C.R.AIPS 3

Proceedings of *the Académie Internationale de Philosophie des Sciences*

Comptes Rendus de l'Académie Internationale de Philosophie des Sciences

Tome III

Models and Representations in Science

Éditeur

Hans-Peter Grosshans

The requirement of total evidence: epistemic optimality and political relevance

Gerhard Schurz

Institut für Philosophie, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany

Abstract. The requirement of total evidence says that one should conditionalize one's degrees of belief on one's total evidence. In the first part I propose a justification of this principle in terms of its epistemic *optimality*. The justification is based on a proof of I. J. Good and embedded into a new account of epistemology based on optimality-justifications. In the second part I discuss an apparent conflict between the requirement of total evidence and political demands of *anti-discrimination*. These demands require, for example, that information about the sex of the applicant for a job should not be included in the relevant evidence. I argue that if one assesses the applicant's qualification in terms of those properties that are directly causally relevant for the job performance, then properties that are merely indirectly relevant, such as sex, race or age, are screened off, i.e., become irrelevant. So, the apparent conflict disappears.

1 Introduction

The requirement of total evidence—henceforth abbreviated as RTE—says the following:

> In order to rationally estimate the epistemic probability (P) of a hypothesis, one should conditionalize this probability on one's total evidence, i.e., all 'relevant' evidence that is available to the epistemic subject. Thus, if E is the subject's total evidence, then $P_{\text{actual}}(H) = P(H|E)$. (RTE)

Thereby the evidence E is assumed to be 'approximately certain'.[1] Among others, the RTE was introduced by Carnap (1950, 211f.). If the hypothesis is a singular prediction, Fa, the RTE coincides with Reichenbach's principle of the narrowest reference class, which says that we should conditionalize Fa's probability on its membership in the narrowest (relevant) reference class for which we possess evidence (Reichenbach 1949, sec. 72). That the evidence can be restricted to relevant evidence is obvious, since irrelevant evidence does not change the probability and can be omitted, i.e., $P(H|E_{\text{rel}} \wedge E_{\text{irr}}) = P(H|E_{\text{rel}})$.

[1] For uncertain evidence, Jeffrey conditionalization has to be applied:

$$P_{\text{actual}}(H) = \Sigma_{\pm E} P(H|\pm E) \cdot P_{\text{actual}}(\pm E).$$

(The notion "$\Sigma_{\pm E}$" is explained in the text.)

Models and Representations in Science, edited by Hans-Peter Grosshans.
Comptes Rendus de l'Académie Internationale de Philosophie des Sciences 3 (2025).
G. Schurz, The requirement of total evidence, pp. 209–225.

Why is the RTE reasonable? It is certainly necessary to *fix* the evidence on which we conditionalize somehow, because otherwise we may end up in contradictions.[2] But why should this be the most comprehensive evidence? Why is it not better to leave evidence out if we do not like it? In what follows we illustrate our problem at hand of a simple *weather* example, as follows:

1. R denotes the prediction that it will rain tomorrow in my area.
2. The probability of R, $P(R)$, is assumed to be implicitly conditionalized on given the general background evidence that we live in a sunny area with a 20% rain chance. So we assume $P(R) = 0.20$ and $P(\neg R) = 0.80$.
3. F denotes the additional evidence that the barometer has fallen, indicating a rain-chance of 95%, even for areas that are normally sunny.

Our assumptions entail that $P(R) = 0.20$, but $P(R|F) = 0.95$. So we must fix the evidence on which we conditionalize our probability of the hypothesis, R, in order to avoid probabilistic incoherence. But why should we conditionalize our prediction on the total or *most specific* evidence, F? Why should we not rather be coherentists and stick with conditionalizing our belief about tomorrow's weather on our general background evidence that we live in an overwhelmingly sunny area, *ignoring* the additional evidence F, so that we are not forced to give up the friendly-weather-belief that we like?

Hempel (1960, 453f.) and Suppes (1966) argued that for a Bayesian probabilist, who identifies her or his degrees of belief with rationally estimated probabilities, the RTE follows already from the probability axioms, or equivalently, from the requirement of probabilistic coherence. For $P(A) = 1$ implies $P(B) = P(B|A)$ (since $P(A) = P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B) = P(A|B) \cdot 1 + P(A|\neg B) \cdot 0 = P(A|B)$). So given the evidence F is taken as certain, then $P(R) = P(R|F)$; so our coherent degree of belief in the hypothesis R must already be conditionalized on all evidence that is taken as certain. Likewise, if F is almost certain, then provided $P(R|F)$ is not close to zero, $P(R)$ must be approximately equal to $P(R|F)$. Roush (forthcoming, 31, fn. 47) considers this argument as an advantage of Bayesian probabilism. From the viewpoint of *applied* epistemology, however, I think this argument is insufficient, since real epistemic agents are far from being probabilistically omniscient. What people really do when estimating the probability of a future events, such as the possibility of tomorrow's rainfall, is retrieving from their memory some known facts that are regarded as relevant cues for this prediction, and then estimating the predicted probability conditional on the conjunction of these cues. For this epistemic practice the RTE is

[2] In application to explanations, Hempel (1965, sec. 3.4) spoke of the "ambiguity of statistical explanations".

highly important, because it requires that instead of confining oneself with just one or a few cues, one should actively retrieve all relevant cues that one knows. For example, if you base your prediction on a weather forecaster on the Internet, but there is a second forecaster that predicts differently (a situation that does not occur unfrequently), then the RTE tells you that you should not just rely on one forecaster and ignore the other. Even for rational Bayesians, the RTE is not self-evident, because Bayesianism does not prescribe how an epistemic agent should mold her or his probabilities. For coherentist Bayesians, ignoring a piece of evidence F when estimating the actual probability of a prediction R just means that they change their probability of F from a value close to 1 to some lower value. Why should such a 'probabilistic suppression' of an unwanted fact not be a legitimate epistemic practice, for the mutual sake of increasing the coherence of our beliefs and desires? Why are we *worse off* if we follow this practice rather than follow the RTE? Moreover, why is searching for new (cheap) evidence better than applying the ostrich-method of avoiding the acquirement of new evidence (putting one's hat in the sand)?

To obtain a positive answer to these questions, we need an explicit *justification* of the RTE. Moreover, recall that according to Reichenbach's ingenious idea the justification of the RTE would at the same time tell us how the statistical (or frequentist) probabilities of repeatable events should be connected with the epistemic probabilities of single instances of these events. The above weather example is nothing but such a connection: the statistical chance of rain (Rx) in some reference class (Cx), abbreviated as $p(Rx|Cx)$, is transferred to a particular day, namely *tomorrow* (a), as the epistemic probability of a rainfall tomorrow: $P(Ra) = p(Rx|Cx)$ (where "Cx" is a condition that refers to the past of x, logically expressed by a functor, $Cx = Gfx$). The reason why we *want* a connection between epistemic and statistical probabilities is simple: only if there is such a connection, will the probabilistically *expected utilities*—which are the central guide for rational decisions—agree with our actually experienced *average utilities* (in the long run); otherwise maximization of expected utilities could fail to be actually utility-increasing. However, there are different possible reference classes Cx—in our example that I live in a sunny area, that the barometer fell yesterday, etc. Which reference class should we choose? According to Reichenbach's "principle of narrowest reference class", we should identify the epistemic probability of a single case hypothesis with its statistical probability conditional on the total (relevant) evidence about the respective individual a; in our example: $P(Ra) = p(Rx|Fx)$.[3] Therefore, a justification

[3]The transfer of $p(Rx|Fx)$ to $P(Fa)$ is also called "direct inference" and is related to the so-called statistical principal principle; thereby "P" must be prior in regard to the involved individual a (see Schurz 2014, 160 and 2024, 58f.).

of the RTE would give us at the same time a justification of transferring statistical probabilities to single cases by means of the RTE.

In the next section, we offer such a justification of the RTE, based on a reconstruction of a seminal proof of Good (1966). The proof demonstrates that for practical as well as predictive success, the best what we can do is to conditionalize on the total available evidence. The proof is an instance of what is called an *optimality justification*. It is part of the account of epistemology based on optimality-justifications developed in Schurz (2024) that grew out from work on the optimality of meta-induction (Schurz 2019).

2 An optimality justification of the requirement of total evidence

In what follows I explain the proof for the simplest case of binary partitions, illustrated at hand of our weather example. So we are interested in predicting the binary variable $\pm R$, where "$\pm$" stands for "unnegated" or "negated", i.e., $\pm R \in \{R, \neg R\}$, in our example, that it will rain (R) or not rain ($\neg R$) = tomorrow. Note that strictly speaking we have to represent the prediction R by the atomic formula Ra_{n+1}, where $a_1, a_2, \ldots$ stands for a sequence of days, a_{n+1} for the day tomorrow and a_n for today. We dispense with this formal complication since the meaning is obvious.

Preceding each day we obtain additional evidence about whether the barometer reading has fallen or not, $\pm F$, where according to our estimation $P(R|F) = 0.95$ and $P(R|\neg F) = 0.15$.

Good's proof of the optimality of the RTE is devised for *success in actions*, whose utility depends on the unknown utility-determining circumstances or predictive targets, in our example $\pm R$. We assume that in our example the possible actions are:

the action(s) of taking an umbrella with us or not, abbreviated as $\pm U$.

The decision concerning $\pm U$ must be made today, for example because we leave today for a mountain tour tomorrow. Concerning the utilities, $u(A|C)$ denotes the utility of action A given the circumstance C.[4] In our example, we assume the following utility values:

$$u(\neg U|R) = 0, \quad u(\neg U|\neg R) = 0, \quad u(U|R) = 3, \quad \text{and} \quad u(U|\neg R) = -1.$$

[4]In causal decision theory (Weirich 2020) one often writes $u(A \wedge C)$ instead of $u(A|C)$. This indicates that also C may contribute to the total utility outcome. This notation is appropriate if the circumstances includes factors that are *effects* of the actions, but we do not assume this (see below). Our utilities express the utility-effect of the action relative to the total utility of the action-independent circumstances; this is reflected in the notation "$u(A|C)$", which is close to Savage's notation (Steele and Stefanson 2020, sec. 3.1). For action-independent circumstances both notations are equivalent, because in this case the decision matrix can be rescaled by adding to each row a row-specific constant without changing the Eu-ordering of the actions (see Jeffrey 1983, 35–37).

Utilities and probabilities are assumed to be reliably estimated.

According to decision theory the *expected utility*, Eu, of the actions $\pm U$ are given as:

$$\begin{aligned} \mathrm{Eu}(U) &= P(R) \cdot u(U|R) + P(\neg R) \cdot u(U|\neg R) \\ &=_{\mathrm{def}} \Sigma_{\pm R} P(\pm R) \cdot u(U| \pm R). \\ \mathrm{Eu}(\neg U) &= P(R) \cdot u(\neg U|R) + P(\neg R) \cdot u(\neg U|\neg R) \\ &=_{\mathrm{def}} \Sigma_{\pm R} P(\pm R) \cdot u(\neg U| \pm R). \end{aligned} \tag{1}$$

In informal words: The Eu of action U is the sum of U's utilities under the different circumstances $\{R, \neg R\}$ multiplied with their probabilities. (Similarly for $\neg U$.)

So in our example, without additional evidence we get:

$$\mathrm{Eu}(U) = 0.2 \cdot 3 - 0.8 \cdot 1 = -0.2 < \mathrm{Eu}(\neg U) = 0.2 \cdot 0 + 0.8 \cdot 0 = 0.$$

So with the above utilities, if all what I know is $P(R) = 0.2$, then my wisest action is not to take an umbrella.

The philosophical assumption behind the decision-theoretic formula (1) is that the choice of action is *free* in the sense of being probabilistically *independent* from those utility-determining circumstances that are *not causally influenced* by the actions. In the formula (1), the cells of the partition of circumstances range over those circumstances, in our example $\pm R$. This assumption justifies that we write $P(\pm R)$ instead of $P(\pm R| \pm U)$, since $\pm U$ has no causal influence on tomorrow's rain. We will defend this assumption below. Here we merely point out that we may include action-dependent circumstances by expanding in equation (1) the term $u(U| \pm R)$ as follows:

$$u(U| \pm R) = \Sigma_i P(D_i|U) \cdot u(U \wedge D_i| \pm R),$$

where $\{D_1, \ldots, D_n\}$ is an additional partition of action-dependent facts. Inserting this equation into (1) gives us

$$\mathrm{Eu}(U) = \Sigma_{\pm R} P(\pm R) \cdot \Sigma_i P(D_i|U) \cdot u(U \wedge D_i| \pm R),$$

which is a version of Skyrms' causal decision theory (Skyrms 1980, sec. IIC; Weirich 2020, sec. 2.3).

The argument of Good's proof in my reconstruction consists of two steps:

Step 1 of Good's proof: The expected utility Eu of a *fixed* action—one that is independent of which additional evidence you observe—is provably *preserved* under conditionalization of the probabilities of the circumstances on new evidence $\pm F$. In other words: the Eu does not change under

refinements of the partition of action-independent circumstances. In our example this means the following:

$$\mathrm{Eu}(\neg U) = \mathrm{Eu}(\neg U|\{F, \neg F\}), \tag{2}$$

where

$$\mathrm{Eu}(\neg U|\{F, \neg F\}) = P(F) \cdot \mathrm{Eu}(\neg U|F) + P(\neg F) \cdot \mathrm{Eu}(\neg U|\neg F),$$

and

$$\begin{aligned}\mathrm{Eu}(\neg U|F) &= \Sigma_{\pm R} P(\pm R|F) \cdot u(U|\pm R) \\ &= \text{the Eu of } \neg U \text{ updated with } P(\pm R|F),\end{aligned}$$

and similarly,

$$\mathrm{Eu}(\neg U|\neg F) = \text{the Eu of } \neg U \text{ updated with } P(\pm R|\neg F).$$

Similarly for $\mathrm{Eu}(U)$.

The proof of (2) is as follows. Analytically it holds that

$$\begin{aligned}\mathrm{Eu}(U|\{F, \neg F\}) = P(F) \cdot \Sigma_{\pm R} P(\pm R|F) \cdot u(U|\pm R \wedge F) + \\ P(\neg F) \cdot \Sigma_{\pm R} P(\pm R|\neg F) \cdot u(U|\pm R \wedge \neg F).\end{aligned} \tag{3}$$

We assume, however, that

$$u(\pm U|\pm R \wedge \pm F) = u(\pm U|\pm R) \quad \text{(utility-neutral additional evidence)}. \tag{4}$$

This holds because the circumstances $\pm R$ determine the utilities of the actions. So the fact expressed by the evidence, $\pm F$, has no further influence on their utility.[5] From (3) and (4) we obtain:

(5) *Proof of (2):*

$$\begin{aligned}&\mathrm{Eu}(U|\{F, \neg F\}) \\ &\quad = P(F) \cdot \Sigma_R P(\pm R|F) \cdot u(U|\pm R) + P(\neg F) \cdot \Sigma_R P(\pm R|\neg F) \cdot u(U|\pm R) \\ &\quad = \Sigma_R [P(\pm R \wedge F) \cdot u(U|\pm R) + P(\pm R \wedge \neg F) \cdot u(U|\pm R)] \\ &\quad = \Sigma_R [P(\pm R \wedge F) + P(\pm R \wedge \neg F)] \cdot u(U|\pm R) \\ &\quad = \Sigma_R P(\pm R) \cdot u(U|\pm R) = \mathrm{Eu}(U). \qquad \text{Q.E.D.}\end{aligned}$$

Step 2 of Good's proof: Now, the point of conditionalization is that the new evidence may *change* the optimal action under a particular observational

[5] It is also possible to prove (2) without assumption (4), by assuming Jeffrey's framework that identifies utilities and expected utilities; his "desirability axiom" (1983, 80, (5–2)) implies for our example that (2) holds. However, Jeffrey's axiom is rather strong.

outcome $\pm F$. If F is observed, this indicates a high chance of rain, and so the F-conditional Eu of U is much higher than that of $\neg U$. In our example we get

$$\mathrm{Eu}(U|F) = 0.95 \cdot 3 - 0.05 \cdot 1 = 2.8 > \mathrm{Eu}(\neg U|F) = 0.95 \cdot 0 + 0.15 \cdot 0 = 0.$$

If $\neg F$ is observed, we should not change the best evidence-independent action $\neg U$; in this case, the surplus of $\neg U$ over U even increases. In our example we get

$$\mathrm{Eu}(U|\neg F) = 0.15 \cdot 3 - 0.85 \cdot 1 = -0,4 < \mathrm{Eu}(\neg U|\neg F) = 0.15 \cdot 0 + 0.15 \cdot 0 = 0.$$

In conclusion, after conditionalization the rational subject performs the conditionalized or *evidence-dependent action*

$$U^* =_{\mathrm{def}} \text{“}U \text{ if } F \text{ and } \neg U \text{ if } \neg F\text{”}.$$

For U^* the Eu is computed as follows:

$$\mathrm{Eu}(U^*|\{F, \neg F\}) = P(F) \cdot \mathrm{Eu}(U|F) + P(\neg F) \cdot \mathrm{Eu}(\neg U|\neg F). \tag{6}$$

$\mathrm{Eu}(U^*|\{F, \neg F\})$ is greater than the Eu of the best fixed action, $\mathrm{Eu}(\neg U|\{F, \neg F\})$, since $\mathrm{Eu}(U|F) > \mathrm{Eu}(\neg U|F)$. To see this, compare equation (6) with the equation below the equation (2): the two equations differ only in the terms $\mathrm{Eu}(U|F)$ respectively $\mathrm{Eu}(\neg U|F)$, and since $\mathrm{Eu}(U|F) > \mathrm{Eu}(\neg U|F)$, $\mathrm{Eu}(U^*|\{F, \neg F\}) > \mathrm{Eu}(\neg U|\{F, \neg F\})$ follows, where $\mathrm{Eu}(\neg U|\{F, \neg F\}) = \mathrm{Eu}(\neg U)$ (as proved in (5)) and $\neg U$ is the best fixed action.

The basic argument is entirely independent of the assumed utilities. Even if the utility of taking an umbrella given rain would be much smaller than given not-rain (for example because of a dictator who punishes people who are taking an umbrella while it rains), the theorem would go through. Either under one of the two evidential outcomes $\pm F$ the evidence-dependent Eu of one of the two actions, say A', becomes greater than the best evidence-independent action, call it A_{ind}, then we switch from A_{ind} to A' under this outcome and this will increase the Eu, or under both evidential outcomes A_{ind} has still maximal Eu, in which case we stay with A_{ind} and (by the proof in (5)) the Eu will be preserved.

This proof generalizes to arbitrary finite partitions of possible actions, circumstances and evidence, leading to the following result:

Theorem (Optimality of the RTE)**.** Assume a partition $\mathbf{C}$ of possible circumstances and a partition of possible actions $\mathbf{A}$ whose Eu is governed by the decision-theoretic formula (1). Then:

(i) Conditionalization of the probabilities of the circumstances $C \in \mathbf{C}$ of the agent's possible actions $A \in \mathbf{A}$ on the cells of a partition $\mathbf{F}$ of additional evidence can only increase but not decrease the Eu of the agent's evidence-dependent action A^* defined as follows:

(U^*) "For all cells $F \in \mathbf{F}$, if F is observed, then choose action A_F", where A_F is the action with the highest F-conditional Eu.

(ii) Moreover: Let A_{ind} be the fixed (evidence-independent) action with highest Eu. Then: If for all $F \in \mathbf{F}$, $A_F = A_{\text{ind}}$, then A^* has the same Eu as A_{ind}, but if for at least one $F \in \mathbf{F}$, $A_F \neq A_{\text{ind}}$, then the Eu of A^* increases.

The general mathematical fact behind this theorem is expressed by Schwartz (2021) as follows: The maximum of a weighted average (which is $\text{Eu}(\neg U|\{F, \neg F\})$ is always smaller than or at most equal to the corresponding average of the maxima (which is $\text{Eu}(U^*|\{F, \neg F\})$ (see also Bradley and Steel 2016, 4).

Three features of this general result are remarkable:

First: The argument holds for *every* utility function. This result is astonishing, in particular in the domain of predictions (see below).

Second: The only essential assumption of the optimality result is that the costs of *acquiring* new information are negligible.[6] If these costs are too high, they could of course offset the benefits gained. Some counterexamples to the RTE are of this sort—for example, the first counterexample in Schwarz (2021).

Third: The result implies two things: (i) That you should take into account all the (relevant) evidence that you actually possess, but also (ii) that you should try to gather new evidence whenever this is easily possible, because by doing so you cannot decrease and will in most cases increase the Eu of your actions.

Horwich (1982, 125–128) objected against Good's proof that it would apply only to practical (non-epistemic) actions. But this is not true: the possible actions in Good's proof may also be purely epistemic actions, for example, *predictions* whose utility is given by a predictive scoring measure. In our example, the actions would be predictions of tomorrow's weather, abbreviated as "pred($\pm R$)" for predicting R or $\neg R$. The optimal fixed prediction in our weather example would be pred($\neg R$). But conditional on observing F the optimal prediction is not $\neg R$ but R. So the rational forecaster predicts R if F was observed and $\neg R$ if $\neg F$ was observed, and this increases the predictive score. Let us designate this evidence-dependent prediction as pred*. Good's proof applies in precisely the same way and our theorem applies: the Eu of pred* can only increase but not decrease the Eu of the best evidence-independent prediction, and this results holds for every scoring function (for details cf. Schurz 2024, sec. 7.3).

[6]Note that the utility of the *acquirement* of evidence is a different matter than the utility of the fact expressed by the evidence.

We have illustrated Good's argument for *qualitative* predictions (predictions of events), but a related argument applies to the predictions of *probabilities* (cf. Thorn 2017). In this case the possible predictions are P-distributions $P : \{e_1, \ldots, e_n\} \to [0, 1]$, where $\{e_1, \ldots, e_n\}$ are the possible events (in our example $\pm R$). The prediction is scored against the truth-value "1" of the true event, e_{true}, among the partition of predicted events, i.e., $\text{score}(\text{pred}) = 1 - \text{loss}(P(e_{\text{true}}), 1)$, where "loss" is a loss function (cf. Cesa-Bianchi and Lugosi 2006, ch. 9). For probabilistic predictions the scoring function is usually assumed to be proper (e.g., quadratic), because only for proper scoring functions is it optimal for the forecaster to predict her (rationally estimated) probabilities of the events (cf. Brier 1950; Maher 1990, 113). In contrast, for linear scoring ($\text{loss}(\text{pred}, 1) = 1 - \text{pred}$), it is optimal to predict the roundings of the event's probabilities to 0 or 1 (the so-called "maximum rule"; cf. Schurz 2019, 103). However, Good's optimality argument for the RTE generalizes also to non-proper scorings, provided the predictions $\text{pred} \in [0, 1]$ are allowed to deviate from one's actual probabilities that are used to compute the Eu.[7]

Let me finally note that the optimality of the RTE has an important consequence for the *externalism-internalism* debate, in the justificational sense of externalism/internalism (cf. Schurz 2024, sec. 3.2). In epistemological externalism, the question of choosing the right reference class in which the reliability of a belief-generating method should be determined is part of what is called the *generality problem* (Conee and Feldman 1998, Matheson 2015). Within externalism this question is largely undecided or at least hard to answer. But within justification-internalism, the question has a straightforward and unique solution: the reliability should be evaluated with regard to the agent's total relevant evidence for the belief in question.

At the end of this section let me return to the presupposition of our decision-theoretic formula (1): that the choice of action is *free* in the sense of being probabilistically *independent* from those utility-determining circumstances C_i that are not causally influenced by the actions. First, note that if we conditionalize our decision on the available evidence E, this independence condition has to be formulated conditionally: C_i and the chosen action A should be independent conditional on E, i.e., $P(C_i|E) = P(C_i|E \wedge A)$. *Second*, the independence condition excludes various versions of Newcomb's

[7] It may happen that conditionalizing on one cell of $\pm E$, say on E, brings the old actual probabilities close to 0.5 (e.g., of $P(R) = 0.2$ and $P(R|E) = 0.3$). In this case Good's strategy with linear scoring would require to predict the old non-actual probabilities conditional on E (and the new conditionalized probabilities conditional on $\neg E$), which is not allowed if one must allows predict one's actual probabilities. Horwich (1982, 128f.) proved that the RTE maximizes the Eu of one's actual probabilities even under linear scoring, which is a second important result. But his proof is specially designed for linear scorings and does not generalize to arbitrary scorings.

paradox, in which some past event X (in Newcomb's paradox the prediction of a perfect or nearly perfect forecaster) determines which action you will choose, or the probability with which you will choose it, already in advance, so that there is a probabilistic dependence between the circumstances C_i (that incorporate $\pm X$) and your choice of action. Newcomb's paradox in its various versions forms a second line of purported counterexamples against Good's proof of the universal rationality of the RTE (the 2nd, 3rd and 4th counterexample in Schwarz 2021 falls under this category). I am inclined to think, however, that the assumption of Newcomb's paradox is in conflict with the fact that decision theory delivers a normative recommendation. It is not possible for me here to go into the extensive literature on the Newcomb paradox[8] and I content myself here with a brief statement of my main argument. Decision theory gives the normative recommendation that you should *always* choose the action with highest expected utility, conditional on the total evidence E. But in in typical Newcomb-type situations, the normatively recommended action is different from that action that is determined or predicted by the past event X. This implies that in many cases it will be *impossible* for you to follow the decision-theoretic recommendation. But this means that the decision-theoretic recommendation will itself be itself unreasonable, because according to the famous *Ought-Can* principle (Ought implies Can), a normative recommendation can only be reasonable if the recommended action *can be* done. But in Newcomb-type situations you know that with considerable probability the recommended action cannot be done, because a past event forces the agent to choose an action different from the recommended one. On the other hand, if the recommended action luckily agrees with the action the agent is forced to do, then the normative recommendation becomes superfluous.

In conclusion, if actions are determined by past circumstances, then normative recommendation either violate the Ought-Can principle or become superfluous. Therefore the freedom assumption seems to be an implicit presupposition of decision-theoretic recommendations.

3 The political relevance of the requirement of total evidence

In the concluding section we discuss an apparent *conflict* of the RTE with political requirements of *anti-discrimination*. Consider the example of sex discrimination in job hiring (Birkelund et al. 2022):

1. According to the RTE, information about the sex (or biological gender) of the applicant should be included in the qualification-relevant evidence *iff* it is statistically relevant.

[8] Cf., e.g., Nozick (1969), Eells (1981), Lewis (1981), Skyrms (1982), Horwich (1985), Weirich (2020).

2. In contrast, politicians of anti-discrimination often require sex to be ignored despite of its statistical relevance, because it would lead to discrimination.

Of course, if the belief about a correlation between sex and job qualification is not statistically supported, but is based on *prejudice* or some other sort of *cognitive bias*, then the RTE does *not* demand sex to be included. Then we should leave out the male/female information simply because the job assigner's beliefs about properties correlated with biological sex is biased, i.e., wrong. There is a rich literature about cognitive prejudice and bias, but here we will not enter these topics. Rather, we make the idealizing assumption that our statistical beliefs are well supported by the statistical evidence. In other words, the assessment procedure of the job assigner is not biased but well calibrated. Then it seems that we have a conflict: For the job assigner, conditionalizing on the additional information about sex *increases the expected qualification* of the chosen candidate(s). But at least for some candidates this seems to be *unfair*, given that fairness means that the job assignment corresponds to the candidates' objective job-relevant qualifications. This understanding of fairness is also called the *meritocratic* understanding (cf. Barocas et al. 2023, ch. 4).

Let us give an *example*: A woodworking factory has to hire a person for a wood chipper job that requires a lot of physical strength. According to statistical evidence, males are physically stronger on average than females. So if sex is a criterion for job hiring, then a female applicant will have less chances *even if* she is physically very strong. If statistics is correct, these cases of unfairness will be in the minority, but they will unavoidably occur, and with significant frequency. Similar examples may be given with sexes switched. For example, assume a nursery school hires a person for early childhood care. According to statistical evidence, females caregivers are better accepted by young children than males. So if sex is used as a criterion, a male person will have less chances to be hired even if children would like him most (cf. Birkelund et al 2022, 347).

The only solution which I see is the following: One should base the decision about the job assignment solely on information about the *directly relevant* properties of the applicants. With this I mean those properties that are most direct causes of the job performance of the candidate (if the candidate would be hired), within the set of evidentially accessible variables. If we do this, then the merely indirectly relevant properties such as sex, race or age are *screened off*, which means that after conditionalization on the directly relevant properties, they become irrelevant. In our example: The wood factory should directly test the candidates for their physical strength and other directly relevant properties, such as social skills, reliability, etc. Given this information, additional information about sex or other merely

indirectly relevant properties of the applicants becomes irrelevant. This is an implication of the so-called *causal Markov* condition, according to which conditionalization on the direct causes screens off indirect causes from their effects, and likewise, conditionalization on the common causes screens off their effects from each other.[9] This means in terms of probabilities:

$$P(\text{qualification} \,|\, \text{physical strength \& sex}) = P(\text{qualification} \,|\, \text{physical strength}). \quad (7)$$

Let us generalize this idea. Assume the following variables (or partitions of their possible values) designated by bold-face letters:

1. **Q** is a partition of degrees of qualification of candidate (e.g., from 1 (best) to 5 (worst), understood as expressions of their *future job performance* which is to be *predicted*.

2. **D** is a partition of evidentially accessible properties of the candidates that are (supposedly) directly causally relevant for **Q** and measured by a score **S** on which the decision is based.

3. **A** is a partition of additional information, for example about sex, race or age (etc.), that is merely indirectly relevant, by being correlated with **S**. In the literature on fairness in machine learning, **A** is often called the (partition of) *sensitive attributes* (Barocas et al. 2023, ch. 3; Mitchell et al. 2021, 149).

Then I propose the following

> *Fairness criterion:* If score is fair, then Indep($\mathbf{Q}, \mathbf{A}|\mathbf{S}$) should hold (where "Indep($\mathbf{X}, \mathbf{Y}|\mathbf{Z}$)" means that if we fix the variable **Z** to a particular value, then the values of **X** and **Y**, respectively, are mutually probabilistically independent). (F)

In the literature on fairness in machine learning, (F) corresponds to an important anti-discrimination criterion that has been called *sufficiency* (Barocas et al. 2023, ch. 3) or *predictive parity* (Mitchell et al. 2021, 154).

The causal model behind the above fairness criterion is illustrated in Figure 1 below. Causal arrows are distinguished into required ones (marked with "r"), admissible but not required ones (marked with an "a"), and excluded ones (marked with a backslash "\"). Thus, the sensitive attribute **A** may (but need not) be relevant for **Q**, the job qualification, but if **A** is relevant for **Q**, then merely indirectly, via the path over the directly relevant

[9] Cf. Lauritzen et al. (1990), 50; Spirtes et al. (2000), sec. 3.4.1–2; Pearl (2009), 16–19; Schurz and Gebharter (2016), sec. 2.3, conditions (6) and (8).

properties **D**, whence **A** is screened off by conditionalization on **D**. This requires that the variable **D** must be complete, in the sense of covering all or almost all properties of the job candidate that are direct causes for **Q**. Moreover, the score **S** must be accurate in the sense of measuring the values of **D** precisely; if this is the case, then not only **D** but also **S** screens off *A* from **Q**—which is the required condition because **S** determines the decision who will get the job. What is excluded is that information about **A** directly influences the score **S** or the decision (independent from **D**), or that **A** has a direct influence on **Q** (relative to the model), which would mean that the scoring variable **S** leaves out important causal information and, thus, fails to screen off indirect causes of **Q**.

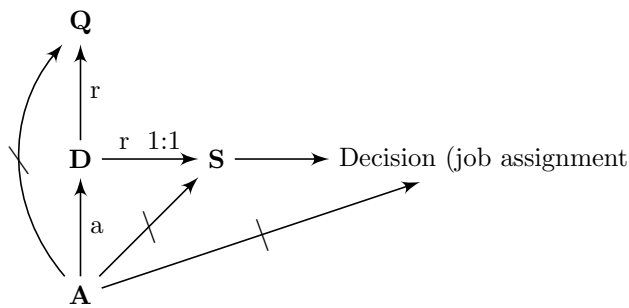

FIGURE 1. The causal model behind the fairness criterion of sufficiency (or predictive parity). Causal arrows are distinguished onto required ones ("r"), allowed ones ("a") and excluded ones ("\").

Summarizing, it seems that by conditionalizing on the directly relevant properties, unfairness can be avoided. Moreover, if we are not sure which of the evidentially accessible variables are the directly relevant ones, then conditionalization on *more* information can reveal possibly discriminating variables that are merely indirectly relevant—by detecting screening-off relations. So it seems that the RTE 'wins': it is not really in conflict with anti-discrimination. Is this true?

I conclude this paper with a brief discussion of three objections to the above fairness criterion.

Objection 1: In the literature on fairness in machine learning, there is a hot controversy about the "right" criterion of fairness (Barocas et al. 2023, ch. 3 & 4). In my view the above criterion is the right one, given the causal model of Figure 1. Let me mention two rival criteria of fairness:

The first rival fairness criterion is called *independence* or *statistical parity* and requires Indep($\mathbf{S}, \mathbf{A}$) (Barocas et al. 2023, ch. 3). This means that on average all **A**-members—in our example both sexes—should achieve

the same qualification score. Obviously, this can only be compatible with meritocratic fairness if on average all **A**-members—in our example both sexes—are equally qualified. Otherwise this criterion leads to some sort of "affirmative action" that is discussed below.

A second rival is called the criterion of *separation* (ibid., ch. 3) which requires Indep($\mathbf{S}, \mathbf{A}|\mathbf{Q}$). In this criterion, the roles of the variables **S** and **Q** are switched, compared to our preferred criterion (F). Thus in the causal model on which the separation criterion is based, **S** is assumed not to express causes but the effects of **Q**. This implies a rather different understanding of **Q** and **S**. It makes sense if **Q** takes the role of **D**, i.e., is identified with actually measurable properties of the candidates that are supposedly relevant for its job qualification, while **S** is a possibly inaccurate score of **Q**.

Objection 2: Some people, politically mainly left-wing oriented, argue for so-called *affirmative action*. This is based on the idea that members of an underrepresented or even discriminated group should be preferred even if they are on average less qualified, because this kind of "compensatory unfairness" is necessary for breaking up historically or socially anchored injustice. An example would be the university policy to hire 50% males and 50% females for a professor job in theoretical philosophy, which is a discipline where we typically have 75% males and 25% females among students, researchers and applicants for the professor job. Affirmative action is controversial—how much unfairness (in the meritocratic sense) is tolerable in this attempt to encourage women's engagement with theoretical philosophy? I do not want to discuss this question here. Rather, I want to emphasize that *even if* one supports affirmative action, the general optimality proof of the RTE stays intact, since RTE's optimality holds for *all* utility functions. All what changes for a selection criterion based on affirmative action is the relevant utility function of the available actions and the partition of utility-determining circumstances. In our example, the utility of the of hired applicant is then not only based on the candidates merits, but also on other desired properties such as the sex of the candidate. So "sex" is no longer merely "indirectly relevant", but becomes a directly relevant property.

Objection 3: In my view this is the hardest objection. It objects that our claims hold only under the idealizing assumption that we possess sufficient information about the directly relevant qualification properties of the candidates. If the job recruiter is uncertain about these properties of the candidates, then the RTE recommends conditionalization of the estimated qualification on evidence about merely indirectly relevant evidence properties. This will increases the expected qualification of the hired candidate, since now his or her qualification is no longer screened off from these indirectly relevant properties. However, the so achieved increase of the average qualification has the cost that it will produces a certain amount of unfairness. This

unfairness can be measured in terms of the numbers of pairs of candidates A, B in which A is preferred over B although A is less competent than B.

In conclusion, in such a situation there is a trade-off between maximizing the expected qualification of the chosen candidate and maximizing meritocratic fairness. What policy would be reasonably fair in such a situation? I cannot go into the details of this question but confine myself with a remark concerning a frequently heard suggestion, namely that without knowledge about the directly job-relevant properties, one should choose the candidate *randomly*. Remarkably, many people find such a random choice as fair. However, if we use our measure of unfairness—the numbers of pairs of candidates A, B in which A is preferred over B although A is less competent than B—then a random choice will in most cases both decrease the expected qualification of the chosen candidate *and* increase the amount of unfairness. So the random-choice strategy is not a truly satisfying solution. I conclude that a true dissolution of the conflict is not possible by the suppression of information, but only by its magnification, by trying to achieve as much information as possible about those properties that are directly relevant for the decision one has to make.

Acknowledgement. Work on this paper was supported by the DFG Grant SCHU1566/9-1 as part of the priority program "New Frameworks of Rationality" (SPP 1516). For valuable comments and inspirations I am indebted to Mario Günther, Johan van Benthem, David Papineau, Otavia Bueno, Michael Rescorla, Corina Strößner, Gila Sher and Sherilyn Roush.

References

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning*. Cambridge/M.: MIT Press.

Birkelund, G. E., Lancee, B., Nergard Larsen, E., Polavieja. J. G., Radl, J., & Yemane, R. (2022). Gender discrimination in hiring: Evidence from a cross-national harmonized field experiment. *European Sociological Review*, 38, 337–354.

Bradley, S., & Steele, K. (2016). Can free evidence be bad? Value of information for the imprecise probabilist. *Philosophy of Science*, 83, 1–28.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Cesa-Bianchi, N., and Lugosi, G. (2006): *Prediction, Learning, and Games*. Cambridge: Cambridge University Press.

Conee, E., & Feldman, R. (1998). The generality problem for reliabilism. *Philosophical Studies*, 89, 1–29.

Eells, E. (1981). Causality, utility, and decision. *Synthese*, 48, 295–329.

Good, I. J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science*, 17, 319-321. Reprinted in Good, I. J. (1983), *Good thinking* (pp. 178–180). Minneapolis: University of Minnesota Press.

Hempel, C. G. (1960). Inductive inconsistencies. *Synthese*, 12(4), 439–469.

Hempel, C. G. (1965): *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York-London: Free Press.

Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge Univ. Press.

Horwich, P. (1985). Decision theory in light of Newcomb's problem. *Philosophy of Science*, 52, 431–450.

Jeffrey, R. (1983): *The Logic of Decision* (2nd ed.). University of Chicago Press.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., & Leimer, H.-G. (1990). Independence properties of directed Markov-fields. *Networks*, 20, 491–505.

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, 59, 5–30.

Maher, P. (1990). Why scientists gather evidence. *British Journal for the Philosophy of Science*, 41, 103–119.

Matheson, J. D. (2015). Is there a well-founded solution to the generality problem? *Philosophical Studies*, 172, 459–468.

Mitchell, S., Potash, E., Barocas, S., Amour, A., & Lum, H. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 141–63.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 114–146). Dordrecht: Reidel.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

Reichenbach, H. (1949). The theory of probability. Los Angeles: University of California Press.

Roush, S. (forthcoming). Epistemic justice and the principle of total evidence.

Schurz, G. (2014): *Philosophy of Science. A Unified Approach*. New York: Routledge.

Schurz, G. (2019). *Hume's Problem Solved: The Optimality of Meta-Induction*. Cambridge/M.: MIT Press.

Schurz, G. (2024). *Optimality Justifications: New Foundations for Epistemology.* Oxford and New York: Oxford University Press.

Schurz, G., & Gebharter, A. (2016). Causality as a theoretical concept: Explanatory warrant and empirical content of the theory of causal nets. *Synthese*, 193, 1071–1103.

Schwarz, W. (2021): Counterexamples to Good's theorem. https://www.umsu.de/blog/2021/740.

Skyrms, B. (1980). *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven, CT: Yale University Press.

Skyrms, B. (1982). Causal decision theory. *Journal of Philosophy*, 79, 695–711.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press.

Steele, K., & Stefánsson, H. O. (2020). Decision theory. In *Stanford Encyclopedia of Philosophy* (Winter 2020).

Suppes, P. (1966). Probabilistic inference and the concept of total evidence. In J. Hintikka and P. Suppes (Eds.), *Aspects of Inductive Logic* (pp. 49–65). North-Holland, Amsterdam.

Thorn, P. (2017). On the preference for more specific reference Classes. *Synthese*, 194, 2025–2051.

Weirich, P. (2020). Causal decision theory. In *Stanford Encyclopedia of Philosophy* (winter 2020).