ASTROBIOLOGY
Volume 20, Number 10, 2020
Mary Ann Liebert, Inc.
DOI: 10.1089/ast.2019.2151

Evaluating Biosignatures for Life Detection

Andrew Pohorille[1] and Joanna Sokolowska[2]

Abstract

Conceptual frameworks are developed for evaluating the ability of different biosignatures to provide evidence for the presence of life in planned missions or observational studies. The focus is on intrinsic characteristics of biosignatures in space environments rather than on their detection, which depends on technology. Evaluation procedures are drawn from extensive studies in decision theory on related problems in business, engineering, medical fields, and the social arena. Three approaches are particularly useful. Two of them, Signal Detection Theory and Bayesian hypothesis testing, are based on probabilities. The third approach is based on utility theory. In all the frameworks, knowledge about a subject matter has to be translated into probabilities and/or utilities in a multistep process called elicitation. We present the first attempt to cover all steps, from acquiring knowledge about biosignatures to assigning probabilities or utilities to global quantities, such as false positives and false negatives. Since elicitation involves human judgment that is always prone to perceptual and cognitive biases, the relevant biases are discussed and illustrated in examples. We further discuss at which stage of elicitation human judgment should be involved to ensure the most reliable outcomes. An example, how evaluating biosignatures might be implemented, is given in the Supplementary Information. Key Words: Biosignatures—Life detection—Decision theory—Signal Detection Theory—Utility theory—Bayesian hypothesis testing. Astrobiology 20, 1236–1250.

1. Introduction

Searching for extraterrestrial life is one of the most exciting human endeavors. Since the Viking missions to Mars designed to detect signs of life *in situ* in the form of active metabolism yielded results that were commonly (Klein, 1978; Dick, 2006; Quinn *et al*., 2013), although not universally (Levin and Straat, 2016), interpreted as negative, subsequent strategy relied primarily on reconstructing geological history of potentially inhabited celestial bodies and on determining their habitability without explicit reference to the search for extant or extinct life.

Only recently, this strategy has changed. Unambiguous evidence of the presence of large quantities of liquid water in Mars' history (Carr, 2006; Di Achille and Hynek, 2010) and the discovery of organic material by the Curiosity rover (Freissinet *et al*., 2015; Eigenbrode *et al.,* 2018) have shifted attention from habitability to searching for evidence of past life on this planet. In parallel, studies on icy moons of Saturn and Jupiter have led to a conclusion that these bodies could support life (Hand *et al.,* 2009; Priscu and Hand, 2012; Tsou *et al*., 2012; Lunine, 2017). Consequently, they have been proposed as targets for future life-detection missions. Other potential targets within the Solar System (Schulze-Makuch and Irwin, 2006; Shapiro and Schulze-Makuch, 2009), such as Titan (Fortes, 2000; Raulin and Owen, 2002; Schulze-Makuch and Grinspoon, 2005) and Venus (Cockell, 1999; Limaye *et al.,* 2018), have been also proposed, although they are considered by some as problematic on fundamental grounds (Pohorille and Pratt, 2012; Hoehler *et al.,* 2020). Further, discoveries of extrasolar planets in habitable zones (Kane *et al*., 2016) have provided strong impetus for observational efforts to detect signs of life beyond the Solar System (Des Marais *et al*., 2008).

Most approaches aimed at detecting life rely on searching for biosignatures. Following other, recent studies (Catling *et al.*, 2018; Kiang *et al.,* 2018), we define biosignatures as chemical species, features or processes that provide evidence for the presence of life. A more precise definition is given in the next section. Examples of biosignatures are chemical compounds or classes of compounds characteristic of biological matter, for example, pigments; specific features of compounds generated biologically, such as chirality; characteristic biology patterns of complexity, for example in the structure and distribution of

[1]Exobiology Branch, NASA Ames Research Center, Moffett Field, California, USA.
[2]SWPS University of Social Sciences and Humanities, Warsaw, Poland.

amino acids or lipids; molecular oxygen on exoplanets; or signs of metabolic activity. The current definitions differ subtly but significantly from older ones (Des Marais *et al.,* 2008), according to which the existence of a biosignature ''specifically requires a biological agent.'' The previous definitions imply the unique relationship between biosignatures and life; finding a qualified biosignature would be equivalent to claiming life detection. In contrast, the updated definitions emphasize that biosignatures contain information about the presence of life but in general may be inconclusive. Since finding different biosignatures may give us a different degree of confidence that they have been produced biologically, they need to be carefully evaluated in this respect. Such evaluation is necessary for both further development of life detection science and planning future missions and remote observations aimed at searching for signs of life.

Evaluation of biosignatures is a complex, multistep process. First, informative biosignatures should be identified, organized, and considered in the environmental context. So far, most activities in the field have been conducted along these lines. In particular, several innovative ideas for biosignatures have been recently proposed (Cronin and Walker, 2016; Benner, 2017; Marshall *et al.,* 2017; Johnson *et al.,* 2018). Their common feature is that they are not bound by what is known about terrestrial life but instead appeal to our current understanding of ''universal biology.'' In other words, they should be associated with all forms of life, independently of their environmental context. For this reason, these biosignatures are often called ''agnostic.'' The significance of this approach is in expanding the concept of biosignatures from what is known to exist to what might exist.

Once the inventory of biosignatures is available, next steps follow. The evaluation criteria have to be formulated, detection requirements have to be identified and confronted with current technical capabilities, and appropriate detection instruments have to be secured. In this respect, the recent paper on the Ladder of Life Detection (LoLD) (Neveu *et al.,* 2018) stands out, as it takes, for the first time, a comprehensive approach aimed at addressing all these steps. The authors do not attempt to be definitive. Instead, in their own words, the study ''is intended as a starting point to stimulate discussion, debate, and further research.''

The main goal of this paper is to contribute to this discussion by focusing on a step that is possibly the most difficult—the evaluation procedure. In the LoLD approach, the authors formulate a set of binary rules (0/1 or true/false) that, once applied according to the Boolean algebra, lead to an unambiguous, binary conclusion whether the collected evidence is sufficient to claim life detection. Even though the rules appear to be *ad hoc* rather than based on a formal theory, and the connection between the assignment of binary values and the domain knowledge is unclear, it was argued that alternative, formally better justified methods are too difficult to implement at this time because of their complexity and our inability to assign reliable probability values. Unfortunately, binary rules are inadequate, independently of the current state of knowledge.

For example, according to Neveu *et al.* (2018), finding a spectral signature of a pigment would be the ''smoking gun'' pointing to the presence of life, as no abiotic, uncatalyzed pigment synthesis is known. However, a possibility of catalyzed, abiotic chemistry that might take place in certain environments incapable of supporting life (such as, for example, Titan), or in specific microenvironments (such as micelles, microemulsions, or water droplets), is not considered. Even if pigments were synthesized only biologically, they still could be poor biosignatures. This would be the case if the likelihood of finding them in living systems present at a target environment were very low. Then, chances would be high that missions or observations would return negative results irrespective of whether life is present or absent. The binary rules do not leave room for such considerations.

To this end, an important advancement was made in the reports on strategies to search for signs of life on exoplanets (Catling *et al.,* 2018; Walker *et al.,* 2018). In these reports, the probabilistic nature of the evaluation problem was explicitly acknowledged by adopting the framework of Bayesian hypothesis testing. A similar approach has been used in many fields of science (see, *e.g.,* Cleophas and Zwinderman, 2018; Fenton and Neil, 2012; Nyberg, 2018) and has proven to be particularly successful if a sufficient amount of data is available. It could, however, lead to ambiguous results if data are scarce, as pointed out further in this paper.

Here, we take a broader perspective by appealing to decision theory, a mature field of science concerned with processes underlying human choices. The main feature of this theory is its universality, as it applies to decisions made in all application domains. From this perspective, evaluating and choosing biosignatures that are the most informative for life detection are not a unique problem but just another application of knowledge from the field of decision making.

The next two sections are devoted to normative decision making, which deals with how choices should be made. In Section 2, we consider two commonly used frameworks that are largely probabilistic in nature—Signal Detection Theory (Green and Swets, 1966; Wickens, 2002; Schonhoff and Giordano, 2006) and Bayesian hypothesis testing (Jeffreys, 1973; Kass and Raftery, 1995; Andraszewicz *et al.* 2015). In Section 3, a different but equally popular framework based on utility theory is described. All three frameworks can be used to evaluate biosignatures, but challenges faced in each case are somewhat different. The frameworks described in Section 2 require assigning probabilities that are associated with events of interest. In many instances, this can be done on the basis of statistical data. For example, medical diagnostics often takes advantage of the known frequencies with which given symptoms are found among sick and healthy patients. For unique or rare events, relevant statistical data are not available and instead probabilities or utilities (see Section 3) of different outcomes are assigned on the basis of the domain knowledge. This situation is often encountered, for example, in evaluating different ecological solutions (Dorazio and Johnson, 2003; Regan *et al*., 2005; Dee and Gerber, 2012). This is also the case with evaluating biosignatures. The role of domain knowledge and the means to organize it are discussed in Section 4. Methods for quantifying uncertainties in assigning probabilities are outlined in the Supplementary Information (SI), part C.

All decisions involve at some stage human judgment, which is often at variance with the assumptions of normative theories. We deal with this issue, which falls into the realm of psychological decision making, in Sections 5 and 6. It has been widely recognized that this aspect of the problem cannot be neglected (Bornstein and Emler, 2001; Elstein and Schwarz, 2002; Budescu *et al*., 2014). In fact, accounting for

human behavior beyond perfectly rational *homo economicus* is the primary reason for the rapid development of behavioral economics that led to Nobel Prizes in 1978, 2002, 2013, and 2017 to Herbert Simon, Daniel Kahneman, Robert Shiller, and Richard Thaler, respectively. Similar concerns motivate, for example, efforts to improve understanding climate change (Budescu *et al.*, 2014) and the development of decision-making software to guide medical diagnostics (Miller and Geissbuhler, 1999; Musen *et al.*, 2014). It is also likely to play an important role in evaluation of biosignatures.

The paper closes with conclusions. Instead of recommending a specific evaluation methodology, we summarize advantages and disadvantages of different strategies. In the SI, part B, we provide an example of how all steps of the evaluation process might be carried out.

Several important issues involved in the complete treatment of biosignatures for life detection fall outside the scope of this paper. We do not provide a list of biosignatures, nor do we identify specific evaluation criteria. Also, we do not address technical requirements and capabilities to detect biosignatures. In our discussion of evaluation strategies, we limit ourselves to the inherent features of biological or abiotic origin present at the target and do not consider the technology-related aspects of the problem, such as contaminants. Issues related to detection and contaminations can be considered in similar frameworks to those discussed below (Lorenz, 2019).

One might ask whether evaluating biosignatures for life detection is at all necessary. The answer to this question has to be positive. Evaluations for scientific or programmatic reasons are always made. The only choice is whether to carry them out *ad hoc* or base them on well-defined, scientific methods.

2. Probabilistic Frameworks for Evaluating Biosignatures

2.1. Signal Detection Theory

A conceptual framework to consider life detection is based on Signal Detection Theory (SDT), which provides a precise language for decision making under uncertain conditions. SDT has roots in both psychology and the military, as it was initially applied during World War II to interpret radar signals. Subsequently, it has found frequent applications in many fields ranging from medicine to telecommunications and artificial intelligence. In SDT, it is assumed that there is a stimulus that, if present, elicits response. For example, a disease causes a response in the form of specific symptoms that are recognized by a physician, prompting appropriate treatment. In the case of life detection, the presence of life (L) is the stimulus and the presence of a biosignature (B) is a response to this stimulus. If the relation between the stimulus and the response was unique, the stimulus would be always followed by the response and thereby correctly identified. In the absence of the stimulus, there would be no response. These two outcomes are called, respectively, ''true positives'' and ''true negatives.'' In most cases of practical interest, however, the response to the signal might be incorrect. Two distinct situations are possible. If there is response even though the stimulus is absent, the outcome is called false positive. If there is no response to the stimulus, the outcome is called false negative. In statistics, these two outcomes correspond to type I and type II errors, respectively. For example, the presence of nodules in the lungs is often used to diagnose lung cancer. However, healthy patients may also have nodules (false positives), and patients with cancer may have no nodules (false negatives). All four possible outcomes, two yielding correct results and two yielding incorrect results, are summarized in Table 1.

Signal Detection Theory (SDT) is most frequently used for repeated decision making. Table 1 is populated with the number of cases corresponding to each outcome. Subsequently, they can be converted to probabilities after proper normalization. In other words, past statistical data define the probabilities in the table. Application to unique events, such as life detection, for which no statistics are available is less common. In these cases, frequencies of different outcomes cannot be determined, and only the probabilistic representation is possible. Instead, probabilities are assigned on the basis of domain knowledge. How to carry out this assignment is the focus of the discussion that follows.

The probabilities in Table 1 are conditional probabilities. $P(B|L)$ abbreviates the probability that biosignature B is present under the condition that life is present, whereas $P(\sim B|L)$ is the probability that B is absent if life is present (the symbol $\sim$ abbreviates negation). These two probabilities are not independent. They correspond to outcomes that are complementary. If life is present, a biosignature is either present or absent; no other outcomes are possible. Therefore, their sum has to be equal to 1:

$$P(B|L) + P(\sim B|L) = 1 \tag{1}$$

In the second column, $P(B|\sim L)$ is the probability of finding biosignature B in the absence of life and $P(\sim B|\sim L)$ is the probability that B is absent when life is absent. In analogy to Eq. 1,

$$P(B|\sim L) + P(\sim B|\sim L) = 1 \tag{2}$$

Thus, only two probabilities in Table 1 are independent variables, one in each column. The remaining two probabilities can be obtained from Eqs. 1 and 2. Depending on the

Table 1. Relation between Stimulus (Life) and Response (Biosignature) in Signal Detection Theory

	Stimulus (Life) present	*Stimulus absent (no Life)*
Reaction (biosignature) present	True positive Biosignature present if there is life Probability $P(B\|L)$	False positive (error I) Biosignature present if no life Probability $P(B\|\sim L)$
Reaction (biosignature) absent	False Negative (error II) No biosignature if there is life Probability $P(\sim B\|L)$	True negative No biosignature if no life Probability $P(\sim B\|\sim L)$

specific focus, different pairs of conditional probabilities can be chosen to characterize outcomes. If the focus is on true and false discovery rates, then P(B|L) and P(B|∼L) are of direct interest. P(B|L) can be interpreted as ''signal'' and P(B|∼L) as ''noise'' that obscures this signal. If the focus is on correct outcomes, then the corresponding probabilities are P(B|L) and P(∼B|∼L), which respectively are called ''sensitivity'' and ''specificity'' in SDT. If one wants to characterize errors due to false negatives and false positives, then the appropriate pair of probabilities is P(∼B|L) and P(B|∼L).

If the probabilities of both false positives and false negatives were equal to zero, biosignature B would be always present if life were present but would be always absent if life were absent. Such B can be considered a perfect biosignature. In the language of SDT, B would be the ideal response to the stimulus, not burdened with errors. In practice, such biosignatures are unlikely to exist. All other biosignatures can be evaluated with respect to this ideal. An index developed for this purpose in SDT is called ''informedness'' or Youden's J statistics and is considered an unbiased accuracy measure (Ruopp *et al.,* 2008; Powers, 2011). It was originally introduced to assess performance of a diagnostic test for cancer (Youden, 1950). The informedness is defined as

$$J = P(B|L) + P(\sim B|\sim L) - 1 \tag{3}$$

which means that J is specified by a sum of sensitivity and specificity. It changes in the range between -1 and 1. Taking advantage of Eq. 2, this equation can be rewritten in terms of true and false positives:

$$J = P(B|L) - P(B|\sim L) \tag{4}$$

It reaches the maximum value of 1 when the probability of true positives, P(B|L), is equal to 1 and the probability of false positives, P(B|∼L), is zero. If J is equal to zero, the probabilities of true and false positives are equal. In this circumstance, J provides no useful information about the presence of life. In general, J is a measure of the distance from perfect biosignatures. The closer this index is to 1 the more likely it is that finding B means that life is present.

With the aid of Eq. 1, J can also be expressed in terms of false positives and false negatives:

$$J = 1 - [P(B|\sim L) + P(\sim B|L)] \tag{5}$$

The sum of false positives and false negatives measures the distance from perfect biosignatures.

In psychological literature on SDT (Swets, 1988; Lynn and Barrett, 2014), the distance between true positive and false positive, usually abbreviated as d', measures the ability of a decision maker to differentiate between signal and noise. In a frequency formalism, both signal and noise are represented as distributions, and d' is the distance between the peaks of these distributions. As d' increases, so does our confidence that we correctly distinguish the signal from the noise. In contrast, no distribution is available in the probabilistic formulation for single or rare events, as considered here, and the equivalent of d' is J calculated as a difference between appropriate probabilities (see Eq. 4).

In Eq. 5, false positives and false negatives are taken with equal weights. However, depending on specific goals, it might be desirable to place more emphasis on either false positives or false negatives. To do so, we can introduce parameter α ($0 \le \alpha \le 1$) and redefine J such that it is still expressed in terms of the same probabilities and remains in the range [-1,1]:

$$J = 1 - [\alpha P(B|\sim L) + (1-\alpha)P(\sim B|L)] \tag{6}$$

Large values of α (close to 1) correspond to avoiding false positives whereas small values (close to zero) are used if the goal is to avoid false negatives. Thus, α can be considered as a ''mission objectives parameter.'' If several missions to a given target are considered, the goal of the initial mission might be to establish whether this target is worth further exploration in search for life. Then, it might be desirable to set α to a small value, not to overlook possible, though not definitive, signs of life. To some extent, this strategy is being pursued in Mars exploration. In contrast, for a single, high-profile mission aimed at life detection at a distant target, for example, Europa, it would be prudent to focus on avoiding false positives and, therefore, set α to a high value. It is, however, not advisable to use extreme values of α. For example, if α is set to 1 to avert false positives, the probability of false negatives is ignored. Since probabilities of false positives and false negatives are often correlated, concentrating on biosignatures that provide a definitive evidence for life may greatly increase probabilities of false negatives, that is, chances that the mission will return null results even if life exists at the target.

An equation that is nearly identical to Eq. 6 can be derived if we assign utilities for detecting life to each probability in Table 1 and calculate the expected global utility. This is done in the SI, part A. Probabilities of true positives and true negatives will have positive utilities, whereas probabilities of false positives and false negatives will be associated with negative utilities. All these utilities will be, in general, different.

In psychological literature, different weighing of false positives and false negatives that reflects goals of a decision maker is called response bias or decision criterion and arises from the agent's motivation. For example, in law an agent wants to avoid false positives whereas in medicine the emphasis is on avoiding false negatives. The response bias is measured as the ratio of true positives to false positives, that is, a signal/noise ratio, and is usually abbreviated β (Stanislaw and Todorov, 1999). Since it allows for assigning different weight to costs of false positives and false negatives, it plays the same role as α in the formulation outlined above (see Eq. 6). Depending on the goals of a decision maker, it might explain over- or under-weighting different outcomes, as proposed by Lopes and Oden (1999), who formulated the choices in terms of security and potential. For example, if true positive outcomes are favorable, people focused on security will avoid false positives whereas those focused on potential will concentrate on reducing false negatives. For review of interdependences between decisional weights and outcomes see Weber (1994) and Weber and Kirsner (1996).

In probabilistic formulations of SDT outside psychology, a measure of signal/noise ratio, K(B), defined as:

$$K(B) = \frac{P(B|L)}{P(B|\sim L)} \quad (7)$$

can be used, along with J, to ascertain utility of a biosignature for detecting life. Although J and K(B) are different, they are monotonically related, which means that they will yield the same ranking of biosignatures. An interesting feature of K(B) is that it connects SDT with the Bayesian hypothesis testing formalism, as described below.

2.2. Bayesian hypothesis testing

In Bayesian hypothesis testing, the quantity to be estimated is the probability P(H|D), called posterior probability, that a hypothesis H is true given data D. In life detection, the hypothesis is that life is present, and the data are biosignatures. Thus, P(H|D) is the conditional probability P(L|B) that life is present if biosignature B is present and has already been found. This is different from the probability P(B|L) of finding B if life is present, which is sought in SDT. On this basis, it might appear that Bayesian hypothesis testing and SDT would lead to different evaluations of biosignatures. As shown below, this is not necessarily the case. According to Bayes' theorem

$$P(L|B) = \frac{P(B|L)}{P(B)}\, P(L) \quad (8)$$

where P(B) is the probability that B is present from either biological or abiotic sources, and P(L) is the prior probability that reflects one's belief about the presence of life before evidence in the form of biosignature B is collected. Bayes' theorem follows directly from the chain rule of probability calculus. This rule connects joint probability, P(L,B), that both life and biosignature B are present with the conditional probabilities P(L|B) or P(B|L).

$$P(L, B) = P(L|B)P(B) = P(B|L)P(L) \quad (9)$$

Bayes' theorem follows immediately from the second equality.

P(L|B) can be compared with the probability, $P(\sim L|B)$, of the alternative hypothesis: there is no life if B is present. In other words, B comes entirely from abiotic sources or contaminations. Analogously to Eq. 8, this probability can be expressed as

$$\frac{P(\sim L|B) = P(B|\sim L)}{P(B)\, P(\sim L)} \quad (10)$$

where $P(\sim L)$ is the prior probability that there is no life. Of course, P(L) and $P(\sim L)$ sum to 1. The ratio, R_{LB}, of the posterior probabilities in Eqs. 8 and 10

$$R_{LB} = \frac{P(L|B)}{P(\sim L|B)} = \frac{P(B|L)}{P(B|\sim L)}\, \frac{P(L)}{P(\sim L)} = K(B)\frac{P(L)}{P(\sim L)} \quad (11)$$

is the relative probability of the hypothesis that life is present compared to the hypothesis that life is absent, both evaluated assuming the presence of biosignature B. In the last equality, we take advantage of the definition in Eq. 7. In the Bayesian framework, K(B) is called Bayes factor. If R_{LB} is less than 1, it means that the presence of B does not give us high confidence in the presence of life, as B is likely to be of abiotic origin. If R_{LB} is larger than 1 it is more likely than not that life is present if B is present. In statistics, a markedly stronger evidence is usually expected to claim confidently that a hypothesis is confirmed. It would be typical to require that $R_{LB} \geq 20$.

An inconvenient feature of Eq. 11 is that it depends not only on the likelihood ratio but also on prior belief about the probability that life is present at the target. The assignment of this probability is subjective and might differ widely even between experts, leading to the correspondingly different evaluation of R_{LB}. If ample data were available, evidence captured in P(B|L) would be sufficiently strong to overcome prior belief even if it disagreed with evidence. In life detection, obtaining such data, especially in a single mission, cannot be expected.

Fortunately, the difficulties due to the influence of the prior disappear when two biosignatures are compared. Consider another biosignature, B'. For this biosignature, one can define the ratio, R_{LB}, in exactly the same fashion as it is done for B in Eq. 11.

$$R_{LB'} = K(B')\, \frac{P(L)}{P(\sim L)} \quad (12)$$

Then, the ratio

$$\frac{R_{LB}}{R_{LB'}} = \frac{K(B)}{K(B')} \quad (13)$$

is independent of the prior and depends only on the Bayes factors for B and B'. If K(B) is larger than K(B'), then B can be considered a better diagnostic biosignature than B'. We would reach the same conclusion if we used Eq. 7 as the evaluation index in SDT. Thus, the ranking of biosignatures obtained from SDT and Bayesian hypothesis testing should be the same.

2.3. Process of assigning probabilities

In many instances, probabilities defined in SDT, for example, P(B|L) and $P(B|\sim L)$, might be difficult to estimate because the ability to observe a biosignature if life is present or absent does not arise from a single process but instead is an aggregate property that depends on a series of processes. Not only must a biosignature be produced in a given context, but also it has to, for example, survive degradation and be detected. It might therefore be beneficial to construct a model that represents all processes leading to different possible outcomes. Each constituent process is associated with a probability. These probabilities are more elemental than the probabilities in SDT and, therefore, might be easier to estimate. Once all of them are assigned, the probabilities required in SDT can be evaluated from the model with the aid of the standard probability calculus.

A number of approaches can be used to represent the connection between the existence of life in a given environment and the possibility of observing a biosignature. One of them is to create a decision tree, a tool commonly used in decision analysis (Koning and Smith, 2017; Sullivan, 2018).

Decision trees are flowchart-like diagrams in which each node represents an outcome of a constituent process that occurs with a certain probability. A decision tree usually starts with a single node. Each node branches to two or more nodes, depending on the number of outcomes of a process under consideration. A simple example of a decision tree that involves only the production and survival of a biosignature generated by biological and abiotic means is shown in Fig. 1. Although this model is not aimed at providing the actual, complete description of a complex process of biosignature detection, it is sufficient to illuminate several key features of a properly defined life-detection model. As seen in Fig. 1, such a model consists of two main branches representing, respectively, possible biological and nonbiological origins of a biosignature. If only the biological branch were included in the model, estimates of false positives or ''noise'' would not be available, and no meaningful index to measure utility of a biosignature could be constructed. The model in Fig. 1 can be readily generalized. For example, to include terrestrial contamination the right branch would be labeled ''indigenous biological'' whereas the left branch would split to ''abiotic'' and ''terrestrial biological,'' both contributing to false positives.

It is highly desirable that the decision tree be universal for all biosignatures. This influences resolution of a model, as a detailed model for one biosignature might not be appropriate for another biosignature. If different models, potentially containing different number of levels, were created and evaluated for different biosignatures, it would not only complicate the structure of the problem but also would make outcomes prone to biases, as discussed in Section 5, and in 5.2.1 in particular. Additional information about decision trees and alternative approaches is available in the SI, part D.

3. Frameworks Based on Utility Theory

Utility theory, which has deep conceptual and methodological roots, provides approaches to the problem of choice that are alternative to probabilistic methods. In these approaches, choices reflect utilities of available options. The concept of utility was introduced by Bernoulli in 1738 (Bernoulli, 1954), who observed that people considered subjective rather than objective values of outcomes. In his Expected Utility (*EU*) model, utility referred to monetary outcomes, that is, payoffs in bets. Later, however, it was broadly recognized that assigning utilities is a general feature of human perception that applies to different stimuli, activities, situations, or outcomes. Here, options are biosignatures; and each of them, once discovered, has utility for detecting life beyond Earth.

The philosophical basis for utility theory lies in utilitarianism formulated by Bentham and Mill. According to them, rational behavior means maximizing utility. Modern, axiomatic versions of utility theory, especially under risk, are largely based on the work of von Neumann and Morgenstern (1947). Four axioms about preferences are usually accepted.

(1) Completeness: The decision maker is always able to define preferences for any two options, A and B. A is preferred to B, B is preferred to A, or the decision maker is indifferent between A and B.
(2) Transitivity: If A is preferred over B and B is preferred over C then A is preferred over C.
(3) Independence: The order of preferences between two options does not change if each of them is subjected to the same linear transformation. It means that if A is preferred to B then $A+C$ is preferred to $B+C$, when C is an additional, third option. Similarly, αA is preferred over αB, where α is a constant.
(4) Continuity: If A is preferred to B and B is preferred to C then there exists a number p, $0 < p < 1$, such that $B = pA + (1 - p)C$.

Although these axioms are quite reasonable, perhaps even obvious, from the mathematical point of view, they are not always preserved in human judgments. Several consequences of deviations from the axiomatic theory will be discussed in Section 5.1.

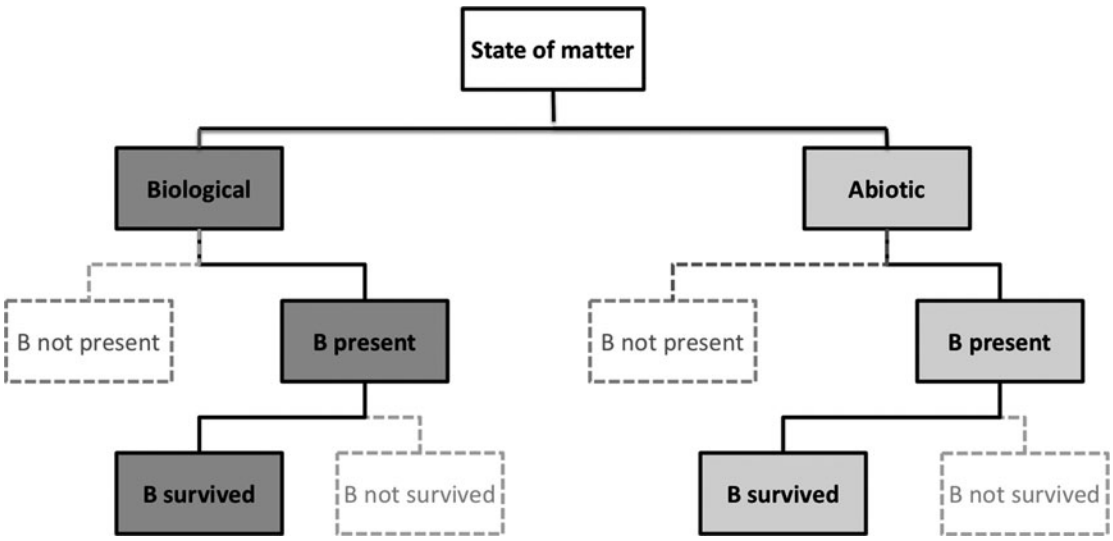

FIG. 1. Schematic of a ''life-detection decision tree'' with two levels (existence and survival). The tree was terminated when biological and abiotic sources of biosignatures become indistinguishable. Bold lines and grey boxes lead to outcomes that are true positives (left side) or false positives (right side).

3.1. Multi-attribute decision approach

The framework broadly applied to decision making in many areas, ranging from policy, business and finance decision making to medicine and consumer choices, is based on Multi-Attribute Utility (*MAU*) approach. This approach is also suitable for evaluating biosignatures. In line with the independence axiom, each option, which in the context of life detection is a biosignature, is evaluated in *MAU* independently of all other options. Each option is evaluated on attributes or criteria related to their utility. For example, a person buying a car might consider such criteria as price, safety, repair record, fuel economy, maintenance costs, performance, comfort, and design esthetics. In life detection, examples of relevant criteria to judge a biosignature might be the likelihood of finding it in biological systems, the ability to survive in the target environment in quantities sufficient for detection, and the existence of abiotic processes producing the biosignature.

To evaluate the overall utility of each option, one has to account for all relevant criteria. Otherwise, important aspects are not included in calculating the overall utility, potentially leading to significant inaccuracies. The criteria must also be disjoint. If the criteria overlap, some aspects are counted more than once, most likely with detrimental effects to the accuracy of the overall utility. To ensure the proper comparison of overall utilities of options, all of them should be evaluated on the same set of criteria.

In *MAU,* each criterion is scored independently of other criteria. Otherwise, the evaluation may be biased even though all relevant criteria that are disjoint and the same for all options are considered. This means, for example, that the likelihood of finding a biosignature in biological systems should be evaluated independently of the existence of abiotic processes producing this biosignature. Similarly, degradation of a biosignature in an aqueous environment due to hydrolysis should be evaluated independently of degradation due to radiation if both modes of degradation were previously selected as evaluation criteria. If utility of a score on one criterion depends on scores on other criteria, one cannot construct a monotonic scale for criteria.

To obtain overall utility, scores on different criteria are combined additively. Since different criteria may be of different importance to the decision maker, that is, influence choices to a different degree, scores on criteria are multiplied by weights assigned to criteria. The weights sum to 1. This leads to the following formula for the overall utility, *U(B)*, of a biosignature (option) *B:*

$$U(B) = w_1 * u_1 + w_2 * u_2 + \ldots + w_n * u_n \qquad (14)$$

where indices 1, …, n count criteria, w_i is the weight assigned to criterion *i,* and u_i is the score of B on this criterion.

In summary, in accord with Eq. 14, evaluation of *U(B)* consists of three steps. First, a set of appropriate criteria is defined. Next, each criterion is assigned a weight. Weights do not depend on specific biosignatures. The third step is to score each biosignature on each criterion. Scores can be positive or negative, depending on whether they describe desired or undesired attributes. For example, criteria related to abiotic sources of biosignatures reduce utility; therefore, they should be associated with negative scores. This provides a set of weighted u_i which allows for evaluating the global utility *U(B)* from Eq. 14.

It is also possible to formulate *MAU* such that utilities for individual criteria are combined multiplicatively rather than additively (Keeney and Raiffa, 1993). Empirical evidence as to which of these two approaches is more accurate is mixed (Jansen, 2011).

The defining characteristics of *MAU* are that evaluations of options are independent, global and compensatory. This means that all criteria have to be considered in evaluating the overall utility, and a poor score on one criterion can be compensated with high scores on other criteria. Even though *MAU* is a method of choice in many decision-making problems, sometimes its characteristics lead to outcomes that are not optimal. Other utility-based models exist that do not adopt some of these characteristics (Payne, 1976; Svenson, 1979).

Specifically, the compensatory nature of *MAU* causes concerns because evaluation criteria are often conflicting. This is the case for cost and quality, cost and safety, or maximizing return and minimizing risk of an investment. For example, a car may have excellent fuel economy but poor performance, or a low price but a questionable safety record. In the medical field, conflicting criteria might be diagnostic accuracy and availability of care. Azar (2000) discussed the application of *MAU* to evaluating four techniques for detecting breast cancer. Their features were found to be in conflict. Magnetic resonance gives accurate, high-resolution images but is costly, whereas mammography is inexpensive but markedly less accurate. In the context of life detection, a biosignature could be highly diagnostic of life but degrade rapidly and therefore score poorly on the survivability criterion. Conflicting criteria might lead to a situation in which the option with the best overall utility scores highly on most criteria but poorly on a few. Such a solution might not be the desired one if a minimum level of performance is required for every criterion. Examples abound of natural or engineered systems that underperform or even fail because just one element functions poorly.

3.2. Conjunction rule

A decision strategy that does not suffer from the potential disadvantage of compensation is based on the Conjunction Rule, which derives from the Principle of Satisficing proposed by Simon (Simon, 1957; Svenson, 1979). In this model of ''rational'' choice, only options that meet aspiration level are considered. First, the minimum score that is considered satisfactory is determined for each criterion. Then, all options are treated separately, one at a time. Any option that does not meet the minimum cutoffs for all criteria is rejected. All options that pass the cutoff test are considered satisfactory, and no further optimization is carried out. In practice it means that the first such option should be chosen. A typical situation to which the strategy applies is when choosing a book to read on vacation. Searching the bookstore for a book with the highest utility would be unnecessary and wasteful.

As a single decision strategy, the Conjunction Rule is clearly not useful for life detection because the goal is to identify the most informative set of biosignatures. It may, however, profitably complement *MAU*. First, the Conjunction Rule is applied to eliminate all biosignatures that do not

meet minimum requirements for all criteria. Then, the remaining biosignatures are evaluated by way of *MAU*.

4. Role of Knowledge in Evaluating Probabilities and Utilities

Evaluation of probabilities or utilities should be based on the complete, current knowledge. This knowledge consists of information and evidence bearing on biological or abiotic sources of biosignatures. It is drawn from a wide range of fields such as organic chemistry, cellular and molecular biology, ecology, biogeochemistry, planetary science, and astronomy. The inherently interdisciplinary, highly diverse nature of knowledge, exceeding expertise of a single scientist, necessitates creating a knowledge base (*KB*) that forms the common, factual basis for the evaluation process. Such a knowledge base should be structured to facilitate efficient and reliable evaluation. In particular, arguments supporting or contradicting the value of a given biosignature as evidence for life should be explicitly stated and organized in groups based on evaluation criteria, as this would assist researchers in comparing different arguments, even if they were derived from fields outside their main expertise. This organization is markedly more useful to researchers than outcomes of literature searches based on key words or other, similar criteria.

Another desired feature of the *KB* is that it should be community based. This means that information can be contributed by all registered members of the community. In contrast to encyclopedic knowledge, the providers of information do not have to be objective. Even if a specific researcher supplies facts that are biased for or against a given biosignature, the overall body of evidence from the scientific community should provide a balanced viewpoint. Curation would additionally protect the integrity of information captured in the *KB*.

A structure that fulfills most of the requirements for the *KB* already exists and has been implemented in the Hypothesis Browser for Astrobiology (Pohorille and Keller, 2010). Below, we briefly outline its main features and describe how a modified version of this structure can be used for life detection. In close analogy to the Hypothesis Browser, scientific information of interest is organized via arguments supporting or contradicting a statement that presence of a given biosignature provides evidence for the presence of life. For example, the argument about near universality of homochirality in terrestrial biology supports treating this feature as a useful biosignature whereas the existence of natural or synthetic structures of mixed chirality provides an opposing argument. The corresponding ontology of the *KB* consists of five concepts: (i) biosignatures, (ii) criteria, (iii) arguments, (iv) evidence, and (v) research source. The relationship between these concepts is shown in Fig. 2. Arguments are grouped according to criteria they address. For example, arguments dealing with the universality of homochirality in biology are related to the presence of this biosignature whereas arguments about racemization rates address its survival. The discovery of nearly homochiral species in abiotic environments provides an argument for assessing the probability of false positives for this biosignature. Each argument is supported by evidence presented in or inferred from scientific literature, which constitutes the primary research source. This ensures that provenance of all knowledge captured in the *KB* is known.

The likelihood of detecting a biosignature depends not only on its intrinsic features but also on the environment. For example, finding a pigment of biological origin is more probable on Mars than in the depth of the ocean world of

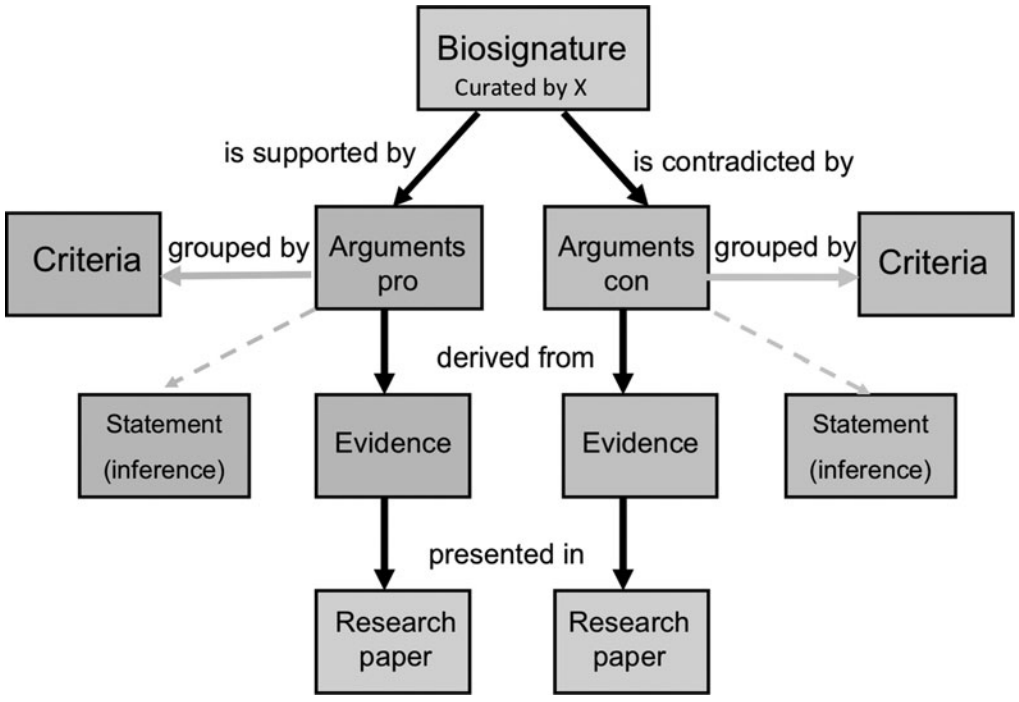

FIG. 2. The ontology of the biosignature knowledge base.

Europa, where a significant amount of light is unavailable. Similarly, degradation pathways and rates of delivery of organic material are strongly dependent on the environment. For this reason, arguments in the *KB* should contain information about environments to which they apply. In fact, the *KB* can be constructed as a two-dimensional biosignature× environment matrix that can be searched in both directions.

5. Elicitation Probabilities and Utilities

To make decisions or evaluations, information about a subject matter has to be translated to the global probabilities listed in Table 1 and/or global utilities. This process is called elicitation. Since statistical data for life detection are not available, elicitation is based on assessing domain knowledge, which means that it has to involve, at some stage, human judgment. This raises a question: At which stage of elicitation should human judgment be involved to ensure the most reliable results? As has already been mentioned in Section 2.3, global probabilities might be difficult to evaluate directly. Instead, it might be simpler and more reliable to evaluate probabilities for individual criteria and then obtain global probabilities. This is analogous to scoring biosignatures on each criterion, which is a required step in utility-based approaches.

Another possibility is to push human involvement to an even more elemental level and ask experts to judge supporting and contradicting arguments collected in the *KB* rather than probabilities or utilities. Arguments should be evaluated on two dimensions—significance and strength. These two traits are separate. Significance is the extent to which an argument, if proven, informs elicitation. Strength measures how well an argument is supported by evidence. For example, the argument that any biological system should contain polyionic molecules as information carriers (Benner, 2017) is highly significant, but one might have doubts whether it has been proven sufficiently strongly. Conversely, the existence of molecules with mixed chirality, such as gramicidin, in biological systems is undoubtedly true, but the significance of this argument for contradicting homochirality of life is uncertain, as known molecules of mixed chirality perform only a very limited range of functions in terrestrial biology. Evaluating arguments has a number of advantages. It is easier to carry out than evaluating probabilities or utilities. Also, it forces experts to judge only the knowledge in the *KB* and leaves less room for unsupported opinions. Finally, this approach is compatible with dynamic evaluation, that is, evaluation that can be readily updated as new facts become available. Its main disadvantage is that it requires an additional elicitation step of integrating evaluations of different arguments and converting them to probabilities or utilities. The rules for doing this can be developed, for example in the spirit of normalized Eq. 14, but their veracity has to be carefully tested.

An important, complicating factor is that human judgment involved in elicitation is always prone to perceptual and cognitive biases. This applies not only to laypersons but also to experts, even though the latter should be protected from such biases. A number of findings document that experts often make judgments based on stereotypes, images (Cooper *et al.,* 2001; Cornell, 2001; Shefrin, 2001), or affect (Savadori *et al.,* 2004; Slovic *et al.,* 1999, 2004). For example, Slovic *et al.* (1999) asked members of the British Toxicological Society to rate 30 chemicals on an affect scale (*bad-good*) and to judge risks associated with the exposure to these chemicals at levels that were markedly below the standards accepted by the regulatory agency. Even though these risks were negligible, they were evaluated as high, but only for the chemicals rated unfavorably on the affect scale. Sometimes, expert knowledge may reinforce rather than reduce biased evaluations because of personal or group interests. Savadori *et al.* (2004) found that perception of risks and benefits from biotechnology was more biased among experts than among non-experts, most likely because of their personal involvement.

Since biases in judgments are inevitable, understanding their sources is essential for assessing and reducing their impact on evaluations. Although biases are of serious concern in other fields (Garthwaite *et al.*, 2005), they have never been discussed in the context of life detection. For this reason, general mechanisms of cognitive biases are briefly discussed, and biases that are most relevant to life detection are described. Some of them apply to both utilities and probabilities, whereas others affect only probabilities. All topics are illustrated with representative experiments in cognitive psychology and relevant situations that might arise during evaluating signs of life.

5.1. Independence axioms and comparative judgment

The chief concern with applying utility and probability theories, discussed in the previous sections, to processes that involve human judgment is related to the axioms that require each option and each criterion to be evaluated independently from each other (see Section 3). In variance with these axioms, it is broadly acknowledged that human judgment is comparative (*e.g.,* Russo and Dosher, 1983; Mellers and Biagini, 1994; Noguchi and Stewart, 2014). This applies to any kind of sensation, such as visual perception, sound, or taste. All judgments are made relative to a reference point. For physical stimuli, the most common reference point is the lack of a stimulus, for example, darkness, when changes in brightness are evaluated. The most general principle of human sensation is diminishing sensitivity to changes in stimulus with the increasing distance from the reference point. It is easier to differentiate between coffee with no sugar and coffee with one teaspoon of sugar than between coffee with 5 and 6 teaspoons of sugar, even though the objective difference in sugar added to the coffee is the same in both cases. In psychology, this is called the Weber-Fechner law. In economics, this is called marginal utility.

Comparative judgments also apply to cognitive evaluations. In these cases, changes, most often evaluated with respect to a status quo or another reference point, are also subject to diminishing sensitivity. Weber (2018) illustrates this in an example in which evaluation of risk defined as variability of outcomes depends on the expected value of a risky option. Risk of a lottery in which one can win $150 or lose $50 is perceived as large. However, the same variability is perceived as small when the expected value of a lottery is 1 million dollars.

A number of common biases stem from comparative evaluations, such as taking stereotypes as reference points.

For example, experts in life detection might consider the ubiquity of a biosignature in terrestrial life instead of the likelihood that it is produced in any biological system. Another common problem is that the assessment of utility of an object depends on the context. For example, the assigned utility of a given biosignature may depend on other biosignatures included in the set presented for evaluation. The psychophysical and cognitive mechanisms of comparative judgments related to reference points, stereotypes, and context are discussed below.

5.1.1. Comparisons with reference points. Diminishing sensitivity with respect to the reference point is a frequent source of biases. If utilities are measured with respect to the same reference point, the evaluated utility of an objective value X is smaller than the sum of utilities of objective values $x_1 \ldots x_n$ that contribute to X: $u(X) < u(x_1) + \ldots + u(x_n)$. This means that the utility of one strong argument in favor of biosignature A with strength of 8 on a scale from 0 to 10 will be evaluated lower than the overall utility of 4 weak arguments in favor of biosignature B, each having strength 2.

Another consequence of diminishing sensitivity is related to overestimating the significance of the first evidence contradicting a commonly accepted view, which typically serves as the reference point. For example, for a long time it was accepted that chiral molecules synthesized abiotically in space existed as nearly perfect racemic mixtures. Significant excess of one enantiomer was interpreted as evidence of biological activity. The discovery of amino acids of abiotic origin in meteorites markedly biased toward one optical isomer (Pizzarello *et al.,* 2012) is assigned high value as evidence contradicting biological origins of homochirality in general, even though the underlying mechanism might be highly specific to a narrow range of meteorites (Burton and Berger, 2018). A similar bias is also observed in evaluating probabilities. When no good medical treatment for HIV was available, initial, poorly effective drugs were perceived as greatly increasing the probability of survival, because the change from no hope, which was the reference point at that time, to a very small probability of survival was perceived as very salient. On this basis, it might be expected that finding an abiotic route to synthesis of a biosignature believed to be produced only by biological means would lead to overestimating the probability of its abiotic origins, even if this route were rare in planetary or space environments.

5.1.2. Comparisons to stereotypes. In contrast to the phenomenon of diminishing sensitivity, which is rooted in psychophysics, comparisons to stereotypes lead to biases that have their source in cognition. Probably the best-known example of such bias is from an experiment by Tversky and Kahneman (1983) in which respondents were given the following description of a young woman: ''Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.'' Next, participants answered which was more probable: (1) Linda was a bank teller or (2) Linda was a bank teller active in the feminist movement. More than 80% chose the latter option, regardless of whether they were novice, intermediate, or expert statisticians, even though a population defined in this option is a subset of a population defined in option 1 and, therefore, has to be less probable. This is called the conjunction fallacy and follows from judging similarity of Linda to stereotypes of a bank teller and a feminist instead of assessing probability. People compare salient features of objects and evaluate ''the degree of correspondence… between an outcome and a model'' (Tversky and Kahneman, 1983, p 295).

Such correspondence is usually evaluated on the basis of similarity and/or diagnosticity, which is defined as essential features/criteria shared by all members of a given class (Rosch, 1975; Tversky, 1977). An object is representative of a class either if it is a prototype or a typical member irrespective of its frequency in the population, for example, a robin is more representative of birds than a chicken, although a robin is less frequent (Rosch, 1978/1988).

A similar situation might arise in evaluating probability of the biological origin of biosignatures, which might be done on the basis of their similarity to and diagnosticity for terrestrial life. For example, nucleic acids are frequently perceived as having higher diagnosticity than polyionic polymers in general, since they serve as universal genetic polymers for life on Earth. This is the case, even though other polyionic polymers are capable of playing the same role (Eschenmoser, 1999) and might have preceded nucleic acids on early Earth (Hud *et al.,* 2013) or be favored under different space conditions. If the probability of finding nucleic acids in extraterrestrial life were evaluated higher than the probability of finding polyionic polymers, which might be an outcome in some expert opinions, this would be an example of conjunction fallacy. Another example of conjunction fallacy was observed in our pilot study described in the SI, part B, empirically demonstrating relevance of this bias to life detection.

5.1.3. Biases related to choice-set (context) dependence. Often, utilities or probabilities of biosignatures are not evaluated independently but instead depend on other biosignatures included in the evaluation set. The ''effect of decoys'' in a choice-set illustrates this phenomenon (*e.g.,* Ariely and Wallsten, 1995). Readers of the *Economist* were given a choice between (A) a 1-year online subscription for $59 and (B) a 1-year subscription to the print edition with online access for $125. Forty percent chose A, and 60% chose B, indicating that B was, on average, slightly more attractive than A. However, when 1-year subscription to the print edition without online access for $125 was added as the third option, 80% chose B and only 20% chose A. Thus, the attractiveness of B increased when a decoy option was added to the choice set.

Another example of choice-set dependence is called asymmetric dominance (*e.g.,* Simonson, 1989). Assume that there are two biosignatures, A and B, evaluated on two criteria, and A is slightly better on the first criterion, whereas B is slightly better on the second criterion. Then, the perceived difference in their utility is expected to be small. Assume further that biosignature C is added to the choice set. C is worse than both A and B on the first criterion but equally good as B and better than A on the second one. Then B is never worse than C, but the same is not true for A (see Table 2, set 1). In other words, B dominates C in this set whereas A does not. As a result, the perceived utility of B is found to increase compared to A. Conversely, adding

Table 2. An Example of Asymmetric Dominance in Evaluation of Biosignatures

	Biosignature	*Rates on Dimension 1*	*Rates on Dimension 2*
Set 1	A	6	4
	B	4	6
	C	2	6
Set 2	A	6	4
	B	4	6
	D	6	2

biosignature D that is worse than A and B on the second criterion and equally good as A on the first one will increase the perceived utility of A (see Table 2, set 2). In this set, D is dominated by A but not by B.

Similar biases were also documented in evaluating probabilities. Windschitl and collaborators (Windschitl and Wells, 1998; Windschitl and Young, 2001; Windschitl *et al.,* 2002) found that people judged probability of a focal outcome not in absolute terms but rather in comparison to the most likely alternative outcomes. For example, respondents evaluated the probability of winning in a raffle in which they held 17 tickets while others held 8, 8, 9, 8, and 9 tickets higher than the probability of winning in a raffle in which others held 7, 6, 16, 6, and 7 tickets (Windschitl and Young, 2001), even though their objective chances were, respectively, 17/58 and 17/53. The biased evaluation resulted from the comparison of the number of own tickets with the highest number of tickets held by rivals instead of calculating the probability of winning. In the first raffle, no rival held more than 9 tickets, whereas in the second raffle one person held 16 tickets. The probability of a focal outcome is judged higher when it is favored in comparison with the strongest alternative. In life detection, judgment of the probability that a biosignature was created in a biological process (focal outcome) might depend on the distribution of arguments in favor of its abiotic origin (residual outcome). Even though arguments may produce the same total probability of abiotic origins of several different biosignatures, these probabilities may be evaluated quite differently if the distribution of argument strengths differs.

5.2. Specific biases in elicitation of probabilities

5.2.1. Subadditivity. A common bias in probability judgment is that the probability of the whole is evaluated lower than the sum of probabilities of its independent parts. This phenomenon is called subadditivity. It is similar to diminishing sensitivity, but the underlying mechanism is different.

Subadditivity was first documented by Fischhoff, Slovic, and Lichtenstein (1978). Car mechanics with an average of 15 years of experience were asked to evaluate the probabilities of different causes of a failure to start a car. Initially, the mechanics evaluated the probability that the reason ''is something other than the battery, the fuel system, or the engine'' as 0.22. When the question was divided into several more specific causes (*e.g.,* the ignition system), this probability increased to 0.44. Similarly, experts asked about degradation of a given biosignature in general are expected to provide lower probability estimates than when asked about degradation due to a number of specific reasons, such as radiation, hydrolysis, pyrolysis, and so on.

This phenomenon was initially interpreted in terms of the availability or salience effect: ''what is out of sight is also out of mind'' (Fischhoff *et al.,* 1978, p 333). Fox and colleagues (Fox and Rottenstreich, 2003; Fox and Clemen, 2005) proposed a different explanation of subadditivity. According to them, people initially assign equal probabilities to all events, what is called the ignorance prior. Then they adjust these probabilities according to their beliefs of how the likelihood of these events differs. Bias arises because the adjustment is usually insufficient. This interpretation is particularly relevant to evaluating poorly constrained probabilities with high uncertainty, which is often the case with biosignatures.

5.2.2. Anchoring and other factors that impact focus of attention. The interpretation given by Fox and Clemen (2005) is related to another bias called anchoring and adjustment heuristic (Tversky and Kahneman, 1974). According to this heuristic, people focus attention on the initial values called anchors and evaluate probabilities in relation to them. Regardless of the source of the starting value, the adjustment is usually too small (Slovic, 1972). For example, experts asked about the probability of survival of a biosignature on Europa, given that the probability of survival on Earth is p, are likely to anchor their attention on p, providing estimates close to this value.

Anchoring applies even when explicit numbers are not given. When asked whether biosignature A is more likely to survive than biosignature B under given conditions, experts mostly look for arguments supporting survival of A. In contrast, when asked whether A is less likely to survive, the same experts will tend to look for arguments contradicting survival of A. This has implication for SDT. According to anchoring and adjustment heuristic, asking questions about true positives or false negatives and about false positives or true negatives is not equivalent, even though it should be, as each pair of probabilities adds to unity.

Similar biases may result from primacy and recency effects because of a stronger focus of attention and better recall of information presented either at the very beginning (primacy effect) or at the very end (recency effect) at the expense of information given in the middle. For example, if both the first and the last argument support biological origin of a biosignature, experts will most likely overestimate the corresponding probability. However, if the order of arguments is changed such as the first and the last arguments support abiotic origins of this biosignature, the probability of its biological origin will be underestimated.

5.3. Advantages and disadvantages of different measurement tools in light of biases

As discussed earlier, human judgments are not in agreement with the independence axioms. There are two ways to deal with this disagreement: either independence is enforced by way of applying noncomparative measurement scales or the independence axioms are relaxed, and measurements are done with the aid of comparative techniques. How both approaches are implemented is discussed in the SI, part E.

Noncomparative scales are in agreement with logical conditions, such as transitivity, but do not reflect the way in which people make judgments. Forced to make noncomparative judgments, respondents frequently use reference points or stereotypes that are unknown to those who collect data. Comparative scales do not require external reference points and are free from numbers and labels. Such scales also have other advantages. They are easy to understand and use and allow for capturing even small differences between evaluated objects. This comes at a price of ordinal instead of interval measurement. To convert ordinal to interval measurement, as required in SDT or utility theory, procedures of uncertain accuracy have to be used.

6. Conclusions

Evaluating biosignatures for their information value is an essential step in developing research, mission, and observational strategies for detecting life beyond Earth. In this paper, we present the key aspects of this process that are firmly based in normative and descriptive decision theory. A number of possibilities exist to carry out evaluation at both formal and practical level. We do not endorse any of them but instead discuss their advantages and disadvantages.

Three formal frameworks broadly used in other application domains based on SDT, Bayesian hypothesis testing, and utility theory have been presented. The SDT framework has some advantages over the Bayesian approach. Most importantly, it does not rely on prior probabilities of finding life that frequently are poorly constrained. Since the assignment of priors, P(L), at different targets is based not only on limited, consensus knowledge but also on subjective opinions, it is likely to differ significantly between experts. This difficulty has been appreciated in the context of exoplanets. As stated by Kiang *et al.* (2018), ''P(life) is truly quantifiable only with large statistics, after more examples of life have already been discovered.'' Since this is unlikely to happen in the near future, it is difficult to take advantage of the most attractive feature of Bayesian statistics—a systematic way to update posterior probabilities in the light of new data. Interestingly, as shown in Eq. 13, prior probabilities disappear for comparative evaluations of biosignatures at the same target, and only Bayes factors remain relevant.

Another advantage of SDT is the ability to incorporate goals in the process of evaluation, for example through asymmetric weighing component probabilities to emphasize avoiding false positives or false negatives. This feature is one of the main reasons why SDT has been widely applied in engineering, medicine, and psychology. Including goals, a feature that SDT shares with the utility framework in which scores are weighed by relative weights assigned by a decision maker, reflects human decision-making often governed by the need to balance desired and undesired features of different action or options. The idea of considering goals in judgments is close to the concept of satisficing introduced by Simon (see Section 3.2). In the SDT formalism presented here, goals are captured in parameter α in Eq. 6, which weighs the relative importance of false positives and false negatives. In medical diagnostics, the emphasis is usually on reducing false negatives, whereas in life-detection missions the focus is likely to be on avoiding false positives. The importance of avoiding false negatives should not be, however, ignored. Otherwise, there is danger of limiting the search to uniquely biological but rare biosignatures, and by doing so greatly increasing the likelihood of finding nothing.

Although probabilistic frameworks are more appropriate from the formal standpoint, utility-based approaches have practical advantages. There is a large body of evidence that people, even expert statisticians, are often confused when they have to deal with probabilities. In contrast, assessing the pros and cons of options, assigning them relative importance and scoring positive and negative aspects of different options is common in everyday decision making. Thus, people have more experience with judgments based on utility. A similar reasoning is behind a proposal to evaluate the strength and significance of individual arguments supporting or contradicting biological origins of a given biosignature rather than probabilities, and subsequently apply an algorithm to turn these evaluations to probabilities. Considering that the three evaluation frameworks have different features, it might be worthwhile to employ all of them.

Every evaluative process involves human judgment, which is inevitably subjected to perceptual and cognitive biases at the evaluation and measurement levels. These biases will be, in general, different when biosignatures are evaluated individually (noncomparative judgments) or through comparison with other biosignatures (comparative judgments) in probabilistic or utility theory frameworks, but they will always exist. For example, biased responses to sensory stimuli associated with the law of diminishing sensitivity cannot be influenced because they arise from psychophysics. However, it might be possible to reduce cognitive biases related to this law through enforcing the desired reference points in carefully designed instructions or comparative measurements. More generally, reducing or eliminating some biases usually happens at a price of introducing other biases. Since different biases are not expected to have equal effect on outcomes, knowledge from psychological decision making can be brought to bear to design the evaluation process such that the overall impact of biases interferes as little as possible with the goals of a study.

Several issues relevant to evaluating biosignatures have not been addressed in this paper. For example, simultaneous evaluation of several biosignatures has not been discussed. If they do not depend on each other, simple probability calculus applies. If they interact synergistically, it might be more profitable to define them as a separate, ''collective biosignature.''

A reason for serious concern is a possibility of ''black swan events'' (Taleb, 2010), that is, findings that are very rare, unpredictable, and extend beyond the realm of expectations based on current knowledge. Admitting such possibility forms the basis for the concept of ''weird life'' (National Research Council, 2007) and motivates the search for agnostic biosignatures, but almost by definition assigning probabilities to such findings is extremely difficult.

This paper is aimed at presenting the basic concepts involved in evaluating biosignatures of life detection. A host of advanced approaches in probabilistic reasoning and decision making (see, *e.g.,* Pearl, 2014) can be brought to bear, but to what extent it can be fruitfully done considering the

nature of the available data is not known. This makes it an interesting and important area for studies in both life detection and decision theory.

Acknowledgments

This work was supported by the NASA Internal Scientists Funding Model (ISFM) supporting Center for Life Detection.

Author Disclosure Statement

No competing interests exist.

Supplementary Material

A. Derivation of informedness measure from the expected utility approach

B. Example of evaluating biosignatures

C. Accounting for uncertainties in estimating probabilities and utilities

D. Tools for representing the relation between stimulus and response

E. Comparative and non-comparative measurements

References

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., and Wagenmakers, E.J. (2015) An introduction to Bayesian hypothesis testing for management research. *Journal of Management* 41:521–543.

Ariely, D. and Wallsten, T.S. (1995) Seeking subjective dominance in multidimensional space: an explanation of the asymmetric dominance effect. *Organ Behav Hum Decis Process* 63:223–232.

Azar, F.S. (2000) *Multiattribute Decision-Making: Use of Three Scoring Methods to Compare the Performance of Imaging Techniques for Breast Cancer Detection,* Technical Report MS-CIS-00-10, University of Pennsylvania, Department of Computer & Information Science, Philadelphia, PA. Available online at http://repository.upenn.edu/cis_reports/119

Benner, S.A. (2017) Detecting Darwinism from molecules in the Enceladus plumes, Jupiter's moons, and other planetary water lagoons. *Astrobiology* 17:840–851.

Bernoulli, D. (1954) Exposition of a new theory on the measurement of risk. *Econometrica* 22:23–26.

Bornstein, B.H. and Emler, A.C. (2001) Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *J Eval Clin Pract* 7:97–107.

Budescu, D., Por, H., Broomell, S.B., and Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nat Clim Chang* 4:508–512.

Burton, A. and Berger, E. (2018) Insights into abiotically generated amino acid enantiomeric excesses found in meteorites. *Life* 8, doi:10.3390/life8020014.

Carr, M.H. (2006) *The Surface of Mars,* Cambridge University Press, New York.

Catling, D.C., Krissansen-Totton, J., Kiang, N.Y., Crisp, D., Robinson, T.D., DasSarma, S., Rushby, A.J., Del Genio, A., Bains, W., and Domagal-Goldman, S. (2018) Exoplanet biosignatures: a framework for their assessment. *Astrobiology* 18:709–738.

Cleophas, T.J., and Zwinderman, A.H. (2018) Bayesian network. In *Modern Bayesian Statistics in Clinical Research,* edited by T.J. Cleophas and A.H. Zwinderman, Springer International Publishing, New York, pp 143–173.

Cockell, C.S. (1999) Life on Venus. *Planet Space Sci* 47:1487–1501.

Cooper, M.J., Dimitrov, O., and Rau, P.R. (2001). A Rose.com by any other name. *J Finance* 56:2371–2388.

Cornell, B. (2001). Is the response of analysts to information consistent with fundamental value function? The case of Intel. *Financial Management* 30:113–136.

Cronin, L. and Walker, S.I. (2016) Beyond prebiotic chemistry. *Science* 352:1174–1175.

Dee, L. and Gerber, L. (2012) Applications of decision theory to conservation planning and management. *Nature Education Knowledge* 3(10):11.

Des Marais, D.J., Nuth, J.A., III, Allamandola, L.J., Boss, A.P., Farmer, J.D., Hoehler, T.M., Jakosky, B.M., Meadows, V.S., Pohorille, A., Runnegar, B., and Spormann, A.M. (2008) The NASA Astrobiology Roadmap. *Astrobiology* 8:715–730.

Di Achille, G. and Hynek, B.M. (2010) Ancient ocean on Mars supported by global distribution of deltas and valleys. *Nat Geosci* 3:459–463.

Dick, S.J. (2006) NASA and the search for life in the Universe. *Endeavour* 30:71–75.

Dorazio, R.M. and Johnson, F.A. (2003) Bayesian inference and decision theory—a framework for decision making in natural resource management. *Ecol Appl* 13:556–563.

Eigenbrode, J.L., Summons, R.E., Steele, A., Freissinet, C., Millan, M., Navarro-González, R., and Archer, P.D. (2018) Organic matter preserved in 3-billion-year-old mudstones at Gale crater, Mars. *Science* 360:1096–1101.

Elstein, A.S. and Schwarz, A. (2002) Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ* 324:729–732.

Eschenmoser, A. (1999) Chemical etiology of nucleic acid structure. *Science* 284:2118–2124.

Fenton, N. and Neil, M. (2012) *Risk Assessment and Decision Analysis with Bayesian Networks,* Chapman and Hall CRC Press, Boca Raton, FL.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978) Fault trees: sensitivity of estimated failure probabilities to problem representation. *J Exp Psychol Hum Percept Perform* 4: 330–334.

Fortes, A.D. (2000) Exobiological implications of a possible ammonia-water ocean inside Titan. *Icarus*146:444–452.

Fox, C.R. and Clemen, T. (2005) Subjective probability assessment in decision analysis: partition dependence and bias toward the ignorance prior. *Management Sci* 5:1309–1448.

Fox, C.R. and Rottenstreich, Y. (2003) Partition priming in judgment under uncertainty. *Psychol Sci* 14:195–200.

Freissinet, C., Glavin, D.P., Mahaffy, P.R., Miller, K.E., Eigenbrode, J.L., Summons, R.E., Brunner, A.E., Buch, A., Szopa, C., Archer, P.D., Jr., Franz, H.B., Atreya, S.K., Brinckerhoff, W.B., Cabane, M., Coll, P., Conrad, P.G., Des Marais, D.J., Dworkin, J.P., Fairén, A.G., François, P., Grotzinger, J.P., Kashyap, S., ten Kate, I.L., Leshin, L.A., Malespin, C.A., Martin, M.G., Martin-Torres, J.F., McAdam, A.C., Ming, D.W., Navarro-González, R., Pavlov, A.A., Prats, B.D., Squyres, S.W., Steele, A., Stern, J.C., Sumner, D.Y., Sutter, B., Zorzano, M.-P., MSL Science Team (2015) Organic molecules in the Sheepbed Mudstone, Gale Crater, Mars. *J Geophys Res* 120:495–514.

Garthwaite, P.H., Kadane, J.B., and O'Hagan, A. (2005) Statistical methods for eliciting probability distributions. *J Am Stat Assoc* 100:680–701.

Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics,* Wiley, New York.

Hand, K.P., Chyba, C.F., Priscu, J.C., Carlson, R.W., and Nealson, K.H. (2009) Astrobiology and the potential for life on Europa. In *Europa*, edited by R. Pappalardo, W. McKinnon, and K. Khurana, University of Arizona Press, Tucson, AZ, pp 589–629.

Hoehler, T.M., Bains, W., Davila, A., Parenteau, N., and Pohorille, A. (2020) Life's requirements: habitability and biological potential. In *Planetary Astrobiology,* edited by V. Meadows, G.N. Arney, B.E. Schmidt, and D.J. Des Marais, University of Arizona Press, Tucson, AZ, pp 37–70.

Hud, N.V., Cafferty, B.J., Krishnamurthy, R., and Williams, L.D. (2013) The origin of RNA and ''my grandfather's axe.'' *Chem Biol* 20:466–474.

Jansen, S.J.T. (2011). The Multi-Attribute Utility method. In *The Measurement and Analysis of Housing Preference and Choice,* edited by S.J.T. Jansen, R.W. Goetgeluk, and H.C.C.H. Coolen, Springer, New York, pp 101–126.

Jeffreys, H. (1973) *Scientific Inference,* 3rd ed., Cambridge University Press, Cambridge, UK.

Johnson, S.S., Anslyn, E.V., Graham, H.V., Mahaffy, P.R., and Ellington, A.D. (2018) Fingerprinting non-terran biosignatures. *Astrobiology* 18:915–922.

Kane, S.R., Hill, M.L., Kasting, J.F., Kopparapu, R.K., Quintana, E.V., Barclay, T., and Hinkel, N.R. (2016) A catalog of Kepler habitable zone exoplanet candidates. *Astrophys J* 830, doi:10.3847/0004-637X/830/1/1.

Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J Am Stat Assoc* 90:773–795.

Keeney, R.L. Raiffa, H. (1993) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs,* Cambridge University Press, New York.

Kiang, N.Y., Domagal-Goldman, S., Parenteau, M.N., Catling, D.C., Fujii, Y., Meadows, V.S., and Walker, S.I. (2018) Exoplanet biosignatures: at the dawn of a new era of planetary observations. *Astrobiology* 18:619–629.

Klein, H.P. (1978) The Viking biological experiments on Mars. *Icarus* 34:666–674.

Koning, M. and Smith, C. (2017) *Decision Trees and Random Forests: A Visual Introduction for Beginners*. Amazon Digital Services LLC - Kdp Print Us.

Levin, G.V. and Straat, P.A. (2016) The case for extant life on Mars and its possible detection by the Viking Labeled Release experiment. *Astrobiology* 16:798–810.

Limaye, S.S., Mogul, R., Smith, D.J., Ansari, A.H., Słowik, G.P., and Vaishampayan, P. (2018) Venus' spectral signatures and the potential for life in the clouds. *Astrobiology* 9:1181–1198.

Lopes, L.L. and Oden, G.C. (1999). The role of aspiration level in risky choice: a comparison of cumulative prospect theory and SP/A theory. *J Math Psychol* 43:286–313.

Lorenz, R.D. (2019). A Bayesian approach to biosignature detection on ocean worlds. *Nat Astron* 3:466-467.

Lunine, J.I. (2017) Ocean worlds exploration. *Acta Astronaut* 131:123–130.

Lynn, S.K. and Barrett, L.F. (2014) ''Utilizing'' Signal Detection Theory. *Psychol Sci* 25:1663–1673.

Marshall, S.M., Murray, A.R., and Cronin, L. (2017) A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philos Trans A Math Phys Eng Sci* 375, doi: 10.1098/rsta.2016.0342.

Mellers, B. and Biagini, K. (1994) Similarity and choice. *Psychol Rev* 10:505–518.

Miller, R.A. and Geissbuhler, A. (1999) Clinical diagnostic decision support systems—an overview. In *Clinical Decision Support Systems,* edited by E.S. Berner, Springer, New York, pp 3–34.

Musen, M.A., Middleton, B., and Greenes, R.A. (2014) Clinical decision-support systems. In *Biomedical Informatics,* edited by E.H. Shortliffe and J.J. Cimino, Springer, London, pp 643–674.

National Research Council. (2007) *The Limits of Organic Life in Planetary Systems,* The National Academic Press, Washington, DC.

Neveu, M., Hays, L.E., Voytek, M.A., New, M.H., and Schulte, M.D. (2018) The Ladder of Life Detection. *Astrobiology* 18: 1375–1402.

Noguchi, T. and Stewart, N. (2014) In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition* 132: 44–56.

Nyberg, S.O. (2018) *The Bayesian Way: Introductory Statistics for Economists and Engineers*. John Wiley and Sons, New York.

Payne, J.W. (1976) Task complexity and contingent processing in decision making: an information search and protocol analysis. *Organ Behav Hum Perform* 16:366–387.

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, Amsterdam.

Pizzarello, S., Schrader, D.L., Monroe, A.A., and Lauretta, D.S. (2012) Large enantiomeric excesses in primitive meteorites and the diverse effects of water in cosmochemical evolution. *Proc Natl Acad Sci USA* 109:11949–11954.

Pohorille, A. and Keller, R. (2010) Hypothesis-based, community supported organization of scientific information in astrobiology [abstract 5659]. In *Astrobiology Science Conference 2010: Evolution and Life: Surviving Catastrophes and Extremes on Earth and Beyond,* Lunar and Planetary Institute, Houston.

Pohorille, A. and Pratt, L.R. (2012) Is water the universal solvent for life? *Orig Life Evol Biosph* 42:405–409.

Powers, D.M. (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2:37–63.

Priscu, J.C. and Hand, K.P. (2012) Microbial habitability of icy worlds. *Microbe* 7:167–172.

Quinn, R.C., Martucci, H.F., Miller, S.R., Bryson, C.E., Grunthaner, F.J., and Grunthaner, P.J. (2013) Perchlorate radiolysis on Mars and the origin of martian soil reactivity. *Astrobiology* 13:515–520.

Raulin, F. and Owen, T. (2002) Organic chemistry and exobiology on Titan. *Space Sci Rev* 104:377–394.

Regan, H.M., Ben-Haim, Y., Langford, B., Wilson, W.G., Lundberg, P., Andelman, S.J., and Burgman, M.A. (2005) Robust decision-making under severe uncertainty for conservation management. *Ecol Appl* 15:1471–1477.

Rosch, E. (1975) Cognitive representation of cognitive categories. *J Exp Psychol Gen* 104:192–233.

Rosch, E. (1978/1988) Principles of categorization. In *Readings in Cognitive Science, a Perspective from Psychology and Artificial Intelligence,* edited by A. Collins and E.E. Smith, Morgan Kaufman Publishers, San Mateo, CA, pp 312–322.

Ruopp, M.D., Perkins, N.J., Whitcomb, B.W., and Schisterman, E.F. (2008) Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50:419-430.

Russo, J.E. and Dosher, B.A. (1983) Strategies for multiattribute binary choice. *J Exp Psychol Learn Mem Cogn* 9: 676–696.

Savadori, L., Savio, S., Nicotra, E., Rumiati, R., Finucane, M., and Slovic, P. (2004) Expert and public perception of risk from biotechnology. *Risk Analysis* 24:1289–1299.

Schonhoff, T.A. and Giordano, A.A. (2006) *Detection and Estimation Theory and Its Applications,* Pearson Education, Upper Saddle River, NJ.

Schulze-Makuch, D. and Grinspoon, D.H. (2005) Biologically enhanced energy and carbon cycling on Titan? *Astrobiology* 5:560–564.

Schulze-Makuch, D. and Irwin, L.N. (2006) The prospect of alien life in exotic forms on other worlds. *Naturwissenschaften* 93:155–172.

Shapiro, R. and Schulze-Makuch, D. (2009) The search for alien life in our solar system: strategies and priorities. *Astrobiology* 9:335–343.

Shefrin, H. (2001) Do investors expect higher returns from safer stocks than from riskier stocks? *J Behav Financ* 2:1–15.

Simon, H.A. (1957) *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting,* John Wiley and Sons, New York, NY.

Simonson, I. (1989) Choice based on reasons: the case of attraction and compromise effects. *J Consum Res* 16:158–174.

Slovic, P. (1972) Psychological study of human judgment: implications for investment decision making. *J Finance* 27:779–799.

Slovic, P., MacGregor, D.G., Malmfors, T., and Purchase, I.F.H. (1999) *Influence of Affective Processes on Toxicologists' Judgments of Risk,* Report No. 99-2, Decision Research, Eugene, OR.

Slovic, P., Fincane, M., Peters, E., and MacGregor, D.G. (2004) Risk analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Analysis* 24:311–322.

Stanislaw, H. and Todorov, N. (1999) Calculation of Signal Detection Theory measures. *Behav Res Methods Instrum Comput* 31:137–149.

Sullivan, W. (2018) *Decision Tree and Random Forest: Machine Learning and Algorithms: The Future Is Here!* CreateSpace Independent Publishing Platform.

Svenson, O. (1979) Process descriptions of decision making. *Organ Behav Hum Perform* 23:86–112.

Swets, J.A. (1988) Measuring the accuracy of diagnostic systems. *Science* 240:1285–1293.

Taleb, N.N. (2010) *The Black Swan: The Impact of the Highly Improbable,* 2nd ed., Penguin, London.

Tsou, P., Brownlee, D.E., McKay, C., Anbar, A.D., and Yano, H. (2012) LIFE: Life Investigation for Enceladus: a sample return mission concept in search for evidence of life. *Astrobiology* 12:730–742.

Tversky, A. (1977) Features of similarity. *Psychol Rev* 84:327–352.

Tversky, A. and Kahneman, D. (1983) Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol Rev* 91:293–315.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–1131.

von Neumann, J. and Morgenstern, O. (1947) *Theory of Games and Economic Behavior,* Princeton University Press, Princeton, NJ.

Walker, S.I., Bains, W., Cronin, L., DasSarma, S., Danielache, S., Domagal-Goldman, S., Kacar, B., Kiang, N.Y., Lenardic, A., Reinhard, C.T., Moore, W., Schwieterman, E.W., Shkolnik, E.L., and Smith, H.B. (2018) Exoplanet biosignatures: future directions. *Astrobiology* 18:779–824.

Weber, E.U. (1994) From subjective probabilities to decision weights: the effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychol Bull* 115:228–242.

Weber, E.U. (2018) ''Risk as feelings'' and ''perception matters:'' psychological contributions on risk, risk taking and risk management. In *The Future of Risk and Risk Management,* edited by H. Kunreuther, R. Meyer, and E. Michel-Kerjan, University of Pennsylvania Press, Philadelphia, PA, pp 30–47.

Weber, E.U. and Kirsner, B. (1996). Reasons for rank-dependent utility evaluation. *J Risk Uncertain* 14:41–61.

Wickens, T.D. (2002) *Elementary Signal Detection Theory,* Oxford University Press, New York.

Windschitl, P.D. and Wells, G.L. (1998) The alternative-outcomes effect. *J Pers Soc Psychol* 75:1411–1423.

Windschitl, P.D. and Young, M.E. (2001) The influence of alternative outcomes on gut-level perceptions of certainty. *Organ Behav Hum Decis Process* 85:109–134.

Windschitl, P.D., Young, M.E., and Jenson, M.E. (2002) Likelihood judgment based on previously observed outcomes: the alternative-outcomes effect in a learning paradigm. *Mem Cognit* 3:469–477.

Youden, W.J. (1950) Index for rating diagnostic tests. *Cancer* 3: 32–35.

Address correspondence to:
Andrew Pohorille
NASA Ames Research Center
Exobiology Branch
MS 239-4
Moffett Field, CA 94035
USA

E-mail: andrew.pohorille@nasa.gov

Submitted 31 July 2019
Accepted 24 June 2020

Abbreviations Used

KB = knowledge base
LoLD = Ladder of Life Detection
MAU = Multi-Attribute Utility
SDT = Signal Detection Theory
SI = Supplementary Information