The Precautionary Principle as a Heuristic Patch

Kaivanto, Kim and Kwon, Winston

Lancaster University, University of Edinburgh

September 2015

The Precautionary Principle as a Heuristic Patch*

Kim Kaivanto†

Lancaster University, Lancaster LA1 4YX, UK

Winston Kwon

University of Edinburgh, Edinburgh EH8 9JS, UK

this version: September 29, 2015

Abstract

In this paper we attempt to recover an integrated conception of the Precautionary Principle (PP). The $\alpha = .05$ inferential-threshold convention widely employed in science is ill-suited to the requirements of policy decision making because it is fixed and unresponsive to the cost trade-offs that are the defining concern of policy decision making. Statistical decision theory – particularly in its Signal-Detection Theory (SDT) variant – provides a standard framework within which to incorporate the (mis)classification costs associated with deciding between intervention and non-intervention. We show that the PP implements preventive intervention in precisely those circumstances where the SDT-based model yields a (1,1) corner solution. Thus the PP can be understood as a *heuristic* variant of the SDT corner solution, which in turn serves to *patch* the incongruity between the inferential practices of science and the inferential requirements of policy decision making. Furthermore, SDT's analytical structure directs attention to a small number of variables – (mis)classification costs and prior probabilities – as determinants of the (1,1) corner solution. Subjective biases impinging upon these variables – omission bias, protected values, and the affect heuristic in particular, moderated by the decision maker's industry-aligned (insider) or industry-opposed (outsider) status – combine within SDT to successfully retrodict features of the PP previously considered puzzling, if not inconsistent or incoherent. These psychological biases do not exclude, and may in part reflect, the decision maker's deontological moral beliefs, or indeed social norms embodied in the nation's legal system (common law vs. civil law).

Keywords: precautionary principle; misclassification costs; scientific uncertainty; omission bias; affect heuristic; significance testing; signal-detection theory; behavioral economics

JEL classification: D81, K32, Q58

†tel +44(0)1524594030; fax +44(0)1524594244; e-mail k.kaivanto@lancaster.ac.uk

1 INTRODUCTION

The Precautionary Principle (PP) is commonly framed as being applicable only to problems that lack reliable quantitative information.(1) Accordingly, contemporary formalizations of the PP are predicated upon Knightian uncertainty, also known as ambiguity, which is distinguishable from risk in being characterized by multiple irreducible priors.(2) Thus it might appear that the PP's domain of applicability excludes the case of risk – that is, uncertainty representable with a unique probability distribution – and the very environmental- and health-hazard questions upon which scientific research is rapidly generating petabytes of quantitative information.[1]

In this paper we develop a complementary framework within which the PP retains a role even in the presence of (i) scientific research generating quantitative information and (ii) uncertainty representable with unique, if possibly high-dispersion probability distributions. The intellectual ancestry of this undertaking may be traced back to early decision analyses of hazard policy.(4–6)

We demonstrate that the PP may be understood as a post-hoc *patch* of the incongruity between the nature of information generated by current scientific practice on the one hand, and the form that this information needs to be processed into for policy decision-making purposes, on the other. Current scientific practice gives pivotal prominence to statistical significance testing and the convention – often a de facto hurdle to publication – of applying inferential procedures that discretize results into either 'significant' or 'non-significant' categories with reference to the fixed statistical significance level $\alpha = 0.05$.

At its simplest, policy action or inaction is also discrete. Policy decision making thereby also requires summative discretization of the evidence, namely into the categories 'intervention required' or 'intervention not required'. However, whereas in science the Neyman-Pearson lemma determines the accepted combination between power $(1 - \beta)$ and test size (α), in policy decision making the costs associated with misclassification – and the trade-offs between different misclassification costs – cannot be ignored. Policy decision making therefore requires incorporation of these trade-offs, which analytically equates to the determination of an optimal combination of test size and power $(\alpha^*, 1 - \beta^*)$ that reflects problem-specific misclassification costs. Signal-Detection Theory (SDT) – a binary classification framework in the tradition of Abraham Wald's statistical decision theory(7,8) – integrates these problem-specific costs by de-

[1]For instance, the US National Oceanic and Atmospheric Administration (NOAA) archives more than a petabyte (a quadrillion, or 10^{15} bytes) of new data each year. NOAA projects that the total volume of environmental data held in its archives will rise to 140 petabytes by 2020.(3)

sign. Where the costs associated with false negative errors are sufficiently large relative to the remaining misclassification costs, optimally determined test size and power $(\alpha^*, 1-\beta^*)$ yield intervention/no-intervention classifications that are observationally equivalent to the *post hoc* application of weak PP. Whereas public policy discourse cannot reliably sustain explicit application of SDT – such are the analytical and complexity limitations of public policy discourse – it can and does support application of the more straightforward, weak-form PP. In this sense, the weak PP serves to patch, rather than remedy, the mismatch between the scientific community's inferential practices and the requirements for policy making.

Yet human decision making under risk and uncertainty does not consist of cold, rational calculation alone. Instead emotions, heuristics and psychological biases are also involved, and these impact upon the way in which the PP is formulated and applied. In this sense, one can view particular PP features as *projections* of these psychological factors. Here we highlight the effects of omission bias,[9–19,23,27] protected values,[22–27] and the affect heuristic[28–34] upon the SDT-based model of the PP. These psychological factors crucially influence (i) which potential targets for PP application fall into policy focus, (ii) the development of PP variants, and (iii) the adoption of these variants by disputing interest groups, leading to sharp discord in public policy discourse.

Cass Sunstein argues that the PP fails to satisfy a basic self-consistency requirement.[35] This reprises and refines John Graham's and Frank Cross' observation that the PP should itself be subject to examination for countervailing risks.[6,36,37] A self-consistent PP application would not only prevent the risk of harm from industry's actions, but would also require prevention of second-round risk of harm *arising from the act of preventive intervention.* But in practice, PP-predicated prevention of harm is truncated after the first-round preventive intervention. In this sense, PP-based preventive intervention is in practice not uniformly deployed across impact-round iterations. Neither, however, is the PP uniformly deployed at the macro level either. With regard to honeybee Colony Collapse Disorder (CCD) for instance, the PP is invoked against neonicotinoid pesticides, but not against other, ostensibly important contributing factors: agricultural intensification, habitat loss and fragmentation, pathogens, parasites, and other environmental changes.[38] These apparent inconsistencies are rendered comprehensible – and indeed predictable – within a behaviorally augmented SDT framework.

More than 20 PP definitions are in use, ranging from weak PP through to strong PP and

super-strong PP. PP-definition variegation reflects the asymmetries of omission bias, protected values, and the affect heuristic. These asymmetries couple with interest-group internal structure as well, whereby each interest group's members coalesce around particular PP variants rather than others.

In the sequel we develop a behaviorally augmented SDT model of the PP, which successfully explains heretofore puzzling features of the PP. First we show how SDT-based optimal cutoff thresholds can be used to bridge the gap between scientific inferences and the inferences required for policy decision making. Then we show how the PP serves as an easily-understood 'patch' that implements the same preventive-intervention decisions as would be implemented under optimally determined SDT corner solutions. Finally, we turn to an investigation of how omission bias, protected values, and the affect heuristic impact upon the SDT model to make the preventive-intervention replicating corner solution more – or less – likely.

2 VARIETIES OF THE PRECAUTIONARY PRINCIPLE

Of the twenty definitions of the PP in existence, we focus here on three key spinal points in an ascending scale of stringency: weak PP, strong PP, and super-strong PP.

The PP emerged from Germany in the late 1970s as part of the country's response to large-scale environmental problems including acid rain, pollution of the North Sea, and climate change.(39) Section VII of the Ministerial Declaration announced in London at the conclusion of the 1987 Second International Conference on the Protection of the North Sea included the following statement of the PP:

> Accepting that, in order to protect the North Sea from possibly damaging effects of the most dangerous substances, a precautionary approach is necessary which may require action to control inputs of such substances even before a causal link has been established by absolutely clear scientific evidence.(40)

But the most widely known variant of the PP was adopted as Principle 15 of the 1992 UNCED Declaration on Environment and Development (the Rio Declaration):

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious

> or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.(41)

This is regarded as the definitive articulation of the *weak* PP. A slightly more verbose restatement of it appears in Article 3 of the United Nations Framework Convention on Climate Change. Under this weak variant of the PP, there is no mention of which party bears the burden of proof.

A *strong* PP variant was articulated in the Wingspread Consensus Statement on the Precautionary Principle (the Wingspread Statement), which was signed by all 32 scientists, philosophers, lawyers and environmental activists who participated in the Science and Environmental Health Network's January 24–26 1998 Conference on the Precautionary Principle held in the Wingspread Conference Center, Racine, WI:

> When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically. In this context the proponent of an activity, rather than the public, should bear the burden of proof.[2]

Unlike the weak PP, the strong PP (i) does not mention costs, (ii) does not acknowledge that different states have different levels of resources ('capabilities') available for environmental protection, and (iii) does not limit preventive intervention to threats of serious or irreversible harm. The strong PP employs the operative word 'should', which can refer to either the moral duty (moral imperative) for action, or the moral desirability of action. Hence from the text of the strong PP alone, it is not clear (a) whether preventive intervention is called for as a moral, categorical imperative, regardless of the direct and indirect (opportunity) costs of implementing preventive intervention, or (b) whether preventive intervention is called for as being desirable, yet subject to the practical direct- and indirect-cost trade-offs within the totality of obligations involved in running a nation state, given its resources and degree of economic development. From context we may infer that Wingspread Conference attendees intended the former, moral-imperative interpretation. But this is not evident from the text of the Wingspread Statement alone. Finally – yet crucially – the strong PP explicitly imposes the burden of proof on the proponent of an activity.

The *super-strong* PP precludes aforementioned ambiguity by specifying not only the burden of proof, but also the standard of proof:

[2] `http://www.sehn.org/wing.html`

> the [PP] mandates that when there is a risk of significant health or environmental damage to others or to future generations, and when there is scientific uncertainty as to the nature of that damage or the likelihood of the risk, then *decisions should be made so as to prevent such activities from being conducted unless and until scientific evidence shows that the damage will not occur.*(42) [emphasis added]

Thus under the super-strong PP preventive intervention is the default condition when (i) there is a risk – any risk – of significant harm and (ii) there is scientific uncertainty over the level or probability of that harm. The burden of proof lies with those who wish to proceed with the potentially harmful activity. The standard of proof required by the super-strong PP is extreme, in that preventive intervention remains in place "until scientific evidence shows that the damage will not occur." This is not the preponderance-of-evidence ($> 50\%$) standard of proof employed in US Common Law. Neither is it full conviction of the judge (90%, 95%, or 99.8%) standard of proof employed in continental European Civil Law.(43) A literal reading of the super-strong PP requires a 100% standard of proof to be achieved before preventive intervention may be withdrawn.

Henceforth, references to 'the PP' shall be read as references to the weak-PP variant, unless separately stipulated otherwise.

3 PP AS A PATCH

In computer science the term *patch* refers to retrospectively installed update code that repairs, improves or adapts the functioning of an existing piece of software. Although not constituted of computer software code, the PP serves as a patch in this sense, adapting the output of scientific inferential conventions to the misclassification-cost-sensitive requirements of policy decision making. For problems satisfying weak-PP applicability criteria, preventive intervention decisions are thus triggered 'as if' they had been taken under optimally determined inferential thresholds.

This role as a patch bridges between (a) fixed-inferential-threshold convention in science, and (b) misclassification-cost-optimal inferential thresholds for policy decisions. We present each of these below in turn before turning to PP as a patch.

3.1 Scientific inferential convention: NHST

Null Hypothesis Significance Testing (NHST) is the workhorse method of statistical inference in modern science. It combines Neyman and Pearson's concept of a critical rejection region[44] with Fisher's formulation of p-values.[45] Although there are basic, pointed philosophical differences between the developers of these two concepts,[3] in modern usage these differences have been glossed over or subsumed within a unified framework.[47,48]

That NHST has become a central preoccupation within empirical science was critically noted already by Yates.[49] Since Yates, criticism of this preoccupation and of NHST per se has been repeated and expanded.[50–52] John Ioannidis' widely cited paper entitled 'Why most published research findings are false' represents one culmination of this stream of criticism.[53] Some of the strongest and most persistent critics of NHST are advocates of Bayesian statistical methodology.[54] Nevertheless NHST remains the prevailing convention – in all but one journal of which we are aware.[4]

3.1.1 The fixed $\alpha = 0.05$ threshold

Fisher introduced significance testing and the concept of a p-value, i.e. the probability that a test-statistic $T = t(X)$,[5] equals or exceeds the observed value $t(x)$ given that the null hypothesis $H_0 : \theta = \theta_0$ is true, i.e. $p = P(t(X) \geq t(x)|H_0)$. In Fisher's approach to significance testing, there is no explicit alternative hypothesis under consideration. This is because there are innumerable different conceivable alternative hypotheses. Fisher views the alternative hypothesis – and therefore any quantities derived from it, such as statistical power – as 'unknown'. Although Fisher believed that p-values require researchers' *subjective* interpretation, his early expositions advocated using $p < 0.05$ (i.e. a 5% significance level) as the standard for concluding that there is evidence against H_0.

> [In 1925:] The value for which P = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered

[3] The distinction between 'inductive inference' as advocated by Fisher, and 'inductive behavior' as advocated by Neyman, was at the heart of of their disagreement. Neyman advocated a theory of mathematical statistics predicated on probability (not subjective likelihood), the basis of which is provided by "the conception of frequency of errors in judgement."[46,47]

[4] In 2015, the editors of *Basic and Applied Social Psychology* announced that they will be removing p-values and other NHST measures from papers published in *BASP*.[55]

[5] computed on observed data drawn from a continuous distribution $X \sim f(x|\theta)$ on support $\mathbb{R}$

> significant or not. ... ...We shall not often be astray if we draw a conventional line at 0.05 ... .[56]
>
> [In 1926:] Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level.[57]
>
> [In 1935:] It is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard... .[58]

Fisher viewed the p-value as an index of the 'strength of evidence' against H_0. Fisher's approach to significance testing thus focuses on controlling type-I error alone. Although in his later work Fisher attacked the notion of a standard or conventional threshold for type-I error, empirical researchers continue to employ the $\alpha = 0.05$ level suggested by Fisher. Fisher's influential texts included tabulations of exact small-sample $\mathcal{X}^2$-, t- and F-test statistics. He economized on page-space and enhanced the usability of his tables by providing only selected quantiles, key among which being the 5% quantile. Neyman and Pearson followed suit in endorsing a fixed 5% level – and in turning their attention to controlling type-I error and in developing their method around a 'rule of behavior' – under the influence of Fisher's 5% and 1% quantile tables.[47]

Neyman and Pearson held that one could only test a null hypothesis *against* an alternative hypothesis. Thus Neyman and Pearson were concerned with type-II error as well as type-I error. Following this concern, they introduced the concept of statistical power. They sought to supplant the subjective element present in Fisher's approach with a formalized decision procedure (a behavioral rule) embodying the frequentist principle: "In repeated practical use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error."[48] Neyman and Pearson sought to distinguish their theory from Fisher's 'significance testing', and did so by referring to their formalized decision rule as 'hypothesis testing'.

Statement 3.1 (Neyman-Pearson hypothesis testing)**.**

(i) Derive type-I and type-II error probabilities $\alpha = P(t(X) \geq c\,|H_0)$ *and* $\beta = P(t(X) < c\,|H_1)$ *for given for simple hypotheses* $H_0 : \theta = \theta_0$ *and* $H_1 : \theta = \theta_1$ *where* $X \sim f(x|\theta_i)$, $i = \{0, 1\}$, $\theta_1 > \theta_0$, *and* c *is a critical threshold in the codomain of* $t(\cdot)$*;*

(ii) Determine the most powerful test (in particular its critical threshold c) and the most appropriate type-I error probability α^* using $\alpha = P(t(X) \geq c\,|H_0)$, $\beta = P(t(X) < c\,|H_1)$, $X \sim f(x|\theta_i)$, and the costs associated with type-I and type-II errors;

(iii) Use the pre-chosen critical value c to reject H_0 if $t(X) \geq c$, else accept H_0.

Notice that there are two components in Part (ii) of this statement. The first is the determination of the most powerful test. This is accomplished with the Neyman-Pearson lemma. The second is the determination of the most appropriate type-I error probability α^*. For this, Neyman and Pearson did not provide a formal procedure, but offered clear verbal guidance. We elaborate the Neyman-Pearson lemma first, followed by α^*, even though the latter is technically a required input parameter for application of the Neyman-Pearson lemma. The following presentation of the Neyman-Pearson lemma is adapted from Lehmann and Romano,[59] which may also be consulted for the associated proof.

Theorem 3.1 (Neyman-Pearson lemma)**.** *Let there be two continuous distributions $X \sim f(x|\theta_i)$, $i = \{0, 1\}$, indexed by the parameters $\theta_1 > \theta_0$.*

(i) Existence. For testing the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$, there exists a test function ϕ and a constant $k > 0$ such that

$$E_{\theta_0}\phi(X) = \alpha \tag{3.1}$$

and

$$\phi(x) = \begin{cases} 1 & \text{if} \quad \frac{f(x|\theta_1)}{f(x|\theta_0)} > k \\ 0 & \text{if} \quad \frac{f(x|\theta_1)}{f(x|\theta_0)} < k \end{cases} \tag{3.2}$$

(ii) Sufficient condition for a most powerful test. If ϕ satisfies (3.1) *and* (3.2) *for some constant k, then ϕ is Most Powerful (MP) for testing H_0 against H_1 at level α.*

(iii) Necessary condition for a most powerful test. If a test ϕ^* is MP at level α, then it satisfies (3.2) *for some k, and it also satisfies* (3.1) *unless there exists a test of size strictly less than α with power 1.*

Although the Neyman-Pearson lemma is framed in terms of simple hypotheses, the test ϕ^* can be shown to be Uniformly MP against a composite alternative hypothesis when the family of distributions indexed by θ_i satisfies the monotone likelihood ratio property.

Neyman and Pearson explicitly acknowledge that the critical threshold c, which demarcates between the null-hypothesis rejection region and the null-hypothesis acceptance region, should be determined by the researcher. This determination is dependent upon the context:

> ...in some cases it will be more important to avoid the first [type-I error], in other the second [type-II error]... ...determining just how the balance should be struck, must be left to the investigator. ... ...we attempt to adjust the balance between the risks [of the two types of error] to meet the type of problem before us.(44)

In this 1933 formulation, consideration of consequences – costs of error – remain implicit. With time Neyman's position shifted, however. In 1950 he articulated the view that controlling type-I errors is 'more important' than controlling type-II errors:

> Because an error of the first kind is more important to avoid than an error of the second kind, our requirement is that the test should reject the hypothesis tested when it is true very infrequently... ...The ordinary procedure is to fix arbitrarily a small number α... ...and to require that the probability of committing an error of the first kind does not exceed α.(60)

From these beginnings, inertia took hold.(61) Today, use of $\alpha = 0.05$ reflects a customary, conventional, common frame of reference:

> It is customary therefore to assign a bound to the probability of incorrectly rejecting $[H_0]$ when it is true and to attempt to minimize the other probability subject to this condition. ... ...The choice of a level of significance α is usually somewhat arbitrary... ...Standard values, such as .01 or .05, were originally chosen to effect a reduction in the tables needed for carrying out various test [sic]. By habit, and because of the convenience of standardization in providing a common frame of reference, these values gradually became entrenched as the conventional levels to use.(59)

The key feature of operating under the Neyman-Pearson lemma is accepting – as given, short of sample-size considerations – the maximum achievable statistical power $1 - \beta = E_{\theta_1}\phi(X)$ associated with level α. This is equivalent to fixing α on the abscissa of the Receiver Operating Characteristics (ROC) space, and accepting as given the associated *power* as indicated by the ordinate of the ROC curve, i.e. the locus of all $(\alpha, 1 - \beta)$ points obtained parametrically by

varying the cutoff threshold, given the distributions $X \sim f(x|\theta_i),\ i = \{0, 1\}$. Neither NHST nor the Neyman-Pearson lemma supports any explicit consideration of *trade-offs* between type-I and type-II errors.

3.1.2 Observations

The present paper is not intended to augment the general critique of NHST. Nevertheless we flag three observations which also feature in that literature.

First, note that the $\alpha = 0.05$ level is, ostensibly, arbitrary.[51,59] Section 3.1.1 traces the broad outlines of how this convention arose, starting with the recommendations and statistical tables of Ronald Fisher. In fact the $\alpha = 0.05$ level is not a sufficiently demanding criterion that it would identify only strong evidence against the null.

Second, modern commentators such as David Cox are in agreement with Ronald Fisher, who held that drawing sharp distinctions between p-values such as 0.051 and 0.049 introduces an artificially sharp dichotomy.[62] Ceteris paribus, the evidential value of a study supplying a p-value of 0.051 is virtually indistinguishable from that of a study supplying a p-value of 0.049. Applying the labels 'non-significant' to the former and 'significant' to the latter facilitates dichotomous thinking – where the underlying evidence does not in itself support such a distinction.

Third, 'statistical significance' is not synonymous with 'scientific significance'.[62] The connection with policy-making relevance is even more tenuous. For instance observational studies can achieve statistical significance by virtue of sample size, but the effect size may be miniscule, contributing little to overall scientific understanding or to the understanding of effective policy levers for decision making.

However, as we show in the following section, these three detractions lose force when a fixed α is abandoned in favor of a contextually optimal inferential threshold α^*.

3.2 Optimal inferential thresholds for policy decisions

Some of the problems inherent in NHST as currently practiced can be addressed through incorporation of a context-dependent loss function into the determination of an appropriate α level to be used within the Neyman-Pearson lemma. Among the numerous approaches to incorporating error costs into statistical inference, the simplest – and one which has the advantage of being consistent with Neyman and Pearson's frequentist approach – is known as Signal Detection The-

ory (SDT).[63–65] The core elements of SDT, in addition to the above-mentioned ROC curve, are (i) the misclassification cost matrix, (ii) the objective function under which the inferential threshold is to be optimized, and (iii) the population prevalence rates of the conditions captured in H_0 and H_1 respectively, i.e. the parameters in frequentist statistics which correspond to Bayesian prior probabilities for H_0 and H_1.

We begin by introducing the confusion matrix, entries of which consist of True Positives (TP), False Negatives (FN), False Positives (FP) and True Negatives (TN) counts obtained from repeated application of a specific threshold x' (see Table 1a). It is common to re-express these entries as row-specific (within-hypothesis) rates: $TPR = TP/(TP + FN)$, $RNR = FN/(TP + FN)$, $FPR = RP/(FP + TN)$, $TNR = TN/(FP + TN)$. Associated with each cell of the confusion matrix is a corresponding misclassification cost, which is independent of the value of the threshold x' employed to generate the confusion matrix (see Table 1b). The essence of 'context' is represented via a particular set of misclassification costs. For the purpose of presenting SDT, misclassification costs are assumed to be measured or estimated in an unbiased manner, reflecting overall societal concerns. This entails unbiased accounting for both atemporal (i.e. generation-specific) as well as intertemporal (i.e. inter-generational) externalities.[6]

Table 1: Classification matrices.

(a) Confusion matrix (counts).

		Inference under x'	
		$\neg H_0$	H_0
Actual	H_1	TP	FN
	H_0	FP	TN

(b) Misclassification cost matrix.

		Inference	
		$\neg H_0$	H_0
Actual	H_1	C_{TP}	C_{FN}
	H_0	C_{FP}	C_{TN}

Letting N denote the total number of observations in the (random) sample $TP+FN+FP+TN = N$, then the sample-based estimates of the population prevalence rates may be written as $P(H_0) = (FP + TN)/N$ and $P(H_1) = (TP + FN)/N$.

With few exceptions,[66] applications of SDT are couched in terms of *minimizing expected misclassification cost*. The central results of classical SDT are all derived under this expected misclassification cost objective function. For present purposes – including those of Section 4 – the parsimony and tractability of this objective function serve well.

The optimally chosen cutoff threshold x^* minimizes expected misclassification costs $E(C)$

[6] Consideration of the consequences flowing from the possibility that different interest groups may face different misclassification costs or hold different priors is deferred to Section 4.

subject to the constrained relationship between the TPR and the FPR, which may be represented with the twice-differentiable function $G : [0,1] \to [0,1]$. This function, written as TPR $= G(\text{FPR})$, captures the ROC curve. As N grows larger, $\lim_{N\to\infty} TPR = 1-\beta$ and $\lim_{N\to\infty} FPR = \alpha$, which in turn are defined by

$$\alpha = P(X > x' \,|\, \theta_0) = \int_{x'}^{+\infty} f(x|\theta_0)\,\mathrm{d}x \tag{3.3}$$

$$1-\beta = P(X > x' \,|\, \theta_1) = \int_{x'}^{+\infty} f(x|\theta_1)\,\mathrm{d}x \ . \tag{3.4}$$

The slope at a point on the ROC curve determined parametrically by x' is given by the derivative at the point x'

$$\left(\frac{\mathrm{d}P(X > x' \,|\, \theta_1)}{\mathrm{d}P(X > x' \,|\, \theta_0)}\right)_{x'} = \frac{-f(x'|\theta_1)}{-f(x'|\theta_0)} = l(x') \ , \tag{3.5}$$

which is the likelihood ratio at x'. We assume $G' > 0$ and $G'' < 0$, ensuring that the monotone-likelihood ratio condition holds.[7]

Solving the constrained minimization problem

$$\min_{x'} E(C) \qquad \text{s.t.} \qquad 1-\beta = G(\alpha) \tag{3.6}$$

gives the optimality condition

$$l(x^*) = \frac{P(\theta_0)}{P(\theta_1)}\left[\frac{C_{\mathrm{FP}} - C_{\mathrm{TN}}}{C_{\mathrm{FN}} - C_{\mathrm{TP}}}\right] = \left(\frac{\mathrm{d}(1-\beta)}{\mathrm{d}\alpha}\right)_{\bar{C}^*} \ , \tag{3.7}$$

which states that the slope of the cost-minimizing iso-expect-cost line at the optimal operating point is given by the ratio of the expected opportunity cost of misclassifying a Negative to the expected opportunity cost of misclassifying a Positive. From (3.5) and (3.7) it is also clear that the optimality condition defines the critical likelihood ratio $l(x^*)$, and that (3.7) is a tangency condition between the least-cost iso-expected-cost line and the ROC curve. From (3.3) and (3.4)

[7] Note that $G'' < 0$ is not satisfied by arbitrary combinations of sampling distributions. When both distributions are Gaussian, $G'' < 0$ is satisfied everywhere in the support of x only when the two sampling distributions have the same variance.[(67)]

we have that

$$\alpha^* = \int_{x^*}^{+\infty} f(x|\theta_0)\,\mathrm{d}x \tag{3.8}$$

$$1-\beta^* = \int_{x^*}^{+\infty} f(x|\theta_1)\,\mathrm{d}x \ . \tag{3.9}$$

When the cutoff threshold is optimally determined by (3.7), the associated optimal level of the test α^* responds to changes in misclassification costs and population prevalence rates $P(\theta_0)$ and $P(\theta_1)$. Setting $\theta_0 = 0$ WLOG and furthermore assuming Gaussian sampling distributions $X \sim N(\theta_i, 1),\ i = \{0,1\},\ \theta_1 > \theta_0$, the optimal cutoff threshold x^* responds to the remaining parameters as follows:

$$x^* \ = \ \frac{1}{\theta_1}\left(\ln(C_{FP}-C_{TN}) - \ln(C_{FN}-C_{TP}) + \ln(P(\theta_0)) - \ln(P(\theta_1)) + \frac{\theta_1^2}{2}\right) \ . \tag{3.10}$$

If misclassification costs are symmetrical in the sense that $C_{FP} - C_{TN} = C_{FN} - C_{TP}$ and the base-rate probabilities are also symmetrical $P(\theta_0) = P(\theta_1)$, then the optimal cutoff threshold x^* falls half-way between θ_0 and θ_1, where the two pdfs intersect $f(x^*|\theta_0) = f(x^*|\theta_1)$. The associated optimal operating point $(\alpha^*, 1-\beta^*)$ is that ROC-curve point that coincides with the minor diagonal, where the slope of the iso-expected-value line is unity $l(x^*) = 1$. Due to the concavity of $\ln(\cdot)$, increasing misclassification-cost increments have a diminishing impact upon x^*. However, the natural logarithm's concavity and limiting value $\lim_{P\to 0^+} \ln(P) = -\infty$ entail that the hypothesis with the smaller base rate has a disproportionately larger impact upon the location of the optimal cutoff threshold. This responsiveness characteristic of x^*, α^* and $(1-\beta^*)$ under SDT sits in contradistinction to their fixed nature under the Neyman-Pearson lemma, i.e. $1-\beta^{\mathrm{NP}} = G(0.05)$.

Whereas the α level is arbitrary under NHST, it is optimally adapted to base-rates and misclassification costs in SDT. Whereas in NHST, the distinction made between p-values 0.051 and 0.049 is artificially sharp, under SDT the distinction made between p-values $\alpha^* + 0.01$ and $\alpha^* - 0.01$ is not artificial, but anchored in real-world consequences (C_{FP}, C_{TN}, C_{FN}, C_{TP}) and base rates ($P(\theta_0)$, $P(\theta_1)$). Finally, whereas statistical significance in NHST is not synonymous with scientific or decision-making significance, rejecting the null hypothesis under SDT's optimal α^* level is, by design, synonymous with decision-making significance.

We conclude this section by noting that the approach embodied in SDT is consistent with David Cox's general exhortations concerning the use of p-values.

> The P-value has, *before action or overall conclusion can be reached*, to be combined with any external evidence available and, *in the case of decision-making*, with *assessments of the consequences of various actions.*[62] [emphasis added]

3.3 PP for 'as if' optimal inference

In this section we show that application of the weak PP is observationally equivalent to a particular corner solution in the SDT framework. Under the weak PP, absence of conclusive evidence and the persistence of uncertainty does not constitute sufficient grounds not to proceed with preventive intervention. Taking $H_0 : \theta_0 = \theta$ to be the status-quo level of the critical index variable and $H_1 : \theta_1 = \theta$ ($\theta_1 > \theta_0$) to be the (irreversible) higher value of the critical index-variable induced by a harmful commercial innovation, then we may note that the PP-based policy decision (preventive intervention) is observationally equivalent to the SDT-based policy decision associated with the corner solution in which $1-\beta^* = 1$ and $\alpha^* = 1$. This corner solution obtains whenever the following condition holds.

Condition 3.1 (PP-mimicking corner-solution condition).

$$\frac{P(\theta_0)}{P(\theta_1)}\left[\frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}}\right] \;\leq\; \lim_{x' \to -\infty} \frac{f(x'|\theta_1)}{f(x'|\theta_0)} \;=\; \lim_{\alpha \to 1} G'_{d'}(\alpha) \;. \tag{3.11}$$

The limit on the right-hand side of this inequality depends on three parameters succinctly summarized by the *discriminability* index:

$$d' \;=\; \frac{\theta_1 - \theta_0}{\sigma} \;. \tag{3.12}$$

A given $\theta_1 - \theta_0$ difference can be consequentially large or consequentially small, depending on the value of σ. Small absolute effect sizes $\theta_1 - \theta_0$ and large standard deviations – whether due to limited precision of scientific measurement or due to explicit gaming of the research process by non-independent researchers (see Section **??**) – are associated with small Area Under the Curve, AUC $= \Phi\left(\frac{d'}{\sqrt{2}}\right)$, where Φ is the standard normal CDF. Along the principal diagonal of the ROC space, where d'=0 and AUC=0.5, the SDT-based inference performs no no better than

chance, as achieved e.g. with the toss of a fair coin. Larger d' and AUC permit improvement, in principle, over mere chance (see Figure 1a).

The limit on the right-hand side of (3.11) may be identified in Figure 1b, which plots ROC-curve slopes for four discriminability-parameter values (.2, .5, 1, 2). The right-hand side vertical intercept of each d'-specific curve gives $\lim_{\alpha\to 1} G'_{d'}(\alpha)$. Applying Condition 3.1 to Figure 1b, it can be seen that a PP-mimicking corner solution obtains when the slope of the iso-expected-cost line falls within the half-open interval between zero and this right-hand side vertical intercept. For $d' = 0.2$ for instance – which corresponds to questions subject to considerable scientific or measurement uncertainty – this interval is $[0, 0.5)$. As d' and AUC grow larger, the upper boundary of this corner-solution supporting interval collapses toward zero. However, for all non-degenerate corner-solution supporting half-open intervals $[0, \lim_{\alpha\to 1} G'_{d'}(\alpha)) \equiv \Gamma_{d'}$, $\Gamma_{d'} \neq \varnothing$,

$$\frac{P(\theta_0)}{P(\theta_1)}\left[\frac{C_{\mathrm{FP}} - C_{\mathrm{TN}}}{C_{\mathrm{FN}} - C_{\mathrm{TP}}}\right] \in \Gamma_{d'} \tag{3.13}$$

is satisfied in the region of the parameter space where the expected cost of misclassifying a True Negative is sufficiently small relative to the expected cost of misclassifying a True Positive. Obviously, if either $P(\theta_1) \to 1$ or $(C_{FN} - C_{TP}) \to \infty$, or both, then (3.13) is satisfied. But these extreme limits are not necessary for the corner solution. It is sufficient for the slope of the iso-expected-value line to fall within $\Gamma_{d'}$.

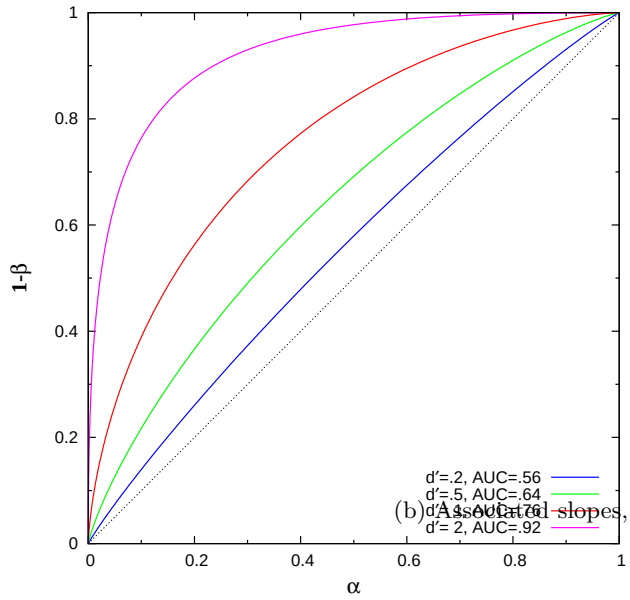
The wording of the PP selects as targets for preventive intervention those commercial innovations that are characterized by a large denominator in Equation (3.13). The Rio declaration, embodying the weak PP, focuses preventive intervention "[w]here there are threats of serious irreversible damage...". Here 'threats' refers to $P(\theta_1)$, while 'serious irreversible damage' refers to $(C_{FN} - C_{TP})$. Unlike the weak PP however, strong and super-strong variants of the PP cannot be rationalized within the SDT framework without making recourse to behavioral effects. These effects are strong and widely felt.

4 PP AS A PROJECTION OF BEHAVIORAL EFFECTS

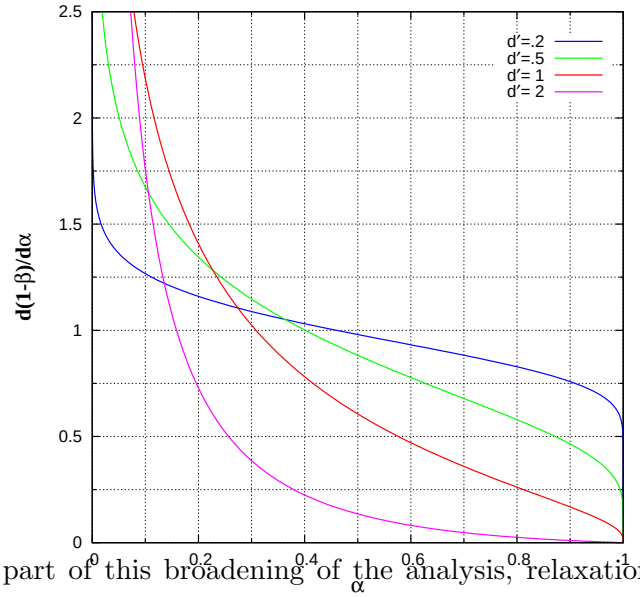
Whereas the analysis in Sections 3.2 and 3.3 is conducted under the assumption that parameters, distributions and misclassification costs are independently identifiable and unique at the societal level, here we introduce the effects of subjectivity, psychology and attendant heterogeneity.

Figure 1: ROC curve and its slope, assuming $\sigma_0 = \sigma_1 = 1$.

(a) ROC curves; four different discriminability parameters.



(b) Associated slopes, $G'_{d'}(\alpha)$.



As part of this broadening of the analysis, relaxation of the frequentist interpretation allows the formal analytical representation to incorporate heterogeneity in the prior-odds component of the optimal cutoff threshold expression. Different people, interest groups and institutions confront societally consequential problems with potentially very different subjective priors and perceptions concerning the costs of different eventualities. Furthermore, psychological effects induce or accentuate asymmetries in perceived, subjective misclassification costs, and these asymmetries are expressed differently in different people, interest groups and institutions.

4.1 Priors

The frequentist framework requires that the terms $P(\theta_0)$ and $P(\theta_1)$ in equation (3.7) be interpreted as base rates or population prevalence rates. Relaxing this requirement, allowing these terms to be viewed as priors, expands SDT's scope to embrace larger, longer-term, prospective societal and planetary challenges. It also permits heterogeneity between individuals and interest groups in the priors they hold. This is the case even without invoking behavioral biases. Conceptual reorientation away from frequentist base rates and toward (possibly subjective) priors is a prerequisite to bringing behavioral considerations to bear upon SDT.

Many of the consequential and controversial long-term societal, environmental, and planetary issues facing humanity are largely prospective in nature. We do not have access to dozens or hundreds of earth replicas that underwent hydrocarbon-fueled industrialization so that we can establish the relative frequencies of those worlds on which catastrophic global warming took place and those worlds on which global warming proved to be benign. Anthropogenic global warming, like many other large- and smaller-scale threats, is a new emergent problem, rather than one that has been experienced many times before, and for which a gold-standard test exists. For this reason, a conception of probability that is restricted to conveying relative frequency information is ill-suited to many of the most challenging problems to which the PP is being applied.

The priors $P(\theta_0)$ and $P(\theta_1)$ are most appositely understood in broadly Bayesian terms, representing the state of knowledge or belief, integrating and summarizing available evidence. Following Leonard Savage, "Probability measures the confidence that a particular individual has in the truth of a particular proposition...".[(68)] Reasonable individuals who conform with Savage's seven postulates (axioms) may hold different degrees of confidence in a particular proposition, even after having viewed precisely the same body of evidence. Thus, another consequence of abandoning the frequentist framework is the need to recognize and embrace the underlying heterogeneity of subjective probabilities. Heterogeneity without discernible differentiation or distinction does not lend itself to enlightening analysis, however. But behavioral factors – specifically 'affect heuristic' effects discussed in Section 4.4 – systematically shape subjective priors in ways that are consequential for an understanding of the discord that attends PP-based preventive intervention.

4.2 Omission bias

Consider a choice setting in which a decision maker is confronted with two alternatives. Both alternatives lead to the same, objectively identical negative outcome. The first alternative involves passively letting nature take its course, i.e. inaction in the sense of omission of separate observable action. The second alternative involves taking an explicit, observable action. *Omission bias* is the tendency to favor the first (omission, inaction) alternative over the second (commission, overt action) alternative.[9–19,23,27]

In certain circumstances, the distinction between harm by omission and harm by commission is not purely moral or psychological. Omission may result from ignorance or an attention budget deployed across other considerations, while a responsible act of commission requires effort and conscious intent, which cannot be predicated on ignorance or limited attention. If on the other hand knowledge, intent and consequences are the same in the case of harm by omission as in the case of harm by commission, there should be no consequentialist grounds for distinguishing between the passively permitted harm and the actively chosen harm. For this reason, some authors add a rider that restricts omission bias to being an overgeneralization of distinctions between commissions and omissions to problem settings in which these distinctions are absent.[9,10]

Omission bias as an empirical regularity has proven robust in replication studies, both in the laboratory and in the field. Empirical studies have been situated in the context of risky medical treatments,[9–11,14,15] financial decisions,[12,13] professional sports refereeing,[16] civil litigation (standard of proof),[17] risky (conditional probability) lotteries,[18] Tax Credit repayment,[19] and human or animal deaths.[23,27] Carefully designed studies have disentangled the omission-bias effect from status-quo bias,[12] and normality bias.[14] This body of evidence notwithstanding, there has been some work critical of the concept,[20] and other work arguing that omission is a strategy choice (with plausible deniability) rather than a psychological bias.[19,21]

Factors ranging from feelings of regret to moral and ethical principles have been offered as explanations for the empirical instantiation and prevalence of omission bias. In consequentialist ethical systems – such as utilitarianism – the personal and moral assessment of choices is based solely on the outcomes yielded by those choices. Consequentialist moral assessment excludes the intent of the decision maker, the moral character traits of the decision maker, the nature of the choice process, and the manner in which choices implement final outcomes (e.g. action

or inaction). Indeed, consequentialist moral assessment excludes all aspects of a decision problem save the final outcomes. In contrast, deontological ethical systems assess choices without reference to final outcomes. Under deontological ethics, proscribed types of action choices are not rendered morally acceptable regardless of any possible positive outcomes – and regardless of their magnitude, whether measured in absolute terms or in relative terms – that they bring about. Instead, choice is guided by moral rules and moral duty. Furthermore, some kinds of action choices are strictly permitted, regardless of whether they are outcome dominated by other potential action choices.

Omission bias would not be observed within a purely consequentialist decision-making framework. The psychology of regret[(69)] and the distinction between direct and indirect causation[(14)] are the primary lenses through which we will understand omission bias in the present paper. Even though we largely eschew moral philosophy – leaving this to those with competence to address the profound philosophical issues involved – it is also clear that empirically observed omission bias and the underlying regret aversion may derive in part from deontological moral duty such as, "Above all, do no harm."[(13,26,27)]

Common law distinguishes between acts, which one is liable for, and omissions, which one is generally not liable for. A manufacturer owes its customers a duty of care, and is liable for harm caused by its products. However, a manufacturer is not liable for the harm that could have been prevented if it had produced a particular product – pharmaceutical companies are not held liable for not producing specific vaccines, or for not producing treatments for particular diseases (e.g. orphan diseases).[(70)] This asymmetry also holds for individuals. Under common law, there is no general legal duty for a bystander to warn, prevent or assist an individual in peril.[8] A bystander can watch a child drown, or a blind man walk into the path of an oncoming vehicle, and not be held to account.[(71)] However, if an individual has created a hazardous situation that has placed another person in peril, then the legal duty to rescue does exist. Whether this legal duty exists or not, the rescuer can be held liable for injuries caused by ill-advised actions during the rescue attempt.[9]

In a classic application of decision analysis, Ronald A Howard, James E Matheson, and D Warner North evaluated whether the U.S. government should seed hurricanes with silver iodide

[8] In contrast, under civil law as in e.g. continental Europe and Quebec, it is a criminal offence not to assist an individual in an emergency.

[9] Good-Samaritan laws, passed in many US states, limit the extent of this liability.

to reduce their intensity and thereby attenuate their destructive force.[4] The analysis reveals that when a hurricane is seeded by the U.S. government, the damage it causes ceases to be purely of a 'natural disaster' variety. Intervening in the development of a storm alters who subsequently suffers damage and losses, and those conducting the intervention (seeding) become responsible – morally and politically, perhaps even legally – for the damage that would not have occurred without the seeding intervention (even though it is not possible to definitively determine who these people are). The decision analysis thus has to factor in the 'government responsibility cost' associated with the seeding. The analysts conclude that there is no firm legal basis for operational seeding of hurricanes, that the sovereign immunity enjoyed by government is only partial and unpredictable protection, and that there are substantial grounds for individuals to recover damages where it can be proven that seeding caused harm.

4.2.1 Effect on misclassification costs

Outsider perspective: From an industry outsider's perspective, the industry's decision to introduce an innovation is seen as a deliberate act. The potential harms from this act of innovation are therefore weighted more heavily than any harms that would occur in the absence of the deliberate introduction of the innovation, i.e. by omission of this innovation. Consequently, the expected cost of misclassifying a positive (i.e. of classifying as non-harmful an innovation that is in fact harmful) is biased upward relative to the expected cost of misclassifying a negative (i.e. of classing as harmful an innovation that is in fact not harmful). This biases the slope of the iso-expected-value line *downward*, making it *more likely* to fall within the interval Γ where the SDT corner solution obtains.

Insider perspective: For industry, periodic if not continual innovation is a natural part of its very existence. Hence the salient act is not innovation, but the decision to implement protective intervention. Just as in both civil litigation[17] and professional sports refereeing,[16] the harm associated with mistakenly undertaking this act is overweighted relative to the harm associated with mistakenly omitting to undertake this act. Consequently, the expected cost of misclassifying a negative (i.e. of classifying as harmful an innovation that is in fact not harmful) is biased upward relative to the expected cost of misclassifying a positive (i.e. of classifying as non-harmful an innovation that is in fact harmful). This biases the slope of the iso-expected-

value line *upnward*, making it *less likely* to fall within the interval Γ where the SDT corner solution obtains.

4.2.2 PP target selection: Which problems?

We suggest that the distinction between omission and commission also has a bearing upon which problems trigger PP-based preventive intervention. Consider Colony-Collapse Disorder (CCD), in which worker bees disappear from the colony, leaving the colony unviable. The United States Department of Agriculture (USDA) has identified a number of important contributory causal factors for CCD: insecticides, agricultural intensification, habitat loss and fragmentation, pathogens, parasites, and other environmental changes.[(38)] Many of these are slow-moving, long-existing background factors that are difficult to attribute to the actions of specific entities in the economy. Only a subset – specific types of new insecticides, such as neonicotinoids – appear as deliberate acts by identifiable agents. Hence, these new insecticides are viewed and evaluated according to the 'outsider perspective' discussed above in Section 4.2.1. Accordingly, the slope of the iso-expected-value line is biased downward, making it more likely to fall within the interval Γ where the SDT corner solution obtains. Conversely, the remaining factors – agricultural intensification, habitat loss and fragmentation, pathogens, parasites, and other environmental changes – are viewed and evaluated according to the 'insider perspective' discussed above in Section 4.2.1. Accordingly, the slope of the iso-expected-value line is biased upward, making it less likely to fall within the interval Γ where the SDT corner solution obtains.

4.2.3 PP target selection: $n \geq 2$ round effects?

When preventive intervention is implemented, why are the possible harms associated with the act of preventive intervention not themselves subject to PP-based preventive intervention? Indeed an infinite regress of such questions can be constructed. Why are all of these $n \geq 2$ round effects not subject to PP-based preventive intervention?

The answer lies again in the distinction between the 'insider perspective' and the 'outsider perspective', and how these become applied across the sequence of consequent harms. From the perspective of the pro-PP faction, the $n = 1$ first-round effect is the result of a conspicuous act (i.e. the industry's innovation) with harms that are biased in accordance with the 'outsider perspective', leading to an SDT corner solution and preventive intervention. The pro-PP faction

sees the $n = 2$ second-round effect not as the result of an explicit preventive-intervention action, but as merely a preservation of the state of affairs that existed prior to the industry introducing its innovation. Being pre-existing, the pro-PP faction applies an 'insider perspective', which biases the expected cost of misclassifying a negative upward and the expected cost of misclassifying a positive downward. The slope of the iso-expected-value line is biased upward, making it less likely to fall within the interval Γ where the SDT corner solution obtains. Hence preventive intervention is also less likely. Without preventive intervention for the $n = 2$ second-round effect, there are no further rounds of consequent harms.

Thus, the self-consistency violation identified by Cass Sunstein[35] is explained with an omission-bias augmented SDT model of the PP.

4.3 Protected values

Investigation of Protected Values (PVs) started within and emerged from the omission-bias literature.[22–27] PVs are rooted in deontological ethical principles, and may reflect personal or social norms. The defining characteristic of PVs is their absolute resistance to trade-offs: they are in this sense 'protected' from being subject to trade-offs with other values or attributes. This means that no amount of compensating benefit will induce an individual to make even a small sacrifice to her PV. For an individual who views ecosystem life as sacrosanct (i.e. a PV), there is no finite amount of compensating economic gain that could justify the extinction of a single species. In terms of utility, PVs are associated with vertical – infinite Marginal Rate of Substitution (MRS) – indifference curves.[10]

The protection in PVs is afforded against acts, not omissions, and against trade-offs with gains in other values, not losses. The protection in PVs is absolute and non-compensatory.[11] Omission bias is stronger in people with PVs.[23,24] Because PVs are seen as personal, agent-relative moral obligations, attempts to forcibly induce diminution of a PV often triggers anger or moral outrage.

The PVs held by individuals not aligned with industry thereby amplify the omission-bias effect in the 'outsider perspective' as elaborated in Section 4.2.1. Furthermore, if the SDT model of the PP is applied to an innovation that threatens to harm a PV, the implicit cost

[10] If the PV is instead represented on the vertical axis, then the associated indifference curves are horizontal, and have zero MRS.

[11] Being non-compensatory, it is inconsistent with the continuity assumption of standard utility theory.

of failing to exercise preventive intervention when it is fact warranted becomes unbounded, i.e. $C_{\mathrm{FN}} \to \infty$. Consequently, the denominator in (3.13) explodes while the numerator remains finite, which together ensure that the SDT corner solution obtains, triggering PP-based preventive intervention to actively protect the PV. The slope of the iso-expected-value lines approach zero as $C_{\mathrm{FN}} \to \infty$, meaning that the corner solution – and preventive intervention – is supported for all $G'(1) > 0$. Whereas omission bias increases the probability of a preventive-intervention supporting corner solution, PVs guarantee it for all ROC curves satisfying $G'(1) > 0$.

For individuals who are aligned with industry, the effect is reversed: PVs – concerning free enterprize, the national importance of an industry, or merely profits and employment – amplify the omission-bias effect in the 'insider perspective' as elaborated in Section 4.2.1. From this perspective, it is preventive intervention that is seen as the overt act which threatens to harm the PV, and the implicit cost of mistakenly undertaking this act becomes unbounded, i.e. $C_{\mathrm{FP}} \to \infty$. Hence the numerator in (3.13) explodes while the denominator remains finite. The slope of the iso-expected-value line approaches infinity, and the (0,0) corner solution obtains for all $G'(0) < \infty$. At this (0,0) corner solution, there is zero probability of implementing preventive intervention.

4.4 Affect heuristic

If the intensity of emotions may be represented on a spectrum, then visceral emotion is located at one extreme, while affect – i.e. the 'faint whisper of emotion' – is located at the other extreme.[34] Affect refers to either the quality of 'goodness' or 'badness' (i) in feelings associated with a stimulus, or (ii) in an experienced-feeling state. The *affect heuristic* in turn refers to reliance on such feelings, which is characteristic of the intuitive, experiential, System-1 pathway in dual-process theories of decision making.[28–34]

Whereas risk and benefit are positively correlated in nature and in the economy,[12] perceptions and judgments of risk and benefit become negatively correlated (inversely related) in the presence of affective valence. Under positive-affect valence, high benefit is associated with low risk. Under negative-affect valence, low benefit – or indeed harm – is associated with high risk. This is part of the 'risk-as-feelings' breakthrough in psychology: that people judge risk by how they feel about it, rather than on the basis of reasoned thought and analysis.[33] Affect influences perception

[12] because the coincidence of low-risk and amble benefit does not persist for long, due to scarcity brought about by exhaustion, competition, predation, or parasitism

and judgment directly and independently, without any pre-requisite prior priming by logical analytical evaluation.[34]

Alhakami and Slovic's pathbreaking study showed the empirical inverse relationship between perceived risk and perceived benefit on a sample of 40 items, including herbicides ($\rho = -.52$), DDT ($\rho = -.5$), asbestos ($\rho = -.48$), vaccinations ($\rho = -.43$), nuclear power ($\rho = -.4$), chemical manufacturing plants ($\rho = -.32$), and pesticides ($\rho = -.29$).[28] A survey of British Toxicology Society members confirmed that even among field experts, the affect valence perceived by the expert mediates the strength of the inverse relationship between the hazard's risk and benefit.[29] Experiments have verified and extended these results. Finucane et al. showed that the inverse relationship is strengthened in individuals subject to time pressure, who have fewer cognitive resources available for System 2 analytical deliberation, and thereby place greater reliance on the resource-efficient System 1 (affect-based) response.[30] And finally, Yoav Ganzach's experiments have shown that affect valence mediates judgments of risk and return – in the manner predicted by the affect heuristic – for financial assets that are not already familiar to the subject.[31]

For the SDT-based model of the PP, the affect heuristic forges a link between the prior-odds term and the misclassification-cost term in the slope expression for iso-expected-value lines.

For positive-affect valence, low prior probability is associated with high benefit – i.e. low harm. Consequently, both the prior-probability term and the misclassification-cost term in the denominator of (3.13) is small, entailing steep iso-expected-value lines with slopes *less likely* to fall within the corner-solution interval Γ.

For negative-affect valence, high prior probability is associated with low benefit – i.e. large harm. Consequently, both the prior-probability term and the misclassification-cost term in the denominator of (3.13) is large, entailing flat iso-expected-value lines with slopes *more likely* to fall within the corner-solution interval Γ.

The language that interest groups adopt to describe and define themselves offers an indication as to the affect valence they are likely to associate with particular industrial innovations, e.g. Greenpeace, Friends of the Earth, World Wildlife Foundation, Save the Whales, Center for Biological Diversity, Royal Society for the Protection of Birds (UK), Woodland Trust (UK), and Frack Off (UK). Through this clear articulation of identity and identification, not only are

particular patterns of positive- and negative-affect valence associations clearly implied, but so is the PV status of the environment and its biological diversity, as is the 'outsider perspective' with regard to applying omission bias to potentially harmful industrial innovations. Although it is not as immediately apparent from the naming conventions employed by corporations and their industry associations, the language used by officers of these associations and firms clearly conveys the pattern of positive- and negative-affect valence associations, the PV status of the industry and its profits, and the 'insider perspective' with regard to applying omission bias to their (potentially harmful) industrial innovations. The language used by firms' and industry-associations' public-relations arms and their legal representation reinforces this pattern.

4.5 Application: PP variants

Behavioral factors also shed light on the strong-form and super-strong-form variants of the PP. The weak PP explicitly incorporates cost and resource considerations, which permits SDT-based optimization of the cutoff threshold. Behavioral factors are external to the weak-PP definition, but may be incorporated into the SDT-based analysis via their impact upon perceived misclassification costs and perceived prior probabilities. In contrast, the strong-PP and super-strong-PP definitions incorporate behavioral factors *directly*. These PP variants hard-code not only the behavioral factors, but the *outsider perspective* in particular, directly into their definitions.

The strong-PP definition places the burden of proof on proponents of the industrial innovation. As such, the default state of affairs is absence (non-introduction) of the industrial innovation. Passive omission entails continued protection from the potential harms of the industrial innovation. In contrast, the overt-action alternative comprises introduction of the industrial innovation. This is precisely the outsider perspective, which entails that omission bias reduces the slope of the iso-expected-value lines, and thereby renders the SDT (1,1) corner solution – i.e. preventive intervention – more likely.

The strong PP's wording falls short of being unambiguous, however. The strong-PP definition excludes all references to cost, but relies on the ambiguous term 'should', which can mean either 'morally imperative' or 'morally desirable'. If 'morally imperative' is the operative interpretation, then cost considerations cannot impinge upon the inferential threshold, and the strong PP *can* be said to embody PVs. Yet if 'morally desirable' is the operative interpretation, then cost considerations cannot be strictly excluded from impinging upon the inferential

threshold, and the strong PP *cannot* be said to embody PVs. Thus, the strong-PP definition hard-codes the outsider perspective of omission bias, but falls short of hard-coding the outsider perspective of PVs.

The super-strong-PP definition incorporates a further standard-of-proof stipulation so strong (100%) that it leaves literally no scope for trade-offs of any non-zero magnitude between harm and any other value. In other words, the super-strong-PP definition hard-codes outsider-perspective PVs into the policy institution. Other studies have noted that a standard of proof requiring that harm "will not occur" is "often an impossible burden to meet."[35] Within the SDT framework, however, outsider-perspective PVs yield horizontal iso-expected-value lines and the (1,1) corner solution for all ROC curves satisfying $G'(1) > 0$. The associated inferential threshold systematically classifies *all cases* as 'intervention required': preventive intervention becomes the policy verdict on all potentially harmful industrial innovations.

5 CONCLUSION

In contrast to the classical normative use of SDT, the present work employs a behaviorally augmented SDT model for descriptive purposes. We show that the existence of the PP need not be understood solely as a response to Knightian uncertainty. Instead, the PP may be understood as a mental or administrative shortcut – i.e. a heuristic – for implementing SDT corner solutions. The NHST inferential practices followed in science are intended to hold type-I error at the fixed, conventional level of $\alpha = .05$. However policy decision making requires an inferential threshold that reflects operative cost trade-offs. SDT solves this problem of bridging between the practices of science and the needs of practical policy decision making. Identifying PP-based preventive intervention with SDT's (1,1) corner solution offers a formalization of the PP that restricts attention to (i) a small set of variables, (ii) the psychological effects that impinge upon subjectively perceived values of these variables, and (iii) the mathematical relationships between these variables within SDT. Hypotheses may be derived from this framework, and its empirical descriptive value may be tested. The behaviorally enhanced SDT model explains (retrodicts) previously puzzling aspects of PP target selection: (i) What kinds of harm sources attract PP-based preventive intervention?, and (ii) At what impact round, and for what reasons, is PP-based preventive intervention truncated? The influence of omission bias on perceived misclassification costs, moderated by insider or outsider perspective, successfully resolves these

puzzles. Our SDT framework also offers a new lens with which to examine and understand different definitions of the PP, upon which previous analytical models of the PP have shed very little light. Strong-form PP and super-strong-form PP are differentiated from weak-form PP in the extent to which behavioral factors are hard-coded directly within the PP policy institution: not at all in weak-form PP, omission bias alone in strong-form PP, and both omission bias and PVs in super-strong-form PP.

The SDT-based model also offers new analytical traction on the interplay between the corner-solution supporting interval Γ and variance-enhancing non-independent research. Due to *Risk Analysis'* manuscript-length conventions, this investigation, incorporating an examination of the PP's effect on incentives to undertake variance-enhancing research, is deferred to the future.

References

1. Peterson M. The precautionary principle is incoherent. Risk Analysis, 2006; 26:595–601.

2. Basili M. A rational decision rule with extreme events. Risk Analysis, 2006; 26:1721–1728.

3. NOAA. Environmental Data Management at NOAA: Archiving, Stewardship, and Access. Washington, DC: The National Academies Press.

4. Howard RA, Matheson JE, North DW. The decision to seed hurricanes. Science, 1972; 176:1191–1202.

5. Page T. A generic view of toxic chemicals and similar risks. Ecology Law Quarterly, 1978; 7:207–244.

6. Graham JD. Decision-analytic refinements of the precautionary principle. Journal of Risk Research, 2001; 4:127–141.

7. Wald A. Contributions to the theory of estimation and testing hypotheses. Annals of Mathematical Statistics, 1939; 10:299–326.

8. Wald A. Statistical Decision Functions. New York, NY: Wiley, 1950.

9. Ritov I, Baron J. Reluctance to vaccinate: Omission bias and ambiguity. Journal of Behavioral Decision Making, 1990; 3:263–277.

10. Spranca M, Minsk E, Baron J. Omission and commission in judgment and choice. Journal of Experimental Social Psychology, 1991; 27:76–105.

11. Asch DA, Baron J, Hershey JC, Kunreuther H, Meszaros J, Ritov I, Spranca M. Omission bias and pertussis vaccination. Medical Decision Making, 1994; 14:118–123.

12. Ritov I, Baron J. Status-quo and omission biases. Journal of Risk and Uncertainty, 1992; 5:49–61.

13. Baron J, Ritov I. Reference points and omission bias. Organizational Behavior and Human Decision Processes, 1994; 59:475–498.

14. Baron J, Ritov I. Omission bias, individual differences, and normality. Organizational Behavior and Human Decision Processes, 2004; 94:74–85.

15. Brown KF, Kroll JS, Hudson MJ, Ramsay M, Green J, Vincent CA, Fraser G, Sevdalis N. Omission bias and vaccine rejection by parents of healthy children: Implications for the influenza A/H1N1 vaccination programme. Vaccine, 2010; 28:4181–4185.

16. Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won. New York, NY: Crown Archetype, 2011.

17. Zamir E, Ritov I. Loss aversion, omission bias, and the burden of proof in civil litigation. Journal of Legal Studies, 2012; 41:165–207.

18. Kaivanto K, Kroll EB, Zabinski M. Bias-trigger manipulation and task-form understanding in Monty Hall. Economics Buletin, 2014; 34:89–98.

19. Hallsworth M, List JA, Metcalfe RD, Vlaev I. The making of homo honorarius: From omission to comission. NBER working paper no. 21210, `http://www.nber.org/papers/w21210`.

20. Connolly T, Rb J. Omission bias in vaccination decisions: Where's the "omission"? Where's the "bias"? Organizational Behavior and Human Decision Processes, 2003; 91:186–202.

21. DeScioli P, Christner J, Kurzban R. The omission strategy. Psychological Science, 2011; 22:442–446.

22. Baron J, Spranca M. Protected values. Organizational Behavior and Human Decision Processes, 1997; 70:1–16.

23. Ritov I, Baron J. Protected values and omission bias. Organizational Behavior and Human Decision Processes, 1999; 79:79–94.

24. Baron J, Leshner S. How serious are expressions of protected values? Journal of Experimental Psychology: Applied, 2000; 6:183–194.

25. Tetlock PE, Kristel OV, Elson SB, Lerner JS, Green MC. The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. Journal of Personality and Social Psychology, 2000; 78:853–870.

26. Tanner C, Medin DL, Iliev R. Influence of deontological versus consequentailist orientations on act choices and framing effects: When principles are more important than consequences. European Journal of Social Psychology, 2008; 38:757–769.

27. Baron J, Ritov I. Protected values and omission bias as deontological judgments. Psychology of Learning and Motivation, 2009; 50:133–167.

28. Alhakami AS, Slovic P. A psychological study of the inverse relationship between perceived risk and perceived benefit. Risk Analysis, 1994; 14:1085–1096.

29. Slovic P, MacGregor DG, Malmfors T, Purchase IFH. Inuence of affective processes on toxicologists judgments of risk. Eugene, OR: Decision Research, 1997.

30. Finucane ML, Alhakami A, Slovic P, Johnson SM. The affect heuristic in judgments of risks and benefits. Journal of Behavioral Decision Making, 2000; 13:1–17.

31. Ganzach Y. Judging risk and return of financial assets. Organizational Behavior and Human Decision Processes, 2000; 83:353–370.

32. Slovic P, Finucane ML, Peters E, MacGregor D. The affect heuristic. In: Gilovich T, Griffin D, Kahneman D (eds). Heuristics and Biases: The Psychology of Intuitive Judgment. New York, NY: Cambridge University Press, 2002:397-420.

33. Slovic P, Finucane ML, Peters E, MacGregor DG. Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk and rationality. Risk Analysis, 2004; 24:311–322.

34. Slovic P, Peters E. Risk perception and affect. Current Directions in Psychological Science, 2006; 15:322–325.

35. Sunstein C. Laws of Fear: Beyond the Precautionary Principle. New York, NY: Cambridge University Press, 2005.

36. Graham JD, Wiener JW. Risk Versus Risk: Tradeoffs in Health and Environmental Protection. Cambridge, MA: Harvard University Press, 1995.

37. Cross F. Paradoxical perils of the precautionary principle. Washington and Lee Law Review, 1996; 53:851–925.

38. USDA. Colony Collapse Disorder Progress Report. Washington, DC: CCD Steering Committee, United States Department of Agriculture, 2010.

39. deFur PL, Kaszuba M. Implementing the precautionary principle. Science of the Total Environment, 2002; 288:155–165.

40. I.L.M. Second international conference on the protection of the North Sea: Ministerial declaration calling for reduction of pollution [London, November 25, 1987]. International Legal Materials, 1988; 27:835–848.

41. I.L.M. United Nations Conference on Environment and Development: Rio Declaration on Environment and Development [Rio de Janeiro, June 14, 1992]. International Legal Materials, 1992; 31:874–880.

42. Blackwelder B. Testimony by Dr. Brent Blackwelder (President, Friends of the Earth) before the Senate Appropriations Committee, Concerning the Cloning of Humans and Genetic Modifications. January 24, 2002. Available from the Institute for Agriculture and Trade Policy, `http://www.iatp.org/files/Testimony-By_Dr_Brent_BlackwelderBefore_the_Se.htm`

43. Schweizer M. Loss aversion, omission bias and the civil standard of proof. In Mathias K (ed). European Perspectives on Behavioural Law and Economics. New York, NY: Springer, 2015:125–145.

44. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London, Series A, 1933; 231:289–337.

45. Fisher RA. *Statistical Methods and Scientific Inference* (2nd ed.) New York, NY: Hafner Publishing, 1959.

46. Neyman J. Discussion of Fisher (1935a). Journal of the Royal Statistical Society, 1935; 98:74–75.

47. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American statistical Association, 1993; 88:1242–1249.

48. Berger JO. Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science, 2003; 18:1–32.

49. Yates F. The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. Journal of the Americal Statistical Association, 1951; 46:19-34.

50. Nickerson RS. Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods, 2000; 5:241-301.

51. Sterne AC, Smith GD. Sifting the evidence – what's wrong with significance tests? British Medical Journal, 2001; 322:226–231.

52. Ziliak ST, McCloskey DN. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives. Ann Arbor, MI: University of Michigan Press, 2008.

53. Ioannidis PA. Why most published research findings are false. PLoS Medicine, 2005; 2:696–701.

54. Kruschke JK. Bayesian data analysis. Wiley Interdisciplinary Reviews: Cognitive Science, 2010; 1:658–676.

55. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology, 2015; 37:1–2.

56. Fisher RA. Statistical Methods for Research Workers. Edinburgh, UK: Oliver & Boyd, 1925.

57. Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture of Great Britain, 1926; 33:503–513.

58. Fisher RA. The Design of Experiments. Edinburgh, UK: Oliver & Boyd, 1935.

59. Lehmann EL, Romano JP. Testing Statistical Hypotheses, 3rd edition. New York, NY: Springer, 2005.

60. Nayman J. First Course in Probability and Statistics. New York, NY: Holt, 1950.

61. Cowles M, Davis C. On the origins of the .05 level of statistical significance. American Psychologist, 1982; 37:553–558.

62. Cox DR. Statistical significance tests. British Journal of Clinical Pharmacology, 1982; 14:325–331.

63. Egan JE. Signal Detection Theory and ROC Analysis. London: Academic Press, 1975.

64. Green DM, Swets JA, Signal Detection Theory and Psychophysics. London: Wiley, 1966.

65. Macmillan NA, Creelman CD. Detection Theory: A User's Guide. Cambridge: Cambridge University Press, 1991.

66. Kaivanto K. The effect of decentralized behavioral decision making on system-level risk. Risk Analysis, 2014; 34:2121–2142.

67. Hills SL, Berbaum KS. Using the mean-to-sigma ratio as a measure of the improperness of binormal ROC curves. Academic Radiology, 2011; 18:143–154.

68. Savage LJ. The Foundation of Statistics. New York, NY: John Wiley and Sons, 1954.

69. Kahneman D, Miller DT. Norm theory: Comparing reality to its alternatives. Psychological Review, 1986; 93:136–153.

70. Inglehart JK. Compensating children with vaccine-related injuries. New England Journal of Medicine, 1987; 316:1283–1288.

71. Feldbrugge FJM. Good and bad Samaritans: A comparative survey of criminal law provisions concerning failure to rescue. American Journal of Comparative Law, 1966; 14:630–657.