Multiverse analysis, Bayesian decision theory and agentic AI: drawing connections

Giovanni Parmigiani[1,2*] and Alessandro Zito[1,2]

[1*]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA.
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

*Corresponding author(s). E-mail(s): gp@ds.dfci.harvard.edu;
Contributing authors: azito@hsph.harvard.edu;

Abstract

Multiverse analysis explores the sensitivity of statistical conclusions to the many choices analysts make in the course of a data analysis, which that literature calls analytic decisions. Decision theory studies more generally the choice of an action whose consequences depend on unknown states of the world. This article draws connections between multiverse analysis and formal frameworks for decision making under uncertainty, robust Bayesian decisions in particular, and connects both to agentic artificial intelligence (AI), the design of artificial agents that make decisions autonomously in uncertain environments. We review Bayesian decision theory, robustness in Bayesian learning and in Bayesian decisions, multiverse analysis, and multi-agent systems built from large language models. We then compare multiverse and robust Bayesian analysis along the elements of a decision problem. We describe multiverses of Bayesian analyses and model averaging, and outline a staged approach from local diagnostics to global robustness, to the robustness of decisions, and to the multiverse. We associate each point of the multiverse with the perspective of an individual agent, and connect this reading to Dawid's notion of a bevy of Bayesians, to multi-Bayesian decision theory, and to agentic AI. In the last, a multi-agent system implements a multiverse wherein agents with different perspectives specify the analytic choices. We apply this concept to an example where twenty language-model agents estimate a clinical referral cutoff based on a psychological test.

Keywords: Multiverse analysis, Robust Bayesian analysis, Bayesian decision theory, Multi-agent systems, Agentic artificial intelligence

1 Introduction

Multiverse analysis, Bayesian decision theory, and agentic AI are three areas of research that have developed largely independently, but which share a common concern with decisions under uncertainty. This paper explores connections between these areas and provides a more comprehensive understanding of how multiverse analysis can inform and be informed by decision theory in practice.

The paper is organized as follows. Section 2 provides concise reviews of the key concepts and methods from each area, highlighting their relevance for the connections drawn in Section 3. Section 3.1 draws parallels between multiverse analysis and robust Bayesian decision theory, highlighting similarities and differences between the two approaches, and discusses how multiverse analysis can inform decision-making in practice in settings such as model averaging and staged approaches to robustness. Section 3.2 discusses how the points in a multiverse can be thought of as coming from analyses of separate agents. Section 3.3 further explores the connections between multiverse analysis and agentic AI, highlighting how multi-agent systems can implement a multiverse by having agents with different perspectives specify the analytic choices. Section 4 illustrates these concepts in a practical setting, using twenty language-model agents each estimating a referral cutoff in a medical application. Section 5 concludes.

2 Reviews

This section briefly reviews the key aspects of each area. Section 2.1 discusses the key elements behind Bayesian decision analysis. Sections 2.2 and 2.3 discuss robustness in Bayesian learning and Bayesian decisions. Section 2.4 introduces multiverse analysis. Section 2.5 presents a primer on agentic AI and its connection with decision theory.

2.1 Bayesian decision analysis

2.1.1 Elements of a decision problem

Bayesian decision theory is a formal framework for making decisions under uncertainty. It combines probability theory and utility theory to provide a systematic approach for choosing among alternative actions based on their expected outcomes. The framework is grounded in the subjective interpretation of probability, where probabilities represent an individual's beliefs about uncertain events, and utility represents the consequences of possible outcomes (Raiffa and Schlaifer 1961; Robert 1994; Parmigiani and Inoue 2009). Formally, a decision maker (DM) chooses an action a within a set of actions, called $\mathcal{A}$, whose consequences depend on an unknown state of the world θ, which belongs to a set of states called Θ. For example, a traveler may have the option to purchase travel insurance before a flight. The consequences of this decision depend on unknowns about the future, namely the events (that is, the states of the world) that would trigger a claim. The basis for choosing among actions $a \in \mathcal{A}$ is a quantitative assessment of their consequences, which are measured by a *utility function* $u(a, \theta)$, that depends on both a itself and the state of the world $\theta \in \Theta$.Since the true θ is unknown, the DM must account for its uncertainty when weighting options.

Table 1 Elements of a subjective Bayesian decision problem.

Symbol	Interpretation
$a \in \mathcal{A}$	an action; $\mathcal{A}$ is the set of available actions
$\theta \in \Theta$	a state of the world; Θ is the set of possible states
$u(a, \theta)$	the utility of the consequence of action a when the state is θ
$\pi(\theta)$	the DM's probability distribution of the state of the world
x, $f(x\|\theta)$	the outcome of an experiment, and its probability distribution given θ
$\pi(\theta\|x)$	the posterior distribution of θ given x
$\delta(x)$	a decision rule, used to turn data into actions
$\mathcal{U}_{\pi_x}(a)$	the expected utility of a under the posterior $\pi_x = \pi(\cdot\|x)$
a^*, δ^*	a Bayes action; a Bayes decision rule

This uncertainty is quantified by a probability distribution $\pi(\theta)$, which is specified by the DM.

From a statistical perspective, interesting questions arise when the outcome x of an experiment whose distribution $f(x|\theta)$ depends on θ is available. In this case, the DM translates data into action using a decision function, or decision rule, which is any function $\delta(x)$ of the data that returns values in $\mathcal{A}$.

The core principle behind Bayesian decision theory is the *expected utility principle*: a rational DM should choose the action that maximizes the expected utility of the consequences, where the expectation is taken with respect to the DM's own probability distribution over the states of the world. When the outcome x of an experiment is available, this probability is the posterior distribution, obtained via Bayes' rule as $\pi(\theta|x) \propto \pi(\theta) f(x|\theta)$. Then, the expected utility of action a is

$$\mathcal{U}_{\pi_x}(a) = \int_{\Theta} u(a, \theta)\, \pi(\theta|x)\, d\theta, \tag{1}$$

and an action a^* that maximizes it is called a Bayes action. Before the experiment, or when there is none, the same expressions hold with $\pi(\theta)$ in place of $\pi(\theta|x)$. Since the Bayes action depends on x, the principle implicitly defines a decision rule, which we call the Bayes decision rule $\delta^*(x)$. Table 1 summarizes the notation for all the elements of a decision problem. The next section presents an illustrative example from the psychology literature.

2.1.2 A practical illustration

van der Linden and Mellenbergh (1977) consider a situation in which a total test score is used to classify examinees into two categories: accepted and not accepted. A latent variable, unknown to the examiner, represents the true ability of each examinee, and a value θ_0 of the latent variable, fixed in advance, separates suitable examinees ($\theta \geq \theta_0$) from not suitable ones. In the notation of Table 1, the DM is the examiner, the two actions are to accept and to not accept the examinee, θ is the examinee's value on the latent variable, x is the observed total score and $f(x|\theta)$ its measurement model, and $\pi(\theta)$ represents the examiner's knowledge, prior to the test, of the examinee's

ability. If nothing is known about such ability, a reasonable choice is to set $\pi(\theta)$ as the distribution of the latent variable in the population of examinees. We assume that the utility of the examiner is linear in θ, and is obtained by the following set of equations:

$$u(\text{accept}, \theta) = \gamma_1 + \beta_1(\theta - \theta_0), \quad u(\text{not accept}, \theta) = \gamma_0 + \beta_0(\theta_0 - \theta), \quad \beta_0 + \beta_1 > 0. \tag{2}$$

Each part of the utility function has two components. The constants γ_1 and γ_0 are the effects on utility of accepting and of not accepting an examinee, irrespective of ability. The terms $\beta_1(\theta - \theta_0)$ and $\beta_0(\theta_0 - \theta)$ are proportional to the distance between the examinee's ability and θ_0: accepting is worth more the further the ability is above θ_0 and less the further it is below, and *vice versa*. Notice that a constant utility would instead assign the same value to not accepting an examinee just above θ_0 and one far above it. The four coefficients are chosen so that the utility represents the psychological, social, and economic consequences of both actions.

By equation (1), the action of accepting has the larger posterior expected utility if and only if

$$E(\theta|x) > \theta_0 + \frac{\gamma_0 - \gamma_1}{\beta_0 + \beta_1}, \tag{3}$$

where $E(\theta|x)$ is the mean of the latent variable under the posterior $\pi(\theta|x)$. The coefficients enter the choice of action only through the ratio in equation (3), which moves the bar for the posterior mean above θ_0 when not accepting carries the larger constant utility and below θ_0 when accepting does. When the two constant utilities are equal, the examinee is accepted when the posterior mean of the ability exceeds θ_0. The Bayes rule $\delta^*(x)$ accepts the examinee when the observed score is at least the smallest x for which equation (3) holds.

2.1.3 Justifications and limitations

The expected utility principle has been justified by translating it into interpretable axioms characterizing rational choice (Savage 1954; Anscombe and Aumann 1963); see Kreps (1988) for a critical discussion. However, in many applications, it is neither desirable nor practical to take the expected utility view in literal terms: some decisions involve aspects that are difficult to quantify reliably along one-dimensional utilities, and the level of quantification required is not always achievable, because of difficulties in relating a DM's values to quantitative utilities or because of gaps in knowledge (Parmigiani 2002). Hence, as a description of behavior, the principle is often violated in practice. Prospect theory (Kahneman and Tversky 1979), for example, was developed to describe some of the most important patterns of deviations from expected utility. A comprehensive review on the topic is presented by Gilboa (2025), who also discusses alternative theories proposed for choices with known and with unknown probabilities, and the more recent evidence that challenges those alternatives.

2.2 Robustness in Bayesian learning

Robust Bayesian methods investigate how sensitive conclusions are to the specification of the prior and the sampling distribution. The goal is to assess whether the conclusions are robust to plausible departures from the operational model, and if not, to identify

what further elicitation or data collection is necessary. The following review is based on Berger (1994) and Watson and Holmes (2016), with notation as in those sources.

2.2.1 Approximate models and global robustness

In a strict Bayesian framework, the DM specifies beliefs through the joint model of data and parameter, namely $p(\theta, x) = f(x|\theta)\pi(\theta)$, conditions on the data, and acts according to equation (1). A difficulty, in the words of Watson and Holmes (2016), is that this "assumes perfect precision in specifying" the joint model, captured by π and f. In reality, both quantities are approximations, as acknowledged by the founders of subjective probability. de Finetti, as quoted by Watson and Holmes, held that it is important to know not only the exact answer for an exactly specified initial problem, but "what happens [when] changing in a reasonable neighborhood the assumed initial opinion". However, eliciting a prior is a difficult problem in practice. The purpose of robust Bayesian methods is to operate with a partial specification, compute the corresponding range of answers, and suggest whether further elicitation is necessary and what to elicit (Berger 1994).

One plausible approach in robust Bayes is to specify an operational model, that is, a prior π_0 and a sampling distribution f_0, to the best of the available time and ability, and then to investigate the effect of departures from it. For example, *local* approaches consider derivatives of the posterior quantities of interest with respect to perturbations of the operational model (Carota et al. 1996). Instead, *global* approaches consider the variation of these quantities as the prior and the sampling distribution vary in classes containing the operational model. These departures are typically taken in the prior, with f held fixed at distribution f_0.

In this regard, any quantity of scientific interest, such as the posterior mean of θ or the posterior probability of an event pertaining to it, can generally be expressed as functionals of the posterior distribution, namely $\psi(\pi, f) = \int h(\theta)\pi(\theta|x)d\theta$. When f is fixed, we write $\psi(\pi)$. Then, global robustness assumes that $\pi \in \Gamma$ and $f \in \mathcal{F}$ for some classes of densities Γ and $\mathcal{F}$ containing π_0 and f_0, respectively, and it computes the following lower and upper bounds:

$$\underline{\psi} = \inf_{\pi\in\Gamma} \inf_{f\in\mathcal{F}} \psi(\pi, f), \qquad \overline{\psi} = \sup_{\pi\in\Gamma} \sup_{f\in\mathcal{F}} \psi(\pi, f). \tag{4}$$

The conclusion is considered robust if the interval $[\underline{\psi}, \overline{\psi}]$ is narrow enough for it to be sufficiently clear. If the range is too wide, further elicitation, data collection, or analysis is recommended (Berger 1994).

The most common choice for the class of sampling distributions is a parametric one. For instance, one may choose Student t densities scaled to share their quartiles with the normal density of the operational model, so that θ keeps the same meaning across $\mathcal{F}$. For the prior, a challenge is to define the nature and size of Γ so as to capture the plausible ambiguity in π_0, while taking into account ease of specification and computational tractability. Berger proposes a set of desirable properties that Γ should maintain. First, it should be easy to elicit and interpret, and extendable to higher dimensions. Second, its distributions should be easy to handle computationally, which

usually means that the extrema in equation (4) are attained on a low-dimensional set of extreme points. Third, the size of Γ should reflect the vagueness of the prior information, neither so small that robustness is concluded erroneously nor so large that it contains clearly unreasonable priors. Classes that satisfy such properties can be defined in various manners. For example, one can specify the shape or smoothness of the density. Alternatively, one can specify the moments directly, or set bounds on the prior probabilities of specified subsets of Θ (e.g., the quantile class) or the density itself (e.g., the density ratio class). A third approach consists of specifying mixtures of a family of densities indexed by a latent quantity λ, namely $\pi(\theta) = \int \pi(\theta|\lambda)\, dG(\lambda)$, with the mixing distribution G in a class $\mathcal{G}$. Finally, one can define a flexible class by applying a contamination of the base prior as

$$\Gamma = \{\pi : \pi = (1-\epsilon)\pi_0 + \epsilon q,\ q \in \mathcal{Q}\}, \tag{5}$$

where ϵ is the perceived possible error in π_0 and $\mathcal{Q}$ is the allowed class of contaminations. Berger's own preference goes to the contamination and quantile classes, because they are the easiest to elicit and interpret, and to mixture classes, because in multidimensional problems the range of answers is typically very wide unless Γ is constrained so as not to contain unreasonable distributions, a behavior which is easily obtainable using mixtures. The key difference with Watson and Holmes is that they place the neighborhood on the posterior rather than on the prior, and on the states that enter the decision: either a ball in Kullback–Leibler divergence around the posterior of the operational model, over which a range as in equation (4) is computed, or a nonparametric distribution over probability measures centered at that posterior, such as a Dirichlet process, from which the variation of ψ is averaged rather than bounded. We further comment on the difference in Section 2.3.

2.2.2 Limitations of global robustness approaches

Not all the points in the classes are equally plausible in view of the data. The marginal density of the data, $m(x|\pi, f) = \int f(x|\theta)\pi(\theta)d\theta$, can be considered to be the marginal likelihood of π and f, and the bounds in equation (4) only consider the extremes, irrespective of $m(x|\pi, f)$. These extremes may be attained at priors that predict the observed data poorly, that is, at π with small $m(x|\pi, f)$. The range is then wide because of priors that the data speak against, and this consideration makes bounds less attractive as a summary. In Berger's example, the class of priors for a vector of normal means is indexed by a common prior mean μ allowed to vary over a wide interval, and the range of the posterior mean of one coordinate is correspondingly wide. However, $m(x|\pi, f)$ with f fixed, as a function of μ, is a normal likelihood centered at the sample mean with variance $1/p$ for p coordinates. This means that only a short interval of values of μ has appreciable likelihood, and over that interval the range of the posterior mean is narrow. A global analysis might therefore indicate a lack of robustness where a full Bayesian analysis, which weights each π by $m(x|\pi, f)$, would find robustness.

Berger's remedies are a higher-level robustness analysis, in which the uncertainty is moved to a class of priors on the quantities that index Γ, and the replacement of Γ by $\{\pi \in \Gamma : m(x|\pi, f) \geq K\}$ for a constant K. The second is related to empirical

Bayes, which retains only the prior maximizing $m(x|\pi, f)$, and can give the wrong answer, so he does not definitively recommend it. Besides global robustness, Berger reviews the construction of inherently robust procedures, through flat-tailed models and noninformative, partially informative or nonparametric priors, and the diagnostics that detect that a robustness problem exists and where it lies. For further reviews on diagnostics for decisions, see Watson and Holmes (2016).

2.2.3 Bayesian reasoning with imprecise probabilities

The class Γ can also take the role that the single prior π has in the standard setting. This approach is explored in the theory of imprecise probabilities (Walley 1991), where a set of probability distributions replaces the single distribution as the basic modeling entity. Augustin (2022) surveys the statistical uses and counts robust Bayesian analysis among the approaches "directly working with sets of probabilities". The set "has to be understood and treated as an entity of its own": its elements are not to be mixed into a precise distribution, and it "can naturally be chosen to reflect the quality/determinacy of prior knowledge: strong prior knowledge leads to 'small' sets; weak prior knowledge to 'large' sets". A set of priors Γ induces the set of posteriors $\{\pi(\theta|x) : \pi \in \Gamma\}$, which Walley's theory of coherence justifies as the generalized Bayes rule. The width of the set of posteriors carries information, as it contracts with the sample size when data and prior agree and is inflated under prior-data conflict at intermediate sample sizes (Augustin 2022). A set of distributions can be updated on data by more than one rule. Gong and Meng (2021) compare three such rules and show that, unlike Bayes' rule with a single prior, each can behave in unexpected ways: the updated set can be wider than the initial set whatever the data show (dilation), or narrower whatever the data show (contraction), or imply bets that lose whatever the outcome (sure loss). Two of the three rules also cannot update at all when the set of priors is uninformative, however strong the data. Caprio and Gong (2023) study a related updating scheme, in which an agent whose initial beliefs are a set of probabilities revises them on partial information, and point to conditions, given in a companion preprint, under which the updated set contracts, dilates, or incurs sure loss.

2.3 Robustness in Bayesian decisions

2.3.1 Stability of the Bayes action

In a decision problem, the quantities whose stability matters are the posterior expected utility of each action and the Bayes action itself. Watson and Holmes hold that the consequence of an approximate model is contextual and should be dealt with in a decision theoretic framework (Watson and Holmes 2016; Parmigiani and Inoue 2009): a model can be adequate for one decision and inadequate for another, as the same normal approximation to an exponential distribution serves the estimation of the mean and fails the estimation of the probability of an interval. The utility function is an input like the prior and the sampling distribution, and imprecision in its specification enters in the same way. Kadane and Chuang (1978) provide an early conceptualization of this issue: they characterize a decision problem by a utility and an opinion and call the pair stable when small changes in both, weak convergence of the opinion and

uniform convergence of the utility, produce small changes in the expected utility of nearly optimal actions. They find that squared error is unstable with any opinion when the parameter space is the real line, while any bounded utility that is continuous in θ uniformly in a is stable with any opinion. Later work treats imprecision in beliefs and in preferences together (Ríos Insua 1990; Ríos Insua and Martín 1994); Berger et al. (2000) review it. Nevertheless, robustness with respect to the utility has received less attention than robustness with respect to the prior.

2.3.2 Measures of robustness

Let Λ be a class of utility functions containing the operational utility u_0, alongside the classes Γ and $\mathcal{F}$ of Section 2.2. Abraham and Cadre (2004) list the usual measures of global robustness for a decision: the range of the posterior expected utility of each action as the inputs vary in their classes,

$$\underline{\mathcal{U}}(a) = \inf_{u\in\Lambda,\, \pi\in\Gamma,\, f\in\mathcal{F}} \mathcal{U}_{\pi_x}(a), \qquad \overline{\mathcal{U}}(a) = \sup_{u\in\Lambda,\, \pi\in\Gamma,\, f\in\mathcal{F}} \mathcal{U}_{\pi_x}(a); \tag{6}$$

the set of Bayes actions obtained as the inputs vary; and the maximum regret of acting on the operational inputs, that is the largest amount, over the classes, by which the posterior expected utility of the operational Bayes action a^* falls short of the largest posterior expected utility attainable under the same inputs. They show that, as the sample size n grows, the range of Bayes actions shrinks at rate $\sqrt{n}$ and the other two measures at rate n. The measures in (6) share the limitation of Section 2.2: they consider the extremes of the classes irrespective of how plausible each member is.

A different view starts from the map that sends each pair of inputs (π, u) in $\Gamma \times \Lambda$ to the action it makes optimal. The sets of inputs that lead to the same action partition the class of inputs: an action is potentially optimal if some inputs in the class lead to it, and it is a competitor of the operational Bayes action if the inputs that lead to it border those that lead to the operational Bayes action. The distance from the operational inputs (π_0, u_0) to the nearest inputs that lead to a different action, in a chosen metric, is an index of the sensitivity of the action, and points to what would have to be elicited further. This is the framework of Ríos Insua (1990) and Ríos Insua and French (1991), developed for a discrete set of alternatives and judgmental inputs, probabilities and utilities together, constrained to a set of plausible values, and organized in four steps: removal of dominated alternatives, identification of the potentially optimal ones, of those adjacent to the current optimum, and distance analysis. Its extension to a continuous set of actions and to statistical decision theory, with the prior and the utility varying together, is in Ríos Insua et al. (1997). The question itself, how far the probabilities can be in error before the optimal act changes, goes back at least to Isaacs (1963).

Watson and Holmes (2016) place the class on the posterior of the operational model, $\pi_0(\theta|x)$, restricted to the states that enter the utility, and take it to be the ball $\Gamma_C = \{\pi : \mathrm{KL}(\pi \,\|\, \pi_0(\cdot|x)) \le C\}$, where KL is the Kullback–Leibler divergence and C its radius. For a bounded utility, the member of Γ_C with the smallest expected utility

for action a has the closed form

$$\pi^{\inf}_{a,C}(\theta) \propto \pi_0(\theta|x) \exp\{-\eta\, u(a,\theta)\}, \tag{7}$$

the posterior tilted exponentially toward low utility, with $\eta \geq 0$ a function of C; the member with the largest expected utility tilts in the opposite direction. Plotting the range of expected utility of each action against C shows the radius at which the smallest expected utility of a^* first falls below the largest expected utility of some other action, that is the size of misspecification that would change the action; the derivative of the smallest expected utility at $C = 0$ is the variance of the utility under $\pi_0(\theta|x)$, with a negative sign, which serves as a local measure. As an alternative to bounds, and closer to the Bayesian standpoint, they place a probability distribution on the members of a neighborhood, a Dirichlet process centered at $\pi_0(\theta|x)$, and obtain for each action a distribution of expected utilities rather than a range.

2.4 Multiverse analysis

Steegen et al. (2016) introduced multiverse analysis for the choices made in constructing a data set from raw data: "instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios". The idea was extended to the choice of the statistical model (Simonsohn et al. 2020; Girardi et al. 2024); Altoè et al. (2025) describe it as "a systematic exploration of all reasonable analytical paths". In the notation of Table 1, a specification $k = 1, \ldots, K$ is one combination of processing and modeling choices; it determines a processed data set x_k and a model f_k, both indexed by the specification since several specifications may share one data set or one model, and yields an estimate or a test, written t_k, of a quantity that the specifications share, for instance the coefficient of a given predictor (Girardi et al. 2024). Four aspects of how the literature conceptualizes the multiverse matter for the parallels drawn later.

2.4.1 The set of specifications

The set is defined by the analyst as the combinations of "reasonable" options at each analytic choice (Steegen et al. 2016); Simonsohn et al. (2020) require specifications that are "theoretically justified, statistically valid and non-redundant". Del Giudice and Gangestad (2021) give criteria for the inclusion of an alternative, and distinguish three types of analytic decisions, in their terminology: Type E, principled equivalence, where alternatives are expected to be practically equivalent and the choice among them "can be regarded as effectively arbitrary"; Type N, principled nonequivalence, where some alternatives are better justified than others and the choice is made rather than varied; and Type U, uncertainty, where there is not enough information to tell, and for which they recommend separate, exploratory multiverses. In their words, Type E decisions "are properly used to populate a homogeneous multiverse". Heyman et al. (2025) propose to define the set by crowdsourcing the specifications from several analysts, and Aczel et al. (2026) let many analysts each follow their own judgment on the same data.

Cantone and Tomaselli (2024b) observe that the specifications are not independent draws and model the set as a network of connected specifications.

2.4.2 Weights and summaries

In the original proposal all specifications are reported, each with the same standing, and the analysis "gives pointers as to which choices are most consequential" (Steegen et al. 2016). Specification curve analysis displays the estimates sorted by size and adds a joint inference across all specifications (Simonsohn et al. 2020); PIMA combines the standardized test statistics of the K models through a function that is nondecreasing in each argument, such as the mean, a weighted mean, or the maximum (Girardi et al. 2024). Del Giudice and Gangestad (2021) note that a median or mean effect across the multiverse summarizes it when the specifications that estimate the effect of interest are the majority. Cantone and Tomaselli (2024a) compare uniform weighting of the estimates with Bayesian model averaging and report that in well-specified models the alternatives did not perform better than uniform weighting; they place this within "the ongoing debate on how to weigh a multiverse" (Cantone and Tomaselli 2024b).

2.4.3 The role of extremes

The variability of the results across specifications is itself a finding. Patel et al. (2015) quantify the vibration of effects by the variance of the effect size and of the p value across all combinations of adjustment covariates; specification curves are read for the "consequential specifications decisions" (Simonsohn et al. 2020), and the chapter describes the analysis as identifying "points of fragility, where results are highly sensitive to specific decisions" (Altoè et al. 2025). In inference the extreme can carry the test: with the maximum statistic, the global test of PIMA rejects when at least one specification does, with strong control of the family-wise error rate, so that the researcher may then select the significant specifications without inflating the error (Girardi et al. 2024). Aczel et al. (2026) report the spread across independent re-analyses of the same claim: for 100 studies, 34% of re-analyses gave the same result as the original within 0.05 in Cohen's d, 74% reached the same conclusion, 24% an inconclusive result, and 2% the opposite conclusion.

2.4.4 Working model

Whether one specification serves as an operational model, in the sense of Section 2.2, varies. The multiverse of Steegen et al. (2016) and the specification curve of Simonsohn et al. (2020) have no operational specification. The chapter contrasts three reporting strategies: a traditional analysis, which considers several branches and reports one selected path; a planned analysis, which follows a single predefined branch; and a multiverse analysis, which reports all plausible branches (Altoè et al. 2025). Götz et al. (2024) describe the multiverse as the background against which "the results from a particular set" of choices, a universe, are examined. In PIMA the researcher is free to choose "the preferred model post hoc, after trying all models and seeing the results", with adjusted p values (Girardi et al. 2024); in the principled multiverse of Del Giudice and Gangestad (2021), Type N decisions fix a preferred option; and in the

multi-analyst design the original report is the reference against which the re-analyses are compared (Aczel et al. 2026).

2.5 Agentic AI in decision analysis

2.5.1 Agents and multi-agent systems

A large language model (LLM) is a neural network language model, trained on a very large corpus of text, often to predict the next token, whose size has crossed the scale at which it can carry out tasks stated to it in natural language. The literature on AI agents is showing a rapid growth in recent months. See for example Zhao et al. (2026b) for a survey of their construction, adaptation, and evaluation. An agent, in this literature, is an LLM equipped to perceive, reason, plan and act. Wang et al. (2024) review the construction, applications, and evaluation of such agents, while Guo et al. (2024) state that "LLM-based agent systems have rapidly evolved from single-agent planning or decision-making to operating as multi-agent systems", review the domains in which these systems operate, the profiling and communication of the agents, and the means by which they develop skills. A first categorization of the multi-agent system is presented in Tran et al. (2025), who classify the collaboration mechanisms of by "actors (agents involved), types (e.g., cooperation, competition, or coopetition), structures (e.g., peer-to-peer, centralized, or distributed), strategies (e.g., role-based or model-based), and coordination protocols". Furthermore, Jin et al. (2025) survey cooperative decision making by multiple agents more broadly, including the simulation environments and the allocation of rewards among agents.

In the notation of Table 1, agents $k = 1, \ldots, K$ work on a shared problem and typically on shared data x. Each agent produces an output, an estimate, a probability statement or a recommended action from its own reading of the problem, that is, with its own implicit prior π_k and utility u_k. Then, a coordinating agent, or a fixed rule, turns the K outputs into the action a of the system. Importantly, this formulation does not assume that the agents' probabilities and utilities match those of the DM.

2.5.2 Multi-agent decision support

Medicine is one field where these systems have been applied to decision-making. Gridach et al. (2026) review agentic AI in healthcare, categorize single-agent and multi-agent frameworks, and "examine the communication, reasoning, and decision-making processes of these AI agents", with applications to electronic health records, clinical triage, medical question answering and diagnosis. Four examples show the architectures in use. MDAgents (Kim et al. 2024) assigns to a team of LLMs a collaboration structure, solo or group, "tailored to the medical task at hand, emulating real-world medical decision-making processes adapted to tasks of varying complexities". MoMA (Gao et al. 2025) uses specialist agents to convert images and laboratory results into text summaries, an aggregator agent to combine them with the clinical notes, and a predictor agent to produce the clinical prediction. DeepRare (Zhao et al. 2026a) is a multi-agent system for rare disease differential diagnosis, with more than 40 tools, that returns ranked diagnostic hypotheses "with transparent reasoning linked to verifiable medical evidence". Wu et al. (2026) build an orchestrator–specialist system for

secondary headache: the orchestrator decomposes the case, routes it to the specialist agents whose red flag it involves, which execute in parallel, each independently, and aggregates their outputs into a unified structured result. The authors ground the design in the multi-agent principle that "subdividing a complex task into well-defined components and assigning them to autonomous agents improves reliability, efficiency, and flexibility".

Two recent evaluations are instructive to understand the breadth of the application. By benchmarking two agent systems on clinical decision tasks, Liu et al. (2026) find that access to tools yielded only modest accuracy gains over baseline LLMs. Zhu et al. (2025) audit systems that "emulate multidisciplinary consultation through specialist roles, peer review and consensus formation" and derive, from 3,600 execution logs, a taxonomy of ten recurrent failures "spanning task comprehension, collaborative discussion, and synthesis and decision-making". Their title names the risk of false consensus, and they hold that in clinical decision support "apparent consensus is not enough", since clinicians "also need to know whether agents checked the evidence, addressed disagreement and kept uncertainty visible". The elements of a decision analysis appear in this literature under other names: the collaboration structure, the aggregator, the orchestrator and the consensus rule are the devices by which the outputs of several agents become one action.

3 Parallels

All the concepts reviewed in Section 2 share the goal of understanding the influence of judgment on conclusions and actions grounded on data. In this section we highlight similarities and differences in these approaches. They are largely complementary and practitioners may also be interested in using them together, so we sketch a few ways in which that can be done.

3.1 Multiverse and robust Bayes

3.1.1 Data processing as part of the model

An important clarification concerns the distinction between judgment on the processing of the data and judgment on the choice of a model. The multiverse literature typically separates the two: a specification first produces a processed data set x_k, by choices about inclusion, transformation, coding and imputation, and then fits a model f_k to it. A Bayesian analysis that follows Section 2.1 to the letter makes no such separation. If there is processing, it is considered part of the model. For example, handling missing values may be considered part of preprocessing or data cleaning in some multiverse analyses. In contrast, missing data in a Bayesian model are unknowns with a distribution that is part of the model and are integrated out together with the parameters (Rubin 1976). Naturally this model involves judgment and will require additional priors, both of which can be varied in $\mathcal{F}$ and Γ. The uncertainty carried by the processing is then propagated to the posterior in a single fit, for instance by one Markov chain Monte Carlo run over parameters and unknown data alike. From this perspective, a processing branch of the multiverse maps to a pair (π_k, f_k) applied to the same

raw data x, and the multiverse of data sets of Steegen et al. (2016) is mapped to a class of models. The choice between excluding the incomplete units and imputing the missing values is an example. One model may group the units by their pattern of missingness and give each pattern its own parameters, as in the pattern-mixture models of Little (1993). Excluding the incomplete units is the Bayesian analysis of the model in which the parameters of the patterns are a priori independent and interest is in estimating the parameter of the complete pattern, so that the incomplete units carry no information about it. Imputing is the analysis of the model in which the patterns share parameters and the missing data mechanism can be modeled (Rubin 1976), so that the incomplete units contribute through their observed components.

In practice an important difference is that the multiverse fixes the data before fitting, in the way an unknown parameter is replaced by an estimate, while the encompassing model carries the uncertainty of the processing through to the conclusion. To digress a little bit, in practice the separation is often kept on purpose also in Bayesian models. In a Bayesian graphical model made of modules, each relating to one source of data, "misspecification of any module can contaminate the estimate and update of others", and practitioners, "particularly when certain modules are trusted more than others", have preferred to restrict the propagation of information between modules (Jacob et al. 2017). A cut model is a graphical model in which the flow of information across a chosen link is blocked in one direction, so that a module informs the rest of the model but is not informed by it. The device is the cut function of the BUGS software, described by Lunn et al. (2009) for sequential pharmacokinetic and pharmacodynamic modeling; Bayarri et al. (2009) study it under the name of modularization, for situations in which the analyst is hoping to avoid that a less reliable source of information may influence other more solid parts of the analysis; Plummer (2015) gives it a formal treatment; and Nott et al. (2024) review it as modular inference, one of the routes to Bayesian inference with a misspecified model.

In computational genomics, imaging and other fields raw data are preprocessed through a pipeline, that is, a sequence of nested algorithms that perform different summarizations, data cleaning or feature extraction. A pipeline can be thought of as a fully cut model: information flows forward only. The choice between a pipeline and an encompassing model is itself a modeling choice. Pipelines are the rule in some fields, and their variation has been studied as a multiverse. In functional neuroimaging, Carp (2012) identified ten analysis steps with two to four strategies each, 6,912 pipelines in all, and applied them to one experiment; Botvinik-Nezer et al. (2020) had seventy teams analyze the same data set; Dafflon et al. (2022) approximate the full set of pipelines by active learning on a low-dimensional representation of their relationships. In single-cell transcriptomics, Vieth et al. (2019) evaluated about 3,000 pipelines over mapping, imputation, normalization and differential expression testing, and found the choice of normalization to have the largest effect. An encompassing model can absorb parts of such pipelines, at a cost in effort and time: Vallejos et al. (2015) model the normalization and the technical noise of single-cell sequencing data jointly with the biological variation of interest.

Table 2 Multiverse analysis and global robustness in Bayesian learning, compared on the elements of Table 1 and on the four aspects of Section 2.4.

	Robust Bayesian analysis	Multiverse analysis
Objects varied	The prior π, the model f, or both, over Γ and $\mathcal{F}$	The processing and the model, over specifications $k = 1, \ldots, K$ with processed data x_k and model f_k
Data	Raw data x; processing written into (π, f), its uncertainty propagated	Processed data x_k; output of each processing choice fixed before fitting
Set construction	Constraints on the density (quantiles, contamination, bands, mixtures); extreme points	Combinations of options at each analytic choice; Type E decisions
Operational model	Typically present; e.g. classes as neighborhoods of (π_0, f_0)	Absent in the original formulation; present in some (universe against multiverse, post hoc preferred model)
Report	The range $(\underline{\psi}, \overline{\psi})$ of a posterior quantity	The set $\{t_k\}$ of estimates or tests, displayed sorted or summarized
Extremes	The main focus	Points of fragility; the maximum statistic in the global test
Testing	Bounds on the Bayes factor over Γ	Adjustments for multiple testing over the specifications
Decisions	Set of Bayes actions, regret, radius at which the action changes (Section 2.3)	Endpoint is an estimate or a test; a decision endpoint was not found in the literature reviewed

3.1.2 Can a multiverse analysis be formally viewed as a robust Bayesian analysis?

We begin by comparing the multiverse analysis as practiced, in which each specification yields a frequentist estimate or test; the case of a multiverse applied to a Bayesian analysis is taken up in the next section. In the notation of Sections 2.2 and 2.4, a multiverse is a finite set of specifications $k = 1, \ldots, K$, each determining a processed data set x_k and a model f_k, and each yielding an estimate or a test t_k of a shared quantity of interest. A global robustness analysis takes classes Γ and $\mathcal{F}$ containing an operational prior and model and reports the range of the posterior quantity $\psi(\pi, f)$ over them. The two share the structure, a set of specifications and the set of answers it yields; they differ in the answers themselves, an estimate or a test on one side and a posterior quantity on the other, and in what is done with the set. Table 2 proposes a simplified summary of where the two agree and where they part. There are many examples where hybrids, combinations and variations are offered. The goal of this table is to define dimensions along which to describe various analyses.

Many similarities exist in the structures. Both replace a single specification by a set, both include in the set the alternatives that are judged reasonable and leave out those that are not, both report what the set yields rather than a single value, and both look for the choices that would change the conclusion. In their original formulations

both give the members of the set equal standing; the weighted extensions of both are taken up below under model averaging.

Differences arise in the treatment of the processing, in the role of an operational model, in the treatment of extremes and in inference. Typically multiverse analyses fix the output of each processing branch and then estimate; a Bayesian analysis, with the processing written into the model, conditions on the raw data and carries the uncertainty of the processing through, as in the previous section. Robust Bayesian analysis is usually built around an operational model, with classes that are neighborhoods of it; several multiverse formulations have no operational specification, and those that have one use it as a reference against which the others are read. The range (4) reports two members of the class and nothing about the others; a multiverse analysis reports every specification, and some of its inferential procedures use the extreme, through the maximum statistic, to control the family-wise error rate over the specifications. In testing, robust Bayesian analyses have reported bounds on the Bayes factor as the prior varies over Γ (Edwards et al. 1963; Berger and Sellke 1987; Berger 1994), whereas multiverse analyses adjust for multiple testing over the specifications. In the multiverse literature reviewed the endpoint is an estimate or a test; the robustness of a decision, the set of Bayes actions and the regret of Section 2.3, has no counterpart there that we found. The correspondence is close for the structure of the set and for what is done with its members one at a time, and approximate for what is done with the set as a whole.

3.1.3 Multiverse analysis of a Bayesian analysis

A multiverse analysis can be applied to a Bayesian analysis, each specification being a Bayesian model. The correspondence of Table 2 then tightens: a specification maps to a pair (π_k, f_k) applied to the raw data, as in Section 3.1, the multiverse is the finite class $\{(\pi_1, f_1), \ldots, (\pi_K, f_K)\}$, the output of specification k is the posterior quantity $\psi_k = \psi(\pi_k, f_k)$, conditional on the same x, the question of multiplicity does not arise, and the questions robust Bayesian analysis asks of a class, the range, the weight $m(x|\pi_k, f_k)$ of each member, the members at which the extremes are attained, can be asked of the multiverse. Three published analyses have this flavor. Hoogeveen et al. (2023) reanalyze a multi-laboratory replication project with Bayesian hierarchical models and, "in a multiverse analysis, we assess the robustness of the results with varying data-inclusion criteria and prior settings"; the specifications vary the prior, a finite Γ, and the inclusion of units, the exclusion models of the previous section, and the report is the set of Bayes factors, which "largely converge to a common conclusion". van Doorn et al. (2021), in guidelines for conducting and reporting a Bayesian analysis, prescribe robustness checks over prior settings as a stage of the analysis, a multiverse over Γ in small form. Schad et al. (2023) study the stability of Bayes factors "against different MCMC draws and sampling variation in the data" and "the variability of decisions based on Bayes factors using a utility function", which reaches the decision question of Section 2.3. Each reports the set of posterior answers and reads it for convergence; the range, the weights and the extremes of Section 2.2 can be computed from the same set.

3.1.4 Model averaging

Both frameworks have an extension in which the members of the set are weighted, and on the Bayesian side the extension is a standard analysis. Bayesian model averaging (Madigan and Raftery 1994; Draper 1995; Raftery et al. 1997; Hoeting et al. 1999) places a prior probability on each member of a finite class, updates it by the marginal density of the data under that member, $m(x|\pi_k, f_k)$ in the notation of Section 2.2, and reports the posterior quantity of interest averaged over the class with the posterior model probabilities as weights. It is the higher-level analysis of Section 2.2 carried to its conclusion: the class becomes a hierarchical prior, and the range of answers becomes a single answer with the model uncertainty propagated into its posterior. Two conditions follow from the definition. The members must condition on the same data, so the processing must be written into the model as in Section 3.1; and the analyst must be willing to mix the members, which the imprecise-probability view of Section 2.2 declines. In the multiverse literature, the comparison of uniform weighting with Bayesian model averaging by Cantone and Tomaselli (2024a), recalled in Section 2.4, and the observation of Cantone and Tomaselli (2024b) that the specifications are not independent draws bear on the same question. In frequentist analyses, the weights are supplied from outside.

3.1.5 Combining robustness and multiverse approaches

These various perspectives can be complementary. For example, they could be combined via a sequence of questions, each answered with the tools of one section above. This sequence is not intended to be prescriptive, but it illustrates how they can be combined to provide a more complete picture of the robustness of conclusions and actions.

1. *Are conclusions robust to modest perturbations around difficult choices?* These are, for example, the Type E decisions of Del Giudice and Gangestad (2021) or the specification of prior distributions when earlier evidence is scarce. This initial challenge can be addressed using diagnostics and local measures, such as the derivatives of Section 2.2, the leave-one-out measures on prior and data and the variance of the utility of Watson and Holmes (2016). If the answer is yes, the analysis may stop here.
2. If diagnostics question this robustness, then the analyst may broaden the scope and ask *are conclusions robust within a range of plausible model assumptions?* An answer may now come from global robustness: the range (4) over classes Γ and $\mathcal{F}$, judged against the conclusion at stake.
3. The lack of robustness of a study conclusion may or may not have practical implications. A decision theoretic perspective then asks: *are downstream actions robust to the choice of assumption within the plausible set?* An answer here may come from the measures of Section 2.3: the set of Bayes actions, the regret, and the radius at which the action changes. A wide range of ψ is compatible with a single optimal action.
4. If by these analyses one is led to question robustness of practical implication of an analysis, then the question arises of *which assumptions support which conclusions*

and actions. A multiverse analysis reports every specification with its result, identifies the choices that drive the differences, and can be used to partition the set of assumptions by the action each leads to, as in the framework of Ríos Insua (1990). The report is then not one action but a map from assumptions to actions, and the reader's own assumptions locate them on it.

Two conceptual hurdles run through all four steps. The first is the definition of the plausible sets, whether a class $\Gamma \times \mathcal{F}$ or a multiverse. Berger's criteria for a class, recalled in Section 2.2, and the criteria of Simonsohn et al. (2020) and Del Giudice and Gangestad (2021), that specifications be justified, valid, non-redundant and equivalent on a priori grounds, address the same question from the two sides. Both literatures record the consequence of a set that is too large: a range too wide to conclude, or a multiverse in which few specifications estimate the effect of interest. The second hurdle is the relative plausibility of the members of the set, and in particular of the sets of assumptions that lead to each action in step 4. The tools for weighing them are those of Section 3.1, model averaging with the marginal density of the data as the weight, with the conditions and reservations stated there.

3.2 Multiverse and multi-agent decision theory

A different view of all these procedures begins by considering the perspective that both robustness and multiverse analysis have the intended role of convincing others that the results are sound. What then if we begin to view the robustness and multiverse sets themselves as representative of multiple viewpoints?

3.2.1 Can a multiverse analysis be formally viewed as a multi-agent Bayesian analysis?

Take again a multiverse of Bayesian specifications, as in Section 3.1, and read each specification as an agent: agent k observes the same data x, holds the prior π_k and the model f_k, and, if a decision is at stake, carries a utility u_k and a Bayes action a_k^*. The set $\{\psi_k\}$ of a multiverse is then the set of posterior opinions of K agents about a shared quantity, and the map from assumptions to actions of Section 3.1 is the collection of their Bayes actions. In a multiverse the agents are the alternative analyses one analyst is willing to entertain; in the multi-analyst studies of Section 2.4 they are persons.

Dawid (1980) gives one condition under which such agents agree on part of the analysis. Write the parameter as (θ, ϕ), with ϕ a nuisance parameter, and consider a collection of Bayesians who agree on the model $f(x|\theta, \phi)$ and on the conditional prior $\pi(\phi|\theta)$, while holding arbitrary priors on θ; Dawid calls it a bevy of Bayesians. All of them share the marginal model $f(x|\theta) = \int f(x|\theta, \phi)\pi(\phi|\theta)d\phi$, so that their posteriors for θ differ only through their priors on θ, and they agree on which reductions of the data are sufficient for θ. For a multiverse the condition separates the specifications: those that differ in the treatment of ϕ but agree on $\pi(\phi|\theta)$ form one bevy, whose members' posteriors are $\pi_k(\theta)f(x|\theta)$ up to a constant, a class Γ over θ alone; those that differ in $\pi(\phi|\theta)$, or in the model, do not share a marginal likelihood, and their disagreement is of a different kind.

When the agents must agree on one action, the setting is that of multi-Bayesian decision theory: several Bayesians with the same data, their own utilities and possibly different beliefs, choosing a joint decision procedure. Weerahandi and Zidek (1981, 1983) propose the Pareto optimal procedures, among which the one maximizing the Nash product of the agents' utilities, and note that the solution is randomized when opinions or preferences are sufficiently diverse; Hylland and Zeckhauser (1979) show that beliefs and values cannot be aggregated separately into a group Bayesian; Seidenfeld et al. (1989) show that the preferences two Bayesians share are those of no single Bayesian; and Genest and Zidek (1986) review the pooling of the agents' probabilities into one. For the multiverse this literature supplies summaries of the set of specifications other than the range and the average: the set of actions that no specification would trade for another, and the pooled opinion. We do not pursue the connection further here.

3.3 Agentic AI for the multiverse

In the notation of Section 2.5, an agent is a point in the multiverse: the agent index and the specification index are the same k. Each agent, working from the shared data x, arrives at a processed data set x_k and a model f_k, and with them at its own prior π_k and, when the analysis serves a decision, its own utility u_k; the multiverse is the set of the agents' analyses, and the coordinating agent of Section 2.5 plays the part of the analyst who assembles and reads it. A multi-agent system can implement a multiverse by asking agents set up with different perspectives and concerns, a methodologist's, a subject-matter specialist's, a skeptic's, to specify the analytic choices independently, in the way the multi-analyst studies of Section 2.4 ask persons. The set of specifications is then defined by the diversity of the agents rather than by the enumeration of options at each choice, and the questions of Section 2.4, how the set is built, whether its members are weighed, what the extremes mean and whether one specification is operational, are asked of the agents' prompts and of the coordinating rule.

At an additional level, each agent can provide several points in the multiverse rather than one. In one implementation, a user familiar with the science but not with the statistical techniques prompts each agent for two analytic plans, a conservative one and the one the agent considers most plausible; the multiverse then has $2K$ specifications, paired by agent, and the pair itself carries information, the distance between an agent's conservative and most plausible plan serving as an indication of how much that agent's judgment is doing. The user's own judgment enters through the choice of agents and through the reading of the set, and not through the statistical choices, which is the division of labor the multiverse literature describes between the analyst who defines the set and the reader who locates their own assumptions on it. Section 4 illustrates a related construction, in which the perspective is varied through the prompt and each agent provides one plan.

4 Illustration

We now illustrate a multiverse analysis in which the specifications are produced by language-model agents working independently from partly shared instructions.

4.1 Context

The MacArthur–Bates Communicative Development Inventory (CDI; Fenson et al. 2007) is a checklist of 680 words used by caregivers in developmental psychology to assess children's expressive language development. A child whose score falls below a low percentile of the age-specific distribution is classified as a late talker and referred for further assessment. The choice of the referral cutoff is discretionary, with published reports ranging from the 2nd to the 30th percentile of the spoken word distribution at 24 months (Desmarais et al. 2008).

Our goal is to illustrate an agentic AI workflow that estimates the referral cutoff and its uncertainty. To simplify the analysis, we set the target cutoff as the 10th percentile of the distribution of CDI word counts for a given set of children who undergo the assessment. Following the notation above, the state of the world is

$$\theta = \inf\big\{k \in \{0, \dots, 680\} : F_{24}(k) \geq 0.10\big\}, \tag{8}$$

where F_{24} is the cumulative distribution function of the words produced at 24 months. The action a is the number of words at or below which a child is referred, so $\mathcal{A} = \{0, \dots, 680\}$ as well, and the utility $u(a, \theta)$ quantifies the consequences of a missed referral and of an unnecessary one.

4.2 Data description

We obtain data x from Wordbank (Frank et al. 2017), an open repository of CDI administrations, each a checklist completed for one child, contributed by laboratories worldwide. We consider the by-child export of the English (American) Words and Sentences form, which includes 9,093 children with records of age in months, sex, birth order, caregiver education, ethnicity, and the number of words produced out of the 680 on the checklist. We also have information on which administrations were collected for the study that produced the instrument's published norms (called the "norming" subsample) and which were contributed afterward by other projects (the "contributed" one). In the first group, the median vocabulary size increases monotonically with age, while in the second, it does not (not shown).

Figure 1 displays the distribution of words produced at 24 months. In total, we have 985 children, 434 of whom are in the norming study. The empirical 10th percentile of the pooled sample is equal to 50 words, though it varies by subset, as shown in the right panel. Hence, choices about how to handle the data can create substantial variability in the final recommendation.

4.3 Design

We consider a simplified scenario in which estimating θ requires three types of choices: which children to include, whether and how to borrow information across neighboring ages (for example 23 and 25 months), and which family of distributions to fit and with which prior. As in Section 3.1, each set of choices corresponds to a pair $(\pi, f) \in \Gamma \times \mathcal{F}$ and yields a posterior distribution π_x. A multiverse analysis enumerates a collection

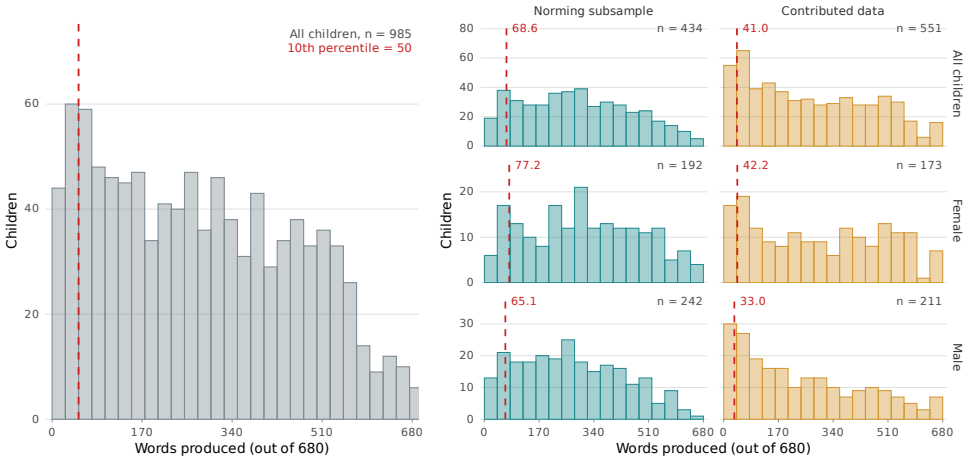


Fig. 1 Words produced at 24 months, as binned counts. Left: all 985 children, with the empirical 10th percentile marked by the dashed line. Right: the same children split by whether the administration belongs to the study that produced the published norms (columns) and by sex (rows), each panel carrying its own 10th percentile. The top row uses every record; the sex rows omit the 167 children whose sex is not recorded, all of them in the contributed data.

of these pairs and reports the resulting variation in the posterior of θ. Here the agents make these choices, while the analyst sees only their reports.

We deploy twenty instances of an LLM, Claude Opus 5. Each instance acts as an independent investigator. Specifically, it receives a copy of the data and three pieces of text: 1) a common brief describing the variables, the required outputs, and the working constraints; 2) an investigator prompt; and 3) a model-class prompt. We require the agents to produce a reproducible R script that draws at least 2,000 posterior samples of θ, along with a written record of each analytic choice with its justification, including how the norming indicator was treated, and the cutoff the agent would recommend in practice. Each agent works in its own directory without access to the other agents' outputs, and is instructed not to search the web, so that no agent can anchor its analysis on a published cutoff. The text of the common brief is reported in Appendix A.

We ran the twenty agents on 15 September 2026 between 01:55 and 04:42 CEST using Claude Opus 5 (model identifier `claude-opus-5`) through the Claude Code command-line interface, versions 2.1.270 and 2.1.272, with reasoning effort set to `xhigh`. We downloaded the Wordbank by-child export for the English (American) Words and Sentences form on 14 September 2026. Analyses were run under R 4.3.1; nine agents fitted their models with Stan (rstan 2.32.6 or brms 2.21.0) and eleven implemented their own Gibbs or Metropolis samplers. Each agent's script, posterior draws, written record of its choices, and full session transcript are deposited at a DOI to be announced.

Figure 2 reports the prompts. The experiment follows a 2×2 factorial design with five replicates per cell. The first factor is the concern communicated to the agent: Investigator A asks for "the most plausible assumptions", whereas Investigator B states a concern about failures to refer a child who needs assessment. Neither states a

Common preamble

Give me the posterior distribution for the 10th percentile for the variable `words_produced` in the dataset (`data/wordbank_by_child_English_American_WS.csv`). The variable records the number of words spoken by a child out of a checklist of 680 words.

I need to estimate the 10th percentile of this distribution for children who are at or around 24 months of age. This is the cutoff to classify a child as a late-talker and is used to signal the need for further clinical assessment.

Investigator A

Make the most plausible assumptions you can make about this dataset and this distribution.

Create a reproducible R file that performs an end-to-end analysis of the dataset and reports the posterior distribution of the 10th percentile. You are free to filter observation (if needed) and make assumptions **based on what is reasonable**. Always justify your choices.

Investigator B

I am very concerned about the possibility of not referring children who need to be referred for further assessment.

Create a reproducible R file that performs an end-to-end analysis of the dataset and reports the posterior distribution of the 10th percentile. You are free to filter observation (if needed) and make assumptions **keeping my concern in mind**. Always justify your choices.

Model class: parametric

Use a **parametric** model for the distribution of `words_produced`. Choose the family yourself and justify it. Fit it in a Bayesian framework so that a posterior for the 10th percentile is available, and state your priors explicitly.

Model class: nonparametric

Use a **nonparametric** model for the distribution of `words_produced`. Choose the class of model yourself and justify it. Fit it in a Bayesian framework so that a posterior for the 10th percentile is available, and state your priors explicitly.

Fig. 2 The prompts, reproduced verbatim; the panel headings are ours. The preamble opens both investigator prompts. Each agent received one investigator prompt and one model-class prompt, together with the common brief of Appendix A. Emphasis as in the original, set in bold; within each pair it marks the main clause that differs.

utility. The second factor is the model class, parametric or nonparametric; the family within the class is left to the agent. Each cell is replicated five times with the prompt held fixed, so the variation within a cell is attributable to the agent rather than to the design.

4.4 Results

Nineteen of the twenty agents returned a posterior, while one failed to reach convergence in its Markov chain Monte Carlo (MCMC) sampler. Within $\mathcal{F}$, the likelihood choice was nearly unanimous. All nine parametric agents fitted a beta-binomial model with 680 trials, and all ten nonparametric agents fitted a Dirichlet process mixture: eight mixed binomial or beta-binomial kernels directly, and two placed the mixture on a latent logit with a binomial observation model. Both groups gave the same justification: the support is $\{0, \ldots, 680\}$ and both endpoints are attained. The right-hand column of Figure 3 records the family chosen and how each agent treated age.

The brief asked every agent to report how it dealt with the norming indicator and why, without recommending a specific treatment, and described the contributed

pool as "more heterogeneous but potentially larger and more representative". All nineteen agents nonetheless restricted the analysis to the norming subsample, and each argued for the restriction based on the data: the age profile of the word counts, which is not monotone in the contributed pool, and the sex missing for three in ten of its administrations. Several agents also checked caregiver education as a competing explanation and rejected it. The restriction is therefore not a design choice, nor one the agents were told to make.

Figure 3 shows the resulting posteriors. The nineteen values of the posterior mean range from 61.4 to 69.6 words, with an average of 67.6 and a standard deviation of 2.2. Investigator B's agents sit 2.4 words above Investigator A's, and the nonparametric agents sit 0.3 words above the parametric ones. The average standard deviation of a single agent's posterior is 5.1, and the nineteen 95% credible intervals share a point. The spread across the whole multiverse is thus smaller than the uncertainty any one member reports. The grey dotted line in the figure marks 68.6, the empirical 10th percentile of the norming administrations at 24 months, which is the value in the upper-left panel of Figure 1.

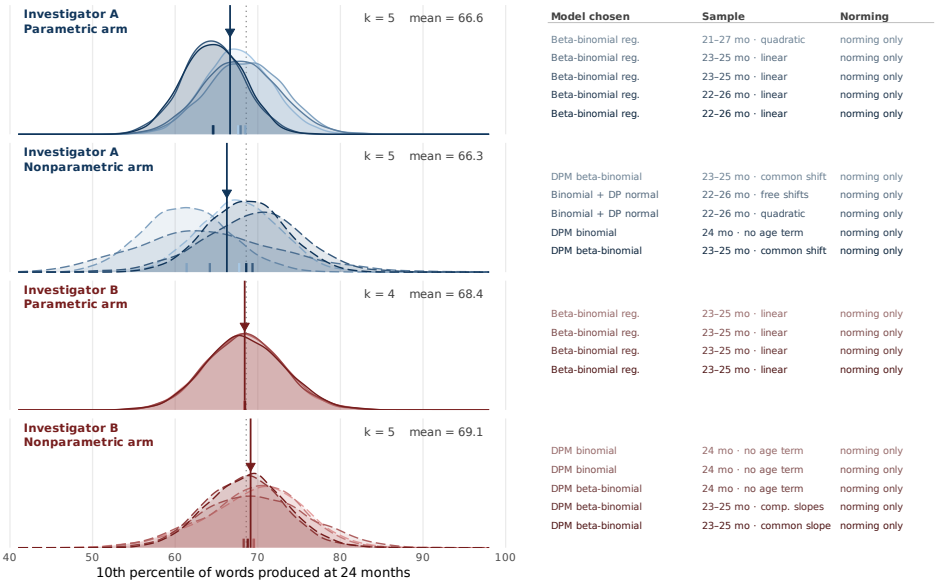


Fig. 3 Posterior distribution of θ, the 10th percentile of words produced at 24 months, for each of the nineteen agents that returned one. Left: one density per agent. Blue marks Investigator A and red marks Investigator B, as do the panel labels; solid curves are parametric agents and dashed curves are nonparametric ones; shades distinguish replicates within a panel. In each panel, the thin vertical line and triangle show the mean of the posterior means for that cell, the ticks below the axis show the individual posterior means, and the grey dotted line shows the empirical percentile in the analysis sample. Right: the modeling choices behind each curve, in the color of that curve: the family or mixture fitted (DPM: Dirichlet process mixture), the age range used and the form in which age enters the model, and the treatment of the norming indicator `is_norming`, which is the same for every agent. The Investigator B parametric panel contains four agents; the fifth did not converge.

This statement concerns the specifications visited, not $\Gamma \times \mathcal{F}$. The range in equation (4) is taken over the whole class, whereas a multiverse displays the range over the pairs it enumerates, and the two coincide only if the enumeration is rich in the directions that matter. Here the enumeration was not. One agent refitted its own model under alternative treatments of the norming indicator and obtained 64.6 words on the norming subsample, 46.6 on the two sources pooled, and 35.9 on the contributed pool alone: a swing of roughly 29 words, against the 8 words that separate the branches the agents actually took. The same agent then fitted all 3,101 administrations it had retained from both sources, with a source effect on both the mean and the dispersion, and read the percentile off at the norming level, recovering 64.6 words. The gap between sources is therefore a level shift between two populations and does not represent information discarded by the restriction.

A second choice that no agent varied is visible in the right panels of Figure 1. The empirical 10th percentile among norming administrations at 24 months is 77.2 words for girls and 65.1 for boys, and the corresponding contributed values are 42.2 and 33.0. No agent conditioned its primary model on sex. One fitted sex-specific norms as a sensitivity analysis and reported them separately. All nineteen therefore estimated and recommended a single sex-pooled cutoff. However, the design did change the agreement among agents. The standard deviations of the posterior means within a cell are 1.90 and 3.37 in Investigator A's two cells, against 0.05 and 0.57 in Investigator B's. Therefore, explicitly stating an asymmetric concern gave the agents a criterion for choosing among defensible options, which made them converge to the same choice.

The referral threshold is an action rather than an estimate, and here the two diverge as decision theory predicts. Several of Investigator B's agents adopted, without being shown it, the utility

$$u_c(a, \theta) \;=\; -\,c\,(\theta - a)^{+} \;-\; (a - \theta)^{+}, \qquad c > 1, \tag{9}$$

where $(z)^{+} = \max(z, 0)$. Under such utility, failing to refer a child who should be referred costs c times as much as referring one who need not be. The associated Bayes action is the $c/(c+1)$ quantile of the posterior of θ. At $c = 9$ this is the posterior 90th percentile. One agent went further: it computed the Bayes action under each specification it judged defensible and reported the largest, the upper end of the set of Bayes actions of Section 2.3, without being asked for it.

The most important effect is visible in the recommendations rather than in the posteriors. Investigator A's agents proposed cutoffs between 61 and 70 words, at or near their posterior medians, whereas Investigator B's proposed between 74 and 80. The two groups' posteriors differ by about two words and their actions by about ten. The concern moved the actions by about ten words while moving the posteriors by about two. The analytic choices enumerated by the moved the posterior means by eight words and the actions within an investigator by a similar amount.

Several of Investigator B's agents recorded a correction they judged defensible and then declined it because of the direction in which it would move the cutoff. One wrote that re-weighting the sample toward representativeness "would lower the cutoff and increase missed referrals", that "given the stated asymmetric concern, that error is in the acceptable direction", and that this was "a judgment about which way to be wrong,

stated openly rather than hidden in a weighting scheme". Under a stated utility, this is the Bayes action for equation (9). Read as a specification search, it is selection on the outcome, one of the problems multiverse analysis is designed to expose. The two readings differ only in whether u was stated in advance.

In summary, at least in this illustration, the agents enumerated a narrow multiverse: their posterior means spanned 8 words, while the choice on which they all agreed moved the estimate by about 29 in one agent's refit. The stated concern moved the recommended cutoff by about ten words and the posteriors by about two. Agents thus supply specifications, justifications, and actions at once, and the reading of Section 3.1, which maps each specification to assumptions and actions, applies to their output as it does to a multiverse built by hand.

5 Conclusions

We have presented an initial exploration of connections among multiverse analysis, Bayesian decision theory and agentic AI. We highlight three to close.

First, a multiverse analysis and a robust Bayesian analysis address the same question, which conclusions survive which choices, from two sides. The robust Bayesian analysis defines a class of inputs and reports a range; the multiverse analysis enumerates specifications and reports each. Once data processing choices are read as choices of model, as in Section 3.1, the two operate on the same objects, and Table 2 records where they agree and where they differ: in how the set is built, whether one specification is operational, how the extremes are read, how multiplicity is handled, and whether the endpoint is a conclusion or an action. The staged sequence of Section 3.1 uses each approach for the question it was built for.

Second, each point of a multiverse can be read as the analysis of one agent with its own prior and utility. This reading connects the multiverse to Dawid's bevy of Bayesians and to multi-Bayesian decision theory, which supply a language for comparing and combining the points of a multiverse.

Third, LLM agents can implement this reading. In Section 4, twenty agents enumerated a multiverse, agreed without being asked on a choice that moved the estimate more than any the design varied, and recommended actions that moved with the concern stated by the user more than their posteriors did. The illustration also shows the limits of the construction: a multiverse of agents is no richer than the directions in which its agents differ, and agreement among agents is not necessarily evidence that a choice is sound.

Many questions remain open in all three areas. Two of them are how to define plausible sets, whether classes of inputs, multiverses or populations of agents, and whether and how to assess the relative plausibility of the specifications that lead to different actions. Agents make the second question easier to pose, because each specification arrives with a justification, and leave the answer to the analyst.

Declarations

Competing interests. The authors declare that they have no competing interests.

Funding. No funding was received for this work.

Ethics approval and consent to participate. Not applicable: the analysis uses de-identified, publicly available data.

Data availability. The data are the by-child export of the English (American) Words and Sentences form from Wordbank (Frank et al. 2017), downloaded on 14 September 2026; the file used is deposited with the code.

Code availability. A complete, reproducible analysis workflow will be deposited in the repository with the final version of the manuscript.

Author contributions. Both authors contributed to every aspect of the work, including conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft and review & editing), visualization.

Appendix A The common brief

The text below was given verbatim to every agent, together with the data file, one investigator prompt and one model-class prompt (Figure 2). Only the typesetting differs from the original.

The dataset
A copy is already in your working directory at
`data/wordbank_by_child_English_American_WS.csv`.

column	type	description
`data_id`	integer	Unique identifier for the administration.
`age`	integer	Child's age in months at administration, recorded in whole months.
`sex`	factor	`Male` / `Female`
`birth_order`	factor	Order of birth for the child
`caregiver_education`	factor	Highest level completed by the reporting caregiver
`ethnicity`	factor	`Hispanic` / `Non-Hispanic`
`is_norming`	logical	`TRUE` if the administration came from a CDI norming study; `FALSE` if it was contributed to Wordbank from another research project. The norming studies sampled diversely but not nationally representatively; the wider contributed pool is more heterogeneous but potentially larger and more representative. Both are present in the file.
`words_produced`	integer	These are MacArthur-Bates CDI: Words & Sentences forms, English (American). Number of checklist words the caregiver reports the child produces, out of 680.

Deliverables
Work **only** inside your own working directory, given below.

1. `analysis.R` — the reproducible end-to-end script. It must run non-interactively from that directory via `Rscript analysis.R`, and you must actually run it and confirm it works.
2. `q10_draws.csv` — written by the script: one column named `q10`, at least 2000 posterior draws of the 10th percentile, one per row, no row names. These must be genuine posterior draws, not a point estimate repeated.
3. `NOTES.md` — a short record containing:
   - every filter you applied and the resulting sample size;
   - how you treated `is_norming`, and why;
   - the model you fitted and why;
   - your priors;
   - the cutoff you would recommend in practice, and why.

Constraints

Do not search the web or fetch web pages. Everything you need is in the dataset and in this brief. Base your modelling on the structure of the data and on your own statistical judgement, not on published figures.

Installing R packages from CRAN is allowed. If you install R packages, install them into `Rlib` inside your working directory: put `.libPaths(c("Rlib", .libPaths()))` at the top of `analysis.R` and pass `lib = "Rlib"` to `install.packages()`.

Do not read, list, or modify anything outside your working directory. Other investigators are working on the same question independently; do not look for their work, and do not look for prior analyses of this dataset in this repository.

When finished, reply with your justification: the choices you made, why you made them, and where they could fail. Be concise.

References

Abraham C, Cadre B (2004) Asymptotic global robustness in Bayesian decision theory. The Annals of Statistics 32(4):1341–1366. https://doi.org/10.1214/009053604000000562

Aczel B, Szaszi B, Clelland HT, et al (2026) Investigating the analytical robustness of the social and behavioural sciences. Nature 652(8108):135–142. https://doi.org/10.1038/s41586-025-09844-9

Altoè G, Gambarota F, Girardi P, et al (2025) Multiverse Analysis in R: Exploratory and Inferential Approaches, chapter 1 (Introduction), sample chapter of a book in preparation for Springer

Anscombe FJ, Aumann RJ (1963) A Definition of Subjective Probability. The Annals of Mathematical Statistics 34:199–205

Augustin T (2022) Statistics with imprecise probabilities: A short survey. In: Aslett L, Coolen FPA, De Bock J (eds) Uncertainty in Engineering: Introduction to Methods and Applications. SpringerBriefs in Statistics, Springer, Cham, p 67–79, https://doi.org/10.1007/978-3-030-83640-5_5

Bayarri MJ, Berger JO, Liu F (2009) Modularization in Bayesian analysis, with emphasis on analysis of computer models. Bayesian Analysis 4(1):119–150. https://doi.org/10.1214/09-BA404

Berger JO (1994) An overview of robust Bayesian analysis (with discussion). Test 3(1):5–124. https://doi.org/10.1007/BF02562676

Berger JO, Sellke T (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. Journal of the American Statistical Association 82(397):112–122. https://doi.org/10.2307/2289131

Berger JO, Ríos Insua D, Ruggeri F (2000) Bayesian robustness. In: Ríos Insua D, Ruggeri F (eds) Robust Bayesian Analysis, Lecture Notes in Statistics, vol 152.

Springer, New York, p 1–32, https://doi.org/10.1007/978-1-4612-1306-2_1

Botvinik-Nezer R, Holzmeister F, Camerer CF, et al (2020) Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582:84–88. https://doi.org/10.1038/s41586-020-2314-9

Cantone GG, Tomaselli V (2024a) Characterisation and calibration of multiversal methods. Advances in Data Analysis and Classification 19(4):989–1021. https://doi.org/10.1007/s11634-024-00610-9

Cantone GG, Tomaselli V (2024b) Theory and methods of the multiverse: an application for panel-based models. Quality & Quantity 58(2):1447–1480. https://doi.org/10.1007/s11135-023-01698-5

Caprio M, Gong R (2023) Dynamic precise and imprecise probability kinematics. In: Miranda E, Montes I, Quaeghebeur E, et al (eds) Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications, Proceedings of Machine Learning Research, vol 215. PMLR, pp 72–83, URL https://proceedings.mlr.press/v215/caprio23a.html

Carota C, Parmigiani G, Polson NG (1996) Diagnostic measures for model criticism. Journal of the American Statistical Association 91:753–762. URL http://www.jstor.org/pss/2291670

Carp J (2012) On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. Frontiers in Neuroscience 6:149. https://doi.org/10.3389/fnins.2012.00149

Dafflon J, Da Costa PF, Váša F, et al (2022) A guided multiverse study of neuroimaging analyses. Nature Communications 13:3758. https://doi.org/10.1038/s41467-022-31347-8

Dawid AP (1980) A Bayesian look at nuisance parameters. Trabajos de Estadística y de Investigación Operativa 31(1):167–203. https://doi.org/10.1007/BF02888351

Del Giudice M, Gangestad SW (2021) A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. Advances in Methods and Practices in Psychological Science 4(1):1–15. https://doi.org/10.1177/2515245920954925

Desmarais C, Sylvestre A, Meyer F, et al (2008) Systematic review of the literature on characteristics of late-talking toddlers. International Journal of Language & Communication Disorders 43(4):361–389. https://doi.org/10.1080/13682820701546854

van Doorn J, van den Bergh D, Böhm U, et al (2021) The JASP guidelines for conducting and reporting a Bayesian analysis. Psychonomic Bulletin & Review 28(3):813–826. https://doi.org/10.3758/s13423-020-01798-5

Draper D (1995) Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society: Series B 57(1):45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psychological Review 70(3):193–242. https://doi.org/10.1037/h0044139

Fenson L, Marchman VA, Thal DJ, et al (2007) MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual, 2nd edn. Brookes Publishing, Baltimore, MD

Frank MC, Braginsky M, Yurovsky D, et al (2017) Wordbank: An open repository for developmental vocabulary data. Journal of Child Language 44(3):677–694. https://doi.org/10.1017/S0305000916000209

Gao J, Rahman MM, Caskey J, et al (2025) MoMA: a mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. npj Digital Medicine 9(1):46. https://doi.org/10.1038/s41746-025-02219-4

Genest C, Zidek JV (1986) Combining probability distributions: A critique and an annotated bibliography. Statistical Science 1(1):114–135. https://doi.org/10.1214/ss/1177013825

Gilboa I (2025) Decision under uncertainty: State of the science. Annual Review of Economics 17:1–30. https://doi.org/10.1146/annurev-economics-090924-041522

Girardi P, Vesely A, Lakens D, et al (2024) Post-selection Inference in Multiverse Analysis (PIMA): An Inferential Framework Based on the Sign Flipping Score Test. Psychometrika 89(2):542–568. https://doi.org/10.1007/s11336-024-09973-6

Gong R, Meng XL (2021) Judicious judgment meets unsettling updating: Dilation, sure loss and Simpson's paradox. Statistical Science 36(2):169–190. https://doi.org/10.1214/19-STS765

Götz M, Sarma A, O'Boyle EH (2024) The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the R package multiverse. International Journal of Psychology 59(6):1003–1014. https://doi.org/10.1002/ijop.13229

Gridach M, Nanavati J, Zine El Abidine K, et al (2026) Agentic AI in healthcare: Opportunities, challenges, and future directions. ACM Computing Surveys 58(13):1–30. https://doi.org/10.1145/3809164

Guo T, Chen X, Wang Y, et al (2024) Large language model based multi-agents: A survey of progress and challenges. https://doi.org/10.48550/arXiv.2402.01680, arXiv preprint, arXiv:2402.01680

Heyman T, Pronizius E, Lewis SC, et al (2025) Crowdsourcing multiverse analyses to explore the impact of different data-processing and analysis decisions: A tutorial. Psychological Methods https://doi.org/10.1037/met0000770, advance online publication

Hoeting JA, Madigan D, Raftery AE, et al (1999) Bayesian model averaging: A tutorial (with discussion). Statistical Science 14(4):382–417. https://doi.org/10.1214/ss/1009212519

Hoogeveen S, Berkhout SW, Gronau QF, et al (2023) Improving statistical analysis in team science: The case of a Bayesian multiverse of Many Labs 4. Advances in Methods and Practices in Psychological Science 6(3). https://doi.org/10.1177/25152459231182318

Hylland A, Zeckhauser R (1979) The impossibility of Bayesian group decision making with separate aggregation of beliefs and values. Econometrica 47(6):1321–1336. https://doi.org/10.2307/1914003

Isaacs HH (1963) Sensitivity of decisions to probability estimation errors. Operations Research 11(4):536–552. https://doi.org/10.1287/opre.11.4.536

Jacob PE, Murray LM, Holmes CC, et al (2017) Better together? Statistical learning in models made of modules. https://doi.org/10.48550/arXiv.1708.08719, arXiv preprint, arXiv:1708.08719

Jin W, Du H, Zhao B, et al (2025) A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. https://doi.org/10.48550/arXiv.2503.13415, arXiv preprint, arXiv:2503.13415

Kadane JB, Chuang DT (1978) Stable Decision Problems. The Annals of Statistics 6:1095–1110

Kahneman D, Tversky A (1979) Prospect Theory: An Analysis of Decision Under Risk. Econometrica 47:263–291

Kim Y, Park C, Jeong H, et al (2024) MDAgents: An adaptive collaboration of LLMs for medical decision-making. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), pp 79410–79452, https://doi.org/10.52202/079017-2522, 2404.15155

Kreps DM (1988) Notes on the theory of choice. Westview Press, Boulder, CO

van der Linden WJ, Mellenbergh GJ (1977) Optimal cutting scores using a linear loss function. Applied Psychological Measurement 1(4):593–599. https://doi.org/10.1177/014662167700100414

Little RJA (1993) Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 88(421):125–134. https://doi.org/10.1080/01621459.1993.10594302

Liu X, Carrero ZI, Jiang X, et al (2026) Benchmarking large language model-based agent systems for clinical decision tasks. npj Digital Medicine 9(1). https://doi.org/10.1038/s41746-026-02443-6

Lunn D, Best N, Spiegelhalter D, et al (2009) Combining MCMC with 'sequential' PKPD modelling. Journal of Pharmacokinetics and Pharmacodynamics 36(1):19–38. https://doi.org/10.1007/s10928-008-9109-1

Madigan D, Raftery AE (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. Journal of the American Statistical Association 89:1535–1546

Nott DJ, Drovandi C, Frazier DT (2024) Bayesian inference for misspecified generative models. Annual Review of Statistics and Its Application 11:179–202. https://doi.org/10.1146/annurev-statistics-040522-015915

Parmigiani G (2002) Modeling in Medical Decision Making: A Bayesian Approach. Wiley, Chichester

Parmigiani G, Inoue L (2009) Decision theory: Principles and approaches. Wiley, Chichester

Patel CJ, Burford B, Ioannidis JPA (2015) Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. Journal of Clinical Epidemiology 68(9):1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Plummer M (2015) Cuts in Bayesian graphical models. Statistics and Computing 25(1):37–43. https://doi.org/10.1007/s11222-014-9503-z

Raftery A, Madigan D, Hoeting J (1997) Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association 92:179–191

Raiffa H, Schlaifer R (1961) Applied Statistical Decision Theory. Harvard University Graduate School of Business Administration, Boston

Ríos Insua D (1990) Sensitivity analysis in multi-objective decision making, Lecture Notes in Economics and Mathematical Systems, vol 347. Springer, Berlin, https://doi.org/10.1007/978-3-642-51656-6

Ríos Insua D, French S (1991) A framework for sensitivity analysis in discrete multi-objective decision-making. European Journal of Operational Research 54(2):176–190. https://doi.org/10.1016/0377-2217(91)90296-8

Ríos Insua D, Martín J (1994) Robustness issues under imprecise beliefs and preferences. Journal of Statistical Planning and Inference 40(2–3):383–389. https://doi.org/10.1016/0378-3758(94)90133-3

Ríos Insua D, Martín J, Proll L, et al (1997) Sensitivity analysis in statistical decision theory: A decision analytic view. Journal of Statistical Computation and Simulation 57(1–4):197–218. https://doi.org/10.1080/00949659708811808

Robert CP (1994) The Bayesian Choice. Springer, New York, https://doi.org/10.1007/978-1-4757-4314-2

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592. https://doi.org/10.1093/biomet/63.3.581

Savage LJ (1954) The foundations of statistics. Wiley, New York

Schad DJ, Nicenboim B, Bürkner PC, et al (2023) Workflow techniques for the robust use of Bayes factors. Psychological Methods 28(6):1404–1426. https://doi.org/10.1037/met0000472

Seidenfeld T, Kadane JB, Schervish MJ (1989) On the shared preferences of two Bayesian decision makers. The Journal of Philosophy 86(5):225–244. https://doi.org/10.2307/2027108

Simonsohn U, Simmons JP, Nelson LD (2020) Specification curve analysis. Nature Human Behaviour 4(11):1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Steegen S, Tuerlinckx F, Gelman A, et al (2016) Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science 11(5):702–712. https://doi.org/10.1177/1745691616658637

Tran KT, Dao D, Nguyen MD, et al (2025) Multi-agent collaboration mechanisms: A survey of LLMs. https://doi.org/10.48550/arXiv.2501.06322, arXiv preprint, arXiv:2501.06322

Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Computational Biology 11(6):e1004333. https://doi.org/10.1371/journal.pcbi.1004333

Vieth B, Parekh S, Ziegenhain C, et al (2019) A systematic evaluation of single cell RNA-seq analysis pipelines. Nature Communications 10:4667. https://doi.org/10.1038/s41467-019-12266-7

Walley P (1991) Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London

Wang L, Ma C, Feng X, et al (2024) A survey on large language model based autonomous agents. Frontiers of Computer Science 18(6):186345. https://doi.org/

10.1007/s11704-024-40231-1, 2308.11432

Watson J, Holmes C (2016) Approximate models and robust decisions. Statistical Science 31(4):465–489. https://doi.org/10.1214/16-STS592

Weerahandi S, Zidek JV (1981) Multi-Bayesian statistical decision theory. Journal of the Royal Statistical Society: Series A 144(1):85–93. https://doi.org/10.2307/2982162

Weerahandi S, Zidek JV (1983) Elements of multi-Bayesian decision theory. The Annals of Statistics 11(4):1032–1046. https://doi.org/10.1214/aos/1176346319

Wu X, Zhang H, Garduno-Rapp NE, et al (2026) Orchestrator multi-agent clinical decision support system for secondary headache diagnosis in primary care. Journal of the American Medical Informatics Association p ocag111. https://doi.org/10.1093/jamia/ocag111

Zhao W, Wu C, Fan Y, et al (2026a) An agentic system for rare disease diagnosis with traceable reasoning. Nature 651(8106):775–784. https://doi.org/10.1038/s41586-025-10097-9

Zhao WX, Zhou K, Li J, et al (2026b) A survey of large language models. Frontiers of Computer Science 20(12):2012627. https://doi.org/10.1007/s11704-026-60308-3, 2303.18223

Zhu Y, Gu L, Wang Z, et al (2025) Auditing medical multi-agent AI reveals risks of false consensus. https://doi.org/10.48550/arXiv.2510.10185, arXiv preprint, last revised 27 May 2026, arXiv:2510.10185