J Philos Logic (2014) 43:283–302
DOI 10.1007/s10992-012-9263-1

# Defeasible Conditionalization

**Paul D. Thorn**

Received: 9 February 2012 / Accepted: 4 December 2012 / Published online: 25 January 2013


**Abstract** The applicability of Bayesian conditionalization in setting one's posterior probability for a proposition, α, is limited to cases where the value of a corresponding prior probability, $P_{PRI}(\alpha|\wedge E)$, is available, where ∧E represents one's complete body of evidence. In order to extend probability updating to cases where the prior probabilities needed for Bayesian conditionalization are unavailable, I introduce an inference schema, *defeasible conditionalization*, which allows one to update one's personal probability in a proposition by conditioning on a proposition that represents a proper subset of one's complete body of evidence. While defeasible conditionalization has wider applicability than standard Bayesian conditionalization (since it may be used when the value of a relevant prior probability, $P_{PRI}(\alpha|\wedge E)$, is unavailable), there are circumstances under which some instances of defeasible conditionalization are unreasonable. To address this difficulty, I outline the conditions under which instances of defeasible conditionalization are defeated. To conclude the article, I suggest that the prescriptions of direct inference and statistical induction can be encoded within the proposed system of probability updating, by the selection of intuitively reasonable prior probabilities.

**Keywords** Conditionalization · Probability updating · Principle of total evidence · Defeasible inference · Direct inference · Induction

## 1 Introduction: Conditionalization and the Principle of Total Evidence

Given a conjunction ∧E representing an agent's *complete* body of evidence, and a prior probability function $P_{PRI}$, standard Bayesian conditionalization prescribes that an agent form a posterior probability function $P_{POS}$, and set the values of $P_{POS}$ according to the equation: $P_{POS}(\alpha) = P_{PRI}(\alpha|\wedge E)$, for all $\alpha$. This approach to probability updating makes the idealizing assumption that rational agents always

P. D. Thorn (✉)
Philosophisches Institut, University of Düsseldorf, Universitätsstr. 1, Düsseldorf 40225, Germany
e-mail: thorn@phil-fak.uni-duesseldorf.de

have access to a prior probability function that is appropriate as a basis for conditionalization (cf. [11]). But, as a matter of psychological fact, an agent's doxastic state rarely encodes all of the priors needed for Bayesian conditionalization, and agents are rarely in a position to fill in all of the needed priors in a way that is justifiable by appeal to acceptable epistemic norms.[1] In order to extend probability updating to cases where the prior probabilities needed for Bayesian conditionalization are unavailable, I will propose a system for probability updating for agents whose prior probabilities are incomplete and/or imprecise. The system is thereby designed to accommodate approaches to rational credence formation that endorse probabilistic representations of uncertainty, but are not prepared to accept the idea that we must invariably represent a rational agent's doxastic state by a complete probability function.[2]

The Bayesian idea that one should update one's probabilities by conditionalization on one's *complete* body of evidence is closely related to Carnap's *principle of total evidence* ([3], 211). The principle of total evidence, as proposed by Carnap, prescribes that one take account of all one's evidence in making judgments of probability, or, more generally, that one take account of all of one's evidence that is *relevant* to a given proposition, in making a judgment about the probability of that proposition. While Carnap maintains that, strictly speaking, agents need only take account of all the relevant evidence, he also held that a proposition, β, is inclusive of an agent's evidence that is relevant to another proposition, $\alpha$, if and only if $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$. Carnap thereby embraced the thesis that agents should update their probabilities by standard Bayesian conditionalization, setting $P_{POS}(\alpha)$ to $P_{PRI}(\alpha|\wedge E)$, for all $\alpha$.

Contrary to Carnap's conception of evidential relevance, a little reflection confirms that the satisfaction of the condition that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$ is *not* sufficient for β being inclusive of an agent's evidence bearing on $\alpha$. Indeed, ∧E may encode much additional evidence, not encoded in β, that is relevant to $\alpha$, where $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$ holds by coincidence. For example, we may know that $P_{PRI}(\alpha|\beta) = 0.9, P_{PRI}(\alpha|\beta\wedge\chi) = 0.1$, and $P_{PRI}(\alpha|\beta\wedge\chi\wedge\delta) = 0.9$, where $\wedge E = \beta\wedge\chi\wedge\delta$, and where β, $\beta\wedge\chi$, and $\beta\wedge\chi\wedge\delta$ describe progressively larger samples that bear on the probability of $\alpha$.[3] It is also reasonable to deny that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$ is a necessary condition for β being inclusive of an agent's evidence bearing on $\alpha$, assuming that we consider agents whose prior probabilities are incomplete or imprecise. Indeed, once we extend our view to consider such agents, then it is possible to imagine cases where β is inclusive of an agent's evidence bearing on $\alpha$, where the value of $P_{PRI}(\alpha|\wedge E)$ is not given, or the range of values given for $P_{PRI}(\alpha|\wedge E)$ is imprecise (where the smallest set containing $P_{PRI}(\alpha|\wedge E)$ is [0, 1], or the range of values that are given for $P_{PRI}(\alpha|\beta)$ is a *proper subset* of the range of values given for $P_{PRI}(\alpha|\wedge E)$). In such cases, we may deny that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$ (or at least

[1] It is assumed here that various 'representation dependent' approaches to selecting probability functions, such as those that apply the principle of indifference ([2, 3, 6, 16], [6, ch. 11]), or the principle of maximum entropy ([12, 29, 30, 51]) cannot be justified by appeal to acceptable epistemic norms. For standard criticisms of such approaches, see [45] and [11].

[2] Such approaches have been widely endorsed. See, for example, [5, 14, 17, 19, 22, 24, 39, 46, 49], and [15].

[3] Although the condition that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$ is insufficient for β being inclusive of an agent's evidence bearing on $\alpha$, it appears that the satisfaction of the condition is sufficient ground for setting $P_{POS}(\alpha)$ to $P_{PRI}(\alpha|\beta)$. For this reason, one should not exaggerate the significance of the fact that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$ is insufficient for β being inclusive of an agent's evidence bearing on $\alpha$.

deny that it is given that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$), while nevertheless accepting: (1) β is inclusive of a respective agent's evidence bearing on $\alpha$, and (2) it is correct to set $P_{POS}(\alpha)$ to $P_{PRI}(\alpha|\beta)$.[4]

One point made in the preceding paragraph is that there are cases where $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$, but β is not inclusive of one's evidence bearing on α. In such cases, it is reasonable to set $P_{POS}(\alpha)$ to $P_{PRI}(\alpha|\beta)$ (i.e., reasonable to update one's probability for α to the value $P_{PRI}(\alpha|\beta)$), despite the fact that β is not inclusive of one's evidence bearing on α. A second point is that there are cases where it is correct to set $P_{POS}(\alpha)$ to $P_{PRI}(\alpha|\beta)$, even though it is not given that $P_{PRI}(\alpha|\beta)$ is identical to $P_{PRI}(\alpha|\wedge E)$. Having accepted these possibilities, I would also like to entertain the possibility that there are cases where it is reasonable to conclude that $P_{POS}(\alpha)\in R$ on the basis of $P_{PRI}(\alpha|\beta)\in R$, even though (1) β is not inclusive of one's evidence bearing on α, and (2) it is not given that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$. For the moment, one need not accept the possibility of such cases. However, I wish to introduce a catch-all expression to describe the relationship between β and α, in cases where it is correct to set $P_{POS}(\alpha)$ to $P_{PRI}(\alpha|\beta)$, or more generally when it is correct to conclude that $P_{POS}(\alpha)\in R$ on the basis of $P_{PRI}(\alpha|\beta)\in R$. In such cases, I will say that β is *sufficiently inclusive* of one's evidence bearing on $\alpha$.

The idea of conditioning on propositions that are sufficiently inclusive of one's relevant evidence (as opposed to conditioning on propositions that encapsulate all of one's evidence) is central to the approach to probability updating proposed in this article. The proposed system thus allows one to assign a posterior probability to a respective proposition α, in cases where the value of $P_{PRI}(\alpha|\wedge E)$ is not given, but the value of another prior probability $P_{PRI}(\alpha|\beta)$ is given, and it is reasonable to accept that β is sufficiently inclusive of one's evidence bearing on α. The proposed system also accommodates the possibility of conditionalizing on imprecise conditional probability statements, thereby allowing one to judge that $P_{POS}(\alpha)\in R$, when $P_{PRI}(\alpha|\beta)\in R$ is given, and it is reasonable to accept that β is sufficiently inclusive of one's evidence bearing on α. In cases where an agent has access to a complete prior probability function, the system that I propose prescribes posterior probabilities that are identical to the ones prescribed by the standard Bayesian approach. But in cases where an agent's prior probabilities are incomplete or imprecise (and the standard Bayesian approach is not applicable), the system still recommends reasonable posterior probability judgments.

Despite some important differences, the system of probability updating that I will propose is a close relative of the standard Bayesian system. Like the standard Bayesian system, the proposed system is very general, promising to reduce the prescriptions of rational credence formation to the prescription that one update one's personal probabilities by (a form of) conditionalization, and to prescriptions

[4] One could, of course, maintain the thesis that one is entitled to *infer* that $P_{PRI}(\alpha|\beta) = P_{PRI}(\alpha|\wedge E)$, in cases where it is reasonable to accept that β is inclusive of one's evidence bearing on $\alpha$ (cf. [37]). I regard this thesis as potentially suspect. Nevertheless, one could augment the system of probability updating that I propose, and adopt the conclusion that $P_{PRI}(\alpha|\wedge E) = P_{PRI}(\alpha|\beta)$, whenever the proposed system permits the conclusion that $P_{POS}(\alpha)\in R$ via conditionalization on a prior $P_{PRI}(\alpha|\beta)\in R$. However, since the inference to $P_{PRI}(\alpha|\wedge E) = P_{PRI}(\alpha|\beta)$, in such cases, is unnecessary and potentially suspect, the system of updating that I propose does not license such inferences.

concerning the choice of prior probabilities.[5] A common variety of the standard Bayesian framework prescribes only that an agent's priors be probabilistically coherent. Later in the article, I will sketch the advantage of accepting additional prescriptions on the choice of prior probabilities. The advantage lies in the possibility of encoding the prescriptions of direct inference and statistical induction within the proposed system via the choice of intuitively reasonable prior probabilities.

## 2 Preliminaries

I will use a, b, c, and d to represent *atoms* of a propositional language $\Phi$ (with the standard truth functional connectives), and use $\alpha$ and $\beta$ as metalogical variables ranging over sentences of $\Phi$. Given this simple language, I will proceed as if the input to the problem of probability updating is a 'knowledge base' K, consisting of a pair $\langle E_K, L_K \rangle$, where $E_K$ is a set of sentences of $\Phi$, and $L_K$ is a set of prior conditional probability statements of the form $P_{PRI}(\alpha|\beta) \in R$, where $\alpha$ and $\beta$ are sentences of $\Phi$, and R is a rigid designator for a set of real numbers. I will also use $\rho$ and $\sigma$ as metalogical variables ranging over prior conditional probability statements of the form $P_{PRI}(\alpha|\beta) \in R$. And, as shorthand, I will occasionally use expressions of the form $P_{PRI}(\alpha|\beta) = r$ in the place of $P_{PRI}(\alpha|\beta) \in \{r\}$, and expressions of the form $P_{PRI}(\alpha|\beta) \geq r$ in the place of $P_{PRI}(\alpha|\beta) \in [r, 1]$. For real life applications, it is clear that $\Phi$ would have to be replaced by a richer language. However, due to limitations in standard axiomatic approaches to probability theory, which treat probability as defined over a Boolean algebra or a propositional language, I will proceed by means of an appropriately simple language, and assume (especially in Section 7) that the account can be generalized to apply to richer languages (cf. [32]).

Facts about the consistency or inconsistency of various sets will play a role in determining which conclusions should be inferred from a knowledge base. I will speak both of the inconsistency of sets of sentences of $\Phi$, and of sets of probability statements. A set of sentences of $\Phi$ is inconsistent if and only if $a \wedge \neg a$ is a logical consequence of the set. A set of posterior probability statements, S, will be regarded as inconsistent if and only if there is no probability assignment, *P*, defined over $\Phi$, such that $\forall s \in S: \; s = P_{POS}(\alpha) \in R \Rightarrow P(\alpha) \in R$.[6]

[5] The standard Bayesian approach usually assumes that the updating of personal probabilities proceeds iteratively, where subsequent to an initial update (by conditionalization by appeal to an *initial* prior probability function), further updates proceed by appeal to a prior probability function that was generated by conditionalization on evidence that was collected at an earlier time (so that non-initial priors encode some of the agent's evidence). In contrast to this 'iterative' approach to probability updating, I assume that defeasible conditionalization always proceeds by conditionalization using the agent's *initial* prior probability function. In the case of Bayesian conditionalization, the usual iterative approach to probability updating is equivalent to the approach where an agent always sets his posteriors by conditionalization on his complete body of evidence using his *initial* prior probability function. This sort of equivalence does not hold for probability updating via defeasible conditionalization (absent some constraints on the set of possible prior probabilities). The equivalence between the two approaches to conditionalization also fails to hold, in general, for Jeffrey Conditionalization [13]. But see [48], where it is shown that Jeffrey's rule does commute across order for identical learning inputs, provided that "identical learning" is appropriately construed and formalized (cf. [28]).

[6] P is a probability assignment defined over $\Phi$ *if and only if* P maps all of the formulas of $\Phi$ into the interval [0, 1], so that Kolmogorov's Axioms are satisfied. That is: $\forall \alpha, \beta \in \Phi$: (1) $P(\alpha) \geq 0$, (2) if $\alpha$ is a tautology, then $P(\alpha) = 1$, and (3) if $\{\alpha, \beta\}$ is inconsistent, then $P(\alpha \vee \beta) = P(\alpha) + P(\beta)$. $P(\alpha|\beta)$ is defined as equal to $P(\alpha \wedge \beta)/P(\beta)$, if $P(\beta) \neq 0$ (otherwise $P(\alpha|\beta)$ is undefined).

Similarly, a set of prior probability statements, S, will be regarded as inconsistent if and only if there is no probability assignment, *P*, defined over Φ, such that $\forall s \in S: s = P_{PRI}(\alpha|\beta) \in R \Rightarrow P(\alpha|\beta) \in R$.

In addition to speaking of the logical consequences of sets of sentences of Φ, I will speak of the logical consequences of sets of probability statements. A probability statement, $P_{POS}(\alpha') \in R'$, will be described as a logical consequence of a set of posterior probability statements, S, if and only if for every probability assignment, *P*, defined over Φ, if $\forall s \in S: s = P_{POS}(\alpha) \in R \Rightarrow P(\alpha) \in R$, then $P(\alpha') \in R'$. Similarly, a prior probability statement, $P_{PRI}(\alpha'|\beta') \in R'$, will be described as a logical consequence of a set of prior probability statements, S, if and only if for every probability assignment, *P*, defined over Φ, if $\forall s \in S: s = P_{PRI}(\alpha|\beta) \in R \Rightarrow P(\alpha|\beta) \in R$, then $P(\alpha'|\beta') \in R'$.

Unless otherwise stated, I will assume that $E_K$ and $L_K$ are consistent and closed under logical consequences (though the system also performs reasonably in the case where $E_K$ and/or $L_K$ are inconsistent).

## 3 Defeasible Conditionalization

In the case where one's knowledge base is K, the principle of total evidence licenses the conclusion that $P_{POS}(\alpha) = r$, when $P_{PRI}(\alpha|\beta) = r \in L_K, \beta \in E_K$, and β is sufficiently inclusive of one's evidence bearing on α. Generalized to set-valued priors, the principle of total evidence licenses the conclusion that $P_{POS}(\alpha) \in R$, in the case where $P_{PRI}(\alpha|\beta) \in R \in L_K$, $\beta \in E_K$, and β is sufficiently inclusive of one's evidence bearing on α. So generalized, we may regard any inference from premises of the form $P_{PRI}(\alpha|\beta) \in R \in L_K$ and $\beta \in E_K$ to a conclusion of the form $P_{POS}(\alpha) \in R$ as a *defeasible inference*, where the inference is defeated if and only if β is not sufficiently inclusive of one's evidence bearing on $\alpha$.[7] The heart of this species of defeasible inference is encoded in the following schema.

**[d-cond] Defeasible Conditionalization**
$P_{PRI}(\alpha|\beta) \in R \in L_K$ and $\beta \in E_K$ provides a defeasible justification (or reason) for inferring (and believing) $P_{POS}(\alpha) \in R$ (for an agent whose knowledge base is K).

A defeasible justification that is generated in accordance with [d-cond] is, of course, defeated in the case where β is not sufficiently inclusive of an agent's evidence bearing on α. What we would like to know are the precise conditions under which an instance of [d-cond] is defeated, or, in effect, the conditions under which β is not sufficiently inclusive of an agent's evidence bearing on α. To begin with, it is clear that a corresponding instance of [d-cond] is defeated, when there exists a proposition, β′, such that β′ is in $E_K$, {β′} entails β, $P_{PRI}(\alpha|\beta') \in R'$ is in $L_K$, and $R \cap R' = \varnothing$. In other words:

[7] The notions of *defeasible reason*, *defeasible justification*, and *defeasible inference* were pioneered within academic philosophy by Hart [7], Chisholm [4], Toulmin [43], Pollock [31], and Rescher [36]. Similar ideas were independently developed by researchers in artificial intelligence, including [25, 26], and [35].

**[s-spec] Simple Specificity Defeat**
The justification for inferring $P_{POS}(\alpha)\in R$ from $P_{PRI}(\alpha|\beta)\in R \in L_K$ and $\beta \in E_K$ is defeated, if $\exists\beta',R'$: $\beta' \in E_K$, $\{\beta'\}$ entails $\beta$, $P_{PRI}(\alpha|\beta')\in R' \in L_K$, and $R\cap R' = \varnothing$.[8]

[s-spec] applies when a more inclusive survey of one's evidence supports a conclusion that conflicts with the one supported by a proposed instance of [d-cond]. For example, suppose that $L_K$ is the deductive closure of $\{P_{PRI}(a|b) = 0.1,\ P_{PRI}(a|b\wedge c) = 0.9\}$, and $E_K$ is the deductive closure of {b, c}. In this case [d-cond] provides a defeasible justification for inferring $P_{POS}(a) = 0.1$ (by appeal to $P_{PRI}(a|b) = 0.1 \in L_K$ and $b \in E_K$), and a defeasible justification for inferring $P_{POS}(a) = 0.9$ (by appeal to $P_{PRI}(a|b\wedge c) = 0.9 \in L_K$ and $b\wedge c \in E_K$). However, since $b\wedge c$ entails b, [s-spec] prescribes the defeat of the inference to $P_{POS}(a) = 0.1$.

While [s-spec] expresses a sufficient condition for the defeat of corresponding instances of [d-cond], [s-spec] does not characterize the full range of cases under which instances of [d-cond] are defeated. Indeed, there are cases where [d-cond] provides a defeasible justification for each element of an inconsistent set, while [s-spec] fails to dictate the defeat of any of the inferences leading to the elements of the inconsistent set. For example, suppose that $L_K$ is the deductive closure of $\{P_{PRI}(a|b) = 0.1,\ P_{PRI}(a|c) = 0.9\}$, and $E_K$ is the deductive closure of {b, c}. In that case, [d-cond] provides a defeasible justification for inferring $P_{POS}(a) = 0.1$ (by appeal to $P_{PRI}(a|b) = 0.1 \in L_K$ and $b \in E_K$), and a defeasible justification for inferring $P_{POS}(a) = 0.9$ (by appeal to $P_{PRI}(a|c) = 0.9 \in L_K$ and $c \in E_K$). While the present knowledge base yields a defeasible justification for two mutually inconsistent conclusions (namely $P_{POS}(a) = 0.1$ and $P_{POS}(a) = 0.9$), [s-spec] fails to dictate the defeat of either of inferences that lead to the inconsistency, since the neither of the propositions that underwrite the two inferences (namely, b and c) entails the other. Since [s-spec] will not always dictate the defeat of at least one instance of [d-cond] in the case where inference by [d-cond] leads to an inconsistent set of conclusions, we know that additional defeat conditions are called for.

## 4 Minimal Inconsistent Sets

In cases where a set is inconsistent and has no inconsistent proper subsets, the set is said to be *minimal inconsistent*. As a prelude to proposing a fully general account of the conditions under which instances of [d-cond] are defeated, I begin by considering hypothetical cases where $S = \{P_{POS}(\alpha_1)\in R_1, \ldots, P_{POS}(\alpha_n)\in R_n\}$ is the complete set of conclusions that are defeasibly justified via instances of [d-cond] from K, and S is minimal inconsistent. Such cases are impossible, on the assumption that $E_K$ and $L_K$ are closed under deductive

[8] Specificity rules are common in theories of default and defeasible reasoning (cf. [37, 38]). In some systems such as Pollock's [32], defeasible inferences are modeled after direct inference, where specificity rules correspond to a preference for narrower reference classes in the selection of a statistical probability statement for use as a premise for a direct inference. Appeal to specificity rules in direct inference traces to Venn [47], though the idea is also associated with Reichenbach [34] and Hempel [8]. The appeal to specificity proposed here is more in line with Carnap's principle of total evidence, since the type of probabilities involved in defeasible conditionalization are not statistical.

consequences. Nevertheless, it is instructive to consider such cases, so in the present section, I waive the assumption that $E_K$ and $L_K$ are deductively closed.

Let $S_{cond} = \{P_{PRI}(\alpha_1|\beta_1)\in R_1, \ldots, P_{PRI}(\alpha_n|\beta_n)\in R_n\}$ be the set of conditional probability statements that led to the elements of S via instances of [d-cond], and let $S_{pre}$ be the set $\{\beta_1, \ldots, \beta_n\}$ of 'preconditions' for the elements of $S_{cond}$. Although [s-spec] does not prescribe the defeat of any of the respective instances of [d-cond] that lead to S (so long as n is greater than two), we know that we must reject at least one element of S (and we know that at least one of the corresponding instances of [d-cond] is defeated).[9] While [s-spec] is unhelpful in addressing such cases, it is still reasonable to consider which elements of $S_{cond}$ have more informative preconditions, in deciding which instances of [d-cond] are defeated. As a means to characterizing the relative informativeness of pairs of propositions, I will say: $\alpha$ *strictly entails* $\beta$ if and only if $\{\alpha\}$ entails $\beta$, and $\{\beta\}$ does not entail $\alpha$. Then, to start with, the following principle is plausible:

(a) If S is the complete set of conclusions defeasibly inferable from K via [d-cond], and S is minimal inconsistent, then the defeasible justification for inferring $P_{POS}(\alpha)\in R$ from $P_{PRI}(\alpha|\beta)\in R \in L_K$ and $\beta\in E_K$ is *undefeated*, if $\exists\beta_i\in S_{pre}$: $\beta$ strictly entails $\beta_i$.

Principle (a) tells us that an instance of [d-cond] is undefeated provided that its associated precondition is more specific than the precondition of at least one of its competitors. The principle is reasonable for the range of cases under consideration. In such cases, the only potential reason for rejecting the conclusion under consideration, $P_{POS}(\alpha)\in R$, is that it is an element of a set of defeasible conclusions S, where S is minimal inconsistent. But then there is no *good* reason to reject the inference to $P_{POS}(\alpha)\in R$ in such cases, since one of the other elements of S (call the element "φ") is obtained without consideration of some evidence upon which the inference to $P_{POS}(\alpha)\in R$ is based, while the conclusion that φ is not based on any evidence that is not also taken account of in the justification for $P_{POS}(\alpha)\in R$. We may thus consider the inference to $P_{POS}(\alpha)\in R$ undefeated, since absent the inference to φ (which is based on strictly less evidence) there would be no problem with the inference to $P_{POS}(\alpha)\in R$.

The guiding assumption that underlies (a) is that the only condition that may result in the defeat of an instance of [d-cond] is that its conclusion is inconsistent with the conclusions of other instances of [d-cond] supported by a respective knowledge base. This guiding assumption is reflected in the fact that (a) deems an instance of [d-cond] as undefeated provided there is some other instance of [d-cond] (based on a less inclusive survey of the available evidence) that can be isolated as the source of the trouble. Given its guiding assumption, one possible objection to (a) cites the phenomena of *undercutting defeaters*. An undercutting defeater for a defeasible inference is a defeater which defeats the defeasible inference without providing a reason for believing the negation of the conclusion of the defeated inference (cf. [33]).

[9] If [s-spec] is applicable in defeating an instance of [d-cond] that is supported by a given knowledge base K, then there is a pair of posterior probabilities $P_{POS}(\alpha)\in R$ and $P_{POS}(\alpha)\in R'$, where $R\cap R' = \varnothing$, and $P_{POS}(\alpha)\in R$ and $P_{POS}(\alpha)\in R'$ are defeasibly justified via instances of [d-cond] from K. But $\{P_{POS}(\alpha)\in R, P_{POS}(\alpha)\in R'\}$ is minimal inconsistent. So if S is the complete set of conclusions that are defeasibly justified via instances of [d-cond] from K, S is minimal inconsistent, and $|S| > 2$, then [s-spec] does not prescribe the defeat of any of the respective instances of [d-cond] that led to S.

According to the proposed objection to (a), an instance of [d-cond] may be defeated for some reason other than that it is jointly responsible in generating an inconsistency.

The positing of the phenomena of undercutting defeaters is typically justified by appeal to examples. According to one example, something's *appearing to be red* provides one with a defeasible reason for thinking that it *is red*. However, if one learns that an object is illuminated by red light, then this defeats, and supposedly *undercuts,* one's reason for thinking that the object is red (assuming one knows that most objects appear red when illuminated by red light). As described, the present example exemplifies the phenomena of undercutting defeat assuming: (1) knowing that an object is illuminated by red light defeats one's reasons for thinking that the object is red, and (2) knowing that an object is illuminated by red light does not provide a reason for denying that the object is red.

As it turns out, the framework of defeasible inference proposed here is equipped to accommodate the sort of examples offered in favor of the phenomena of undercutting defeat. Crucially, such examples can be represented within the proposed framework as cases that do not involve undercutting defeat, thereby eliminating the justification for positing undercutting defeaters, and the plausibility of the proposed objection to (a). For example, we can represent the situation of an object illuminated by red light via a knowledge base containing the following priors:

$P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object}) \in [0.01, 0.02]$
$P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object} \wedge \phi \text{ appears to be red}) \in [0.99, 0.995]$
$P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object} \wedge \phi \text{ appears to be red} \wedge \phi \text{ is illuminated by red light}) \in [0.01, 0.02]$

In the case where it is given that $\phi$ is an object (and it is not given that $\phi$ appears to be red), the preceding priors support the conclusion that $\phi$ is not red (or, more precisely, that the probability is very low that $\phi$ is red). In the situation where it is given that $\phi$ is an object and $\phi$ appears to be red (and it is not given that $\phi$ is illuminated by red light), the priors support the conclusion that $\phi$ is red (with high probability). On the other hand, the latter conclusion is defeated, in the situation where it is also given that $\phi$ is illuminated by red light. In the latter situation, we may assume that $E_K$ is the deductive closure of $\{\phi$ is an object, $\phi$ appears to be red, $\phi$ is illuminated by red light$\}$. In that case, the inference to $P_{POS}(\phi \text{ is red}) \in [0.01, 0.02]$ based on $P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object}) \in [0.01, 0.02]$ is defeated via [s-spec] by appeal to $P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object} \wedge \phi \text{ appears to be red}) \in [0.99, 0.995]$, and the inference to $P_{POS}(\phi \text{ is red}) \in [0.99, 0.995]$ based on $P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object} \wedge \phi \text{ appears to be red}) \in [0.99, 0.995]$ is defeated via [s-spec] by appeal to $P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object} \wedge \phi \text{ appears to be red} \wedge \phi \text{ is illuminated by red light}) \in [0.01, 0.02]$. Thus, we are permitted to infer that $P_{POS}(\phi \text{ is red}) \in [0.01, 0.02]$ by appeal to $P_{PRI}(\phi \text{ is red} \mid \phi \text{ is an object} \wedge \phi \text{ appears to be red} \wedge \phi \text{ is illuminated by red light}) \in [0.01, 0.02]$.

The preceding reconstruction illustrates the manner in which the present framework accommodates the type of examples that motivate the positing of undercutting defeaters. The examples are accommodated via the expressive power of probability assignments and judgments about probabilities. The sort of example that the framework does not accommodate (as it stands) is one where an instance of [d-cond] is defeated in the absence of a prior probability that supports an instance of [d-cond] to a

conflicting conclusion. But the demand to accommodate this kind of example has no intuitive purchase, since the proposed account of probability updating already accommodates the kinds of example that have been offered in favor of the phenomena of undercutting defeat.[10]

As it turns out, I think that (a) characterizes the full set of instances of [d-cond] that are undefeated, in the case where S is the complete set of conclusions defeasibly inferable from K via [d-cond], and S is minimal inconsistent. To see the reason, consider, in the abstract, the set of instances of [d-cond] that are not classified as *undefeated* by (a). It appears that there is no sound means by which we may exonerate any of these inferences. For this reason, it is advisable to conclude that each instance of [d-cond] that is not classified as *undefeated* by (a) is defeated. The present conclusion is encapsulated by the following principle:

(b) If S is the complete set of conclusions defeasibly inferable from K via [d-cond], and S is minimal inconsistent, then the defeasible justification for inferring $P_{POS}(\alpha)\in R$ from $P_{PRI}(\alpha|\beta)\in R \in L_K$ and $\beta \in E_K$ is defeated if and only if $\forall\beta_i\in S_{pre}$: $\beta$ does not strictly entail $\beta_i$.

It may be objected here that, among the inferences that are classified as *defeated* by (b), there is an intuitive difference between those that are based on a prior, $P_{PRI}(\alpha|\beta)\in R$, for which there exists a $\beta_i$ in $S_{pre}$ such that $\beta_i$ strictly entails $\beta$, and those for which there exists no such $\beta_i$. The former inferences have the shortcoming that their precondition is less inclusive of the available evidence than the precondition of another instance of [d-cond] that leads to an element of S. Despite this shortcoming of the former instances of [d-cond], there is little to be said in favor of the latter instances of [d-cond], and there is no positive argument of the sort that justified (a) that can be invoked in their favor. For this reason, I recommend a cautious approach in classifying instances of [d-cond] as undefeated, and assume that (b) is correct for the range of cases under consideration.[11]

[10] Nevertheless, it would not be difficult to accommodate *genuine* instances of undercutting defeat. To do so, we would allow that $L_K$ include statements of a three place relation of the form $\otimes(\chi, \alpha, \beta)$, which asserts that $\chi$ is an undercutting defeater for instances of [d-cond] based on priors of the form $P_{PRI}(\alpha|\beta)\in R$ (cf. [33]). In the case where $\chi$ is in $E_K$, instances of [d-cond] based on such priors, $P_{PRI}(\alpha|\beta)\in R$, would thereby be defeated, and removed from consideration prior to the application of the defeat conditions proposed in the body of this paper. The defeat of such priors would be handled by modifying the definition of *Triggered*(K) (in Section 6) as follows: $Triggered(K) = \{P_{PRI}(\alpha|\beta)\in R \mid P_{PRI}(\alpha|\beta)\in R \in L_K \;\&\; \beta \in E_K \;\&\; P_{PRI}(\alpha|\beta)\in R \notin Redundant(K) \;\&\; \forall\chi\in E_K: \otimes(\chi,\alpha,\beta) \notin L_K\}$.

[11] A less cautious system of defeasible conditionalization (that accepts the importance of the distinction between the former and latter inferences) would accept the following defeat condition: If S is the complete set of conclusions defeasibly inferable from K via [d-cond], and S is minimal inconsistent, then the defeasible justification for inferring $P_{POS}(\alpha)\in R$ from $P_{PRI}(\alpha|\beta)\in R \in L_K$ and $\beta \in E_K$ is undefeated *if and only if* $\exists\beta_i\in S_{pre}$: $\beta$ strictly entails $\beta_i$, or ($\forall\beta_i\in S_{pre}$: $\beta_i$ does not strictly entail $\beta$, and $\exists\beta_j,\beta_k\in S_{pre}$: $\beta_j$ strictly entails $\beta_k$). This less cautious approach to probability updating could then be generalized in a manner analogous to the cautious approach to probability updating proposed in Section 6, so that $\rho$ is preferred in $\Gamma \Leftrightarrow$ ($\exists\sigma$: $\sigma\in\Gamma$ & *Precondition*($\rho$) strictly implies *Precondition*($\sigma$)) or ($\forall\sigma\in\Gamma$:*Precondition*($\sigma$) does not strictly entail *Precondition*($\rho$) & $\exists\sigma_j,\sigma_k\in\Gamma$: *Precondition*($\sigma_j$) strictly entails *Precondition*($\sigma_k$)). For all knowledge bases, the set of conclusions licensed by the presently described system of defeasible conditionalization is a superset of the set of conclusions licensed by the system of defeasible conditionalization described in the body of this article.

The principles (a) and (b) are intended to apply to cases where S is the complete set of conclusions inferable from K via [d-cond], and S is minimal inconsistent. For such cases, the two principles are complete in the sense that they partition the set of inferences leading to the elements of S into those that are defeated, and those that are undefeated. For such cases, adherence to the two principles also guarantees that the use of defeasible conditionalization will not generate mutually inconsistent conclusions, since the two principles insure the defeat of at least one of the instances of [d-cond] leading to the elements of S.

Although principles (a) and (b) do not appear among the general defeat conditions that I propose in Section 6, the principles are instructive, since they capture some of the key intuitions that underlie the general defeat conditions. A central idea implemented within the general defeat conditions is that of evaluating the defeat and non-defeat of instances of [d-cond] *relative to sets of priors* that support inferences to minimal inconsistent sets of conclusions. In particular, the general defeat conditions treat an instance of [d-cond] as defeated if it is defeated in the manner of (b) *relative to any set of priors* that support instances of [d-cond] to a minimal inconsistent set of conclusions.

Before proceeding to propose general defeat conditions for instances of [d-cond], the following section addresses two types of prior probability statement that require special treatment.

## 5 Derivative and Partially Derivative Priors

Notwithstanding the preceding discussion, the system of probability updating proposed in the present article assumes that $E_K$ and $L_K$ are deductively closed. This introduces some issues not faced by many standard systems of defeasible inference. Consider the following example:

**Example 1**
Let $E_K$ be the deductive closure of {a, b}, and let $L_K$ be the deductive closure of $\{P_{PRI}(c|a) \geq 0.9,\ P_{PRI}(d|b) \geq 0.9,\ P_{PRI}(\neg c|a \wedge b) = 1\}$.

In the present case, the set of conclusions (defeasibly) licensed via [d-cond] includes: $P_{POS}(c) \geq 0.9$, $P_{POS}(d) \geq 0.9$, and $P_{POS}(\neg c) = 1$. The conclusion that $P_{POS}(c) \geq 0.9$ is inconsistent with the conclusion that $P_{POS}(\neg c) = 1$, and for this reason it is clear that we should regard the inference to $P_{POS}(c) \geq 0.9$ as *defeated* (since the inference to $P_{POS}(\neg c) = 1$ is based on the broadest possible survey of the available evidence). On the other hand, the inference to $P_{POS}(d) \geq 0.9$ is unproblematic. So K appears to provide an undefeated justification for inferring $P_{POS}(d) \geq 0.9$. But there is a difficulty here, arising from the assumption that $L_K$ is closed under logical consequences. The difficulty derives from the fact that, for all $\alpha$, $L_K$ entails $P_{PRI}(\alpha \vee c|a) \geq 0.9$. On the basis of such priors, and in the absence of special provisions, [d-cond] would license inference to conclusions of the form $P_{POS}(\alpha \vee c) \geq 0.9$, for all $\alpha$. And if such inferences are licensed, K would thereby generate a defeasible justification for inferring $P_{POS}(\alpha) \geq 0.9$, for all $\alpha$ (since K also licenses acceptance of $P_{POS}(\neg c) = 1$). It is plausible to think that the chains of inference leading to conclusions of the form $P_{POS}(\alpha) \geq 0.9$ (for the respective $\alpha$) are mutually defeating. But absent some

special measures, the inference to the conclusion that $P_{POS}(d) \geq 0.9$ would also be defeated, since the inference to $P_{POS}(d) \geq 0.9$ would be in conflict with a host of conclusions inferable from various instances of $P_{PRI}(\alpha \vee c|a) \geq 0.9$ (such as the instance $P_{PRI}(\neg d \vee c|a) \geq 0.9$, for example).

Inference to the conclusion that $P_{POS}(\alpha \vee c) \geq 0.9$, for varied $\alpha$, in Example 1, proceeds from the prior $P_{PRI}(\alpha \vee c|a) \geq 0.9$. But the conclusion also *indirectly* derives from the prior $P_{PRI}(c|a) \geq 0.9$. In this case, I propose that the deductive connection between $P_{PRI}(c|a) \geq 0.9$ and $P_{PRI}(\alpha \vee c|a) \geq 0.9$ vitiates [d-cond] based on the latter prior, since [d-cond] based on the former prior is defeated. And, in general, I propose that an instance of [d-cond] is defeated when (1) the prior from which it proceeds is 'derivative' of another prior, and (2) the instance of [d-cond] proceeding from the latter prior is defeated. The following definition specifies when one prior counts as derivative of another:

**Definition** $P_{PRI}(\alpha|\beta) \in R$ *is derivative of* $P_{PRI}(\alpha'|\beta) \in R'$ if and only if (1) $\alpha'$ strictly implies $\alpha$, and $R = R'$, or (2) $\alpha'$ is $\alpha$, and $R' \subset R$.

The application of condition (1), of the preceding definition, is used in handling the sort of problem generated by Example 1. The application of condition (2) yields the result that instances of [d-cond] based on priors of the form $P_{PRI}(\alpha|\beta) \in R$ are defeated, when another instance of [d-cond] based on a prior $P_{PRI}(\alpha|\beta) \in R'$ is defeated, and $R' \subset R$. The latter policy is correct, since the defeat of an instance of [d-cond] based on a prior $P_{PRI}(\alpha|\beta) \in R'$, and evidence $\beta$, undermines all instances of [d-cond] to conclusions regarding $\alpha$ based on evidence $\beta$ (assuming that correct instances of [d-cond] are meant to be underwritten by the principle of total evidence).[12]

Within the general defeat conditions proposed in Section 6, I treat all instances of [d-cond] based on priors that are derivative of some other element of $L_K$ as *void*, where such void instances of [d-cond] are removed from consideration *prior* to considering what conclusions to draw on the basis of other instances of [d-cond] (thereby permitting the inference to $P_{POS}(d) \geq 0.9$ to go through unmolested, in Example 1). The policy of treating all derivative priors as void is reasonable, since 'voided' priors are indirectly represented by the priors from which they are derivative, within the 'normal' procedure for determining which instances of [d-cond] are defeated. The set of priors that are derivative of some other element of $L_K$ (of a respective knowledge base K) are collected according to the following definition.

**Definition** $Redundant(K) = \{P_{PRI}(\alpha|\beta) \in R \mid P_{PRI}(\alpha|\beta) \in R$ *is derivative of* $P_{PRI}(\alpha'|\beta) \in R' \,\&\, P_{PRI}(\alpha'|\beta) \in R' \in L_K\}$.

In addition to derivative priors, there are priors that I call "partially derivative". Consider the following example:

[12] A significantly stronger system of defeasible conditionalization results, if we eliminate condition (2) of the proposed definition. While I believe that the resulting system is reasonable, I will not defend it here.

**Example 2**
Let $E_K$ be the deductive closure of {a, b}, and let $L_K$ be the deductive closure of $\{P_{PRI}(c|a) \geq 0.8,\ P_{PRI}(c \vee d|a) \geq 0.9, P_{PRI}(\neg c|a \wedge b) \geq 0.8\}$.

In this case, the set of conclusions licensed by [d-cond] includes: $P_{POS}(c) \geq 0.8,\ P_{POS}(c \vee d) \geq 0.9$, and $P_{POS}(\neg c) \geq 0.8$. While we should regard the inference to $P_{POS}(c) \geq 0.8$ as defeated, and the inference to $P_{POS}(\neg c) \geq 0.8$ as undefeated, it is less clear what one should believe regarding the range of possible values for $P_{POS}(c \vee d)$. While $P_{PRI}(c \vee d|a) \geq 0.9$ permits a defeasible inference to $P_{POS}(c \vee d) \geq 0.9$, and $P_{POS}(c \vee d) \geq 0.9$ is *consistent* with the conclusion that $P_{POS}(\neg c) \geq 0.8$, we know that the closely related inference to $P_{POS}(c) \geq 0.8$ is defeated. In fact, had $L_K$ contained only $P_{PRI}(c \vee d|a) \geq 0.8$ (and not $P_{PRI}(c \vee d|a) \geq 0.9$), we would have concluded that it was impossible to draw a justified conclusion regarding the range of possible values for $P_{POS}(c \vee d)$, save that $P_{POS}(c \vee d) \in [0, 1]$ (since $P_{PRI}(c \vee d|a) \geq 0.8$ is derivative $P_{PRI}(c|a) \geq 0.8$, and our reason for accepting $P_{POS}(c \vee d) \geq 0.8$ should stand or fall with our reason for accepting $P_{POS}(c) \geq 0.8$).

I call priors such as $P_{PRI}(c \vee d|a) \geq 0.9$, within Example 2, "partially derivative", since their content is, in some sense, *partially* captured by some other prior (as the content of $P_{PRI}(c \vee d|a) \geq 0.9$ is partly captured by the content of $P_{PRI}(c|a) \geq 0.8$). The following definition specifies when one prior counts as partially derivative of another:

**Definition** $P_{PRI}(\alpha|\beta) \in R$ *is partially derivative of* $P_{PRI}(\alpha'|\beta) \in R'$ if and only if $\alpha'$ strictly implies $\alpha$, and $R \subset R'$.

In determining what may be inferred from a partially derivative prior, I propose that one assume *as far as possible* that the content of the partially derivative prior is derivative of the prior that partially captures its content. I thereby propose, in cases where [d-cond] based on a 'partially deriving' prior is defeated, that the greatest lower posterior probability bound that one accepts on the basis of a corresponding partially derivative prior be $r_L - s_L$, where $r_L$ is the greatest lower bound for the partially derivative prior (in $L_K$), and $s_L$ is the greatest lower bound for the partially deriving prior (in $L_K$). In Example 2, this means that the conclusion that $P_{POS}(c \vee d) \geq 0.1$ is undefeated. Similarly, in cases where [d-cond] based on a 'partially deriving' prior is defeated, I propose that the least upper posterior probability bound that one accepts on the basis of a corresponding partially derivative prior be $1 - (s_U - r_U)$, where $r_U$ is the least upper bound for the partially derivative prior (in $L_K$), and $s_U$ is the least upper bound for the partially deriving prior (in $L_K$).

In the next section, I provide a full and precise treatment of derivative and partially derivative priors, while articulating general defeat conditions for instances of [d-cond].

## 6 General Defeat Conditions

For the purpose of describing general defeat conditions for instances of [d-cond], it is convenient to have an adjective that describes those elements of $L_K$

that may serve as a premise to [d-cond], in a given situation. To fill the desired role, I will say that $P_{PRI}(\alpha|\beta)\in R$ is "triggered" in K just in case $P_{PRI}(\alpha|\beta)\in R \in L_K$, $\beta \in E_K$, and $P_{PRI}(\alpha|\beta)\in R \notin \mathit{Redundant}(K)$ (cf. [9, 10]). Note that by treating a prior as triggered only if that prior is not redundant, we thereby treat derivative priors as untriggered and as inappropriate bases for instances of [d-cond]. For ease of reference, it is also convenient to have notation that refers to the set of conditional probability statements that are triggered (relative to a given K).

**Definition** $\mathit{Triggered}(K) = \{ P_{PRI}(\alpha|\beta)\in R \mid P_{PRI}(\alpha|\beta)\in R \in L_K \;\&\; \beta \in E_K \;\&\; P_{PRI}(\alpha|\beta)\in R \notin \mathit{Redundant}(K)\}$

In assessing instances of [d-cond], in the remainder of this article, I will treat the triggered elements of $L_K$ as the bearers of praise and blame, i.e., as being undefeated or defeated. Use of this shorthand is possible, since a prior conditional probability statement, $P_{PRI}(\alpha|\beta)\in R$, that is used as a premise for an instance of [d-cond], encodes all of the features of that instance of [d-cond] (as the next two definitions illustrate).

In the case where a conditional probability, $P_{PRI}(\alpha|\beta)\in R$, is triggered, it is useful to have special notation to refer to the element of $E_K$ that is responsible for 'triggering' that conditional probability.

**Definition** $\mathit{Precondition}(P_{PRI}(\alpha|\beta)\in R) = \beta$.

It is also useful to have notation for referring to the conclusion for which a triggered prior provides a defeasible reason.

**Definition** $\mathit{Conclusion}(P_{PRI}(\alpha|\beta)\in R) = P_{POS}(\alpha)\in R$.

The set of conclusions corresponding to a set, $\Gamma$, of prior conditional probabilities will be denoted as follows.

**Definition** $\mathit{Conclusions}(\Gamma) = \{P_{POS}(\alpha)\in R | P_{PRI}(\alpha|\beta)\in R \in \Gamma\}$.

Building on these definitions, I introduce notation in order to refer to the subsets of the set of triggered priors whose conclusions form minimal inconsistent sets.

**Definition** $\mathit{min\text{-}incon}(K) = \{\Gamma \mid \Gamma \subseteq \mathit{Triggered}(K)\ \&\ \mathit{Conclusions}(\Gamma)\ \text{is minimal inconsistent}\}$.

Note that (for all K) the defeat of at least one element of each element of *min-incon* (K) is both *necessary* and *sufficient* for insuring that the set of conclusions of undefeated instances of [d-cond], based on K, is consistent.

The defeat conditions that I will shortly propose reflect the idea that prior conditional probabilities with logically weaker preconditions receive priority when determining which instances of [d-cond] are defeated. The following notion of *preference* is central to the mechanism used for classifying instances of [d-cond] as defeated and undefeated. (Recall that $\rho$ and $\sigma$ are metalogical variables ranging over statements of the form $P_{PRI}(\alpha|\beta)\in R$.)

**Definition** $\rho$ is preferred in $\Gamma \Leftrightarrow \exists\sigma : \sigma \in \Gamma$ & *Precondition*$(\rho)$ strictly implies *Precondition*$(\sigma)$.

**Definition** *Preferred*$(\Gamma) = \{\rho | \rho$ is preferred in $\Gamma\}$.

Setting aside the treatment of partially derivative priors for the moment, a prior is deemed to be *1-defeated* relative to an element of *min-incon*(K) according to the following definition, which generalizes the ideas of Section 4 in the obvious way.

**Definition** $\rho$ is 1-defeated relative to $\Gamma \Leftrightarrow \rho \in \Gamma$ & $\rho \notin$ *Preferred*$(\Gamma)$.

Generalizing ideas from Section 5, the notion of *2-defeat* is introduced in order to properly handle partially derivative priors.

**Definition** $P_{PRI}(\alpha|\beta) \in R$ is 2-defeated relative to $\Gamma \Leftrightarrow \exists \rho = P_{PRI}(\alpha'|\beta) \in R'$: $P_{PRI}(\alpha|\beta) \in R$ is partially derivative of $\rho$, $\rho$ is 1-defeated in $\Gamma$, and

$R \subset [r_L - s_L,\ 1 - (s_U - r_U)]$, where

$r_L = \sup\{r | \exists S : P_{PRI}(\alpha | \beta) \in S \in L_K\ \&\ S \cap [0,\ r] = \emptyset\}$,

$s_L = \sup\{s | \exists S : P_{PRI}(\alpha' | \beta) \in S \in L_K\ \&\ S \cap [0,\ s] = \emptyset\}$,

$r_U = \inf\{r | \exists S : P_{PRI}(\alpha | \beta) \in S \in L_K\ \&\ S \cap [r,\ 1] = \emptyset\}$, and

$s_U = \inf\{s | \exists S : P_{PRI}(\alpha' | \beta) \in S \in L_K\ \&\ S \cap [s,\ 1] = \emptyset\}$.[13]

By combining the definitions of 1-defeat and 2-defeat, we have the conditions under which a prior probability statement (understood as representing a corresponding instance of [d-cond]) is defeated relative to a knowledge base:

**Definition** $\rho$ is defeated relative to K $\Leftrightarrow \exists \Gamma \in$ *min-incon*(K): $\rho$ is 1-defeated in $\Gamma$ or $\rho$ is 2-defeated in $\Gamma$.

The set of prior conditional probabilities that are not defeated in K is defined as follows.

**Definition** *Undefeated*$(K) = \{\rho | \rho \in$ *Triggered*$(K)$ & $\rho$ is not defeated in $K\}$.

Among other virtues, it is guaranteed that the set of conclusions of instances of [d-cond] based on elements of *Undefeated*(K) is consistent.

[13] A stronger system of defeasible conditionalization (complementary to the suggestion made in footnote 11), would restate the third conjunct of the proposed definiens as: $\{R \subset [r_L - s_L + t_L,\ 1 - (s_U - r_U) - (1 - t_U)]$, where $t_L = \sup\{t | \exists S : P_{PRI}(\alpha' | \beta) \in S$ is not 1-defeated in $(\Gamma - P_{PRI}(\alpha' | \beta) \in R\}) \cup \{P_{PRI}(\alpha' | \beta) \in S\}$ & $S \cap [0,\ t] = \emptyset\}$, and $t_U = \inf\{t | \exists S : P_{PRI}(\alpha' | \beta) \in S$ not 1-defeated in $(\Gamma - \{P_{PRI}(\alpha' | \beta) \in R\}) \cup \{P_{PRI}(\alpha' | \beta) \in S\}$ & $S \cap [t,\ 1] = \emptyset\}$.

**Theorem** ∀K: *Conclusions*(*Undefeated*(K)) is consistent.

*Proof* The preceding theorem follows immediately from the definition of *1-defeat*. That definition guarantees (for all K) that at least one element of each element of *min-incon*(K) is *1-defeated* (meaning that *Conclusions*(*Undefeated*(K)) has no minimal inconsistent subsets, and thus no inconsistent subsets). (Note that the present theorem holds even if $E_K$ and/or $L_K$ are inconsistent.)

The following establishes that the proposed system assigns the correct posterior probabilities to the deductive consequences of $E_K$.

**Theorem** ∀K,α: $E_K$ entails $\alpha \Rightarrow P_{POS}(\alpha) \in \{1\} \in$ *Conclusions*(*Undefeated*(K))).

*Proof* It is sufficient to see that (for all α in $E_K$) $P_{PRI}(\alpha|\wedge E_K) \in \{1\} \in L_K$ (since $L_K$ is deductively closed), and $P_{PRI}(\alpha|\wedge E_K) \in \{1\} \in$ *Undefeated*(K) (assuming that $E_K$ and $L_K$ are consistent).

More generally, it is clear that if $P_{PRI}$ is complete, then the system delivers the same posteriors as the standard Bayesian approach, for in such cases we have, for all α in Φ, that there exists some r such that $P_{PRI}(\alpha|\wedge E_K) \in \{r\}$ is $L_K$, and that the instances of [d-cond] based on such priors are undefeated.

The set of logical consequences of *Conclusions*(*Undefeated*(K)) is also a superset of the set of conclusions licensed by what may be regarded as the most obvious generalization of the standard Bayesian approach (where the generalization is given for the purpose of updating on incomplete and/or imprecise priors). Given a knowledge base, K, the obvious generalization of the Bayesian approach proposes that an agent accept $P_{POS}(\alpha) \in R$ just in case $P_{PRI}(\alpha|\wedge E_K) \in R$ is entailed by $L_K$. I will call this approach to probability updating "Simple Generalized Bayesianism", and call the set of posterior probability statements licensed by Simple Generalized Bayesianism, for a given knowledge base K, SBayes(K).

**Definition** $\text{SBayes(K)} = \{P_{POS}(\alpha) \in R | P_{PRI}(\alpha|\wedge E_K) \in R \text{ is entailed by } L_K\}$.

**Theorem** ∀K: $\text{SBayes(K)} \subseteq \{P_{POS}(\alpha) \in R | P_{POS}(\alpha) \in R$ is a logical consequence of *Conclusions*(*Undefeated*(K))$\}$.

*Proof* It is sufficient to see, for all posteriors, $P_{POS}(\alpha) \in R$, that are elements of SBayes (K), that $P_{PRI}(\alpha|\wedge E_K) \in R$ is in $L_K$ (since $P_{PRI}(\alpha|\wedge E_K) \in R$ is entailed by $L_K$), and that instances of [d-cond] based on priors of the form $P_{PRI}(\alpha|\wedge E_K) \in R$ are always undefeated (assuming that $E_K$ and $L_K$ are consistent).

The system of probability updating proposed here consists in [d-cond], an inference schema, along with a specification of the conditions under which instances of [d-cond] are defeated. I propose that undefeated instances of [d-cond] provide undefeated reasons for belief for agents possessed of corresponding knowledge bases, and that agents with sufficient deductive abilities should believe the conclusions of

instances of [d-cond] that are undefeated relative to their respective knowledge bases.[14]

The principle of total evidence prescribes that one take account of all one's relevant evidence in making judgments of probability. Applied to the problem of arbitrating between conflicting reasons for belief, the proposed system of probability updating utilizes a principle that generalizes the principle of total evidence. The generalized principle prescribes that one favor defeasible inferences that take account of more of one's available evidence. Armed with this 'specificity principle', the proposed system of probability updating represents a cautious approach to defeasible inference.[15] The system licenses acceptance of a conclusion only if that conclusion is defeasibly justified, and no jointly conflicting conclusions of equal or greater status are defeasibly justified (where the status of such conclusions is determined by the specificity principle). While the proposed system represents a cautious approach to probability updating, the system also licenses inferences that are bolder than the ones licensed by Simple Generalized Bayesianism. The proposed system thereby represents an attractive approach to probability updating with incomplete priors.[16]

## 7 Direct Inference and Statistical Induction

In addition to representing an attractive approach to probability updating with incomplete priors, the proposed system of updating offers a promising framework for systematically representing the prescriptions of direct inference and statistical induction within a very general framework of rational credence formation.

An account of statistical induction codifies and explains the justificatory basis of inferences that move from a premise, describing the incidence of some characteristic among a sample, to a conclusion that states that the incidence of the chosen characteristic among a respective population (from which the sample was drawn) is likely to be very similar to its incidence among the sample. Schematically, such defeasible inferences may be represented as follows:

$$\text{From } S \subseteq G \text{ and } \text{freq}(F|S) = r \text{ infer that } P_{POS}(\text{freq}(F|G) \approx r) \text{ is high.}$$

An underappreciated fact about statistical induction is its 'reducibility' to so called "direct inference" (cf. [18–20, 27, 40, 50]). A direct inference proceeds from a

[14] Defeated instances of [d-cond] do not provide undefeated reasons for belief for agents possessed of corresponding knowledge bases. But, in some cases, there may be distinct instances of [d-cond] supporting the same conclusion, and while one such instance may be defeated, another may be undefeated. So it is possible for an agent to have a defeated reason for accepting a posterior probability, while at the same time possessing an undefeated reason for accepting the very same posterior probability.

[15] I use the term "cautious" here, in order to avoid confusion. I would have liked to use the term "skeptical", in order to make an explicit connection to the distinction between *skeptical* and *credulous reasoners*, as discussed in [44].

[16] I leave the possibility open that a suitably ideal agent (with a knowledge base K) might have reason to accept a proper superset of the set of logical consequences of *Conclusions*(*Undefeated*(K)). One possibility is that there are cases where an inference to $P_{POS}(\beta)\in R$ is defeated, but it is still reasonable to accept that $P_{POS}(\beta)\in R'$, for some set of values of R′, where $R \subset R'$.

premise stating that the frequency[17] with which members of a given reference class, F, are members of a respective target class, G, is r, and a premise stating that a given object, ϕ, is an element of F, and yields the conclusion that the probability that ϕ is a member of G is r. Schematically, this sort of defeasible inference may be represented as follows:

$$\text{From } \phi \in \text{F and } \text{freq}(\text{G}|\text{F}) = \text{r infer that } \text{P}_{\text{POS}}(\phi \in \text{G}) = \text{r}.$$

The reduction of statistical induction to direct inference proceeds from a theorem that describes the propensity of subsets of a set to resemble the set regarding the incidence of any characteristic.

**Theorem** $\forall \text{F,G} : \forall \text{u,v} > 0 : \exists \text{n} : |\omega| > |\text{F}| > \text{n} \Rightarrow \text{freq}(\text{freq}(\text{G}|\text{x}) \approx_{\text{u}} \text{freq}(\text{G}|\text{F})|\text{x} \subseteq \text{F}\ ) > 1 - \text{v}.$[18]

The present theorem is applicable in generating the major premises for direct inferences of the following sort.

$$\text{From S} \subseteq \text{F and } \text{freq}(\text{freq}(\text{G}|\text{x}) \approx \text{freq}(\text{G}|\text{F})|\text{x} \subseteq \text{F}) \approx 1 \text{ infer that}$$

$$\text{P}_{\text{POS}}(\text{freq}(\text{G}|\text{S}) \approx \text{freq}(\text{G}|\text{F})) \approx 1.$$

And from the conclusion that $\text{P}_{\text{POS}}(\text{freq}(\text{G}|\text{S}) \approx \text{freq}(\text{G}|\text{F})) \approx 1$, one may deduce that $\text{P}_{\text{POS}}(\text{freq}(\text{G}|\text{F}) \approx \text{r}) \approx 1$, assuming that one knows that freq(G|S) = r (where S is one's sample of observed Fs). If we suppress mention of the premise $\text{freq}(\text{freq}(\text{G}|\text{x}) \approx \text{freq}(\text{G}|\text{F})|\text{x} \subseteq \text{F}) \approx 1$, then the preceding yields the following form of statistical induction:

$$\text{From S} \subseteq \text{F and } \text{freq}(\text{G}|\text{S}) = \text{r infer that } \text{P}_{\text{POS}}(\text{freq}(\text{G}|\text{F}) \approx \text{r}) \approx 1.$$

So far I have described the manner in which statistical induction is, in some sense, reducible to direct inference. The next task is to show how and why the account of probability updating proposed in this article is apt for encoding the prescriptions of direct inference (and thereby the prescriptions of statistical induction). The proposal is to encode the prescriptions of direct inference via prior conditional probability statements that are akin to Lewis's principal principle [23]. In order to correctly express the content of such priors, I now proceed upon the assumption that the language, Φ, is rich enough to express relative frequencies, statements of set membership, and statements of set inclusion:

(a) $\text{P}_{\text{PRI}}(\phi \in \text{G}|\phi \in \text{F} \wedge \text{freq}(\text{G}|\text{F}) = \text{r}) = \text{r}.$

Assuming that the preceding prior is an element of $\text{L}_{\text{K}}$, one may apply [d-cond] to infer that $\text{P}_{\text{POS}}(\phi \in \text{G}) = \text{r}$, in the case where $\text{E}_{\text{K}}$ contains $\phi \in \text{F}$ and freq(G|F) = r. Of course, such direct inferences are defeasible. Typically, accounts of direct inference prescribe that one judge the probability that a given object is an element of a

[17] While many, including Venn [47], Reichenbach [34], Kyburg [20], and Kyburg and Teng [21], have assumed that the major premises for direct inference are statements of frequency or limiting frequency, other proposals have been made, in [1, 32], and [42]. I here officially leave the question open, concerning what sorts of statistical statements may serve as major premises for direct inference.

[18] The present theorem is from ([32], p. 71). Similar theorems that are also relevant to the reduction of statistical induction to direct inference are found in other sources, including [27].

respective target class by making a direct inference based on the narrowest relevant reference class for which one has reliable frequency information (cf. [1, 20, 21, 32, 34, 42, 47]). This idea is easily represented within the proposed system, by the inclusion of priors of the following form:

(b) $P_{PRI}\left(\phi \in G | \phi \in F' \wedge F' \subset F \wedge freq(G|F) = r \wedge freq\left(G|F'\right) = s\right) = s.$

In the case where $L_K$ contains (a) and (b), and $E_K$ contains $\phi \in F'$, $F' \subset F$, $freq(G|F) = r$, and $freq(G|F') = s$, the proposed system prescribes the defeat of [d-cond] based on (a), while permitting an inference, by appeal to (b), to the conclusion that $P_{POS}(\phi \in G) = s$. The proposed system also yields what is generally regarded as the correct prescription concerning cases where one has relevant statistics for two reference classes, and neither reference class is narrower than the other. Suppose, for example, that $L_K$ contains (a), along with the following prior:

(c) $P_{PRI}\left(\phi \in G | \phi \in F' \wedge freq\left(G|F'\right) = s\right) = s.$

Now in a case where $E_K$ contains $\phi \in F$, $\phi \in F'$, $freq(G|F) = r$, and $freq(G|F') = s$, [d-cond] generates a defeasible reason for inferring $P_{POS}(\phi \in G) = r$, and a defeasible reason for inferring $P_{POS}(\phi \in G) = s$. But in the case where $E_K$ contains neither $F' \subset F$ nor $F \subset F'$ (and $r \neq s$), the inferences based on (a) and (c) are mutually defeating, and no informative conclusion about the value of $P_{POS}(\phi \in G)$ is licensed by appeal to (a) or (c).

The preceding is intended as a *sketch* of how the prescriptions of direct inference could be encoded within the proposed system of probability updating.[19] It is noteworthy that such an approach to direct inference is not possible within the standard Bayesian system of probability updating. Within that approach, conditionalization always proceeds from a prior, $P_{PRI}(\alpha|\beta) = r$, where $\beta$ incorporates the agent's complete body of evidence, thereby making it practically (if not in principle) impossible to represent the prescriptions of direct inference within the system.

## 8 Conclusion

In the present article, I proposed a generalization of the standard Bayesian approach to probability updating. Unlike the standard Bayesian approach, the proposed system is applicable in cases where an agent's prior probabilities are incomplete or imprecise. The inferences licensed by the proposed system are justifiable by appeal to a modest elaboration of Carnap's principle of total evidence. Like the standard Bayesian framework, the proposed framework is very general, promising to reduce the prescriptions of rational credence formation to the prescription that one update one's personal probabilities by *a form of* conditionalization, and to prescriptions concerning the choice of prior probabilities. Because the proposed system permits conditionalization on propositions

[19] At present it would be a mistake to claim that anyone has succeeded in producing an adequate characterization of the correct principles of direct inference. While the problem of characterizing the correct principles of direct inference is unsolved (perhaps due to insufficient attention), I am optimistic that it is possible to articulate such principles. For recent work, see [41] and [42].

that do not encode an agent's complete body of evidence, the proposed system allows for the possibility of encoding the prescriptions of direct inference and statistical induction by the selection of intuitively reasonable prior probabilities.[20]

## References


1. Bacchus, F. (1990). *Representing and reasoning with probabilistic knowledge*. Cambridge: MIT Press.
2. Bacchus, F., Grove, A., Halpern, J., & Koller, D. (1996). From statistical knowledge bases to degrees of belief. *Artificial Intelligence, 87*, 75–143.
3. Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago: University of Chicago Press.
4. Chisholm, R. (1957). *Perceiving*. Ithaca: Cornell University Press.
5. Good, I. (1962). Subjective probability as the measure of a non measurable set. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and the philosophy of science* (pp. 319–329). Stanford: Stanford University Press.
6. Halpern, J. (2003). *Reasoning about uncertainty*. Cambridge: MIT Press.
7. Hart, H. (1948). The ascription of responsibility and rights. *Proceedings of the Aristotelian Society*.
8. Hempel, C. (1968). Lawlikeness and maximal specificity in probabilistic explanation. *Philosophy of Science, 35*(2), 116–133.
9. Horty, J. (2002). Skepticism and floating conclusions. *Artificial Intelligence, 135*, 55–72.
10. Horty, J. (2007). Defaults with priorities. *Journal of Philosophical Logic, 36*, 367–413.
11. Howson, C., & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach* (3rd ed.). Chicago: Open Court Publishing.
12. Jaynes, E. (1968). Prior probabilities. *IEEE Transactions On Systems Science and Cybernetics, 4*(3), 227–241.
13. Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago: The University of Chicago Press.
14. Kaplan, M. (1983). Decision theory as philosophy. *Philosophy of Science, 50*, 549–557.
15. Kaplan, M. (2010). In defense of modest probabilism. *Synthese, 176*, 41–55.
16. Keynes, J. (1921). *A treatise on probability*. London: Macmillan and Company.
17. Koopman, B. (1940). The bases of probability. *Bulletin of the American Mathematical Society, 46*, 763–774.
18. Kyburg, H. (1956). The justification of induction. *Journal of Philosophy, 53*, 394–400.
19. Kyburg, H. (1961). *Probability and the logic of rational belief*. Middletow: Wesleyan University Press.
20. Kyburg, H. (1974). *The logical foundations of statistical inference*. Dordrecht: Reidel Publishing Company.
21. Kyburg, H., & Teng, C. (2001). *Uncertain inference*. Cambridge: Cambridge University Press.
22. Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy, 71*, 391–418.
23. Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in inductive logic and probability, Vol II*. Berkeley and Los Angeles: University of California Press.
24. Maher, P. (1993). *Betting on theories*. Cambridge: Cambridge University Press.
25. McCarthy, J. (1980). Circumscription - a form of non-monotonic reasoning. *Artificial Intelligence, 13*, 27–31.
26. McDermott, D., & Doyle, J. (1980). Non-monotonic logic I. *Artificial Intelligence, 13*, 41–72.
27. McGrew, T. (2001). Direct inference and the problem of induction. *The Monist, 84*, 153–174.
28. Osherson, D. (2002). Order dependence and Jeffrey conditionalization. Unpublished paper available at: http://www.princeton.edu/~osherson/papers/jeff3.pdf.
29. Paris, J., & Vencovská, A. (1990). A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning, 4*, 183–223.
30. Paris, J., & Vencovská, A. (1997). In defence of the maximum entropy inference process. *International Journal of Approximate Reasoning, 17*, 77–103.
31. Pollock, J. (1967). Criteria and our knowledge of the material world. *Philosophical Review, 76*, 28–60.
32. Pollock, J. (1990). *Nomic probability and the foundations of induction*. Oxford University Press.
33. Pollock, J. (1995). *Cognitive carpentry: a blueprint for how to build a person*. Cambridge: MIT Press.


[20] Work on this paper was supported by the DFG financed EuroCores LogiCCC project *The Logic of Causal and Probabilistic Reasoning in Uncertain Environments*, and the DFG project *The Role of Meta-Induction in Human Reasoning* (SPP 1516). For valuable comments, I am indebted to audiences at ProbNet10 in Salzburg, the European Epistemology Network Meeting in Lund 2011, and the Heinrich-Heine-Universität Düsseldorf, and also to Ludwig Fahrbach, Gerhard Schurz, Matthias Unterhuber, Ioannis Votis, and in memoriam to John L. Pollock.

34. Reichenbach, H. (1935). Wahrscheinlichkeitslehre: eine Untersuchung über die logischen Und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung. English translation: (1949). The theory of probability, an inquiry into the logical and mathematical foundations of the calculus of probability. University of California Press.
35. Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence, 13*, 81–132.
36. Rescher, N. (1977). *Dialectics*. New York: SUNY Albany Press.
37. Schurz, G. (1997). Probabilistic default reasoning based on relevance and irrelevance assumptions. In D. Gabbay et al. (Eds.), *Qualitative and quantitative practical reasoning* (pp. 536–553). Berlin: Springer.
38. Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint: ontic, logical and cognitive foundations. *Synthese, 146*(1–2), 37–51.
39. Smith, C. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series B, 23*, 1–37.
40. Stove, D. (1986). *The rationality of induction*. Oxford: Clarendon.
41. Thorn, P. (2011). Undercutting defeat via reference properties of differing Arity: a reply to Pust. *Analysis, 71*(4), 662–667.
42. Thorn, P. (2012). Two problems of direct inference. *Erkenntnis, 76*(3), 299–318.
43. Toulmin, S. (1958). *The place of reason in ethics*. Cambridge University Press.
44. Touretzky, D., Horty, J., & Thomason, R. (1987). A clash of intuitions: the current state of monotonic multiple inheritance systems. *In Proceedings of the Tenth international Joint Conference on Artificial Intelligence* pp. 476–482.
45. Van Fraassen, B. (1989). *Laws and symmetry.* Oxford University Press.
46. Van Frasssen, B. (1990). Figures in a probability landscape. In J. M. Dunn & A. Gupta (Eds.), *Truth and consequences* (pp. 345–356). Dordrecht: Kluwer Academic Publishers.
47. Venn, J. (1866). *The logic of chance*. New York: Chelsea Publishing Company.
48. Wagner, C. (2002). Probability kinematics and commutativity. *Philosophy of Science, 69*(2), 266–278.
49. Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. London: Chapman and Hall.
50. Williams, D. (1947). *The ground of induction*. Cambridge: Harvard University Press.
51. Williamson, J. (2007). Motivating objective Bayesianism: from empirical constraints to objective probabilities. In W. L. Harper & G. R. Wheeler (Eds.), *Probability and inference: essays in honor of Henry E. Kyburg Jr.* London: College Publications.