rule says that the joint distribution can be written as the product of the conditional distribution of X given Y and the marginal distribution over Y. Another consequence is Bayes theorem which relates two complementary conditional probabilities.

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} = \frac{p(Y|X)p(X)}{\sum_{X'} p(Y|X')p(X')} \tag{A.1}$$

It is however not specific to the Bayesian interpretation of probability (see *Bayesian probability*). In principle, the normalization in the denominator can always be obtained by summation (or integration). Bayes theorem A.1 forms the basis for statistical learning as it can be used to update unobserved (latent) variables of a model.

Bayesian probability

The question to what quantity in the world the mathematical concept of a probability refers has been subject to philosophical debate. This problem of reference is reminiscent of the above division of the external state of the world and the internal surrogate that is constructed through perception. Proponents of the Bayesian interpretation take a subjectivist stance. Probability is regarded as a measure of the degree of belief that an agent has towards a proposition. It does not require a random process to be present. This evidential interpretation is different from the physical interpretation which relates probabilities to random physical processes. The latter are roughly subdivided into the empirical, frequentist explanation and the causal propensity account.

The Bayesian interpretation assigns probability to each statement even in the absence of prior evidence. The Bayesian interpretation may itself be further divided into an objective and a subjective variant. Proponents of the objective Bayesian interpretation argue that the choice of the prior cannot be arbitrary as requirements of rationality and consistency should impose common bounds on agents to share essential features - measuring the plausibility of a proposition across agents. Subjectivists on the other hand do not view the *a priori* belief space so strongly constrained.

A.2 Comparison with a model-free learner

The agent is supposed to provide an accurate estimate of the probability that its decision will turn out to be correct (Sec. 2.1.2). In other words, it should give calibrated confidence judgments that correspond to fraction of correct decisions in the long run.

The difficulty of making choices crucially depends on the sample size. For each trial, we chose it to be a draw from an independent uniform random variable over small sample sizes from six to 12. The task consists of independent instantiations of this problem in each trial whereby the latent proportion of blue and red items in the urn varies according to a constant and symmetric Beta$(4, 4)$ distribution.

Agent A1 uses a Beta-Binomial model for inference (Sec. 3.5.6) and computes confidence as expected accuracy (Eq. 2.1). It estimates a blue majority when its corresponding decision confidence is higher than vice versa (ties are broken at random). Agent A2 on the other hand learns a mapping $c_N(q, \boldsymbol{\omega}_N)$ by tuning the parameters $\boldsymbol{\omega}_N$ for each sample size N encountered.

$$c_N(q, \boldsymbol{\omega}_N) = 1/\left(1 + \exp\left[-\left(\omega_{1,N} \cdot |q - 0.5| + \omega_{2,N} \cdot |q - 0.5|^3\right)\right]\right) \quad \text{(A.2)}$$

It is endowed with knowledge of the symmetry of the problem, i.e. decision confidence for a sample with the sufficient statistics $(N_R = 1, N_B = 3)$ should be the same as for $(N_R = 3, N_B = 1)$. This allows the agent A2 to reduce the number of effective observable conditions to the decision-aligned sample proportion $\tilde{q}$ and thus to pool data for more efficient learning.

Nevertheless, batch learning requires agent A2 to memorize the number of correctly made decisions out of all the decisions that were made under this condition $(\tilde{q}, N)$. The objective is to minimize the error of the sum over the contributions from all previous decisions $J = (c(\tilde{q}, \boldsymbol{w}) - y)^2$. The feedback variable is $y = 1$ for correct and $y = 0$ for false decisions.

On-line learning through stochastic gradient descend uses small updates proportional to the gradient computed from the preceding data point only.

$$\boldsymbol{\omega}_t = \boldsymbol{\omega}_t - \eta \nabla_{\boldsymbol{\omega}_t} J_t \quad \text{(A.3)}$$

Differentiation of Eq. A.2 with respect to both weights yields

$$\begin{aligned} \frac{\delta J}{\delta \omega_1} &= 2\left(c(q, \boldsymbol{\omega}) - y\right) \cdot c(q, \boldsymbol{\omega})(1 - c(q, \boldsymbol{\omega})) \cdot |q - 0.5| \quad \text{(A.4)} \\ \frac{\delta J}{\delta \omega_2} &= 2\left(c(q, \boldsymbol{\omega}) - y\right) \cdot c(q, \boldsymbol{\omega})(1 - c(q, \boldsymbol{\omega})) \cdot |q - 0.5|^3 \end{aligned}$$

The learning rate was chosen to be $\eta = 2$ which trades off speed of convergence and volatility. In both learning schemes, random initialization of weights ω_N were chosen by draws from a zero-centered Gaussian $N(0, 2)$.

Appendix B

Study 1: Inductive biases

This is a variation of the main experiment reported in Chapter 4. Since the overall rationale is the same, only differences relating to task design, modeling and results are reported and the reader is encouraged to revisit the corresponding sections above.

B.1 Variations of the task design

In the main experiment, the target capture percentage was fixed at $65\ \%$. Here, the target capture percentage for every trial is an independent and identically distributed sample from the uniform distribution over the interval of $[30, 90]\ \%$. Additionally, for every trial, the number of the observations is independently and uniformly sampled from a categorical distribution over the sample sizes $N \in \{3, ..., 9\}$. As before, participants performed a short training session (30 trials) to familiarize with the task. For this training only, the capture percentage was fixed to $65\ \%$ and the sample size to $N = 4$. As opposed to the main experiment, feedback in the training session (captured percentage and deviation score) was given only after every two trials as an average of the two preceding trials. In total 8 participants (5 female, 3 male, average age 28.3 years) were recruited mainly among students from the Pompeu Fabra University in Barcelona.

Rationale of the extension

These alterations lead to a considerable complication of the task, unless a probabilistic generative model is used for inference. Any inferred probability distribution is independent of the target percentage. The latter merely imposes a different optimization objective (Eq. 4.4) from which to derive the response. While this task is straight-forward for an agent with a probabilistic generative model, it is hard

for an agent learning a suitable end-to-end mapping because there is no sensory representation independent of the objective (target percentage).

Learning input-dependent adjustments from feedback over the task is virtually impossible now. The averaged feedback in the training session poses an assignment problem, as it is not clear which trial contributed how strongly. Had participants nevertheless somehow learned some stimulus-action plan from the short training session, it would be of limited use in the subsequent experimental session. Here, participants have to generalize to different sample sizes and target percentages which may be seen as an instance of transfer learning [195]. Whereas an internal model would easily allow for this generalization, a stimulus-action plan would have to be expanded while there would be no principled way of choosing its parameters. The weighting model used above for instance (Sec. 4.6.3) would have to be equipped with a different set of weights for each sample size. Moreover, this sample size dependence is complicated and follows a normalization rule lowering the influence of each sample's position for larger samples. Hence, even if feedback were given, learning a calibrated response mapping would require a substantial amount of data (Sec. 2.1.2). We can be confident that the participant did not acquire such a mapping over the course of this experiment in the laboratory.

Computational models

Because of the trial-by-trial variation of the target capture percentage p_t, the model output is not just a single fixed mapping of the statistics onto the response anymore but dependent on the target percentage. Explicit density models allow to derive the output in a straightforward way by imposing a different optimization objective. However, as the results of this optimization cannot be derived in simple analytic terms, Eq. 4.4 needs to be numerically solved for every trial and its corresponding target percentage. As before, we identify the result with the dispersion measure $\hat{S}$ as before and then endow it with some more flexibility through the linear response mapping $\hat{y} = \beta_0 + \beta_1 \cdot \hat{S}$ used before (Eq. 4.5). Table B.1 provides an overview of all models and its parameters used for fitting. In the remainder of this section, I will briefly comment on selected models. First, the maximum model (max) tested before can be regarded as inference assuming a centered and symmetric uniform distribution. Maximizing the likelihood of all presented sample points then just yields a distribution extended symmetrically up to the most excentric point. A certain target percentage in this task then merely corresponds to reporting its p_t-th fraction

$$\hat{S} = p_t \cdot \max(|\boldsymbol{d}|) \tag{B.1}$$

Second, a more general form of the generalized normal model that drops the assumption of a zero-centered distribution by inferring its central tendency from the

Table B.1: Overview of model parameters

Model	Abbreviation	Fitting parameters				
Maximum	max	β_0	β_1			
Normal	nm	β_0	β_1			
Generalized normal	gnm	β_0	β_1	p		
Tiling	tlg	β_0	β_1			
Kernel density estimation	kde	β_0	β_1	a		
Nonparametric Bayesian mixture	bnp	β_0	β_1	a	α_0	κ_0

sample was tested. As it was not found to yield significant improvement over the generalized normal model (gnm) assuming a centered distribution, it is not reported. Third, in the limiting case of vanishing kernels in the main experiment, the response was found to collapse onto the third most excentric point. Here, sample sizes and target percentages vary so that the response collapses onto the point of the nearest integer fraction of all points that is closest to the target percentage. This "rounding"-model was tested but is not reported due to inferior results. Finally, the most important addition is an explicit implementation of a hybrid model between pure instance-based generalization and causal inference of a centered Gaussian distribution which is described next.

B.2 Bayesian nonparametric mixture model of Gaussians

Here we provide one possible implementation of the idea of instance-based modulation to causal inference by using a Bayesian nonparametric mixture model of Gaussians (BNP) that determines the adequate structural complexity from data (number of components). Because it is nonparametric, it can adapt to the observed instances to adjustable degrees. Because it is Bayesian, it allows to incorporate prior knowledge about the task's context such as a zero centered distribution. Settings of the prior distribution(s) trade-off these two competing influences. We use a custom adaptation of a variational inference approach to a BNP borrowing ideas from [41, 61].

Every data point is assumed to originate from one of K basis distributions of the mixture model.

$$p(\boldsymbol{d}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \prod_{n=1}^{N} \prod_{k=1}^{K} N(d_n|\mu_k, \tau_k^{-1})^{z_{nk}} \tag{B.2}$$

The conditional distribution of the observed data vector $\boldsymbol{d} = (d_1, ..., d_N)$ depends on the component identity $\boldsymbol{Z} = (Z_1, ..., Z_K)$ and the parameters specifying the

basis distributions. As we are assuming normal distributions, the mean $\boldsymbol{\mu}$ and precision parameters $\boldsymbol{\tau}$ need to be specified. The latent variable $\boldsymbol{Z}$ governs the assignments to the mixture components and follows a categorical distribution parameterized by the mixing coefficients $\boldsymbol{\pi}$.

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{B.3}$$

The assumption of normal distributions for the individual mixture components turns out to considerably simplify the algorithm and allows for a computationally efficient implementation.

$$N(d|\mu, \tau^{-1}) = \sqrt{\frac{\tau}{2\pi}} \exp\left[-\frac{\tau}{2}(d-\mu)^2\right] \tag{B.4}$$

For convenience it is formulated in terms of the precision $\tau = 1/\sigma^2$.

For a fully Bayesian formulation, priors are introduced over the parameters $\boldsymbol{\mu}$, $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$. The *a priori* distribution over the mixing coefficients is a Dirichlet distribution for which the same parameter α_0 is chosen for each component.

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_0}) = C(\boldsymbol{\alpha_0}) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{B.5}$$

$C(\boldsymbol{\alpha}_0)$ is the normalization constant for the Dirichlet distribution. This prior governs the sparsity preference and can be adjusted by the concentration parameter $\alpha_0 \cdot \mathbf{1} = \boldsymbol{\alpha}_0$ which is fitted to our experimental data. For the mean $\boldsymbol{\mu}$ and the precision $\boldsymbol{\tau}$ a Gaussian-Gamma prior is chosen.

$$p(\boldsymbol{\mu}, \boldsymbol{\tau}) = p(\boldsymbol{\mu}|\boldsymbol{\tau})p(\boldsymbol{\tau}) = \prod_{k=1}^{K} N(\mu_k|m_0, (\kappa_0\tau_k)^{-1}) \cdot \mathrm{Gam}(\tau_k|a_0, b_0) \tag{B.6}$$

The factor over μ_k pushes posterior estimates closer to the value of m_0 which we use to incorporate knowledge of a zero-centered distribution. The selectivity of the prior is governed by its precision $\kappa_0\tau_k$ which is a function of the precision τ_k of the Gaussian components. As a consequence, prior knowledge of a zero-centered distribution has a larger influence if the precision of the Gaussian basis distributions is high. More importantly, the strength of the prior can be controlled through the κ_0 parameter which can be freely chosen and which was fit to the experimental data. A Gaussian-Gamma distribution was chosen to obtain conjugate prior distributions. The Gamma distribution over the precision is defined in terms of the

parameters a_0 and b_0 which are symmetric across all K components.

$$\mathrm{Gam}(\tau_k|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \tau_k^{a_0-1} \exp(-b_0\tau_k) \tag{B.7}$$

For the KDE model we assumed that the kernel width scales proportionally to some global dispersion measure which we took to be $\sigma_{ML}(\boldsymbol{d})$. We make a similar assumption for the BNP model but we have to formulate it in terms of the prior distributions over the precision τ. To begin with, all data points are scaled in units of $\sigma_{ML}(\boldsymbol{d}) \cdot q_\tau$ with the proportionality parameter q_τ being determined by fitting. For the scaled problem the expectation of the precision is set to one $\mathbb{E}[\tau] = a_0/b_0 = 1$. For simplicity, the shape parameter a_0 of the Gamma prior distribution (B.7) is clamped to the value of $a_0 = 4$ resulting in a more concentrated density with a single mode. Because of the nonlinear relationship between τ and σ, the expectation of the corresponding σ-distribution is close to, but not exactly, one. For our purposes this is considered sufficient. From the constraint $a_0/b_0 = 1$ the remaining parameter is chosen to be $b_0 = a_0 = 4$. All the above defines the joint distribution over all unknown variables:

$$p(\boldsymbol{d}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}) = p(\boldsymbol{d}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\boldsymbol{Z}, |\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}|\boldsymbol{\tau}) p(\boldsymbol{\tau}) \; . \tag{B.8}$$

Because it is a Bayesian treatment this includes the variables that define the Gaussian mixture distribution and not just the mixture component assignments $\boldsymbol{Z}$. Parameters that define the prior distribution are left implicit.

Variational Bayesian inference approaches draw on a decomposition of the marginal likelihood of the data given a model. Approximation of the posterior distribution such as factorization assumptions can be introduced. The formalism allows to iteratively improve a lower bound of the model evidence making use of variational calculus that may permit to derive tractable iterative update formula. The only approximative assumption we make here is

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}) = q(\boldsymbol{Z}) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}) \tag{B.9}$$

The approximative (variational) posterior distributions are then derived by taking the log expectation of (B.9) with respect to all the other latent variables [41].

In analogy to the EM-algorithm we refer to the iterative updating steps as E- and M-steps. We start with an initial guess for the parameters $(a_k, b_k, \kappa_k, \alpha_k)$ by setting them to the corresponding values of their prior distributions. As for KDE, a Gaussian component is centered on every data point setting $m_k = d_k$, with $k = n$. Contrary to KDE, overly redundant components will be pruned by the algorithm depending on the sparsity constraint. We first perform an E-step to

obtain the responsibilities, i.e. the probabilistic assignments of the data points to the latent variables $\boldsymbol{Z}$ based on the current parameter estimates.

$$\begin{aligned} r_{nk} &= \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}} \\ \rho_{nk} &\propto \tilde{\pi}_k \tilde{\tau}_k^{1/2} \exp\left[-\frac{1}{2\kappa_k} - \frac{1}{2}\frac{a_k}{b_k}(d_n - m_k)^2\right] \end{aligned} \tag{B.10}$$

The following abbreviations were used.

$$\begin{aligned} \tilde{\tau}_k &= \exp(\Psi(a_k) - \ln(b_k)) \\ \tilde{\pi}_k &= \exp(\Psi(\alpha_k) - \Psi(\hat{\alpha})), \qquad \hat{\alpha} = \sum_k \alpha_k \end{aligned}$$

In the subsequent M-step, the responsibilities r_{nk} together with the parameters of the prior distributions $(a_0, b_0, \kappa_0, \alpha_0, m_0)$ are used to calculate revised parameter estimates.

$$\begin{aligned} \alpha_k &= \alpha_0 + N_k \\ \kappa_k &= \kappa_0 + N_k \\ m_k &= \frac{1}{\kappa_k}\left(\kappa_0 m_0 + N_k \bar{d}_k\right) \\ a_k &= a_0 + \frac{1}{2} N_k \\ b_k &= b_0 + \frac{1}{2} N_k S_k + \frac{\kappa_0 N_k}{2\kappa_k}\left(\bar{d}_k - m_0\right)^2 , \end{aligned} \tag{B.11}$$

making use of the abbreviations

$$\begin{aligned} N_k &= \sum_{n=1}^{N} r_{nk} \\ \bar{d}_k &= 1/N_k \sum_{n=1}^{N} r_{nk} d_n \\ S_k &= 1/N_k \sum_{n=1}^{N} r_{nk} (d_n - \bar{d}_k)^2 . \end{aligned}$$

The whole algorithm alternates between E- and M-steps and is repeated until convergence is reached which we determined by a threshold on the changes of the parameter estimates. The inferred density is then used by the optimizer to deter-

mine the response r_t corresponding to the desired target percentage. The result r_t is then transformed back into the original observation space and identified with $\hat{S}_t = r_t \cdot \sigma_{ML}\, q_\tau$.

B.3 Supplementary results

In this version of the task, participants had to demonstrate a more flexible use of their sensory representation of uncertainty by adapting the read-out to the momentary task-objective. It is virtually impossible to learn beneficial behavior from experience alone so that participants must rely solely on internal mechanisms to make quantitatively accurate judgments of uncertainty.

Internal objective can well adapt to varying target percentage

As for the main experiment, behavior is tightly correlated with the maximum likelihood estimator of the Gaussian σ_{ML} (Fig. B.1 A, Pearson correlation coefficient, median across participants: 0.74, 95 %-CI $(0.72, 0.77)$). Beyond that, the capture

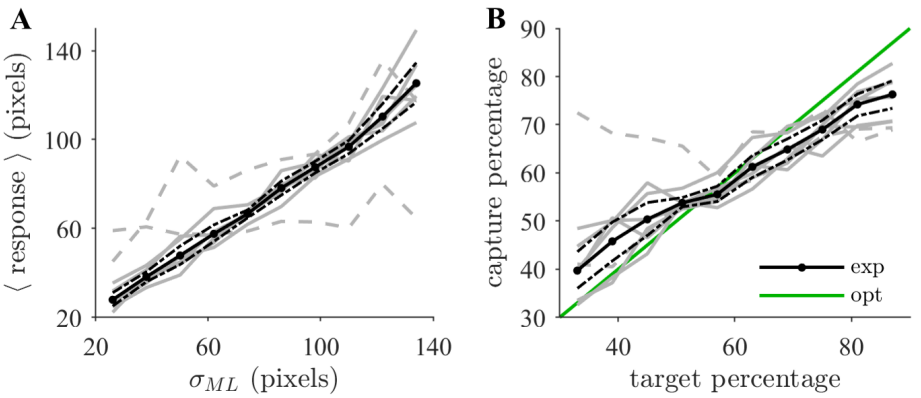

Figure B.1: Behavior is consistent with an internal trial-by-trial objective adapting flexibly to the changing target percentage (A) The behavioral response (black, mean across participants, with 95 %-CI) is an increasing function of the objective dispersion measure σ_{ML}. Individual response curves of all 8 participants tested (gray lines). Two participants displaying poor compliance with the instructed task (dashed) were excluded from the analysis. **(B)** Participants give estimates closely corresponding to the target percentage. Optimal responses (green) result in the unity line. Individual participant responses as in panel A.

percentage of future darts closely corresponds to the target percentage (Fig. B.1 B, Pearson correlation coefficient of binned values across participants $\rho = 0.94$, $p = 8.5 \cdot 10^{-29}$). Hence, participants manage to give good estimates of the future percentage on a trial-by-trial basis despite the fact that a different summary statistic must be read out on each trial. Even though the task complexity increased compared to the main experiment (see Fig. 4.4A), the accuracy of the participants has hardly suffered. The per participant median of the per-trial capture deviation is close to zero percent as for the main experiment (mean across participants, -0.45,

95 %-CI, $(-2.21, 1.34)$ %). Likewise, the median of the absolute value of the per-trial capture deviation is on a comparable level as for the main experiment (here $8.47\,\%$, ($95\,\%$-CI $(7.69, 9.24)\,\%$ vs. $6.54\,\%$ before, medians across participants shown). Consequently, the per-trial deviations do not strongly depend on task complexity. Hence, participants can adapt their trial-by-trial objective well to the target percentage. This is what would be expected from an agent making flexible use of its inferred distributional estimate.

Behavior features a systematic deviation in that small objective target percentages are overstated while large ones are understated (Fig. B.1B) as evidenced by a slope that is considerably smaller than one (fitted linear function, 0.66, 95 %-CI $(0.60, 0.72)$). This is reminiscent of descriptions of probability distortion in prospect theory where lower probability is typically reported to be over-weighted and vice versa [154]. Notably, this finding is independent of any prior distribution over the dispersion that participants might hold as the latter is independent of the target percentage by construction of the task.

Behavior tends to instance-based generalization

As in the main experiment, behavior is inconsistent with inference of a centered Gaussian distribution. Evidence for instance-based generalization is found from the strong support for all methods based on basis distributions such as the tiling model (tlg), the KDE model (kde) and the Bayesian nonparametric mixture model (bnp) (Fig. B.2). All are superior to the normal model on the group-level which is evidenced by the significant exceedance probability (asterisk) of the corresponding pairs, (bnp-nm, kde-nm, tlg-nm) in the model comparison matrix (Fig. B.2A).

The individual differences in terms of the absolute value of the cross-validation log likelihood (CVLL) with respect to the normal model (nm, row 3) are mostly large and individually significant (green shaded background) (Fig. B.2B). The normal model is only found to predict better than the maximum model (max) which is equivalent to maximizing the likelihood of a centered uniform distribution over all sample points (Fig. B.2).

As before, behavior is typically not consistent with inference of a near-Gaussian distribution evidenced by the performance of the generalized normal distribution model (gnm). Despite giving better predictions than the normal model (nm), it is clearly inferior to all approaches based on basis distributions implementing instance-based generalization. Furthermore, models with Gaussian basis distributions (bnp, kde) predict better than the tiling model with spatially confined basis distributions. This confirms the notion that sample points have a non-local effect on the inferred density beyond the neighboring point.

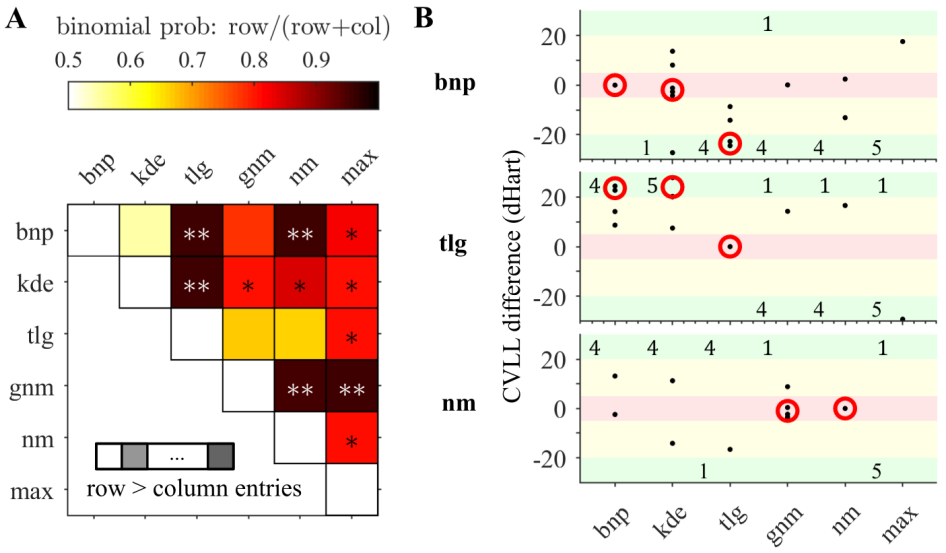

Figure B.2: Behavior is consistent with a Bayesian approach to sparsity-constrained mixture modeling (A) Summarized results of a hierarchical Bayesian model comparison procedure that estimates probability distributions over models. The color code over each square shows estimates of the parameter of the binomial distribution governing the probability by which the model indexed by the row is more likely than the one indexed by the column. Superimposed (asterisk) are large differences of the exceedance probability (Methods). **(B)** Many models can be rejected for individual participants despite group heterogeneity. The plot shows differences of the cross validation log likelihood (CVLL) for each participant (black dot) with respect to the model indicated in each row (bnp, tlg, nm). The median across participants indicates trends on the group level (red circles). The number of decisive individual CVLL differences (green shaded background) is additionally indicated as a number.
Model abbreviations: bnp=Bayesian nonparametric mixture, kde=kernel density estimation, tlg=tiling, gnm=generalized normal, nm=normal, max=maximum

Conceiving behavior as an instance-based modulation to causal inference

In the main experiment, we found evidence that the inferred densities based on the KDE model constitute a redundant representation because of spatially extended and overlapping kernel functions. On the level of raw responses, human estimates are not far from causal inference of a centered Gaussian distribution which was suggested by the framing of the task (see Fig. 4.7). This indicates that participants considerably generalize beyond the instances observed. That raised the question if behavior might not be better conceived an instance-based modulation to causal inference pushing the KDE model closer to a regime where it must approximate inference of a bell-shaped distribution.

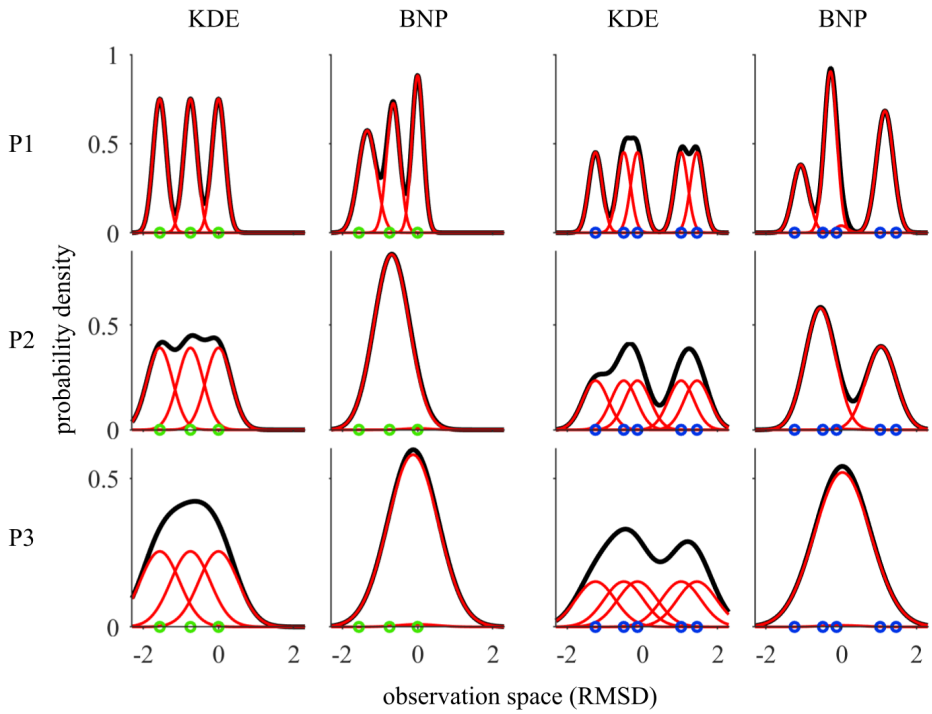

Figure B.3: Bayesian nonparametric mixture of Gaussians explicitly incorporates prior knowledge and sparsity constraints
Comparison of the inferred distributions of the purely instance-based KDE model and the BNP model. The latter is a specific implementation of the concept of instanced based modulation of causal inference which imposes differently strong prior structural knowledge (P1-weak, P2-moderate, P3-strong) on the distribution to be inferred. Inferred internal distributions (black) of three selected participants (P1-P3, rows) for two example samples (green, blue) together with the underlying additive basis distributions (red).

As a better description of the internal representation of human participants, we suggested Bayesian nonparametric methods allowing to incorporate prior knowledge and gradual sparsity constraints. Here we tested a specific implementation through the Bayesian nonparametric mixture model of Gaussians (bnp) (Sec. B.2) to investigate the experimental plausibility of this idea. We found that the explicitly enforced sparsity constraint compared to the KDE model (kde) does not impede predictive performance (Fig. B.2A). It predicts on a similar level and is even found to be decisively superior for one participant (Fig. B.2B).

The inferred densities are illustrated for both models (bnp, kde), for two samples (green, blue) and for three selected participants (P1, P2, P3) (Fig B.3). As in the main experiment, the probability densities inferred by the KDE model feature redundant representations through overlapping kernels. The Bayesian nonpa-

rametric mixture model demonstrates that very similar densities can be more sparsely represented by a smaller number of mixture components for which the overlap is reduced. The number of components varies across participants but shows fewer components than KDE (averaged across trials, mean across participants, 3.0505, $95\ \%$-CI $(2.10, 4.17)$, KDE: 7). Participant P1 (Fig. B.3) more closely follows instance-based generalization while P2 and especially P3 assign observations to fewer components. Because of normalization, the reduction in the number of components, through effectively zero-valued mixture coefficients, strengthens others. The strongest mixture component accounts for several sample points, especially for larger samples (Fig. B.3, blue) (largest component in multiples of the KDE model's (equal) weight(s), 3.20, $95\ \%$-CI $(2.24, 4.14)$, average across participants and trials). Correspondingly, we find a smaller number of components for participants who feature a strong maximum mixture component (Spearman correlation of trial averages, $\rho = -1$, $p = 0.0014$).

The Bayesian nonparametric mixture model (bnp) allows for the incorporation of prior knowledge of a zero-centered distribution. The effect is strongest for small sample sizes (Fig. B.3, green sample) and depends on how strongly the participant's prior belief suggests a centered distribution (Sec. B.2, parameter κ_0). A strong prior belief, such as for participant P3, results in inference of an almost centered distribution despite a negative sample mean (green). A comparison with the inferred density for participant P2 illustrates the difference in the resulting representation as his representation also just consists of one Gaussian component but his prior belief of a centered density is weaker. The parameter governing the strength of this prior knowledge correlates negatively with the number of inferred components of the Bayesian nonparametric mixture model (Spearman correlation, $\rho = -1$, $p = 0.014$). The extent to which participants attribute sample points to fewer mixture components, the more they appear to align their inferred density with the expectations of a single zero-centered cluster of events stemming from a single cause.

B.4 Summary & Conclusions

This additional experiment fully underscores the principal claims made by the first. It shows that participants do not depend on feedback or training beyond understanding the task (objective). Before, it was highly unlikely, though not strictly impossible, that participants learned some behavioral stimulus-response mapping by associating stimuli with beneficial outcomes through training and feedback, even though we found evidence that suggested they did not exploit this option. This task prevented this opportunity by construction. Instead, it required the flexible use of

sensory representations of uncertainty by adapting the read-out to the momentary task-objective. Despite the increase in complexity, participants can flexibly adapt to the varying task objective almost without performance loss. Constructing a distribution over future events is crucial to have a principled trial-by-trial objective, and participants indeed appear to quantify uncertainty in a similar way demonstrating that the match with the fixed target percentage of the main experiment is no coincidence. Despite featuring systematic probability distortion which might stem from instance-biased generalization, these results underpin the claims that participants derive their responses from estimated distributions.

We have demonstrated that a Bayesian nonparametric mixture model of Gaussians could be used to implement instance-based modulation of causal inference. It allows to incorporate prior knowledge about the generative process of the observations. Sparsity constraints enforce a more efficient, less redundant representation connected to contextual knowledge of the causal structure. The particular formulation showed that these ideas are consistent with the experimental data. Our model is at least as good as KDE and we believe that these ideas merit further investigation. To go beyond this proof of concept, more data needs to be collected and the task should be adapted to richer causal structures.

Appendix C

Study 2: Empirical priors

C.1 Calibration of confidence judgments

In Experiment 1, the correlation with the actual probability of deciding correctly is $\rho = 0.81$ and can be considered high (linear correlation on bin values for all participants in Fig. C.1a, $p = 1.27 \cdot 10^{-45}$). Similarly, in Experiment 2, the

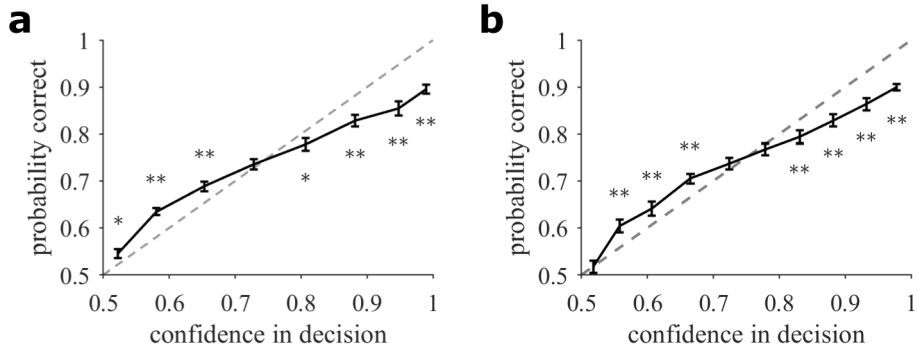

Figure C.1: Human confidence judgments (black, mean ± SEM) correspond to the probability of deciding correctly (optimal model). Responses are grouped in approximately equally filled bins for Experiment 1 (a) and Experiment 2 (b).

correlation with the actual probability of deciding correctly is $\rho = 0.85$ (linear correlation on bin values for all participants in Fig. C.1b, $p = 2.66 \cdot 10^{-63}$). Decision confidence shows systematic deviations from calibrated responses in that the participants are under-confident for difficult decisions of low expected accuracy and overconfident for easy decisions. Significant signed differences of the group median against calibrated responses are computed from a signed rank test and indicated for each bin ($*: \ 0.01 < p \leq 0.05$ and $**: \ p \leq 0.01$).

Table C.1: Overview of fitted models used for prediction

Experiment 1				
Model	Confidence estimate	Response map	Parameters	Comment
opt	$c(B)$	sigmoid	$\boldsymbol{\omega}$	Sec. 3.5.6
ratio	N_B/N	sigmoid	$\boldsymbol{\omega}$	
diff	$N_B - N_R$	sigmoid	$\boldsymbol{\omega}$	
Experiment 2				
Model	Block estimate	Response map	Parameters	Comment
opt	M	zmap	$\boldsymbol{\omega}_Z$	3.5.6
tly	$M^{\pm}$	zmap	$\boldsymbol{\omega}_Z$	3.5.6
avg	M^{q}	zmap	$\boldsymbol{\omega}_Z$	3.5.6
diff	M^{d}	zmap	$\boldsymbol{\omega}_Z$	3.5.6
Model	Confidence estimate	Response map	Parameters	Comment
Beta prior	$c(B;\boldsymbol{\nu})$	sigmoid	$\boldsymbol{\omega}, \nu_1, \nu_2$	C.4

C.2 Overview of fitted models

The models listed in Table C.1 under Experiment 1 were used for model comparison in Fig. C.2. The respective results for Experiment 2 were reported in the main text in Fig. 3.6d. For all models, we used a nonlinear response mapping to account for distortions. The parameters of the one-dimensional sigmoidal mapping (Sec. 3.5.6) are abbreviated by $\boldsymbol{\omega}$ while those of the zmap (Sec. 3.5.6) are referred to as $\boldsymbol{\omega}_Z$.

Construction of stimuli

All sample points to be displayed were separated by color and arranged along a horizontal line. The horizontal extent of the grid has a random number of entries, but always more than the maximum number of samples used over the entire session. We randomly sampled two sub-regions along the horizontal direction which are large enough to accommodate both the red and blue sample circles. Within each sub-region, the grid entries are randomly populated by the respective subsample. In the vertical direction, we linearly divided a randomly chosen range by the same amount of grid entries as determined for the horizontal direction. We then randomly assigned the circles to these positions. The circle density is not preserved over different sample sizes, but roughly for each subsample. Across trials, blue samples are randomly chosen to be either to the left or right side of the red samples.

Experiment 1: Instructions

We emphasized in colloquial terms that the sampling of the passengers is independent and identically distributed (i.i.d.) and that it does not favor either group. In addition, it was stressed that the sample positions are irrelevant to the task. Our participants were asked to report their decision and their confidence in the correctness of that decision. Specifically, a higher decision confidence should lead to a placement farther from the center whereas for guessing, it should be in the middle. They were specifically advised to rely on their intuition while we discouraged any explicit mental arithmetic. We made it clear that for Experiment 1, there is no relationship between the airplanes (trials).

Regarding the base rates, we mentioned that there are just as many airplanes with a red than with a blue majority arriving at the airport. And that most airplanes are known to have a roughly equal number of passengers of the two kinds on board. Apart from the instructions, the participants could ask any questions to the experimenter they deemed necessary to understand the task.

Experiment 2: Instructions

At the beginning of session 2, each participant read further written instructions which introduced the block-wise design. We explained that there is an event in the city (e.g. a concert, a football match, etc.) that tends to attract many more red than blue passengers. That the airplanes would be presented one after the other in consecutive trials grouped together in a block separated by pauses. To make that clear, we additionally added visual indication of the in-block trial by presenting five horizontally equidistantly spaced open circles which turned to solid circles one-by-one as the participant progresses through the trials within a block.

We mentioned that the tendency for 'red' or 'blue' airplane majorities changes unpredictably from city to city and does not favor either group. Moreover, even though red passengers might preferably travel to a particular city, occasionally there might be airplanes with a blue majority. We attempted to make it very clear that the decision is still about a given airplane majority (trial) and not about the overall tendency of one kind to travel to that city (or airport).

Robust estimation of variation of the response distribution

For robustness, we estimated a trimmed SD, i.e. we removed values below or above three interquartile ranges from the lower or upper quartile respectively. On the remaining (non-outlying) trials, the ML estimator for the normal distribu-

tion, $\theta = \sqrt{1/N \sum_t (y_t - \hat{y}_t)^2}$, corresponding to the root mean squared deviation (RMSD) of the residual responses is used. A more ideal solution would be to set θ so as to strictly maximize the likelihood of the responses for the truncated Gaussian. For the sake of faster computations however, we resort to this approximate approach which is justified by the relatively low behavioral response noise (appendix C.3).

Cross-validation splits

As cross validation is a computationally expensive method, we use a random 5-fold split of the data into training and test sets such that each training point is used four times for training and once for testing. However, to avoid splits that are highly unrepresentative of the response distribution, we used a stratified version of CV by ensuring that the mean response $\langle y \rangle$ is approximately equal in all folds. For this purpose, we assigned the data points to one of the q cumulative quantiles of the response distribution. We then constructed slices that contain one value from each cumulative quantile. Subsequently, we sampled the slices to create the 5-fold CV splits. The number of quantiles q is chosen from suitable multiples of the factors of the number of trials close to eight.

To improve the reliability of the per participant estimates of the model evidence (CVLL), we repeated this procedure five times with different random splits and aggregated the output so that in total 25 CV folds are performed for each participant and model. For the prior learning task (Experiment 2), only blocks of trials are split. We basically applied the same logic as before to blocks and attempted to achieve an approximately equal amount of trials from all quantiles of the experimental distribution of the decision confidence $|y_t - 0.5| + 0.5$.

C.3 Experiment 1

Model comparison to evidence sample size effects

The large confidence intervals in Fig. C.2 point to high variability across participants. We do not claim that all are best described by the probabilistic inference model (opt) but acknowledge that few probably follow a rather heuristic approach. One participant had to be excluded due to numerical problems caused by an extremely high model evidence for the difference model.

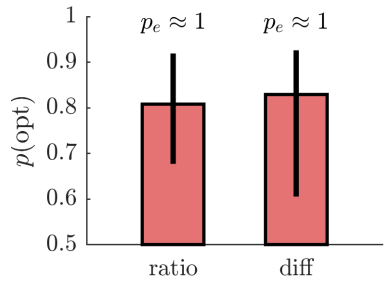

Figure C.2: Sample size is crucial to predict confidence judgments in Experiment 1 Binomial probability of the optimal model to account for the data of a randomly chosen participant (error bars are 95 %-CI, Sec. 3.5.6). Pairwise comparisons to the models (ratio, diff) show that probabilistic information integration yields better predictions on the group level. Additionally, the exceedance probability p_e is used to quantify how much more likely the optimal model is.

Sensory noise

The task design results in low levels of perceptual noise which may obscure accurate perception of the sufficient statistics (N_R, N_B). In the basic task (Experiment 1), optimal decisions should always follow the sample majority. If there is sensory

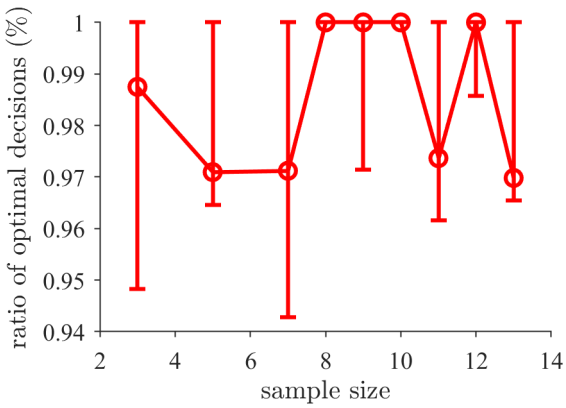

Figure C.3: High ratio of optimal decisions evidences low sensory noise levels The percentage of correctly made decisions is plotted as a function of sample size (median across participants, 95 %-CI).

noise, the internal estimate would deviate from the sample proportion. In particular, trials whose sample proportion is close to $q = 0.5$ would lead to suboptimal choices. However, participants are rarely found to make suboptimal choices (Fig. C.3) which suggests very low levels of perceptual noise.

Task-intrinsic noise

To estimate non-input related response noise that is intrinsic to the task, we searched for all trials with the same sufficient statistic. The only assumption we made is symmetry, i.e. decision confidence should only be a function of the absolute distance from the decision boundary (e.g. $q = 0.25$ and $q = 0.75$ result in the same decision confidence). Consequently, with respect to the sample majority, trials of the same sample proportion can be pooled. If there are ten or more trials for a particular sufficient statistic, we computed their squared deviation from the mean. Subsequently, to estimate the variance for fixed inputs, all squared deviations calculated this way were pooled and the mean is taken individually for each participant. The standard deviation parameter of a corresponding Gaussian distribution is estimated by taking the square root. A median value across participants of 0.104 (95 %-CI, 0.089, 0.126) indicates low to moderate noise levels in Experiment 1.

C.4 Experiment 2

Compliance with hierarchical task

Even though our participants performed the hierarchical task remarkably well, few of them showed such little dependence on previous trials within a block that one may doubt whether they properly understood the hierarchical nature of the task. To evidence this, we fitted a linear function $y = a_q(b)q + a_0(b)$ of the sample proportion q to their responses y conditional on the actual block tendency (see Fig. 3.5a). The separation of the offset $\Delta a = a_0(b = 1) - a_0(b = 0)$ should be significantly positive. We repeated this fit 10000 times with a randomly shuffled assignment of the b-variable for every participant. To derive the p-value that Δa is significantly larger than chance, we compute the fraction that Δa is larger than the surrogates from the shuffling test (see Table C.2). Generally, one should only discard participants on justified grounds. Based on this measure, we chose to leave out only the first participant such that the remaining analysis is based on 23 participants in total.

Table C.2: Estimate of the compliance with the hierarchical task of the least engaged participants. Participants are ordered from left to right according to decreasing p-values.

	1	2	3	4	5	6
Δa	0.0037	0.0201	0.0258	0.0301	0.0224	0.0556
p-value	0.4372	0.2154	0.1415	0.0969	0.0894	0.0190

Inferential patterns for for fitted block tendency

The probabilistic model assumes that the block tendency from which the trial-by-trial (airplane) majorities μ are drawn is given by one of two skewed Beta-distributions (see Sec. 3.5.5). By convention a 'blue' context is characterized by the block tendency $\mathrm{Beta}(\mu|\nu_1 = 14, \nu_2 = 9)$ and $b = 1$ while the 'red' context is correspondingly denoted by $\mathrm{Beta}(\mu|\nu_2, \nu_1)$ and $b = 0$. The two distributions are symmetric with respect to the block aligned trial majorities, $\tilde{\mu}_b = b \cdot \mu + (1 - b) \cdot (1 - \mu)$, which immediately follows from the property of the Beta distribution: $\mathrm{Beta}(\tilde{\mu}_{b=1}|\nu_1, \nu_2) = \mathrm{Beta}(\tilde{\mu}_{b=0}|\nu_2, \nu_1)$. A variation of the optimal inference routine (Eqs. 3.6-3.10) is used that allows for different values of the parameters ν_1, ν_2 governing the block tendency with the restriction that $\nu_1 \geq \nu_2$. In addition, the sigmoidal response mapping (Eq. 3.14) is used to allow for nonlinear distortions of the output.

The model output for the fitted parameters determined by maximum likelihood are plotted together with the experimental data as in the main text (Fig. C.4). As concluded in the main text, the qualitative match with behavior improves but systematic deviations remain. Remarkably, we tried a related model that similarly estimates the block tendency from a differently skewed Beta distribution $M(b; \nu_1, \nu_2)$ but which then uses the zmap response mapping (Eg. 3.5.6), i.e. the integration of the block tendency estimate M with the sample is different. Even though the latter has more parameters and may in this sense be considered more flexible, it did not yield better predictive performance which is why we chose to report the former probabilistic model. Consequently, most of the deviations that the flexible zmap response mapping can account for can also be captured by the more constrained probabilistic model under the assumption of a differently skewed Beta distribution of the block tendency.

Proper normalization of messages

Here we will focus on finding the normalization ψ of Eq. 3.10. Marginalizing out all random variables (integration over the full range of μ_T) must result in the expression being equal to one. Because of independence, the categorical distribution over N factorizes and separately integrates to one. Compact expressions can be found for the μ_T-terms as the product of the distributions $p(D_T|\mu_T)p(\mu_T|b)$ in the integrand is a product between a Binomial distribution and a Beta-distribution. Hence, the resulting distribution is of Beta-shape again but is not normalized. If we drop the index T, the expression for $b = 1$ can be re-written in terms of the

gamma distribution Γ:

$$\begin{aligned} p(D|\mu)p(\mu|b=1) &= \mu^{N_B+\nu_1-1}(1-\mu)^{N_R+\nu_2-1} \\ &= \mathrm{Beta}(\mu|N_B+\nu_1, N_R+\nu_2)\frac{\Gamma(N_B+\nu_1)\Gamma(N_R+\nu_2)}{\Gamma(N_B+\nu_1+N_R+\nu_2)} \end{aligned} \tag{C.1}$$

To determine ψ, we enforce the normalization condition $1/\psi \sum_b M(b) \int p(D|\mu)p(\mu|b)\ \mathrm{d}\mu = 1$. Together with the probability distribution $M(b)$, which can be easily normalized, we arrive at:

$$\psi = \frac{\Gamma(N_B+\nu_1+N_R+\nu_2)}{M(1)\Gamma(N_B+\nu_1)\Gamma(N_R+\nu_2) + M(0)\Gamma(N_B+\nu_2)\Gamma(N_R+\nu_1)} \tag{C.2}$$

The messages to update the belief $M(b)$ about the block tendency Eq. 3.8 can be normalized analogously.

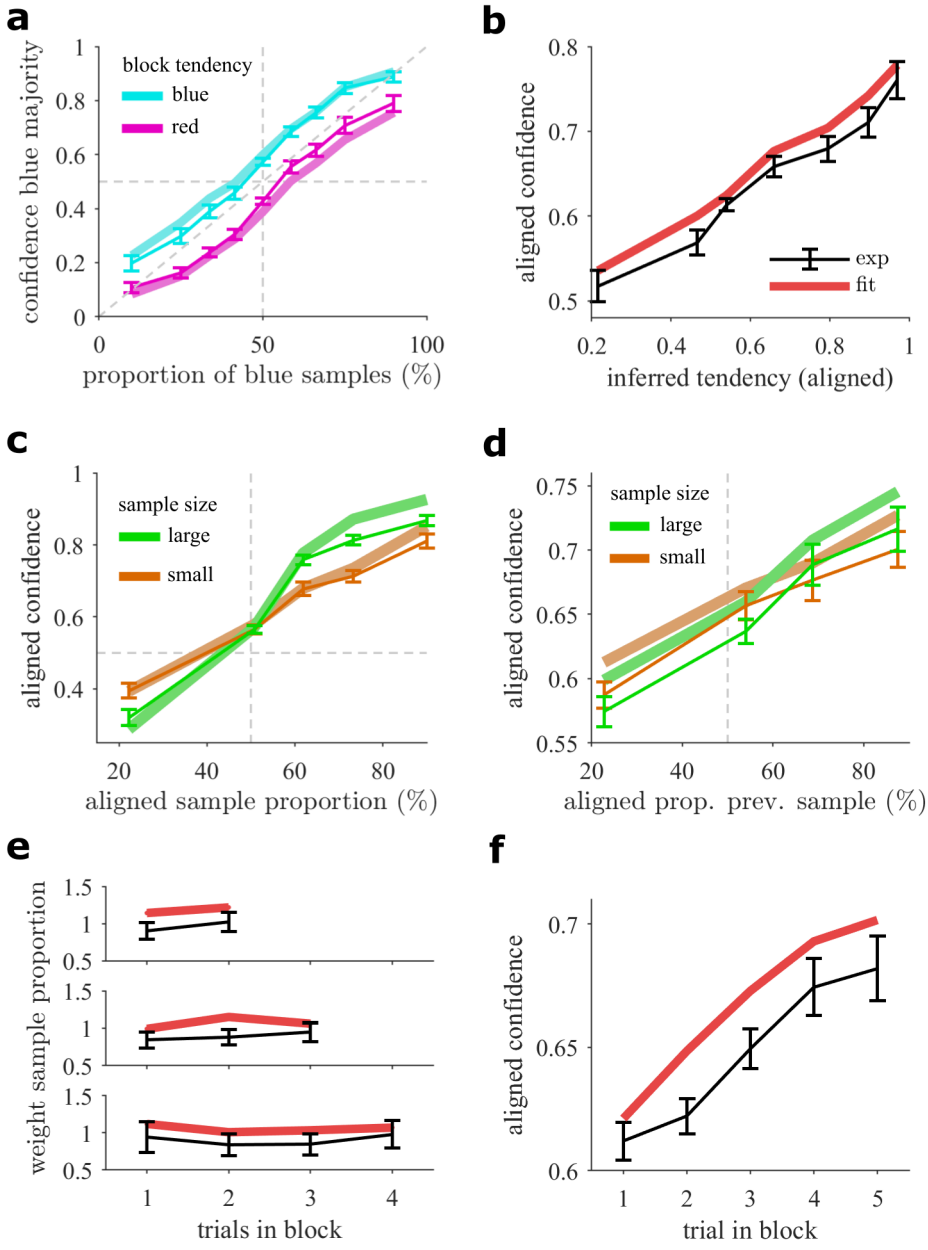

Figure C.4: Behavioral patterns in the hierarchical inference task compared to a fitted model assuming a differently parameterized block tendency. Compare to the theoretical patterns in Fig. 3.4 and the match with the optimal model reported in the main text Figs. 3.5-3.7.

Bibliography

1. Russell, S. & Norvig, P. *Artificial Intelligence: Pearson New International Edition - A Modern Approach* New international edition. ISBN: 978-1-292-02420-2 (Pearson Education, Limited, 2013).
2. Rieke, F. *Spikes: Exploring the Neural Code* ISBN: 9780262681087 (MIT Press, 1999).
3. Schacter, D., Gilbert, D., Wegner, D. & Hood, B. *Psychology - Second European Edition* ISBN: 978-1-137-40673-6 (Palgrave Macmillan, 2015).
4. Davis, R., Shrobe, H. & Szolovits, P. What is a knowledge representation? *Ai Mag.* **14,** 17 (1993).
5. Goldstein, E. *Sensation and Perception* ISBN: 9780495601494 (Cengage Learning, 2009).
6. Rubin, E. *Synsoplevede figurer* 1915.
7. Boring, E. G. A new ambiguous figure. *Am. J. Psychol.* **42,** 444–445 (1930).
8. Carbon, C.-C. Understanding human perception by human-made illusions. *Front. Hum. Neurosci.* **8,** 566 (2014).
9. *The Caribbean Current* Accessed: 15.01.2018. https://www.thecaribbeancurrent.com/ambiguous-laugh/.
10. Schank, R. C. & Wilks, Y. The goals of linguistic theory revisited. *Lingua* **34,** 301–326 (1974).
11. Chomsky, N. in *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky* (1980).
12. Kersten, D. & Yuille, A. Bayesian models of object perception. *Curr. Opin. Neurobiol.* **13,** 150–158 (2003).
13. *Wikimedia Commons* Accessed: 30.03.2018. https://commons.wikimedia.org/wiki/File:Maze_simple.svg.
14. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9,** 292–303 (2008).

15. Hume, D. *A Treatise of Human Nature* reprinted from the Original Edition (1739) in three volumes (ed Selby-Bigge, L.) (Oxford: Clarendon Press, 1896).

16. Helmholtz, H. v. *Handbuch der physiologischen Optik* - (Voss, 1867).

17. Barthelmé, S. & Mamassian, P. Flexible mechanisms underlie the evaluation of visual confidence. *Proc. Natl. Acad. Sci.* **107,** 20834–20839 (2010).

18. De Gardelle, V. & Mamassian, P. Does Confidence Use a Common Currency Across Two Visual Tasks? *Psychol. Sci.* **25,** 1286–1288 (2014).

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

20. Jaynes, E. T. *Probability Theory - The Logic of Science* ISBN: 978-1-139-43516-1 (Cambridge University Press, 2003).

21. Cox, R. T. Probability, Frequency and Reasonable Expectation. *Am. J. Phys.* **14,** 1–13 (1946).

22. Vineberg, S. *Dutch Book Arguments* The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.).

23. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27,** 712–719 (2004).

24. Doya, K., Ishii, S. & Pouget, A. *Bayesian Brain - Probabilistic Approaches to Neural Coding* 1. Aufl. ISBN: 978-0-262-04238-3 (MIT Press, 2007).

25. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16,** 1170–1178 (2013).

26. Penny, W. Bayesian Models of Brain and Behaviour. *ISRN Biomathematics* **2012,** 19 (2012).

27. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9,** 1432–1438 (2006).

28. Thagard, P. *Cognitive Science* The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.).

29. Anderson, J. R. *The adaptive character of thought* (Erlbaum, 1990).

30. Anderson, J. R. Is human cognition adaptive? *Behav. Brain Sci.* **14,** 471–485 (03 1991).

31. Oaksford, M. & Chater, N. Précis of Bayesian Rationality: The Probabilistic Approach to Human Reasoning. *Behav. Brain Sci.* **32,** 69–120 (2009).

32. Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* **14,** 357–364 (2010).

33. Marr, D. *Vision : a computational investigation into the human representation and processing of visual information* – (Freeman, 1982).

34. Poggio, T. The Levels of Understanding Framework, Revised. *Perception* **41,** 1017–1023 (2012).

35. Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Top. Cogn. Sci.* **7,** 217–229 (2015).

36. Griffiths, T. L., Vul, E. & Sanborn, A. N. Bridging Levels of Analysis for Probabilistic Models of Cognition. *Curr. Dir. Psychol. Sci.* **21,** 263–268 (2012).

37. Han, F. & Zhu, S.-C. *Bottom-up/Top-Down Image Parsing by Attribute Graph Grammar* in *Proceedings of the Tenth IEEE International Conference on Computer Vision* (2005), 1778–1785.

38. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10,** 301–308 (2006).

39. Lin, H., Tegmark, M. & Rolnick, D. Why Does Deep and Cheap Learning Work So Well? *J. Stat. Phys.* **168,** 1223–1247 (2017).

40. Craik, K. J. W. *The Nature of Explanation* - (University Press, 1943).

41. Bishop, C. M. *Pattern Recognition and Machine Learning* - 1st ed. 2006. Corr. 2nd printing 2011. ISBN: 978-0-387-31073-2 (Springer, 2006).

42. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* **331,** 1279–1285 (2011).

43. Friston, K. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences* **13,** 293–301 (2009).

44. Ghahramani, Z. Bayesian nonparametrics and the probabilistic approach to modelling. *Phil. Trans. R. Soc. A* **1** (2011).

45. Jacobs, R. A. & Kruschke, J. K. Bayesian learning theory applied to human cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **2,** 8–21 (2011).

46. Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. *Bayesian models of cognition* (ed Sun, I. R.) (Cambridge University Press, 2008).

47. Meyniel, F. & Dehaene, S. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl. Acad. Sci.* **114,** E3859–E3868 (2017).

48. Holroyd, C. & Coles, M. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109,** 679–709 (2002).

49. Logan, G. D. & Crump, M. J. C. Cognitive Illusions of Authorship Reveal Hierarchical Error Detection in Skilled Typists. *Science* **330,** 683–686 (2010).

50. Botvinick, M. M. Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* **12,** 201–208 (2008).

51. Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y. & Botvinick, M. M. Optimal Behavioral Hierarchy. *Plos Comput. Biol.* **10,** 1–10 (2014).

52. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).

53. Pearl, J. *Graphical Models for Probabilistic and Causal Reasoning* 1997.

54. Pearl, J. *Causality* - 2nd Revised edition. ISBN: 978-0-521-89560-6 (Cambridge University Press, 2009).

55. Rottman, B. M. & Hastie, R. Reasoning about Causal Relationships: Inferences on Causal Networks. *Psychol. Bull.* **140,** 109–139 (2013).

56. Meder, B. *Seeing versus Doing: Causal Bayes Nets as Psychological Models of Causal Reasoning* PhD thesis (Georg-August-Universität zu Göttingen, 2006).

57. Vilares, I. & Kording, K. Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Ann. Ny. Acad. Sci.* **1224,** 22–39 (2011).

58. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36,** 181–204 (03 2013).

59. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl. Acad. Sci.* **105,** 10687–10692 (2008).

60. Payzan-LeNestour, E. & Bossaerts, P. Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *Plos Comput. Biol.* **7,** e1001048 (2011).

61. Gershman, S. J. & Blei, D. M. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology* **56,** 1–12 (2012).

62. Jern, A. & Kemp, C. A probabilistic account of exemplar and category generation. *Cognitive Psychology* **66,** 85–125 (2013).

63. Barlow, H. B. Possible principles underlying the transformations of sensory messages. *Sensory Communication,* 217–234 (1961).

64. Barlow, H. Redundancy reduction revisited. *Network: Computation in Neural Systems* **12,** 241–253 (2001).

65. Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S. & Lerchner, A. Early Visual Concept Learning with Unsupervised Deep Learning. *ArXiv e-prints* (2016).

66. Zhang, H., Daw, N. D. & Maloney, L. T. Human representation of visuomotor uncertainty as mixtures of orthogonal basis distributions. *Nat. Neurosci.* **18,** 1152–1158 (2015).

67. Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* **10,** 309–318 (2006).

68. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350,** 1332–1338 (2015).

69. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *Ieee T. Pattern Anal.* **35,** 1798–1828 (2013).

70. Bellman, R. & Bellman, R. *Adaptive Control Processes: A Guided Tour* (Princeton University Press, 1961).

71. Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The Helmholtz Machine. *Neural Comput.* **7,** 889–904 (1995).

72. Cooper, G. F. The computational complexity of probabilistic inference using bayesian belief networks. *Artif. Intell.* **42,** 393–405 (1990).

73. Bossaerts, P. & Murawski, C. Computational Complexity and Human Decision-Making. *Trends Cogn. Sci.* **21,** 917–929 (2017).

74. Heeger, D. J. Theory of cortical function. *Proc. Natl. Acad. Sci.* **114,** 1773–1782 (2017).

75. Pearl, J. *Heuristics: Intelligent Search Strategies for Computer Problem Solving* ISBN: 0-201-05594-5 (Addison-Wesley Longman Publishing Co., Inc., 1984).

76. Suchow, J. W., Bourgin, D. D. & Griffiths, T. L. Evolution in Mind: Evolutionary Dynamics, Cognitive Processes, and Bayesian Inference. *Trends Cogn. Sci.* **21,** 522–530 (2017).

77. Pitkow, X. & Angelaki, D. E. Inference in the Brain: Statistics Flowing in Redundant Population Codes. *Neuron* **94,** 943–953 (2017).

78. Murphy, K. P. *Machine Learning - A Probabilistic Perspective* ISBN: 978-0-262-01802-9 (MIT Press, 2012).

79. Barber, D. *Bayesian Reasoning and Machine Learning* (Cambridge University Press, 2012).

80. Fleming, S. M., Maloney, L. T. & Daw, N. D. The Irrationality of Categorical Perception. *J. Neurosci.* **33,** 19060–19070 (2013).

81. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14,** 119–130 (2010).

82. Sanborn, A. N. & Chater, N. Bayesian Brains without Probabilities. *Trends Cogn. Sci.* **20,** 883–893 (2016).

83. Gershman, S. J., Vul, E. & Tenenbaum, J. B. Multistability and Perceptual Inference. *Neural Comput.* **24,** 1–24 (2011).

84. Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. One and Done? Optimal Decisions From Very Few Samples. *Cognitive Sci.* **38,** 599–637 (2014).

85. Moreno-Bote, R., Knill, D. C. & Pouget, A. Bayesian sampling in visual perception. *Proc. Natl. Acad. Sci.* **108,** 12491–12496 (2011).

86. Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the Origins of Suboptimality in Human Probabilistic Inference. *Plos Comput. Biol.* **10,** e1003661 (2014).

87. Denison, S., Bonawitz, E., Gopnik, A. & Griffiths, T. L. Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition* **126,** 285–300 (2013).

88. Alday, P. M., Schlesewsky, M. & Bornkessel-Schlesewsky, I. Commentary on Sanborn and Chater: Posterior Modes Are Attractor Basins. *Trends Cogn. Sci.* **21,** 491–492 (2017).

89. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron* **92,** 530–543 (2016).

90. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. *Plos Comput. Biol.* **7,** 1–22 (2011).

91. Pouget, A., Dayan, P. & Zemel, R. Information processing with population codes. *Nat. Rev. Neurosci.* **1,** 125–132 (2000).

92. Beck, J. M., Ma, W., Kiani, R., Hanks, T., Churchland, A., Roitman, J., Shadlen, M., Latham, P. & Pouget, A. Probabilistic population codes for Bayesian decision making. *Neuron* **60,** 1142–1152 (2008).

93. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* **13,** 572–586 (2012).

94. Van Atteveldt, N., Murray, M. M., Thut, G. & Schroeder, C. E. Multisensory Integration: Flexible Use of General Operations. *Neuron* **81,** 1240–1253 (2014).

95. Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. Canonical Microcircuits for Predictive Coding. *Neuron* **76,** 695–711 (2012).

96. George, D. & Hawkins, J. Towards a Mathematical Theory of Cortical Micro-circuits. *Plos Comput. Biol.* **5,** e1000532 (2009).

97. Yu, A. J. & Dayan, P. Uncertainty, Neuromodulation, and Attention. *Neuron* **46,** 681–692 (2005).

98. Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiol. Rev.* **95,** 853–951 (2015).

99. Berridge, K. C. & Kringelbach, M. L. Pleasure systems in the brain. *Neuron* **86,** 646–664 (2015).

100. Körding, K. Decision Theory: What "Should" the Nervous System Do? *Science* **318,** 606–610 (2007).

101. Körding, K. P. & Wolpert, D. M. Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* **10,** 319–326 (2006).

102. Neumann, J. v. & Morgenstern, O. *Theory of Games and Economic Behavior* - ISBN: 140-0-829-461- (Princeton University Press, 2007).

103. Kahnemann, D. & Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47,** 263–292 (1979).

104. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. English. *J. Risk Uncertainty* **5,** 297–323 (1992).

105. Gershman, S. J. & Daw, N. D. in (eds Rabinovich, M., Friston, K. & Varona, P.) chap. Perception, Action and Utility: The Tangled Skein (MIT Press, 2012).

106. Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. Decision making and the avoidance of cognitive demand. *J. Exp. Psychol. Gen.* **139,** 665–682 (2010).

107. Hagura, N., Haggard, P. & Diedrichsen, J. Perceptual decisions are biased by the cost to act. *Elife* **6,** – (2017).

108. Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N. & Pouget, A. The Cost of Accumulating Evidence in Perceptual Decision Making. *J. Neurosci.* **32,** 3612–3628 (2012).

109. Kahneman, D. *Thinking, fast and slow* ISBN: 9780374275631 0374275637 (Farrar, Straus and Giroux, 2011).

110. Simon, H. A. *Rational choice and the structure of the environment.* 1956.

111. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349,** 273–278 (2015).

112. Gershman, S. & Wilson, R. in *Advances in Neural Information Processing Systems 23* 712–720 (2010).

113. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8,** 1704– (2005).

114. McGuire, J. T. & Botvinick, M. M. Prefrontal cortex, cognitive control, and the registration of decision costs. *Proc. Natl. Acad. Sci.* **107,** 7922–7926 (2010).

115. Boureau, Y.-L., Sokol-Hessner, P. & Daw, N. D. Deciding How To Decide: Self-Control and Meta-Decision Making. *Trends Cogn. Sci.* **19,** 700–710 (2015).

116. Schwartenbeck, P., FitzGerald, T., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M. & Friston, K. Optimal inference with suboptimal models: addiction and active Bayesian inference. *Med. Hypotheses* **84(2),** 109–17 (2015).

117. Rahnev, D. & Denison, R. Suboptimality in Perceptual Decision Making. *bioRxiv* (2017).

118. Russell, S. J. & Subramanian, D. Provably Bounded-Optimal Agents. *J. Artif. Intell. Res.* **2,** 575–609 (1995).

119. Mosterín, J. *Lo mejor posible. Racionalidad y acción humana.* (ed Editorial, M. A.) (2008).

120. Mele, A. & Rawling, P. *The Oxford Handbook of Rationality* (Oxford University Press, 2004).

121. Dayan, P. Rationalizable Irrationalities of Choice. *Top. Cogn. Sci.* **6,** 204–228 (2014).

122. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5,** 598– (2002).

123. Ma, W. J. Organizing probabilistic models of perception. *Trends in Cognitive Sciences* **16,** 511–518 (2012).

124. Ma, W. J. & Jazayeri, M. Neural Coding of Uncertainty and Probability. *Annu. Rev. Neurosci.* **37,** 205–220 (2014).

125. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* **88,** 78–92 (2015).

126. Moreno-Bote, R. Decision Confidence and Uncertainty in Diffusion Models with Partially Correlated Neuronal Integrators. *Neural Comput.* **22,** 1786–1811 (2010).

127. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19,** 366–374 (2016).

128. Kiani, R. & Shadlen, M. N. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science* **324,** 759–764 (2009).

129. Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 1322–1237 (2012).

130. Fetsch, C. R., Kiani, R., Newsome, W. T. & Shadlen, M. N. Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron* **83,** 797–804 (2014).

131. Shadlen, M. N. & Kiani, R. Decision Making as a Window on Cognition. *Neuron* **80,** 791–806 (2013).

132. Kiani, R., Corthell, L. & Shadlen, M. N. Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron* **84,** 1329–1342 (2014).

133. Oaksford, M. & Hall, S. On the Source of Human Irrationality. *Trends Cogn. Sci.* **20,** 336–344 (2016).

134. Dienes, Z. & Seth, A. Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Conscious. Cogn.* **19,** 674–81 (2010).

135. Fleming, S. M. & Dolan, R. J. Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and Cognition* **19,** 352–363 (2010).

136. Massoni, S., Gajdos, T. & Vergnaud, J.-C. Confidence Measurement in the Light of Signal Detection Theory. *Front. Psychol.* **5** (2014).

137. Kording, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427,** 244–247 (2004).

138. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415,** 429–433 (2002).

139. Meyniel, F., Schlunegger, D. & Dehaene, S. The Sense of Confidence during Probabilistic Learning: A Normative Account. *Plos Comput. Biol.* **11,** e1004305 (2015).

140. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *Plos Comput. Biol.* **11,** e1004519 (2015).

141. Smith, K. & Vul, E. *Prospective uncertainty: The range of possible futures in physical predictions* in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (2015).

142. Sutton, R. S. & Barto, A. G. *Reinforcement Learning - An Introduction* ISBN: 978-0-262-19398-6 (MIT Press, 1998).

143. Dayan, P. & Niv, Y. Reinforcement learning: The Good, The Bad and The Ugly. *Curr. Opin. Neurobiol.* **18,** 185–196 (2008).

144. Newell, B. R. Re-visions of rationality? *Trends Cogn. Sci.* **9,** 11–15 (2005).

145. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron* **92,** 1398–1411 (2016).

146. Tversky, A. & Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science* **185,** 1124–1131 (1974).

147. Griffin, D. & Tversky, A. The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychol.* **24,** 411–435 (1992).

148. Tversky, A. & Kahneman, D. Belief in the law of small numbers. *Psychol. Bull.* **76,** 105–110 (1971).

149. Kareev, Y., Arnon, S. & Horwitz-Zeliger, R. On the misperception of variability. *J Exp Psychol Gen.* **131,** 287–97 (2002).

150. Kahneman, D. & Tversky, A. Subjective probability: A judgment of representativeness. *Cognitive Psychol.* **3 (3),** 430–454 (1972).

151. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2,** 175 (1998).

152. Oswald, M. & Grosjean, S. in (ed Pohl, R.) chap. Confirmation bias (Taylor & Francis, 2012). ISBN: 9781135844950.

153. Klayman, J. in, 385–418 (1995).

154. Gonzalez, R. & Wu, G. On the Shape of the Probability Weighting Function. *Cognitive Psychol.* **38,** 129–166 (1999).

155. Gigerenzer, G. & Gaissmaier, W. Heuristic Decision Making. *Annu. Rev. Psychol.* **62,** 451–82 (2011).

156. Knill, D. C. Learning Bayesian priors for depth perception. *J. Vision* **7,** 13 (2007).

157. Knill, D. C. & Saunders, J. A. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research* **43,** 2539–2558 (2003).

158. Trommershäuser, J., Landy, M. S. & Maloney, L. T. Humans Rapidly Estimate Expected Gain in Movement Planning. *Psychol. Sci.* **17,** 981–988 (2006).

159. Battaglia, P. W. & Schrater, P. R. Humans Trade Off Viewing Time and Movement Duration to Improve Visuomotor Accuracy in a Fast Reaching Task. *J. Neurosci.* **27,** 6984–6994 (2007).

160. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10,** 1214–1221 (2007).

161. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* **90,** 499–506 (2016).

162. Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B. & Bonatti, L. L. Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference. *Science* **332,** 1054–1059 (2011).

163. Orbán, G., Fiser, J., Aslin, R. N. & Lengyel, M. Bayesian learning of visual chunks by human observers. *Proc. Natl. Acad. Sci.* **105,** 2745–2750 (2008).

164. Van Bergen, R. S., Ji Ma, W., Pratte, M. S. & Jehee, J. F. M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18,** 1728– (2015).

165. Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F. & Kepecs, A. Orbitofrontal Cortex Is Required for Optimal Waiting Based on Decision Confidence. *Neuron* **84,** 1–12 (2014).

166. Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *J. Neurosci.* **30,** 12366–12378 (2010).

167. Summerfield, C. & Tsetsos, K. Do humans make good decisions? *Trends Cogn. Sci.* **19,** 27–34 (2015).

168. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron* **74,** 30–39 (2012).

169. Steup, M. *Epistemology* Metaphysics Research Lab, Stanford University, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.).

170. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2,** 79– (1999).

171. Friston, K. A theory of cortical responses. *Philos. Trans. Royal Soc. B* **360,** 815–836 (2005).

172. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in Neural Circuits with Divisive Normalization. *J. Neurosci.* **31,** 15310–15319 (2011).

173. Braun, D. A., Mehring, C. & Wolpert, D. M. Structure learning in action. *Behav. Brain Res.* **206,** 157–165 (2010).

174. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. **40,** e253– (2017).

175. Green, C. S., Benson, C., Kersten, D. & Schrater, P. Alterations in choice behavior by manipulations of world model. *Proc. Natl. Acad. Sci.* **107,** 16401–16406 (2010).

176. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1,** 0064– (2017).

177. Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci.* **110,** 18327–18332 (2013).

178. Bigelow, J. C. Possible Worlds Foundations for Probability. *J. Philos. Logic* **5,** 299–320 (1976).

179. Kok, P., Mostert, P. & de Lange, F. P. Prior expectations induce prestimulus sensory templates. *Proc. Natl. Acad. Sci.* **114,** 10473–10478 (2017).

180. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* **20,** 1434–1448 (2003).

181. Fleming, S. M., Dolan, R. J. & Frith, C. D. Metacognition: computation, biology and function. *Philos. Trans. Royal Soc. B* **367,** 1280–1286 (2012).

182. Jarvstad, A., Hahn, U., Rushton, S. K. & Warren, P. A. Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proc. Natl. Acad. Sci.* **110,** 16271–16276 (2013).

183. Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. & Shams, L. Causal Inference in Multisensory Perception. *Plos One* **2,** e943 (2007).

184. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46,** 1004–1017 (2009).

185. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng-t. Asme.* **82,** 35–45 (1960).

186. Trommershaeuser, J., Gepshtein, S., Maloney, L. T., Landy, M. S. & Banks, M. S. Optimal Compensation for Changes in Task-Relevant Movement Variability. *J. Neurosci.* **25,** 7169–7178 (2005).

187. Purcell, B. A. & Kiani, R. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc. Natl. Acad. Sci.* (2016).

188. Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B. & Griffiths, T. L. in (eds Busemeyer, J., Townsend, J., Wang, Z. & Eidels, A.) 187–208 (Oxford University Press, 2015).

189. Lucas, C. G., Griffiths, T., Williams, J. & Kalish, M. A rational model of function learning. *Psychon Bull Rev.* **22(5),** 1193–215 (2015).

190. Shen, S. & Ma, W. A detailed comparison of optimality and simplicity in perceptual decision making. *American Psychological Association* **123(4),** 452–80 (2016).

191. Bowers, J. S. & Davis, C. J. Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* **138,** 389–414 (2012).

192. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).

193. Lochmann, T. & Deneve, S. Neural processing as causal inference. *Current Opinion in Neurobiology* **21,** 774–781 (2011).

194. Baxter, J. in (eds Thrun, S. & Pratt, L.) 71–94 (Springer, Boston, MA, 1998).

195. Torrey, L. & Shavlik, J. Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* **1,** 242 (2009).

196. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning* (MIT press Cambridge, 2016).

197. Deroy, O., Spence, C. & Noppeney, U. Metacognition in Multisensory Perception. *Trends Cogn. Sci.* **20,** 736–747 (2016).

198. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8** (2014).

199. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. Royal Soc. B* **367,** 1310–1321 (2012).

200. Fetsch, C. R., Kiani, R. & Shadlen, M. N. Predicting the Accuracy of a Decision: A Neural Mechanism of Confidence. *Cold Spring Harb. Sym.* (2015).

201. Acuna, D. & Schrater, P. R. in *Advances in Neural Information Processing Systems 21* 1–8 (2009).

202. Dasgupta, I., Schulz, E. & Gershman, S. J. Where do hypotheses come from? *Cognitive Psychol.* **96,** 1–25 (2017).

203. Balaguer, J., Spiers, H., Hassabis, D. & Summerfield, C. Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. *Neuron* **90,** 893–903 (2016).

204. Krynski, T. R. & Tenenbaum, J. B. *The role of causality in judgment under uncertainty.* 2007.

205. Hilbert, M. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychol. Bull.* **138,** 211–237 (2012).

206. Costello, F. & Watts, P. Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychol. Rev.* **121(3),** 463–480 (2014).

207. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **95,** 245–258 (2017).

208. Kumaran, D., Hassabis, D. & McClelland, J. L. What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends Cogn. Sci.* **20,** 512–534 (2016).

209. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron* **69,** 1204–1215 (2011).

210. Lieder, F., Hsu, M. & Griffiths, T. L. *The high availability of extreme events serves resource-rational decision-making* in *Proc. 36th Ann. Conf. Cognitive Science Society* (2014).

211. Van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N. & Wolpert, D. M. Confidence Is the Bridge between Multi-stage Decisions. *Curr. Biol.* **26,** 3157–3168 (2016).

212. Fletcher, P. C. & Frith, C. D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.* **10,** 48–58 (2009).

213. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10,** 186–(2009).

214. Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U. & Bullmore, E. Functional Connectivity and Brain Networks in Schizophrenia. *J. Neurosci.* **30,** 9477–9487 (2010).

215. Ackerman, R. & Thompson, V. A. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends Cogn. Sci.* **21,** 607–617 (2017).

216. Oaksford, M. & Chater, N. A rational analysis of the selection task as optimal data selection. *Psychol. Rev.* **101,** 608 (1994).

217. Hertwig, R. & Gigerenzer, G. The'conjunction fallacy'revisited: How intelligent inferences look like reasoning errors. *J. Behav. Decis. Making* **12,** 275 (1999).

218. Stenning, K. & Lambalgen, M. V. Semantics as a Foundation for Psychology: A Case Study of Wason's Selection Task. *Journal of Logic, Language, and Information* **10,** 273–317 (2001).

219. Hartmann, S. & Meijs, W. Walter the banker: the conjunction fallacy reconsidered. *Synthese* **184,** 73–87 (2012).

220. Mandel, D. R. The psychology of Bayesian reasoning. *Front. Psychol.* **5** (2014).

221. Martí, L., Mollica, F., Piantadosi, S. & Kidd, C. *What Determines Human Certainty?* in *CogSci 2016* (2016).

222. Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E. & Lau, H. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* **1,** 0139– (2017).

223. Jones, M. & Love, B. C. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* **34,** 169–188 (04 2011).

224. Hahn, U. The Bayesian Boom: Good Thing or Bad? *Front. Psychol.* **5** (2014).

225. Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W. & De Martino, B. Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron* **96,** 348–354.e4 (2017).

226. Binda, P. & Murray, S. O. Keeping a large-pupilled eye on high-level visual processing. *Trends Cogn. Sci.* **19,** 1–3 (2015).

227. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. **8,** 14637– (2017).

228. Selimbeyoglu, A., Keskin-Ergen, Y. & Demiralp, T. What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision. *Clin. Neurophysiol.* **123,** 1158–1167 (2012).

229. Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E. & Bahrami, B. The idiosyncratic nature of confidence. *Nat. Hum. Behav.* **1,** 810–818 (2017).

230. Oechssler, J., Roider, A. & Schmitz, P. W. Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization* **72,** 147–152 (2009).

231. Kool, W. & Botvinick, M. *A labor/leisure tradeoff in cognitive control.* 2014.

232. Legg, S. & Hutter, M. *A Collection of Definitions of Intelligence* in *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006* (2007), 17–24.

233. Legg, S. & Hutter, M. Universal Intelligence: A Definition of Machine Intelligence. *Mind. Mach.* **17,** 391–444 (2007).

234. Peterson, C. R. & Beach, L. R. Man as an intuitive statistician. *Psychol. Bull.* **68(1),** 29–46 (1967).

235. Sustein, C. & Thaler, R. *Nudge: Improving decisions about health, wealth, and happiness* 2008.

236. Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G. & Frith, C. D. Optimally Interacting Minds. *Science* **329,** 1081–1085 (2010).

237. Frith, C. D. The role of metacognition in human social interactions. *Philos. Trans. Royal Soc. B* **367,** 2213–2223 (2012).

238. Massoni, S. & Roux, N. Optimal group decision: A matter of confidence calibration. *J. Math. Psychol.* **79,** 121–130 (2017).

239. Ash, R. & Doléans-Dade, C. *Probability and Measure Theory* ISBN: 9780120652020 (Harcourt/Academic Press, 2000).