

# Truthlikeness and the inclusion fallacy\*

Gustavo Cevolani<sup>†</sup>, Davide Coraci<sup>‡</sup> & Matteo De Benedetto<sup>§</sup>

Preprint  
Forthcoming in *Thinking and Reasoning*

## Abstract

In this paper, we offer a novel normative analysis of the inclusion fallacy. This refers to the robust tendency of experimental participants to reason about categories in a way that violates the laws of probability. In contrast to the received view, we argue that participants' preferences in this kind of task might be rationally defended. To this purpose, we employ the philosophical notion of truthlikeness as a normative benchmark for category-based induction. Our analysis complements the main empirical findings and theoretical predictions of the psychological literature and it explains the difficulties that probabilistic models have in treating cases like the inclusion fallacy.

Keywords: inclusion fallacy, category-based induction, truthlikeness, verisimilitude, truth approximation, probabilistic reasoning, rationality

## 1 Introduction

From its beginnings, the psychological study of human reasoning always had to wrestle between the normative and the descriptive levels. That is, studying a specific form of reasoning, such as deduction, induction, abduction, categorization, and the like, could mean trying to understand how people should reason, how people actually reason, and why people often do not reason as they should. Depending on the specific form of reasoning one focuses on, the study of the normative and the descriptive dimensions of reasoning might evolve together, or, instead, there could be a significant gap between the two.

---

\*All the authors contributed equally to the manuscript. We thank Matias Osta-Vélez and the members of the MInD research group at the IMT School for Advanced Studies Lucca for useful discussions and feedback on the manuscript. We acknowledge financial support from the Italian Ministry of University and Research (MUR) through the PRIN 2022 grant n. 2022ARRY9N funded by the European Union (Next Generation EU) and from the John Templeton Foundation (grant number 62265).

<sup>†</sup>IMT School for Advanced Studies Lucca; gustavo.cevolani@imtlucca.it

<sup>‡</sup>IMT School for Advanced Studies Lucca; davide.coraci@imtlucca.it

<sup>§</sup>IMT School for Advanced Studies Lucca; matteo.debenedetto@imtlucca.it

In the case of category-based induction (CBI), our understanding of the descriptive dimension of this form of reasoning is much more advanced than our understanding of its normative dimension. In CBI, people reason from the premise that the members of a category (e.g., robins) have a given property to the conclusion that the members of another category (e.g., birds) also have that property. In the last fifty years, psychologists have explored an impressive array of possible conceptual mechanisms by virtue of which people project a certain property from one category to another. In contrast, the normative dimension of CBI has received less attention. The only available normative explanations of how people ought to project properties between different categories can be found within the Bayesian paradigm. However, there are instances of CBI that clearly eschew a Bayesian treatment. An important example concerns the inclusion fallacy, i.e., the robust tendency of experimental participants in projecting a given property to a general category like birds, rather than to a proper subset of that category like penguins, despite the plain fact that it is just less likely that all birds, instead of just penguins, have a given property.

Since few alternatives to the Bayesian account have been put forward in the literature, a proper normative explanation of why people commit such a blatant violation of the laws of logic and probability is still lacking, and the inclusion fallacy remains a central open problem in the discussion on CBI.

In this paper, we seek to fill this gap, by offering a novel normative analysis of the inclusion fallacy. We will do that by building upon the philosophical theory of truthlikeness or verisimilitude, which clarifies how one should reason in order to approach the truth about a given domain. The notion of truthlikeness has been systematically studied in twentieth-century philosophy of science in the context of discussions over scientific progress and realism, but it has remained virtually unexplored within cognitive science. We will show how this notion can provide a useful normative benchmark for inductive inferences such as category-based induction. In particular, we will see that the notion of truthlikeness, by licensing strongly informative conclusions in inductive inferences, provides a normative explanation of why people commit the inclusion fallacy. Actually, when judged against the benchmark of truthlikeness, participants' reasoning can be construed as a rational attempt to approach the truth about the relevant domain. This stands in stark contrast with the traditional probabilistic normative benchmark of such inferences.

The goal of this paper is then two-fold. First, we aim to offer a novel normative analysis of the inclusion fallacy, arguing that, in contrast to the received probabilistic view, people's preferences in this alleged fallacy might actually be rationally defended. The second, more general, aim of the paper is to highlight the usefulness of the philosophical notion of truthlikeness as a possible normative benchmark for the study of inductive inferences in the psychology of reasoning.

In Section 2, we present the phenomenon at the center of this paper, i.e., the inclusion fallacy in CBI. In Section 3, we briefly present the idea of truthlikeness in the context of the philosophical theory of cognitive decisions. Section 4 is devoted to applying the theory of truthlikeness as a normative benchmark to

analyze CBI and the inclusion fallacy, and Section 5 discusses how our analysis relates to alternative psychological models of category-based induction. Section 6 concludes the discussion, while Appendix A offers all the formal details for the interested reader.

## 2 Category-based induction and the inclusion fallacy

How plausible is that whales have lungs, given that cows have lungs? To answer questions like this, people engage in a specific form of inductive reasoning known as category-based induction. This kind of inductive inference uses conceptual information to estimate how plausibly a given property may be projected from one category to another. In the last fifty years, the study of CBI has proceeded mostly in the footsteps of the pioneering work of Rips (1975). In this seminal paper, Rips proposed an experimental design that involves asking participants to project an unfamiliar, new property (what is technically called a “blank” predicate) from a base category (robins, in Rips’ first example) to other target categories, more or less similar to the base category. The two main findings of Rips were that the likelihood of projecting a property from one category to another is a positive function of the typicality of the categories involved and of the similarity between them. The central role of typicality and similarity in CBI has been repeatedly highlighted and confirmed in many subsequent studies (cf. Feeney, 2017; Feeney and Heit, 2007; Heit, 1998; Osherson et al., 1990; Sloman, 1993; Smith et al., 1993).

In this connection, experimental studies on CBI demonstrated that people can compare categorical arguments involving different categories based on their strength and that they consistently judge some arguments as stronger than others (e.g., Osherson et al., 1990; Shaffi et al., 1990; Smith et al., 1993). In particular, Smith et al. (1993) noted that people tend to evaluate arguments with more general categories in their conclusions as stronger, contrary to the predictions of formal logic and probability theory. The clearest illustration of this tendency is the phenomenon that Smith and colleagues dubbed the *inclusion fallacy*. In their original study, they observed that people robustly judge an argument like (Osherson et al., 1990, p. 188):

$$\frac{P \quad \text{All robins have an ulnar artery}}{C_1 \quad \text{All birds have an ulnar artery}} \quad (1)$$

to be stronger than an argument like:

$$\frac{P \quad \text{All robins have an ulnar artery}}{C_2 \quad \text{All penguins have an ulnar artery}} \quad (2)$$

This judgment is traditionally considered to be fallacious because, given that penguins are birds, conclusion  $C_1$  necessarily implies conclusion  $C_2$ . Accordingly,  $C_1$  cannot be more probable than  $C_2$  and, as far as argument strength

should depend on the probability of the conclusion, this makes the second argument logically stronger—or at the very least, as strong as the first argument. Thus, people’s categorical reasoning in such context appears to violate sound probabilistic reasoning.

The inclusion fallacy highlights how people struggle to use their knowledge about inclusion relations between categories while reasoning under uncertainty in the context of CBI. Indeed, factors beyond inclusion relations—such as the similarity between categories and their typicality—appear to override categorical knowledge and play a more direct role in guiding inductive reasoning (Sloman & Lagnado, 2005). After these findings, the literature has predominantly focused on developing predictive mathematical models of CBI and testing them against qualitative experimental data of how people project properties across categories. The three main kinds of models proposed in the psychological literature are category-based models (e.g., Osherson et al., 1990), feature-based models (e.g., Sloman, 1993), and probabilistic models (e.g., Heit, 1998; Tenenbaum et al., 2006). Although there are significant conceptual differences between them, it is important to stress that they all postulate, consistent with Rips’ (1975) seminal findings, that the two most important predictors of the strength of a CBI argument are the typicality and the similarity of the categories involved. Thus, all the models above mostly agree in predicting most of the experimental results including the inclusion fallacy. Despite all the empirical and theoretical progress that, in the last decades, has been made in unveiling the main factors affecting how people reason with categories, the normative dimension of this kind of inductive reasoning is still a puzzle. In particular, traditional normative benchmarks of inductive reasoning, such as probability theory, appear to be clearly untenable in light of phenomena like the inclusion fallacy. In fact, probabilistic models of CBI (e.g., Heit, 1998; Tenenbaum et al., 2006), which understand this kind of reasoning as a specific type of Bayesian updating, cannot explain this fallacy. More generally, one cannot find in the literature a satisfactory explanation of why people commit this fallacy. To be clear, some of the aforementioned similarity-based and feature-based psychological models (e.g., Osherson et al., 1990; Sloman, 1993) predict indeed that people would commit the inclusion fallacy; however, they do not provide a satisfactory normative explanation of why people commit this mistake, remaining entirely non-committal on whether and why people’s preferences in the inclusion fallacy are rational or not. This situation contrasts with the discussion of other reasoning fallacies, such as, for instance, the conjunction fallacy (Tversky 1983), for which a number of explanations and solutions have been proposed (gcLindaSienapaper; see Crupi et al., 2008; Tentori et al., 2013, for references and discussion). In the rest of the paper, we will try to fill this gap by offering a normative (dis)solution of the inclusion fallacy. To do that, we need first to make a brief detour through twentieth-century philosophy of science. This will be the task of the next section.

### 3 Truthlikeness as approximation to the truth

Since at least the 1950s, philosophers of science and logicians started debating the role of different “epistemic” or “cognitive utilities” in human reasoning and cognitive decision-making. The discussion was typically framed in terms of the inferences and choices performed by ideal agents—like ideally rationally scientists. However, many of the contributions which followed were essentially independent from this assumption, and largely applicable to reasoning and decision-making as performed by real people. Scholars like Popper, Carnap, Hempel, Levi, Hintikka, and many others introduced in the debate a number of formally well-defined concepts which should clarify how we reason, typically under uncertainty, when we have to decide what to believe choosing among competing hypotheses, theories or explanations. These included, among others, truth, probability, information, accuracy, confirmation, corroboration, explanatory power, and sometimes combinations of them (Niiniluoto2011). A full “cognitive decision theory” was developed in order to analyze and explain the reasoning and cognitive choices of rational agents in terms of their attempt at maximizing one or more of such cognitive utilities, construed as the goals of rational inquiry.<sup>1</sup> With some relevant exceptions (Stroner2020; e.g., Crupi et al., 2008; Tentori et al., 2013), such philosophical research programs remained unrelated to experimental work in psychology and cognitive science. This is unfortunate, as the sophisticated models of reasoning and decision making developed by philosophers can both inform empirical research and receive empirical confirmation from experimentally observed phenomena. In this section we show how the notion of truthlikeness, as developed within cognitive decision theory, might be fruitfully applied as a normative benchmark for the study of category-based induction.

PopperCR<empty citation> introduced the idea of truthlikeness, or verisimilitude, in his attempt to defend a realist but fallibilist view of scientific methodology. According to Popper, scientific theories and hypothesis aim at approaching “the whole truth” about a given domain; however, most of them are false (or will be proven false in the future) and, as such, are at most truthlike or verisimilar conjectures about the truth. To make sense of this idea, one needs a defensible notion of “closeness” or “similarity to the truth”, according to which, say, hypothesis  $H1$  may be said to be closer or more similar to the truth than an alternative hypotheses  $H2$ . While Popper’s original attempt to define such a notion famously failed, different adequate accounts of truthlikeness were later developed (Oddie1986; Niiniluoto1987; Kuipers2019; Schurz2010; gcSchurzFestschrift). Despite relevant differences, all such theories show how truthlikeness subtly balances considerations of truth and information in assessing competing hypotheses or statements (**sep-truthlikeness**). To see how this works, let us start from a simple toy example.

---

<sup>1</sup>The philosophical research program on epistemic utility theory (Pettigrew2016; Leitgeb, 2017), aiming at justifying probabilism as the norm of belief and providing foundations for so-called accuracy-first epistemology, can be viewed as an instance of such approach (Oddie2017).

Suppose a young researcher, call her Eve, is interested in studying birds and their features. Then, the idea of truthlikeness is roughly the following: the more things Eve believes about birds, and the more of such things are in fact true, the closer is Eve to fulfilling her cognitive goal. In other words, we can characterize such goal as follows: Eve aims at having many true beliefs, and few false beliefs, about birds. The idea resounds with William James' famous maxim in *The Will to Believe*: "We must know the truth; and we must avoid error—these are our first and great commandments as would-be knowers" (James, 1897). For instance, Eve would of course prefer to believe, say, that robins fly than to believe the opposite—since the former statement, but not the latter, is true. Also, she would prefer to believe that robins fly and have an orange-red chest, than to only believe that robins fly. This is because the former belief is plainly more informative than the latter—and it is still true: the more true information, the better. Ideally, Eve would like to come to accept all possible true beliefs about robins (that they fly, that they have an orange-red chest, that they sing, and so on), while avoiding accepting any falsehood about them. Even better, Eve would like to know "the whole truth" not just about robins, but also about all other species of birds—sparrows, blackbirds, penguins, and so on. In this sense, Eve's cognitive goal is the complete, true description of all existing species of birds, in terms of all possible relevant features: a sort of encyclopedic, full knowledge of all ornithology. The more truthlike (or verisimilar) the set of her actual beliefs, i.e., the more informative and at the same time mostly true they are, the closer is Eve to reaching this goal.

As simple as it sounds, the above idea has some interesting consequences. In real-life contexts, cognitive agents have to reason and decide under conditions of, more or less severe, uncertainty. This means that the whole truth is typically unknown and that, in order to follow James' maxim, Eve cannot simply *decide* to accept many truths and avoid to believe any falsehoods. Instead, she needs a way to estimate, possibly on the basis of the available evidence, what is true and what is false. The standard way to do this is probabilistic reasoning. Interestingly, however, a crucial consequence of the idea of truthlikeness is that probability cannot be the only guide in Eve's cognitive life. In other words, if Eve aims at the whole truth about birds, she cannot merely accept highly probable beliefs about them. This is simply because, as far as logically stronger beliefs are more informative than logically weaker ones, the former are bound to be no more probable than the latter. It follows that, while more informative beliefs tend to be less probable than less informative beliefs, they can well be cognitively preferable to them.<sup>2</sup> In particular, a logically stronger and more informative belief may well be more verisimilar than a weaker and less informative one, and hence preferable to it in terms of truthlikeness.

To illustrate, suppose  $F(r)$  denotes that "robins fly",  $C(r)$  that "robins have an orange-red chest", and  $G(r)$  that "robins are green". Consider the following three claims or hypotheses about robins:

---

<sup>2</sup>This is the main lesson we take from the conjunction fallacy in probabilistic reasoning. For a discussion of the conjunction fallacy in terms of truthlikeness, see [gcSILFS07paper](#); [gcLindaSienapaper](#)<empty citation>.

$$\begin{aligned}
H1 & F(r) \\
H2 & F(r) \wedge C(r) \\
H3 & F(r) \wedge G(r)
\end{aligned}$$

Let's first compare  $H1$  and  $H2$ . Of course,  $H2$  cannot be more probable than  $H1$ , since  $H2$  logically implies  $H1$  (and not vice versa) and hence  $H2$  is strictly more informative than  $H1$ . However, if Eve aims at approaching the whole truth about robins,  $H2$  may well be a better guess than  $H1$ , especially if Eve has some evidence in favor of the fact that robins may have an orange-red chest. Indeed, given what we know about robins,  $H2$  is actually more truthlike than  $H1$ , since it provides more true information about robins. But even pretending not to know this, it is possible that, given her evidence, Eve estimates the truthlikeness of  $H2$  as higher than the truthlikeness of  $H1$ . In more formal term, while the probability of  $H2$  must be lower than that of  $H1$ , what we will call the "expected truthlikeness" of  $H2$  may be higher than that of  $H1$ . For this reason, Eve may well prefer  $H2$  over  $H1$  in her attempt to approach the whole truth about this kind of birds. Now, let's compare  $H1$  and  $H3$ . Again,  $H3$  cannot be more probable than  $H1$ , for the same reason seen above. Moreover, since it is false that robins are green,  $H3$  is arguably less truthlike than  $H1$ , even if it is more informative. This shows that greater informative content is not enough for truth approximation; what is important for truthlikeness is the trade-off between truth and information: more information may be good, if true, but may also be bad, if false. Similarly, expected truthlikeness balances the probability (of being true) of an hypothesis and its information content. Thus, even if Eve didn't know the color of robins, she may have evidence against  $G(r)$ , and she may estimate that the expected truthlikeness of  $H3$  is lower than that of  $H1$ . As the two examples just considered show, estimates of probability and of expected truthlikeness may both agree and disagree, depending on the specific hypotheses compared.

In short, as **Popper** <empty citation> forcefully argued many years ago, science and human knowledge aims at informative, "strong" truths, whereas probability rewards less informative, "weak" hypotheses about the world. Stronger hypotheses better serve our cognitive aim since, if true or at least highly truthlike, they allow us to approach the whole truth about the domain under investigation. Should we only aim at probability, we should instead eschew such strong hypotheses, preferring weaker ones, which are more likely to be plainly true. In the limit, we should only entertain tautological beliefs, thus being certain of what we believe. This would amount to obey one of James' commandments—"avoid error"—at the price of ignoring the other—"know the truth." Thus, from the point of view of truthlikeness, a rational preference for possibly truthlike, but improbable, beliefs is perfectly justified. As we shall see, this leads to an interesting take on the inclusion fallacy and related issues.

## 4 Truthlikeness in category-based induction

It is now time to put the pieces back together. In this section, we will apply the philosophical notion of truthlikeness to bear on the issue of the inclusion fallacy. More specifically, we will employ truthlikeness as a normative benchmark for category-based induction, i.e., as defining the cognitive utility people should maximize in projecting a predicate from one category to the other. We will see that, through the normative lenses of truthlikeness, the inclusion fallacy is not necessarily a fallacy anymore.

Let us go back to our running example of the inclusion fallacy from Section 2, which we now take in the following more general form (where  $X$  represents some blank predicate, like “having an ulnar artery”):

$$\frac{P \quad \text{All robins have feature } X}{C_1 \quad \text{All birds have feature } X} \qquad \frac{P \quad \text{All robins have feature } X}{C_2 \quad \text{All penguins have feature } X}$$

As we reported in Section 2, most people will judge the argument on the left as stronger than that on the right. What could justify such a preference? In our view, the idea of truthlikeness provides a simple answer. If people are guided by the goal of approaching the whole truth about the underlying domain (in our cases, that of birds), “All birds have feature  $X$ ” may well provide a better guess than “All penguins have feature  $X$ ” in terms of truth approximation. This is because, as a hypothesis about the domain, the former is plainly more informative than the latter; and, if it is likely that most birds actually have  $X$ , it would also be much closer to the truth. Let see how this idea works in detail.

For the sake of clarity, let us focus on a simple toy case involving our bird-enthusiast friend Eve from Section 3 (a more general treatment is provided in the Appendix). Suppose Eve only knows about three species of birds—robins ( $r$ ), sparrows ( $s$ ), and penguins ( $p$ )—and about three possible binary features characterizing them: fly ( $F$ ), have wings ( $W$ ), and have beak ( $B$ ). We assume that Eve aims at the complete, true description of the three species in terms of their features. Then, the best she can do is to specify all the features that characterize each species of bird. That is, she can state that robins and sparrows fly, have wings, and have beaks—in symbols,  $FWB(r)$  and  $FWB(s)$ , respectively—and that penguins don’t fly, have wings, and have beaks—in symbols,  $\overline{FWB}(p)$  (where a bar over a letter means negating the corresponding feature). This amounts to provide a complete, correct classification system for all the species involved, as displayed in fig. 1. The table reads as follows: given the three features  $F$ ,  $W$ , and  $B$ , there are eight possible kinds of birds (those who have  $F$ ,  $W$ , and  $B$ ; those who have  $F$ ,  $W$ , but not  $B$ ; etc.), corresponding to the eight cells of the table. A classification system assigns each species to its correct kind, by placing it in the correct cell: in our case, robins and sparrows belong to the kind of flying, winged, and beaked birds, while penguins belong to that of non-flying, winged, beaked birds. Note that, in this simple toy case, all remaining cells in the table are empty, meaning that no bird in the domain satisfies the corresponding set of properties (in particular, no bird is wing-less or beak-less).

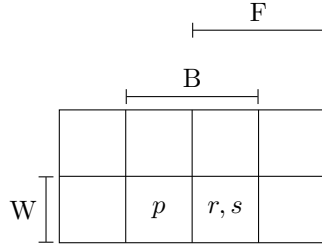


Figure 1: A simple classification system of three species of birds (*robins*, *sparrows*, and *penguins*) according to three possible features (Fly, have Wings, have Beak).

The table in fig. 1 represents the knowledge of an ideally well-informed and perfectly correct agent. Of course, Eve’s actual knowledge may well fall short of this ideal. This typically happens because Eve does not know all the features of all species, meaning that she will be unable to assign some birds to the correct kind (cell). For instance, suppose that Eve, after correctly classifying the three species of birds as in fig. 1, learns that robins also have some new feature  $X$ , besides  $F$ ,  $W$ , and  $B$ . If  $X$  is a blank predicate (e.g., “having an ulnar artery”), it may well be that Eve does not know how it is distributed among the other birds. The new situation is now represented by the larger table in fig. 2: Eve assigns robins to the kind of  $FWBX$ -birds, but does not know where to place, in her classification system, sparrows and penguins. Each of these latter birds may or may not have  $X$ , and hence may belong to two possible kinds (cells), but Eve cannot tell to which one. The best she can do is an informed guess, based on what she knows about robins and other birds.

This brings us back to the inclusion fallacy. In the typical experimental design of Rips (1975), participants are required to choose among two arguments with the same premise and different conclusions. In our running example, the premise informs them that robins have a blank predicate  $X$ . In line with the literature (cf. Heit, 1998; Sloman, 1993), we assume that this piece of information works as the evidence  $E$  on which participants assess competing hypotheses about the domain, represented by the conclusions of the arguments: “All birds have feature  $X$ ” ( $H1$ ) and “All penguins have feature  $X$ ” ( $H2$ ). If, as we assume, the cognitive goal of participants is approaching the whole truth about the domain, they are in Eve’s situation as represented in fig. 2.

How should Eve evaluate the two hypotheses? Given her background knowledge of birds (represented in fig. 1) and the available evidence about robins, the two hypotheses may be represented as follows:

$$\begin{aligned}
 H1 & \quad FWBX(r) \wedge FWBX(s) \wedge \overline{FWBX}(p) \\
 H2 & \quad FWBX(r) \wedge \overline{FWBX}(p)
 \end{aligned}$$

Both hypotheses specify the kind or cell to which different species belong. Note that both are uncertain, since Eve does not know whether sparrows and penguins

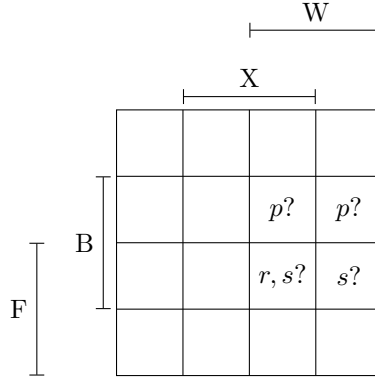


Figure 2: A simple classification system of three species of birds (*robins*, *sparrows*, and *penguins*) according to three common features (Fly, have Wings, have Beak) and a blank predicate  $X$ . Question marks denote that the placement of species in the cell is uncertain or unknown.

have  $X$  or not. Moreover,  $H1$  entails  $H2$ , and hence cannot be more probable. In terms of truthlikeness, however,  $H1$  can well be preferable to  $H2$ . This is because  $H1$  is more informative than  $H2$ , since it specifies the kind of all the three species in the domain, whereas  $H2$  is silent on sparrows. Moreover, if it were true that sparrows have  $X$ , the additional information provided by  $H1$  would be true. This means that, independently of whether penguins have  $X$  or not,  $H1$  would provide more true information about the domain than  $H2$ , and hence would be preferable to it because it is more verisimilar, i.e., it is closer to the truth.

Thus, Eve’s task reduces to evaluating how probable it is that sparrows have  $X$ , given that robins have  $X$ : if this probability is sufficiently high, Eve would act reasonably in “betting” on  $H1$  as a better approximation to the truth than  $H2$ . Slightly more formally, let  $p(H|E)$  denote the probability of some hypothesis  $H$  as estimated on the available evidence  $E$ ; and let  $Evs(H|E)$  denote its expected truthlikeness, again as estimated on  $E$ . Here,  $Evs(H|E)$  represents Eve’s rational estimation of the actual truthlikeness, which cannot be known since the truth is not known. In appendix A, we prove a result to the following effect:

**Lemma 1** *Given evidence  $E$ ,*

$$Evs(H1|E) > Evs(H2|E) \text{ iff } p(X(s)|E) > \frac{1}{k+1}$$

In words, the expected truthlikeness of  $H1$  is greater than the expected truthlikeness of  $H2$  just in case the probability that sparrows have  $X$ , given that robins have  $X$ , is “sufficiently high” (i.e., higher than threshold  $\frac{1}{k+1}$ , where  $k$  is the number of relevant features). Since in our case the number  $k$  of features

is 4, it is sufficient that the probability that sparrows have  $X$  is greater than 0.2 for  $H1$  having a greater expected truthlikeness than  $H2$  and hence being preferred by the agent.

This is where the notions of typicality and similarity come into play, and our analysis gets in touch with the classical results on CBI. In fact, Eve may assume that, since sparrows are typical birds, i.e., they are already quite similar to robins, it is quite probable that they will also share with them the additional feature  $X$ , besides the other three ones already shared (i.e.,  $F$ ,  $W$ , and  $B$ ). Thus, Eve may use the perceived similarity between robins and sparrows as a proxy for the probability that sparrows also have  $X$ . One way to do this would be, for instance, to judge  $p(X(s)|E)$  as approximately equal to .75, since sparrows share already 3 out of 4 features of robins. As a consequence, the threshold provided by lemma 1 would be easily met and the result would follow.

It is also worth noting that lemma 1 above is a consequence of the following more general result (theorem 1), that we prove in appendix A. Let Eve employ  $k$  features to describe  $n$  species of birds. Assume that, on the basis of her background knowledge, Eve knows the relevant  $k - 1$  features of each species, with the exception of one blank predicate  $X$ . Moreover, assume that she receives evidence  $E$  according to which species  $a$  has feature  $X$ . Now let  $H1$  be the hypothesis that all birds have  $X$  and  $H2$  the hypothesis that only species  $a$  and  $b$  have  $X$ . Of course,  $H1$  logically entails  $H2$  (and not vice versa), so  $p(H1|E) < p(H2|E)$ . However, if the probability that each species  $c$  appearing in  $H1$  but not in  $H2$  has  $X$  is greater than  $\frac{1}{k+1}$ , then  $Evs(H1|E) > Evs(H2|E)$ , i.e., the expected truthlikeness of  $H1$  will be greater than that of  $H2$ . For this reason, Eve will prefer  $H1$  over  $H2$  if her cognitive aim is truth approximation. Note that the two main determinants of this effect are typicality and similarity. In fact, if the species  $c$  (in our case, sparrows) is highly similar to  $a$ , then Eve will judge that the relevant probability that  $c$  has  $X$  as well is high. Moreover, if  $a$  is highly typical (as in our case robins are), most of the species appearing in  $H1$  will be similar to  $a$ , thus raising the probability of each of them having feature  $X$ . In this way, we can see how the expected truthlikeness of our hypothesis depends on the underlying typicality and similarity of the species involved.

In sum, the idea of truthlikeness, applied as a normative benchmark for CBI, provides a simple explanation of the inclusion fallacy. People’s robust preferences in favor of a universal conclusion are not puzzling anymore when looked at through the lenses of truthlikeness theory. In fact, from this perspective, by projecting the blank predicate “having an ulnar artery” to the whole category of birds, agents seek to approach the whole truth about birds, justified by the fact that typical birds such as robins possess such a feature. The stronger, more informative hypothesis that “all birds have an ulnar artery” is then expected to be closer to the truth than the weaker, less informative hypothesis that “all penguins have an ulnar artery”, even if, of course, the latter remains more probable than the former. The participants’ preference for the universal conclusion can be seen as a rational bet for maximizing the balance of true and false beliefs over the domain of birds, in view of the evidence that they are given (i.e., that typical birds like robins have some feature). The key to dissolve the fallacy

is then to understand category-based induction as an inference that seeks to maximize the expected truthlikeness of our beliefs over a given domain.

## 5 Discussion

In this section, we will discuss how the verisimilitudinarian analysis of the inclusion fallacy, and the underlying idea of truthlikeness as normative benchmark for category-based induction, relates to the more descriptive models available in the psychological literature and with Bayesian, normative analyses of this type of inference.

For the purpose of this discussion section, we can divide psychological models of CBI in three main groups: category-based models, feature-based models, and probabilistic models.<sup>3</sup> We will first give a brief description of how the three kinds of models work, one by one, and whether and how they predict that people commit the inclusion fallacy. Then, we will discuss how these predictions relate to our analysis.

**Category-based models.** Category-based models conceptualize CBI as primarily based on the relationships between the different concepts involved in the inference. A paradigmatic example of category-based models is the seminal “coverage model” developed by Osherson et al. (1990).<sup>4</sup> According to the coverage model, the argument strength of an instance of category-based induction is determined by two factors: (i) the similarity between the category in the premises and the one in the conclusion; (ii) the degree of “coverage” between the premises’ category and the lowest super-ordinate category that includes all the categories involved in the inference (both in the premises and in the conclusion). Applied to the specific case of the inclusion fallacy (Osherson et al., 1990, p. 195), the coverage model predicts that the strength of the two arguments (1) and (2) from Section 2 depends on two factors: (i) the similarity, respectively, of robins to birds (for argument 1) and of robins to penguins (for argument 2); and (ii) the similarity between the category in the premise (i.e., robins) and the lowest-level category including all the categories appearing in the argument, i.e., robins, penguins, and birds; in our case, this category is “birds” itself. As this second factor is common between the two arguments, the assessment of their relative strength is entirely driven by the comparison between the degree of similarity of robins to birds and the degree of similarity of robins to penguins. Since penguins are very atypical members of the category of birds, while robins are highly typical, the former is greater than the latter. It follows that argument (1) is perceived as stronger than argument (2), thus explaining the inclusion fallacy.

---

<sup>3</sup>The distinction between these three kinds of models is mostly pragmatic and should not be considered as a sharp one. Indeed, in the literature, there are also hybrid accounts of category-based induction that try to combine different kinds of models (**bright2014engine**; cf. Feeney, 2007).

<sup>4</sup>Recently, a generalization of the coverage model has been proposed by Osta-Vélez and Gärdenfors (2020) within the theory of conceptual spaces.

**Feature-based models.** Feature-based models conceptualize CBI as primarily based on the relationships between the features characterizing the categories in the premises and those in the conclusion of the argument. In such models, similarity between different concepts (like “robin” or “bird”) depends on the similarity between the features characterizing the relevant concepts. A paradigmatic example of a feature-based model of CBI is Sloman’s “connectionist” model (Sloman, 1993, 1998). In Sloman’s model, the relative strength of arguments doesn’t depend on the hierarchical relationship among the concepts appearing in the arguments, as in category-based models like Osherson and colleagues’ coverage model. Instead, the model only relies on how such concepts relate to each other in terms of their shared features.<sup>5</sup> More precisely, categories are represented as vectors of features and features may be connected to each other as the nodes of a network. In the case of the inclusion fallacy, the blank predicate (e.g., “having an ulnar artery”) is treated as a feature that, initially, is not connected to any other feature. When the premise “all robins have an ulnar artery” is given, however, the units representing the features of the premise category (“robins”) are connected to the blank predicate unit. Similarly, the blank predicate unit will be connected to the features of the concepts appearing in the conclusions of the arguments (“birds” and “penguins”). The level of “activation” of the blank predicate unit (essentially, the number of established connections with other features) will then predict the strength of the corresponding argument. In particular, since robins are typical birds and hence share many features of birds, the feature vector of robins will be much more similar to the vector of birds than to that of penguins, which are atypical (Sloman, 1993, p. 259). Accordingly, the blank predicate “having an ulnar artery” will be more activated for birds than for penguins and argument (1) will be perceived as stronger than argument (2), thus explaining the inclusion fallacy. Note that, despite the differences in how they represent concepts and describe the mechanisms driving the assessment of argument strength, the connectionist model and the coverage model agree on the main determinant of the fallacy: i.e., the similarity between the categories appearing in the premise and in the conclusion of the arguments.

**Probabilistic models.** Probabilistic models construe CBI as a special case of Bayesian reasoning, i.e., in terms of belief updating. The central idea is that the strength of an argument depends on the posterior probability of its conclusion given its premises (and possibly some background information), as assessed by a Bayesian agent. A good example of a probabilistic model is the one developed by Heit (1997, 1998), who first proposed to view CBI as a case of standard Bayesian updating.<sup>6</sup> On the basis of some reasonable assumptions

<sup>5</sup>Other feature-based, connectionist models have been proposed in the literature (cf. Medin et al., 2003; Rogers and McClelland, 2004). These more recent models propose different measures of feature-matching, but keep fixed Sloman’s core idea of argument strength as a function of premises-conclusion feature-matching.

<sup>6</sup>More recently, Tenenbaum et al. (2006) generalized Heit’s model by representing the priors of the agent as complex structures that represent in turn (the relevant part of) the conceptual

about what people’s priors could be, Heit’s model is able to predict several qualitative phenomena related to CBI and highlighted in experimental research, such as similarity, typicality, diversity, and homogeneity effects. However, the model cannot properly account for the inclusion fallacy, which is our topic here. This is not surprising, since purely probabilistic models are necessarily inconsistent with non-monotonic effects, in the sense that they cannot favor stronger (more informative) hypotheses over weaker (less informative) ones. As we already saw, in the inclusion fallacy participants tend to favor arguments (like argument (1) from Section 2) with a conclusion which is strictly more general, and hence stronger, than a more specific, and hence weaker, conclusion of another argument. As the probability of “all birds have an ulnary artery”, even if conditionalized on the premise “all robins have an ulnary artery”, cannot be greater than the probability of “all penguins have an ulnary artery” (since all penguins are birds), a probabilistic model cannot account for participants’ preference for the former over the latter. Accordingly, argument (1) cannot be stronger than argument (2) as far as the probability of their conclusions is concerned.

How do the psychological models of category-based induction briefly surveyed above relate to our present proposal? Two points are especially worth mentioning. First, it is important to note a crucial aspect uniting our analysis and the psychological models above. Despite their differences, all three kinds of models agree on the most robust prediction of the empirical literature, i.e., that category similarity and typicality are the two main determinants of the strength of category-based arguments.<sup>7</sup> The coverage model is based on a similarity function to assess the degree to which premise categories resemble the conclusion category and the similarity between categories is affected by typicality. Then, the inclusion fallacy is explained by the higher similarity of robins and birds as compared to the similarity between robins and penguins, given that penguins are highly atypical as birds and are highly dissimilar to robins (Osherson et al., 1990, p. 195). Similarly, the feature-based model accounts for the inclusion fallacy by appealing to the number of shared features between the different categories in the arguments. Specifically, Sloman (1993, p. 259) proposes to estimate the similarity between categories as proportional to their common features and inversely proportional to their distinctive features. This measure is closely related to the original one proposed by Tversky (1977) that, in turn, shows relevant formal connections with the basic feature approach to truthlikeness (**gcSchurzFestschrift**; **gcSILFS07paper**) assumed here (see appendix A).

These findings are also crucially employed in our truthlikeness-based analysis of the inclusion fallacy, as the discussion of Lemma 1 shows. Indeed, the key factor which determines the assessment of the expected truthlikeness of the

---

knowledge of the agent (see also Kemp and Tenenbaum, 2009).

<sup>7</sup>It should be noted that another important determinant is the nature of the blank predicate under consideration (cf. **heit1994similarity**). Moreover, in certain specific contexts, other factors have been observed to affect the strength of category-based induction, such as, for instance, agents’ general background knowledge (cf. Murphy and Ross, 2010) and causal knowledge (cf. Bright and Feeney, 2014; Rehder, 2006).

relevant hypotheses is the probability the agent assigns to sparrows having the blank predicate  $X$ . In turn, such probability will depend on how sparrows are similar (in terms of shared features) to robins, and on the fact that robins are typical birds. In this sense, our analysis complements existing descriptive models, providing an adequate normative framework to properly understand the role of similarity and typicality in shaping people’s preferences and explaining the inclusion fallacy.

Second, our truthlikeness-based analysis of the inclusion fallacy sheds new light on the role of probability in human judgment. In particular, it helps assessing the prospects of probabilistic models of CBI (e.g., Heit, 1998; Kemp and Tenenbaum, 2009; Tenenbaum et al., 2006). As discussed in detail above, such models cannot account for phenomena like the inclusion fallacy, where participants exhibit a robust preference for stronger and less probable hypotheses. Our approach helps understanding why probabilistic models, despite their many virtues, lack the conceptual resources for properly analyzing such kind of preference. The essential reason is that the probability of different hypotheses is not enough to determine a cognitive choice among them. To be sure, such probability is a crucial component of the rational assessment of different hypotheses, as we just illustrated above, but it is insufficient alone. In other words, probability is just one ingredient of a more complex recipe to evaluate competing hypotheses. The other one is information content, which is as important as a cognitive utility as probability of being true is. (Expected) truthlikeness encodes precisely this trade-off between probability (expected truth) and content of hypotheses and, as such, it eschews a purely probabilistic treatment. For this reason, maximizing expected truthlikeness, while still involving a probabilistic assessment of the relevant hypotheses, can at the same time reward contentful, strong guesses, thus producing clearly anti-probabilistic effects like the inclusion fallacy. Within such a verisimilitudinarian approach, the failure of probabilistic models to make sense of the inclusion fallacy echoes Popper’s old complaint against purely inductive-probabilistic accounts of scientific inferences: probability cannot hold the keys to scientific inference because it does not reward contentful inferences. And, we can now add, for the same reason, it cannot be a good normative benchmark for content-seeking inferences like category-based induction.

In this way, our verisimilitudinarian analysis of the inclusion fallacy shows why it might be rational to prefer a less probable, but more contentful generalization over another hypothesis that is at least as likely, but less contentful. By rewarding content, the notion of truthlikeness can rationalize content-driven inferences and shed a new light on known non-probabilistic effects in human reasoning. To conclude this section, we briefly discuss two examples of analogous content-driven effects in human reasoning, that might be amenable to a truthlikeness-based normative explanation analogous to the one we offered in this paper for the inclusion fallacy. The first effect, closely associated with the inclusion fallacy, is known as “inclusion similarity” (Sloman, 1993, 1998; Sloman & Lagnado, 2005), “inverse conjunction fallacy” (Connolly et al., 2007; Jönsson & Assarsson, 2016; Jönsson & Hampton, 2006), or the “modifier effect”

(jonsson2012modifier; Stroner2020). This amounts to the fact that people exhibit inconsistent reasoning by attributing a greater probability to more general statements, such as “All sofas have backrests”, rather than to more specific ones, e.g., “All uncomfortable handmade sofas have backrests,” even though the inclusion relation between the categories is transparent and people readily agree that all uncomfortable handmade sofas are sofas (Jönsson & Hampton, 2006). An application of the notion of truthlikeness as a normative benchmark for this effect might successfully reconsider people’s preferences in this case as rational, stressing the higher content of the more general statements in comparison to the more specific ones.

A second, well-known non-probabilistic effect that might be analogously defended by a verisimilitudinarian account could be the *generic overgeneralization effect* (leslie2011all), in which people incorrectly endorse false universal statements if the related generic is true. Thus, for instance, people overgeneralize from a true generic such as “ducks lay eggs” to the related universal statement such as “all ducks lay eggs” (which is false, because of the existence of male ducks). Again, (at least) some instances of this effect could be defended as rational by a truthlikeness account, stressing the high portion of true content of the (technically false) universal statements that people endorse. That said, we leave to future work the exploration of how our present approach may deal with this other kinds of reasoning, as a full application of the notion of truthlikeness to cases like these would need to involve a generalization of the present account and a careful account of the maximization of truthlikeness as a cognitive utility related to the similarity, typicality, representativeness, and causal structure of the categories involved.

## 6 Conclusion

Let us recap the main steps of the present article. We started with noticing the discrepancy between our descriptive and normative understanding of category-based induction. We recalled how, despite all the progress made by psychologists in the last fifty years in uncovering the mechanisms behind this important form of inference, its normative dimension has remained under-studied. In particular, we highlighted how the puzzling phenomenon of the inclusion fallacy is still lacking a proper normative explanation. To fill this gap, we resorted to the philosophical notion of verisimilitude or truthlikeness as studied in twentieth-century philosophy of science. Applying this notion as a normative benchmark for CBI, we showed how people’s tendency to commit the inclusion fallacy can be explained as an attempt to maximize truthlikeness, i.e., to approach the whole truth about the domain under investigation. Finally, we discussed how our analysis complements the main empirical findings and theoretical predictions of the psychological literature on CBI and it explains the difficulties that probabilistic models have in treating cases like the inclusion fallacy.

The idea of truthlikeness highlights the importance, for human cognition, of the trade-off between the content of different beliefs and their probability

of being true. Our analysis deployed such trade-off to illuminate the specific case of CBI, and of the inclusion fallacy in particular. More work is needed to test the robustness of truthlikeness as a general normative benchmark for CBI. For instance, it would be interesting to apply truthlikeness to bear on all the phenomena discussed in the psychological literature on CBI and to investigate experimentally its implications for people’s preferences. Another, more general, future line of research would investigate whether and how the trade-off between truth and information encoded by truthlikeness provides an alternative to the dominant probabilistic inferential paradigm in the psychology of reasoning. In this connection, as we have briefly sketched at the end of the last section, the many fallacies and alleged Bayesian blindspots that recent works in the psychology of reasoning have highlighted—such as, for instance, the modifier effect, the generic overgeneralization effect, or the essentialist bias—represent a promising territory for future work extending the present article.

## References

- Bright, A. K., & Feeney, A. (2014). Causal knowledge and the development of inductive reasoning. *Journal of Experimental Child Psychology*, *122*, 48–61.
- Carnap, R. (1950). *Logical foundations of probability*. Chicago University of Chicago Press.
- Connolly, A. C., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2007). Why stereotypes don’t even make good defaults. *Cognition*, *103*(1), 1–22.
- Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*(2), 182–199.
- Feeney, A. (2007). How many processes underlie category-based induction? effects of conclusion specificity and cognitive ability. *Memory & cognition*, *35*(7), 1830–1839.
- Feeney, A. (2017). Forty years of progress on category-based inductive reasoning. *International handbook of thinking and reasoning*, 167–185.
- Feeney, A., & Heit, E. (2007). Inductive reasoning: Experimental, developmental, and computational approaches. *Fifth International Conference on Thinking, Jul, 2004, University of Leuven, Belgium; Many of the chapter authors for this book talked at the aforementioned symposium*.
- Heit, E. (1997). Features of similarity and category-based induction. *Proceedings of the Interdisciplinary Workshop on Categorization and Similarity*, 115–121.
- Heit, E. (1998). A bayesian analysis of some forms of inductive reasoning. In *Rational models of cognition* (pp. 248–274). Oxford University Press.
- James, W. (1897). *The will to believe: And other essays in popular philosophy* (F. Burkhardt, F. Bowers, & I. K. Skrupskelis, Eds.). Cambridge University Press.
- Jönsson, M. L., & Assarsson, E. (2016). A problem for confirmation theoretic accounts of the conjunction fallacy. *Philosophical Studies*, *173*, 437–449.

- Jönsson, M. L., & Hampton, J. A. (2006). The inverse conjunction fallacy. *Journal of Memory and Language*, 55(3), 317–334.
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, 116(1), 20–58.
- Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford University Press USA - OSO.
- Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, 10(3), 517–532.
- Murphy, G. L., & Ross, B. H. (2010). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*, 63(1), 1–17.
- Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, 97(2), 185–200.
- Osta-Vélez, M., & Gärdenfors, P. (2020). Category-based induction in conceptual spaces. *Journal of Mathematical Psychology*, 96, 102357.
- Rehder, B. (2006). When similarity and causality compete in category-based property generalization. *Memory & Cognition*, 34, 3–16.
- Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, 14(6), 665–681.
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.
- Shaffi, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, 18(3), 229–239.
- Slooman, S. A. (1993). Feature-based induction. *Cognitive psychology*, 25(2), 231–280.
- Slooman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35(1), 1–33.
- Slooman, S. A., & Lagnado, D. (2005). The problem of induction. *The Cambridge handbook of thinking and reasoning*, 95–116.
- Smith, E. E., Shafir, E., & Osherson, D. N. (1993). Similarity, plausibility, and judgments of probability. *Cognition*, 49(1-2), 67–96.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7), 309–318.
- Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, 142(1), 235.
- Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4), 327.

## A Formal appendix

In the following, we define the notions of truthlikeness and of expected truthlikeness, and offer a proof of lemma 1 and of the more general result behind it. Our approach is inspired to the so-called basic feature approach to truthlikeness (**gcSchurzFestschrift**; **gcSILFS07paper**). See also **Niiniluoto1987**<empty citation> for relevant terminology and for a more general approach.

We assume that the domain under investigation includes a finite set  $a_1, \dots, a_n$  of  $n$  individuals, described by a finite set of  $k$  binary features  $F_1, \dots, F_k$ . If  $x$  is an individual, a conjunction  $Q(x)$  of  $k$  statements of the form  $\pm F_1(x) \wedge \dots \wedge \pm F_k(x)$ , where “ $\pm$ ” may be replaced by the negation sign or nothing, describes a possible “kind of individual”: i.e., it is a complete description of individual  $x$  in terms of all the  $k$  features. In the tradition of inductive logic (**Niiniluoto2011**; Carnap, 1950), conjunctions of such form are called  $Q$ -predicates; there are  $q = 2^k$  of them. Note that  $Q$ -predicates provide a complete classification system for the individuals in the domain: each  $Q$ -predicate corresponds to a cell in fig. 1 or fig. 2, and each individual belongs to exactly one cell. Moreover, the (dis)similarity between different kinds of individuals can be defined in terms of the distance between the corresponding  $Q$ -predicates. A natural measure for such distance is  $\delta(Q_i, Q_j)$ , defined as the number of disagreements between  $Q$ -predicates  $Q_i$  and  $Q_j$ , divided by  $k$  (corresponding to the normalized Hamming or city-block distance, see **Niiniluoto1987**). Conversely, similarity or closeness between kinds of individuals is defined as  $1 - \delta(Q_i, Q_j)$ , i.e., in terms of (the normalized number of) their shared features. Note that  $\delta$  corresponds to the number of steps needed to move from cell  $Q_i$  to cell  $Q_j$  in fig. 1 and fig. 2 (divided by  $k$ ).<sup>8</sup>

The cognitive goal of a rational agent interested in classifying the  $n$  individuals in the domain consists in assigning each individual  $x$  to its correct kind  $Q$ . A conjunction  $Q_{i_1}(a_1) \wedge \dots \wedge Q_{i_m}(a_m)$  of  $m$   $Q$ -predicates, with  $1 \leq m \leq n$ , is a possible hypothesis  $H$  about such classification. Each conjunct of  $H$  amounts to the claim that individual  $x$  belongs to kind  $Q_i$ —i.e., that the  $Q$ -predicate  $Q_i(x)$  is true. If  $H$  correctly assigns each of the  $n$  individuals in the domain to its correct kind, then  $H$  is “the whole truth” about the domain. Note that  $H$  may fall short of the whole truth in two ways. First,  $H$  can assign some individuals to the wrong kind or cell, i.e.,  $H$  can entail a false  $Q$ -predicate. Second, if  $m < n$ ,  $H$  will fail to assign some individuals to any kind or cell, i.e.,  $H$  will make no claim about the  $Q$ -predicate of such individuals. The former is an error of falsity, i.e., a proper mistake; the latter is an error of ignorance, i.e., an omission or lacuna of  $H$ .

The truthlikeness or verisimilitude of  $H$  expresses how close  $H$  is to the (whole) truth. Intuitively, the more the “matches” (true  $Q$ -predicates) of  $H$  and the less the mistakes (false  $Q$ -predicates) and the lacunae (omissions) of  $H$ , the greater its truthlikeness. More generally, we define the “prize” or “penalty” received by each  $Q$ -predicate  $Q_i(x)$  via the following payoff function:

$$\pi(Q_i(x)) = \begin{cases} 1 & \text{if } Q_i(x) \text{ is true} \\ -\delta(Q_i, Q_j) & \text{if } Q_i(x) \text{ is false and } Q_j(x) \text{ is true} \end{cases} \quad (3)$$

Intuitively, the above definition is justified as follows: if  $H$  assigns  $x$  to its correct

<sup>8</sup>We assume that the diagram appearing in the figures is a bi-dimensional representation of a toroidal surface, meaning that, for instance, the extreme right cells are actually adjacent to the extreme left cells, and the top cells are adjacent to the bottom cells. In the computer science literature, such a diagram is known as a “Karnaugh map” after **Karnaugh1953**<empty citation>.

kind, it receives a prize equal to 1, which is the self-similarity  $1 - \delta(Q_i, Q_i)$  of  $Q_i$  to itself. If instead  $H$  assigns  $x$  to the wrong kind  $Q_i$ , it receives a penalty equal to the dissimilarity  $\delta(Q_i, Q_j)$  between  $Q_i$  and the correct kind  $Q_j$ . In this way, the seriousness of mistakes is taken into account. The degree  $vs(H)$  of truthlikeness of  $H$  is then computed as the normalized sum of the payoffs received by  $H$  due to each of its  $m$   $Q$ -predicates:

$$vs(H) = \frac{1}{n} \sum_{i=1}^m \pi(Q_i) \quad (4)$$

Note that the lacunae of  $H$  do not appear in eq. (4), meaning that they receive an implicit payoff of 0, smaller than the payoff of a match but bigger than that of a mistake. Note also that  $H$  has the maximum degree of truthlikeness 1 when it is the whole truth itself; and it has the minimum degree of truthlikeness  $-1$  when it assigns to each individual the farthest  $Q$ -predicate from the true one. From eqs. (3) to (4) it follows that:

$$\begin{aligned} vs(H) &= \sum_{i=1}^m vs(Q_i(x)) \\ &= \sum_{i=1}^m \frac{\pi(Q_i(x))}{n} \end{aligned} \quad (5)$$

i.e., that the truthlikeness of  $H$  is just the sum of the truthlikeness of each of its  $Q$ -predicates, where the latter is just the normalized payoff received by each  $Q$ -predicate (note that each  $Q$ -predicate is, by definition, also an hypothesis, with just one conjunct).

The definition of truthlikeness in eq. (4) allows the agent to assess the closeness of each hypothesis  $H$  to the truth, assuming the latter is known. If this is not the case, we assume that the agent can rationally assess the probability  $p$  that a given individual belongs to a given kind. This (epistemic) probability distribution is relativized to the evidence  $E$  available to the agent, which may be empty or include some background information about the various kinds of individuals. The expected truthlikeness  $Evs(H|E)$  of  $H$  given  $E$  can then be defined as the expected value of  $vs(H)$  on the basis of  $p$ , which, given eq. (5), amounts to:

$$\begin{aligned} Evs(H|E) &= \sum_{i=1}^m vs(Q_i(x)) \times p(Q_i(x)|E) \\ &= \sum_{i=1}^m Evs(Q_i(x)|E) \\ &= \sum_{i=1}^m \frac{1}{n} E\pi(Q_i(x)|E) \end{aligned} \quad (6)$$

In other words, the expected truthlikeness of  $H$  is just the normalized sum of the expected payoffs of its  $Q$ -predicates, where, from eq. (3)

$$\begin{aligned} E\pi(Q_i(x)|E) &= 1 \times p(Q_i(x)|E) - \delta(Q_i, Q_j) \times p(\neg Q_i(x)|E) \\ &= (1 + \delta(Q_i, Q_j))p(Q_i(x)|E) - \delta(Q_i, Q_j) \end{aligned} \quad (7)$$

where  $Q_j$  is the true kind of  $x$ .

Measure  $Evs$  allows the agent to rationally, if fallibly, estimate the truthlikeness of  $H$  in the light of available evidence  $E$ . Coming back to the inclusion fallacy, we propose that the agent evaluates the relative strength of different arguments by estimating the expected truthlikeness of their conclusion given the evidence provided by their premises. The conclusions of the two arguments provide the relevant hypotheses to be assessed. In the standard experimental design, they are as follows:  $H1$  states that all individuals have  $X$ ;  $H2$  that just one individual (penguins in our running example) have  $X$ . Given the intended application, we shall assume that the domain contains  $n > 2$  individuals, described by at least  $k \geq 2$  features. Moreover, we assume that, given some background knowledge, the agent knows the relevant  $k - 1$  features of all individuals in the domain (cf. fig. 1), with the exception of feature  $X$  (the blank predicate). It follows that each individual can belong to just two possible kinds: one in which it has  $X$ , and one it has not  $X$  (cf. fig. 2). Note that the distance between these two kinds is  $\frac{1}{k}$ .

The premise of the arguments in the standard task provides the evidence  $E$  according to which one individual (robins in our running example) does have feature  $X$ . Without loss of generality, assume this individual is  $a_1$ ;  $E$  assigns  $a_1$  to the corresponding  $X$ -kind, call it  $Q_i$ . According to hypothesis  $H2$ , another individual, say  $a_2$  (penguins in the example), belongs to the  $X$ -kind  $Q_j$  (in general different from  $Q_i$ , as in fig. 2). According to hypothesis  $H1$ , all other individuals also belong to a  $X$ -kind. In short, our two hypotheses are as follows:

$$\begin{aligned} H1 & Q_i(a_1) \wedge Q_j(a_2) \wedge Q_{i_3}(a_3) \wedge \dots \wedge Q_{i_{n-2}}(a_{n-2}) \\ H2 & Q_i(a_1) \wedge Q_j(a_2) \end{aligned}$$

with the understanding that all kinds  $Q_{i_3} \dots Q_{i_{n-2}}$  are  $X$ -kinds. The only difference is that  $H1$  states that all other individuals, besides  $a_1$  and  $a_2$ , have also feature  $X$ . Of course, we have that  $p(H1|E) < p(H2|E)$ . However, it follows from eqs. (6) to (7) that  $Evs(H1|E)$  will be greater than  $Evs(H2|E)$  if the expected truthlikeness of each  $Q$ -predicate  $Q_{i_1}(a_3), \dots, Q_{i_{n-2}}(a_{n-2})$  is positive. Recall from eq. (6) that such expected truthlikeness is just the normalized expected payoff of the corresponding  $Q$ -predicate, where the payoff is either 1, if the assigned kind is correct, or  $-\frac{1}{k}$ , if it is not. Focusing on one individual at time, say  $s$  (for sparrows, as in our running example), one can check that (we write just  $Q$  to simplify notation):

$$\begin{aligned} Evs(Q(s)|E) > 0 & \quad \text{iff} \\ \frac{1}{n}E\pi(Q(s)|E) > 0 & \quad \text{iff} \\ (1 + \frac{1}{k})p(Q(s)|E) - \frac{1}{k} > 0 & \quad \text{iff} \\ p(Q(s)|E) > \frac{1}{k+1} & \end{aligned} \tag{8}$$

It follows that:

**Theorem 1** *Let  $H1$ ,  $H2$ , and  $E$  be as specified above. If  $p(Q_{i_j}(a_j)|E) > \frac{1}{k+1}$  for each  $a_j$ ,  $3 \leq j \leq n - 2$ , then  $Evs(H1|E) > Evs(H2|E)$ .*

Recalling that the probability  $p(Q(s)|E)$  of  $s$  belonging to  $Q$  is the same as the probability  $p(X(s)|E)$ , since  $Q$  is the  $X$ -kind to which  $s$  may belong, lemma 1 follows. Note that, even if, for simplicity, in the text we focused on the case of three individuals described by four features, the result doesn't depend on the number of individuals in the domain, nor on the number of features describing them.