How can decision models decide to not decide? Modeling suspension in fast-and-frugal trees (FFTs)

Hansjörg Neth and Jelena Meyer

2025-05-20

Abstract

The phenomena of indecision and suspension loom large in both philosophy and psychology. Whereas psychology discusses related phenomena in practical tasks and mostly pathological terms, philosophy strives for conceptual clarification and emphasizes the ubiquity and variety of suspension. In this chapter, we use *fast-and-frugal trees* (FFTs) as a drosophila model for developing a positive account of suspension in decision-making. Being designed for handling binary classification tasks, FFTs seem particularly ill-suited for accommodating a third stance. But by replacing one decision outcome by a *do not know* category or adding it as a third option, we can adapt and extend the FFT framework to explore the causes and consequences of suspension. Considering the distributions of decision outcomes and contrasting the performance of alternative models in terms of cost-benefit trade-offs illustrates the power of this methodology. Overall, a model-based approach provides surprising insights into the functions and mechanisms of suspension and serves as a productive tool for thinking.

Authors Hansjörg Neth and Jelena Meyer
ORCID: 0000-0001-5427-3141 and 0009-0005-2762-8548
Social Psychology and Decision Sciences, *University of Konstanz*

Keywords fast-and-frugal trees (FFTs), judgment and decision making (JDM), binary classification, cost-benefit trade-offs, indecision, computer modeling, machine learning

Reference This is the authors' final manuscript version of the published text.

Please do not cite this draft, but the published book chapter as follows:

- Neth, H., & Meyer, J. (2025). How can decision models decide to not decide? Modeling suspension in fast-and-frugal trees (FFTs). In V. Wagner & A. Zinke (Eds.), *Suspension in epistemology and beyond* (pp. 286–303). New York, NY: Routledge. DOI 10.4324/9781003474302-20

There is nothing as practical as a good theory.
Lewin (1943)

Introduction

Philosophy and psychology share a concern for what should be believed or done, especially when facing risks or uncertainty. But when philosophers ponder questions like "Is the number of stars in the universe even or odd?", decision theorists and psychologists specializing in *Judgment and Decision Making* (JDM) are either puzzled or fail to see a problem. Suspending one's judgment on this issue seems so simple. In fact, responses like "I don't know" or "This cannot be known" seem straightforward and perfectly sensible answers to such questions. But when philosophers then proceed to map out the mental states to which these answers correspond, JDM researchers soon realize that they lack clear concepts of suspension and indecision. Instead, the field of JDM and behavioral economics is littered by a plethora of phenomena and terms that somehow denote an individual's unwillingness or inability to form a judgment or make a decision (see Anderson, 2003, for a review). For instance, individuals prone to withhold certain actions are diagnosed with omission bias (Ritov & Baron, 1990) or inaction inertia (Tykocinski, Pittman, & Tuttle, 1995). Alternatively, a tendency for discounting future options and deferring decisions can be attributed to an endowment effect (Thaler, 1980) or a status-quo bias (Samuelson & Zeckhauser, 1988). As the number of available options increases, a penchant for maximizing information search is linked to unhappiness with decision outcomes (Schwartz et al., 2002) and a paradoxical too-much-choice effect (Iyengar, Wells, & Schwartz, 2006). When inquiring *why* people should exhibit such seemingly irrational behavior, researchers refer to general tendencies for relying on error-prone heuristics (Tversky & Kahneman, 1974), misperceiving risks (e.g., Kahneman & Tversky, 1979), or avoiding regret (e.g., Zeelenberg, 1999).

By mostly addressing variants of being overwhelmed or getting stuck in some form of analysis paralysis, the JDM literature on suspension and indecision is unsatisfying in at least two ways: First, as the bounds and scope of key concepts often remain vague, purely narrative theories only identify and link related phenomena, but fail to provide mechanistic explanations for them. Second, the common denominator of most existing accounts is that individuals deferring a decision or forgoing an action are evaluated negatively — as lacking some important quality or suffering from a bias. As a result of both limitations, the current literature suffers from conceptual confusion and promotes a lopsided and pathological view of indecision and suspension.

While decision theorists may be bewildered by lofty questions regarding the parity of stars, they take pride in addressing questions like "Does this patient suffer from a heart-attack?" (e.g. Green & Mehr, 1997). Intuitively, this diagnostic problem appears to be of a more solid and practical nature. Given some medical data, an agreement on the relevant symptoms and the criteria for a diagnosis, the question seems to have two simple answers with real-world consequences. Beyond its applied context, a key feature of the medical question is that any emergency doctor or unit will repeatedly ask, answer, and act on it. Shifting from single-case considerations to a population of decision instances locates the question in the realm of risk. The defining feature of *risk* is that known unknowns are somehow quantified and can be reckoned with (Gigerenzer, 2002): Under risk, we can apply mathematical or computational methods that estimate probabilities and optimize some criterion. Only when acknowledging that we often cannot escape the reach of unknown unknowns and thus operate on a continuum of uncertainty, the situation gets more complicated and epistemically opaque again (e.g., Neth & Gigerenzer, 2015). Thus, by lacking different types of information, situations of risk and uncertainty correspond to distinct levels of ignorance.

In this chapter, we address our shared concern for dealing with ignorance by arguing for a more positive and precise treatment of indecision and suspension. Our approach combines philosophical and psychological considerations with computational methods. To avoid the pitfalls of purely narrative accounts, we develop our ideas within the framework of fast-and-frugal trees, which provide a formal model for making binary decisions. While our objectives are theoretical in nature, we illustrate our model in a concrete context with real-world consequences that are measured in terms of decision accuracy and error costs.

 DOI 10.4324/9781003474302-20

Fast-and-frugal trees (FFTs)

Fast-and-frugal trees (FFTs) are simple and transparent decision rules that use data to assign a population of cases to one of two outcome categories. Historically, FFTs have been discovered by researchers of simple heuristics (Martignon, Vitouch, Takezawa, & Forster, 2003) and developed as supervised learning algorithms for solving binary classification problems (Martignon, Katsikopoulos, & Woike, 2008). As binary classifications are formalized by signal-detection theory (SDT, Green & Swets, 1966), FFTs have been analyzed in terms of discriminating signals from noise and adopting different biases for making conservative or liberal classification decisions (Luan, Schooler, & Gigerenzer, 2011).

Using an FFT for classifying a case traverses a tree-like structure of nodes that are considered in a fixed sequence until an exit node is reached. Each node evaluates a cue and bifurcates the population of cases (data rows) by comparing the cue's current value to a threshold value. To ensure that FFTs are more frugal than other decision trees, they rarely include more than 3 or 4 cues, and their construction algorithms honor the constraint that every level must contain an exit node. From a data analysis perspective, an FFT considers one or more cues (i.e., variables arranged in data columns) to classify cases (i.e., observations described by a row of data) into a binary outcome variable (e.g., `TRUE` or `FALSE`).[1] From a machine learning perspective, FFTs are binary classification algorithms that chase certain criteria (e.g., some measure of accuracy or of information value). An important issue when developing FFTs is the distinction between *explaining* existing data and *predicting* new data. As with other predictive modeling methods, separating model training from model testing (e.g., by using cross-validation techniques) is a popular way of avoiding over-fitting. Statistically, FFTs rival binomial regression models by using data to explain or predict a binary outcome variable. Cross-tabulating each case's true state with its predicted category yields a 2x2-matrix of classification outcomes that contains the frequencies of two types of correct combinations (with corresponding categories, usually denoted as *hits* and *correct rejections*) and two types of incorrect combinations (with non-corresponding categories, denoted as *misses* and *false alarms*). Depending on the scientific discipline and current interest, the counts of these four outcome types are further organised and quantified in many different ways (see Neth, Gradwohl, Streeb, Keim, & Gaissmaier, 2021). The performance of FFTs is typically evaluated in terms of classification accuracy and costs. A popular software tool for creating and evaluating FFTs is provided by the R package **FFTrees** (Phillips, Neth, Woike, & Gaissmaier, 2017).

Figure 1 illustrates an FFT that used the **FFTrees** package to diagnose a population of 303 patients that were suspected of suffering from heart disease. The binary criterion variable to be predicted provides the true status of their `diagnosis` (i.e., each patient either suffered or did not suffer from heart disease). Possible predictors for this status are 13 cues that include variables like `age` or `sex`, as well other categorical or numeric variables that contain diagnostic information. To avoid over-fitting, the FFT shown in Figure 1 was developed for a subset of the data (containing a random sample of 150 patients) and then used to predict the criterion variable of the other subset (containing the remaining 153 patients). The top panel of Figure 1 provides baseline information and indicates that the values of the criterion variable in this sample are well-balanced: 48% of the patients suffer from heart disease, 52% of them do not. The FFT shown in the center of Figure 1 resulted from evaluating all potential predictors, but selected only three of them: The categorical variables `thal` and `cp` note types of heart rate defect and chest pain, respectively, and the numerical variable `ca` indicates the major vessels colored by flouroscopy (as a value from 0 to 3). The icons on the left and right exits show when and how each of the 153 patients is being classified by this FFT. Predicting a binary criterion typically allows for two correct and two incorrect cases: When predicting *heart disease* (i.e., the signal category on the right), the outcome is either a *hit* (i.e., a correct classification of a patient with heart disease) or a *false alarm* (i.e., an incorrect classification of a patient without heart disease). When predicting *No heart disease* (i.e., the noise category on the left), the outcome is either a *correct rejection* (i.e., a correct classification of a patient without heart disease) or a *miss* (i.e., an incorrect classification of a patient with heart disease). The bottom panel of Figure 1 summarizes the counts of these combinations in a 2x2-matrix, provides common performance metrics (e.g., the overall classification accuracy is 82%), as well as the trade-offs of related FFTs with alternative exit structures.

[1] As the binary outcomes of FFTs usually entail actions, they are commonly described as *decisions*. But given corresponding tasks, the outcome categories can also be described as binary *choices* or *judgments*. In the following, we use *suspension* as a neutral term and label the corresponding outcome category as *do not know*.

 DOI 10.4324/9781003474302-20

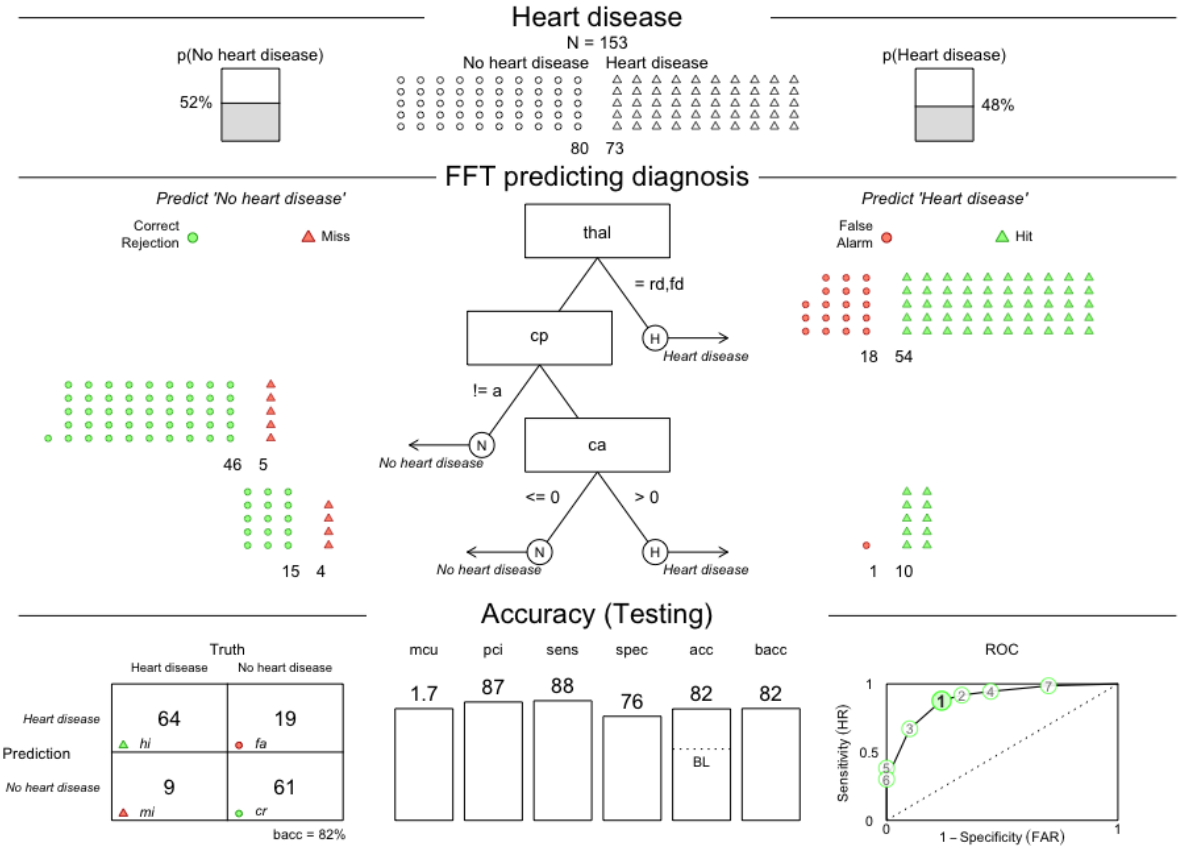

Figure 1: A fast-and-frugal tree (FFT) predicting the presence or absence of heart disease.

 DOI 10.4324/9781003474302-20

Indecision or suspension in FFTs

Where can we locate indecision or suspension in FFTs? *A priori*, there are two possible answers, both of which are disappointing. The first is that FFTs are simply not made for representing an outcome that is neither signal nor noise. As the nature of a binary classification task determines the design of an FFT, allowing for a third stance would effectively call for a different structure and construction algorithm. Alternatively, we could argue that an FFT already entails multiple aspects of indecision or suspension. For instance, when philosophers point out that suspension serves as a motivation for inquiry (for discussions, see Douven, this volume; McGrath (2021), this volume; Wagner (2022), this volume; Zinke (2021)), a state of suspension may be the original impetus for employing an FFT. Similarly, when viewing suspension as a transitory state in which new evidence is sought and being considered (e.g., Staffel, 2019), reaching any node of an FFT provides a concrete instance of suspension. Both of these extreme answers are inflating opposite claims: Whereas the first flat-out denies that FFTs could capture phenomena related to suspension, the second merely points at parts of FFTs and describes them in corresponding terms. While outright dismissals or opportunistic re-framings have their merits, we believe these answers to be unwarranted here. On the one hand, if indecision and suspension are genuine concepts worthy of investigation, it should be possible to accommodate a generic JDM model to capture the corresponding phenomena. On the other hand, any serious treatment of a concept by a model should illuminate and enrich the phenomena that are being explained, rather than just change the model's narrative. In the context of a decision model, we must either show how a change in its assumptions or goals leads to different decisions, or change the model so that it can yield different outcomes.

In the next two sections, we use and modify FFTs to illustrate both of these points. Whereas the first example will add suspension within the binary framework of FFTs, the second example will expand the model by allowing for a third decision outcome.

Suspension as a binary decision outcome

A minimally invasive way of introducing indecision to a binary classification framework is to re-conceptualize the categories that are being explained or predicted. Assuming that our FFT aims to predict *heart disease* as its signal, our interpretation of the noise category allows for some flexibility. In Figure 1, we denoted the signal's absence as *no heart disease*, to reflect the true state of patients with a `diagnosis` value of `FALSE`. Rather than just negating the signal category, we could be tempted to re-conceptualize the noise category as deciding that a patient is *healthy* or that we are uncertain and *do not know*.[2] Having acknowledged that merely re-labeling our decision outcome categories would make a cheap parlor trick, we must not stop here. Instead, our goal is to reflect on the consequences of such a change and incorporate the semantics of each concept into the design of our decision tree.

What changes if we re-conceptualized our task as predicting that a patient is either suffering from *heart disease* or is *healthy*? Clearly, predicting that someone is *healthy* would be a stronger statement than merely predicting the absence of *heart disease*. But how can this semantic shift matter for our FFT? The key construct to be considered here is the 2x2-matrix of outcome type combinations that cross-tabulates the counts of true vs. predicted states for all cases. Any diagnostic tool aims for both high *sensitivity* (i.e., the probability of correctly classifying signals) and high *specificity* (the probability of correctly classifying non-signals). Unfortunately, we usually cannot optimize both, but instead must trade-off one of these measures against the other. Thus, an important question to ask whenever developing a decision tree is: Which of the two types of error is worse?

Changing the noise category of our predictions from *no heart disease* to *healthy* changes the meaning of the cells in the lower row of the 2x2-matrix. In the present context, a *miss* would incorrectly diagnose a patient with heart disease as *healthy*, whereas a *false alarm* would incorrectly diagnose a patient without heart disease as *diseased*. Although the costs of over-treating false alarms can be considerable, missing a *diseased* patient has become even less acceptable by re-framing our noise category from *no heart disease* to *healthy*. In diagnostic terms, reducing the likelihood of a *miss* at the expense of increasing the likelihood

[2]Such creative re-labeling would not do justice to our current data, since the absence of diagnosis of *heart disease* neither implies that a patient is *healthy* nor that we *do not know* the diagnosis. But rather than conflating these differences, the goal of this section is to show that they matter and — *if* the implications were valid — would yield different decision models.

 DOI 10.4324/9781003474302-20

of a *false alarm* corresponds to a preference for increasing the test's sensitivity at the expense of reducing its specificity. When using **FFTrees** to create FFTs, this preference can be accommodated by adjusting a `sens.w` parameter that specifies the weight of sensitivity, relative to the weight of specificity (with both weights summing to 1). When generating the FFT shown in Figure 1 (above), both were weighted equally (`sens.w = .50`) and the algorithm for FFT-creation was optimized for balanced accuracy (`bacc`). But our reasoning that *misses* should be avoided at risk of increasing *false alarms* suggests that we should increase the weight of sensitivity (e.g., `sens.w > .50`) and optimize our FFT for weighted accuracy (`wacc`).

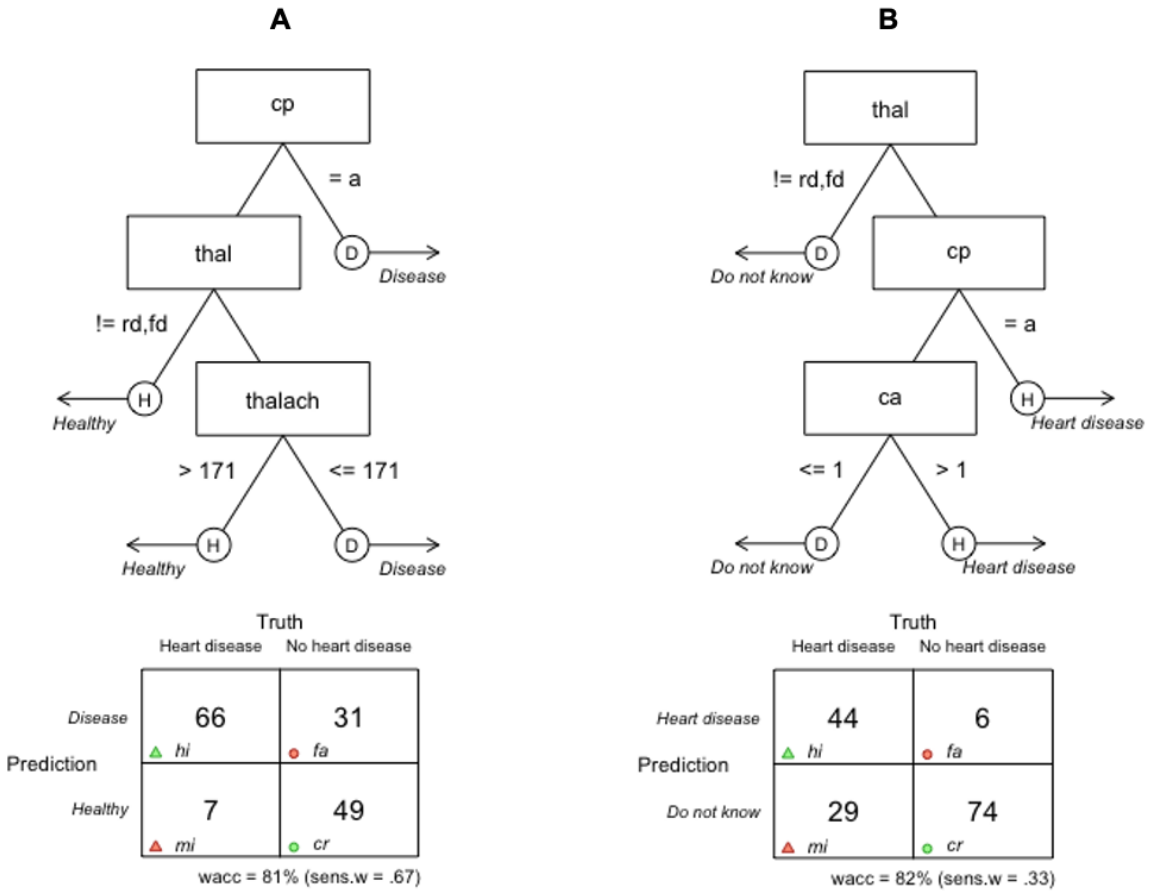

Figure 2: FFTs optimizing for different binary decision outcomes. Whereas FFT A discriminates between 'heart disease' and 'healthy', FFT B discriminates between 'heart disease' and 'do no know'. Considering the consequences of these semantic shifts leads to different FFTs and outcome distributions.

Figure 2A shows an FFT obtained by prioritizing sensitivity (`sens.w = 2/3`) and its classification outcomes. Most importantly, the FFT shown is no longer identical to Figure 1. The new best FFT not only uses a different cue (i.e., by dropping `ca` in favor of `thalach`, which measures the maximum heart rate achieved), but also by swapping the order and exit structures of the two retained cues (`thal` and `cp`). In terms of performance, the 2x2-matrix for the FFT of Figure 2A shows that our emphasis on *sensitivity* has indeed reduced the counts of *miss* cases (from 9 to 7) and increased the counts of *false alarm* cases (from 19 to 31). As a side-effect, we also reduced the number of patients in the *correct rejection* category (from 61 to 49), which is welcome, since patients with a true state of *no heart disease* may actually not be *healthy*. Overall, a slight increase of sensitivity (from 88% to 90%) came at the expense of a marked reduction of specificity (from 76% to 61%). Although the level of balanced accuracy decreased (from 82% to 75%), the weighted accuracy measure for which we optimized stayed almost constant (at 81%).

What are the consequences of re-labeling the noise category of our predictions as *do not know*? The severity

of this change hinges on the practical consequences of acknowledging one's ignorance. As a *miss* no longer declares a diseased patient as *healthy*, it now may motivate further inquiry and call for more diagnostic tests.[3] This would render the consequences of a *miss* much more benign, but *false alarm* cases (i.e., declaring patients without heart disease as *diseased*) become relatively less acceptable. Thus, the binary suspension case of predicting either *heart disease* or *do not know* yields the opposite preference for the distribution of outcomes than declaring patients without heart disease as *healthy*. This can be incorporated into the generation of an FFT by reducing the emphasis on sensitivity (`sens.w = 1/3`) when chasing weighted accuracy (`wacc`).

Again, the best resulting FFT (shown in Figure 2B) changes. Although it contains the same cues and cue order as the FFT of Figure 1, it uses different exit structures for `thal` and `cp` and alters the criterion of the final `ca` cue. In terms of prediction performance, preferring specificity to sensitivity increased the number of *miss* cases (from 7 to 29) and reduced the number of *false alarm* cases (from 31 to 6). As a consequence, we obtain a large increase in specificity (from 61% to 92%), at the expense of reducing sensitivity by a similar margin (from 90% to 60%). While the overall level of balanced accuracy was reduced (at 77%, whereas our original FFT achieved 82%), the weighted accuracy measure reached the same level (at 82%).

The FFT shown in Figure 2B also illustrates an unwelcome side-effect of increasing specificity. The successful reduction of *false alarm* cases (from 31 to 6) came at the cost of increasing the number of *correct rejection* cases (from 49 to to 74). In fact, our new conceptualization classified 92% (i.e., 74 of 80) of the patients without heart disease as *do not know*. But re-framing the noise category as suspension renders the correctness of this combination somewhat dubious. The inflation of cases without a positive diagnosis illustrates two shortcomings of our binary suspension FFT: First, a high level of future diagnostic scrutiny required for *do not know* cases may imply costs that could outweigh the savings achieved by lowering the number of over-treatments (by reducing *false alarm* cases). This quantitative problem could potentially be addressed by re-balancing the distribution of cases by adjusting the sensitivity weight parameter. However, a more serious defect of any FFT that discriminates only between *heart disease* and *do not know* is the fact that it no longer can predict the positive *absence* of heart disease. Thus, accommodating suspension in the binary structure of an FFT is not just a matter of fine-tuning the distribution of outcomes, but also comes at the cost of sacrificing qualitative outcome combinations.

Rather than arguing for a particular FFT, the examples of this section have shown that we can explore the benefits and costs of different conceptualizations. When working with a model, re-labeling the decision outcome of *no heart disease* as either *healthy* or *do not know* is not just idle talk, but has consequences on different levels. As different outcome combinations call for different trade-offs (e.g., between sensitivity and specificity), we generated three distinct FFTs (with different cues, cue orders, thresholds, and exit structures). Despite their substantial differences, their overall level of weighted accuracy remained relatively constant (at 81 or 82%).

But demonstrating that our model can react to different conceptualizations also revealed its qualitative limitations. Specifically, when we want to distinguish cases of suspension from both the presence and the absence of a diagnosis, a binary FFT will not suffice. If we were forced to address this problem within our binary framework, we could generate and combine multiple FFTs. For instance, if a first FFT distinguished positive cases from suspension, a second FFT could aim to identify negative cases in the suspended cases. Such a solution would still aim to tackle a non-binary problem in a risk-based model of binary decision making. A more drastic change would be to change the mechanism or structure of the model to reflect a theoretical insight. This insight is that our FFT-model currently affords no positive role for the notion of uncertainty. Reflecting on this notion will reveal some facets that can be accommodated by extending the model. This route will be pursued in the next section.

Suspension as a third decision outcome

How can we include uncertainty in a decision model? In the decision sciences, uncertainty is usually defined in negative terms — as the absence of certainty — or a lack of some desirable information (e.g., options, probability, or outcome values). A fundamental reason for excluding uncertainty from any model is that

[3]The precise semantics of a *do not know* prediction do not matter here, but we assume that the category implies some notion of non-finality and openness to further evidence.

any strong notion of uncertainty prevents us from optimizing a criterion. If uncertainty implies the inability of computation, the very definition of a model would transform "unknown unknowns" into the "known unknowns" of risk. Nevertheless, we can still try to incorporate some aspects of uncertainty into a model. As the representational constraints of a model force us to precisely explain the corresponding changes and their consequences, a model may help us gain a better understanding of uncertainty. In this section, we distinguish and address two aspects of uncertainty: First, *uncertain inputs* imply that cues or cue values required for making a decision may be missing from our data. By contrast, allowing for *uncertain outcomes* would postulate a novel end result of a decision process. Ontologically, missing data would suggest an environmental source of uncertainty, whereas an uncertain decision outcome implies a state of suspension (or corresponding action) by a decision maker. In the following, we will explore both of these aspects in the context of FFTs.

Missing data is a ubiquitous problem in the applied decision sciences. A common response for handling situations with incomplete data is to impute missing values, with different methods being used based on their random or systematic nature. Whereas such imputation fills gaps in data, we add missing values to our `heartdisease` data to assess their effects on the diagnostic performance of FFTs. In the following analyses, we replace 10% of all predictive cue values by missing (`NA`) values in a completely at random fashion (aka. MCAR). To prevent accidental effects, we repeat this procedure 100 times and average over the results.

Instead of repairing the missing values in our data, we use them to motivate suspension as a decision outcome. But rather than replacing one of the binary outcome categories by a suspension category (as in the previous section), we now extend our model by considering a third decision outcome. Thus, the three possible outcomes of our diagnostic scenario are *heart disease*, *no heart disease*, and the suspension category *do not know*.

Adding a third option abandons the binary structure of the FFT. To evaluate the effects of such a drastic change, we contrast three modified versions of our FFT model:

1. *No suspension* model: A maximally conservative model for dealing with missing data in an FFT could completely avoid choosing the suspension option. When missing data input at a non-final node, the model proceeds to the next node (analogous to asking for the next question in an exam or quiz when lacking the answer to a current one). When missing data input at the final node, however, we cannot ignore this by proceeding further. But rather than choosing to suspend, we can always take a gamble and guess one of the two other outcome categories. A heuristic of always opting for the more frequent category ensures that a guessing strategy will not fall below this baseline. As this model would even refuse to suspend when it lacks all information, it is effectively preventing outcome uncertainty.

2. *Liberal suspension* model: At the other extreme, a liberal policy for translating uncertain inputs into uncertain outputs would choose to suspend whenever it encounters any missing data. This model handles all cases with complete information exactly as the original model, but reacts maximally sensitive to any missing element. As this model seizes any occasion for choosing to suspend, it can be thought of as embracing or seeking outcome uncertainty.

3. *Conservative suspension* model: A third model combines the mechanisms of the two previous models. When missing data input at a non-final node, the model proceeds to the next node (as the *no suspension* model). However, when missing data input at the final node, the model chooses the suspension option (as the *liberal suspension* model). As this model classifies cases at its non-final nodes and only opts for suspension when it evaluates its final node and lacks the value of the corresponding cue, it provides a conservative suspension policy that finally — almost reluctantly — opts for suspension when all earlier classification attempts have failed.

Given our simulation setup, we can compute the outcome distributions for an FFT and a given percentage of missing values (see Table 1). The *no suspension* model (whose distribution of classification outcomes is shown in Panel A1 of Table 1) performs surprisingly close to our original FFT (shown in Figure 1 above). As the `heartdisease` data contains more negative than positive criterion cases (i.e., fewer patients with a diagnosis of heart disease than without it) the baseline guessing mechanism increases the number of *miss* and *correct rejection* cases. However, at 10% of missing data values, the model classifies cases almost on par with the original FFT, with a balanced accuracy value of 80%, while the original model achieved 82%. Thus, the cost of skipping cues and guessing in the face of missing information are very marginal. By contrast,

 DOI 10.4324/9781003474302-20

the lower suspension threshold of the *liberal suspension* model (shown in Panel A2) reduces all four of the previous outcome combinations by assigning almost 16% of its cases (on average) to the suspension category. As the counts of correct predictions (i.e., *hit* and *correct rejection* cases) are reduced more rapidly than the erroneous ones (i.e., *false alarm* and *miss* cases), the balanced accuracy of this model drops to 68%. The late suspension policy of the *conservative suspension* model (shown in Panel A3) substantially reduces the number of suspension choices to 2.5% of its cases (on average). The main reason for this reduction is that the vast majority of cases has been classified before reaching the final node. As a consequence, the balanced accuracy of this model reaches 79% again.

	1 *No suspension*		2 *Liberal suspension*		3 *Conservative suspension*	
A	**Truth**		**Truth**		**Truth**	
Prediction	Heart disease	No heart disease	Heart disease	No heart disease	Heart disease	No heart disease
heart disease	60.5	17.8	55.9	16.8	60.5	17.8
do not know	0.0	0.0	10.0	14.3	1.7	2.2
no heart disease	12.5	62.2	7.1	48.9	10.9	60.0
	bacc = 80%		bacc = 68%		bacc = 79%	
B						
Cost-benefit ratio		Utility		Utility		Utility
1:1		**92.3**		80.9		91.8
2:1		62.0		57.0		**63.1**
3:1		31.7		33.0		**34.5**
4:1		1.4		**9.1**		5.8

Table 1. Distributions of *decision outcomes* (**A**) and corresponding *utility values* (**B**) for three alternative FFT models. (Outcome counts provide averages over 100 simulations with 10% of missing data values. Maximum utility values are shown in bold.)

Does this imply that the *no suspension* model wins, with the *conservative suspension* model scoring a close second place? It would, if balanced accuracy was the only criterion used to evaluate model success. But a key rationale for the incorporation of suspension into a model is that not all errors are alike. From the perspective of the binary criterion variable, whose values describe the true status of a patient as either suffering or not suffering from heart disease, choosing a suspension option always constitutes an error. However, we could argue that predicting *do not know* for a patient with heart disease (i.e., choosing suspension for a signal case) is generally less severe than a *miss* (i.e., erroneously predicting *no heart disease*). Similarly, predicting *do not know* for a patient without heart disease (i.e., choosing suspension for a noise case) is generally less severe than a *false alarm* (i.e., erroneously predicting *heart disease*). Thus, when accounting for the cost of errors, choosing one the suspension categories is less costly than the corresponding error that predicts the incorrect category. But as opting for suspension can also reduce the number of correct classifications, we need to explicate the benefits and costs of all cases in order to compute and compare the outcome distributions.

To summarize each model's performance in a single value, we compute their overall utility by combining their classification accuracy with cost-benefit considerations of outcome types. The abstract utility value can be quantified by assigning a benefit or cost to each outcome type and weigh the frequency of outcomes by these values. To reflect our theoretical considerations, we initially assume that any correct case (i.e., a *hit* or *correct rejection* case) has a positive utility (benefit) of +1, any incorrect case (*false alarm* or *miss* case) has a negative utility (cost) of −1, and any suspension case (*suspension for signal* or *suspension for noise*) has an intermediate utility (or neutral value) of 0. Given these settings, higher utility values indicate better performance.[4] As this setup closely reflects the computation of balanced accuracy — with the only difference

[4]Utility can be expressed in various currencies. Whereas absolute values can be chosen in an arbitrary fashion, the relations between values matter for our model comparisons.

 DOI 10.4324/9781003474302-20

being that an error is assigned a negative utility (or cost) of −1, while suspensions are being ignored — it is not surprising that the three models closely mirror (and rank in the same order as) their balanced accuracy. Given a cost-benefit ratio of 1:1, the outcome distribution of the three models score overall utilities of 92.3 (*no suspension*), 91.8 (*conservative suspension*), and 80.9 (*liberal suspension*).

The advantage of shifting from classification accuracy to utility becomes apparent when exploring different cost-benefit ratios. As opting for suspension ultimately reduces the number of errors, it is clear that models favoring suspension will benefit as error costs increase. In Table 1B, we provide the three model's overall utility values when the cost of an error relative to a correct classification rises from 1:1 to ratios of 2:1, 3:1, and 4:1 (while suspension cases retain their neutral value of 0). While the overall utility of all models decreases (due to the negative utility of errors), the rank order between models at the same cost-benefit ratio shifts: Whereas a balanced cost-benefit ratio of 1:1 favors the *no suspension* model, doubling the cost of errors to a 2:1 ratio favors the *conservative suspension* model. At a ratio of 3:1, the *conservative suspension* model still wins, but is closely followed by the *liberal suspension* model. From a ratio of 4:1 onward, the *liberal suspension* model always wins. Based on our outcome distributions, we can even compute the inflection points at which the rank order of models changes: The *conservative suspension* model overtakes the *no suspension* model when the costs of an error exceed the benefits of a correct classification by 1.3 or more, and the *liberal suspension* model overtakes the *conservative suspension* model when the error costs are 3.3 as high as the benefits of a correct prediction.

Overall, we have no single best model, but ranges of cost-benefit ratios in which each type of model outperforms its competitors.[5] The lesson to be learned here is that the best way for including uncertainty in a decision model depends on the relative benefits and costs of outcome combinations. When errors have low costs, the case for suspension in favor of guessing is surprisingly weak. And although it is analytically true that avoiding errors becomes more important as they get more expensive, our comparison between a conservative and a liberal suspension model allowed to qualify and quantify this insight.

Discussion

In situations of risk or uncertainty, we often may want to suspend judgment or decide to not decide. How these phenomena can be captured by formal JDM models remains a territory largely unexplored (but see Malhotra, Leslie, Ludwig, & Bogacz, 2017; McElfresh et al., 2021; Reynolds et al., 2021, for exceptions). Having acknowledged that FFTs — by being designed for binary classification tasks — may be ill-suited for this purpose, we explored and extended their features to consider suspension in the context of a medical diagnosis task.

Let's briefly take stock of what we have done and what we can learn from it. At first, the simple structure of FFTs, with only two outcomes and a required exit on every level, seemed so rigid that it appeared to prohibit any nuanced account of suspension. Our primary candidate for a role of suspension in this framework was to propose a new outcome category that could represent a judgment or a decision and was labeled as *do not know*. Replacing one of the two original outcome categories by this suspension category and reflecting on the consequences triggered a range of non-trivial changes in the FFT. Realizing that insisting on a binary tree structure deprived us of desirable outcomes, we then extended the FFT-design by allowing for suspension as a third outcome category. To understand the consequences of this change, we contrasted three versions that varied how easily and frequently the suspension option was chosen. Beyond constructing FFTs with different surface structures, we encountered representational constructs of distinct notions of suspension. Aiming to add a modicum of uncertainty to our risk-based model, we observed that missing data could either prompt a leap to the next question or result in an uncertain outcome. Note that this distinction between transitional and final suspension differs from merely re-labeling some parts of a generic FFT as "suspension" insofar as it emerged as a response to a real issue (missing data) and showed that uncertain inputs must

[5] In all considerations of this section, signal and noise cases were weighted equally (i.e., balancing sensitivity and specificity), but asymmetrical situations could be computed in the same way. Note also that the three model variants of this section were derived from our initial FFT (created for the complete data and shown in Figure 1), rather than optimized under the modified conditions. If a specific cost-benefit matrix was given first, we could optimize each of the three model types and would probably discover variants of the original FFT with a higher overall utility (as in the previous section).

 DOI 10.4324/9781003474302-20

not necessarily lead to a state of uncertain outcomes. Embedding these notions in the framework of FFTs forced us to explicate their mechanisms and allowed us to measure the consequences of all structural changes. Overall, exploring suspension in the context of FFTs enriched our model (e.g., by now being able to handle missing data) and our understanding of what it means to decide that we *do not know*.

Before we conclude, we should briefly discuss the status of our FFTs. In the past, FFTs have been touted both as descriptive models of how people *do* make decisions and as prescriptive models of how they *ought to* make decisions. However, our FFTs for diagnosing heart disease were neither designed as descriptive nor as prescriptive tools. Instead, they were created as the results of data inputs, outcome cost considerations, and an optimization algorithm built into the **FFTrees** software package. Nevertheless, our examples of FFTs also illustrate the close proximity of descriptive and prescriptive aspects: Even a purely descriptive model acquires normative implications when the costs of outcomes are being measured or the model is compared to other models.

Conclusion

We began this chapter by juxtaposing philosophers' lofty puzzles with practical concerns addressed by decision scientists. But as the introductory maxim by Kurt Lewin (1943) already suggested, the duality of theory and praxis is artificial. Regarding issues of indecision and suspension, the interplay of conceptual issues and their real-world consequences is too intricate and too important to divide them along disciplinary boundaries.

Whereas psychological research has discussed matters of indecision mostly in vague and pathological terms, philosophical thinkers emphasize the ubiquity and variety of suspension. Despite different goals and perspectives, we believe that a combination of forces offers more opportunities than obstacles. The joint challenge lies in developing a positive theory of suspension — a theory that identifies its functions and mechanisms and illuminates mental states and their environmental correlates.

A key instrument for merging both approaches is of a methodological nature: Our explorations with FFTs demonstrate the general advantages of modeling. As a model imposes qualitative and quantitative constraints, working within its framework may initially seem limiting. But specifying a representational structure and the process by which inputs are transformed into outputs cultivates an attention to detail that can also be liberating. In contrast to purely narrative accounts, a good model explicates its mechanisms and enables precise predictions. For instance, our examples have demonstrated that different assumptions about the types and costs of decision outcomes justify different FFT designs.

Ideally, a useful model should serve as a tool for thinking that is both flexible and rigid: Flexible enough to represent new ideas (like adding a *do not know* option), but rigid enough to assess their consequences. Within a well-specified model, even the subtle effects of semantic shifts (e.g., from *no heart disease* to *healthy*) can be measured (e.g., in terms of accuracy or costs) and evaluated (e.g., in terms of trade-offs or utilities). While simulation models guarantee no surplus of insights over conceptual or mathematical approaches they are simpler to design and can still surprise us (e.g., by showing how a suspension option makes sense when error costs exceed some limit).

Although FFTs — by virtue of their transparency — turned out to be a good drosophila for exploring the causes and consequences of suspension, we encourage others to extend other types of models in similar ways. Our hope and trust for advancing both theoretical and practical insights on suspension by such models is sustained by another aphorism commonly attributed to Kurt Lewin (e.g., in Stam, 1995, p. 31):

If you want truly to understand something, try to change it.

 DOI 10.4324/9781003474302-20

References

Anderson, C. J. (2003). The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin*, *129*(1), 139–167. doi:10.1037/0033-2909.129.1.139

Gigerenzer, G. (2002). *Reckoning with risk: Learning to live with uncertainty*. London, UK: Penguin.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York, NY: Wiley.

Green, L., & Mehr, D. (1997). What alters physicians' decisions to admit to the coronary care unit? *Journal of Family Practice*, *45*(3), 219–226.

Iyengar, S. S., Wells, R. E., & Schwartz, B. (2006). Doing better but feeling worse: Looking for the "best" job undermines satisfaction. *Psychological Science*, *17*(2), 143–150. doi:10.1111/j.1467-9280.2006.01677.x

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391.

Lewin, K. (1943). Psychology and the process of group living. *The Journal of Social Psychology*, *17*(1), 113–131. doi:10.1080/00224545.1943.9712269

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*(2), 316–338. doi:10.1037/a0022684

Malhotra, G., Leslie, D. S., Ludwig, C. J. H., & Bogacz, R. (2017). Overcoming indecision by changing the decision boundary. *Journal of Experimental Psychology: General*, *146*(6), 776–805. doi:10.1037/xge0000286

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*(6), 352–361. doi:10.1016/j.jmp.2008.04.003

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In &. L. M. D. Hardman (Ed.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making* (pp. 189–211). Chichester: John Wiley; Sons.

McElfresh, D. C., Chan, L., Doyle, K., Sinnott-Armstrong, W., Conitzer, V., Schaich Borg, J., & Dickerson, J. P. (2021). Indecision modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 5975–5983. doi:10.1609/aaai.v35i7.16746

McGrath, M. (2021). Being neutral: Agnosticism, inquiry and the suspension of judgment. *Noûs*, *55*(2), 463–484. doi:10.1111/nous.12323

Neth, H., & Gigerenzer, G. (2015). Heuristics: Tools for an uncertain world. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences*. New York, NY: Wiley Online Library. doi:10.1002/9781118900772.etrds0394

Neth, H., Gradwohl, N., Streeb, D., Keim, D. A., & Gaissmaier, W. (2021). Perspectives on the 2x2 matrix: Solving semantically distinct problems based on a shared structure of binary contingencies. *Frontiers in Psychology*, *11*, 567817. doi:10.3389/fpsyg.2020.567817

Phillips, N. D., Neth, H., Woike, J. K., & Gaissmaier, W. (2017). FFTrees: A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgment and Decision Making*, *12*(4), 344–368. doi:10.1017/S1930297500006239

Reynolds, A., Garton, R., Kvam, P., Sauer, J., Osth, A. F., & Heathcote, A. (2021). A dynamic model of deciding not to choose. *Journal of Experimental Psychology: General*, *150*(1), 42–66. doi:10.1037/xge0000770

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, *3*(4), 263–277. doi:10.1002/bdm.3960030404

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*, 7–59. doi:10.1007/BF00055564

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*(5), 1178–1197. doi:10.1037//0022-3514.83.5.1178

Staffel, J. (2019). Credences and suspended judgments as transitional attitudes. *Philosophical Issues*, *29*(1), 281–294. doi:10.1111/phis.12154

Stam, H. J. (1995). Theory and practice. In C. W. Tolman, F. Cherry, R. van Hezewijk, & I. Lubek (Eds.), *Problems of theoretical psychology* (pp. 24–31). International Society for Theoretical Psychology.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60. doi:10.1016/0167-2681(80)90051-7

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi:10.1126/science.185.4157.1124

Tykocinski, O. E., Pittman, T. S., & Tuttle, E. E. (1995). Inaction inertia: Foregoing future benefits as a result of an initial failure to act. *Journal of Personality and Social Psychology*, *68*(5), 793–803. doi:10.1037/0022-3514.68.5.793

Wagner, V. (2022). Agnosticism as settled indecision. *Philosophical Studies*, *179*(2), 671–697. doi:10.1007/s11098-021-01676-3

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of Behavioral Decision Making*, *12*(2), 93–106. doi:10.1002/(SICI)1099-0771(199906)12:2%3C93::AID-BDM311%3E3.0.CO;2-S

Zinke, A. (2021). Rational suspension. *Theoria*, *87*(5), 1050–1066. doi:10.1111/theo.12320