Economics Working Paper Series

2025/003

The Precautionary Principle and the Innovation Principle:
Incompatible Guides for AI Innovation Governance?

Kim Kaivanto

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

LUMS home page: http://www.lancaster.ac.uk/lums/

The Precautionary Principle and the Innovation Principle: Incompatible Guides for AI Innovation Governance?

Kim Kaivanto[*]

Lancaster University, Lancaster LA1 4YX, UK

this version: May 9, 2025

Abstract

In policy debates concerning the governance and regulation of Artificial Intelligence (AI), both the Precautionary Principle (PP) and the Innovation Principle (IP) are advocated by their respective interest groups. Do these principles offer wholly incompatible and contradictory guidance? Does one *necessarily* negate the other? I argue here that provided attention is restricted to *weak-form* PP and IP, the answer to both of these questions is "No." The essence of these weak formulations is the requirement to fully account for type-I error costs arising from erroneously preventing the innovation's diffusion through society (i.e. mistaken regulatory red-lighting) *as well as* the type-II error costs arising from erroneously allowing the innovation to diffuse through society (i.e. mistaken regulatory green-lighting). Within the Signal Detection Theory (SDT) model developed here, weak-PP red-light (weak-IP green-light) determinations are optimal for sufficiently small (large) ratios of expected type-I to type-II error costs. For intermediate expected cost ratios, an amber-light 'wait-and-monitor' policy is optimal. *Regulatory sandbox* instruments allow AI testing and experimentation to take place within a structured environment of limited duration and societal scale, whereby the expected cost ratio falls within the 'wait-and-monitor' range. Through sandboxing regulators and innovating firms learn more about the expected cost ratio, and what respective adaptations — of regulation, of technical solution, of business model, or combination thereof, if any — are needed to keep the ratio out of the weak-PP red-light zone.

---

[*]e-mail: k.kaivanto@lancaster.ac.uk

1 INTRODUCTION

At one extreme of the popular imagination, unfettered development of Artificial Intelligence (AI) promises previously inconceivable betterment of the human condition. At the other extreme, unfettered development of general-purpose AI is associated with extinction-level subjugation of humanity by superintelligence (Bostrom, 2014). In-between, entire categories of labor — both blue-collar (e.g. drivers of taxis, public transport, farm equipment, and road haulage) and white-collar (e.g. accountants, lawyers, and investment advisors) — are at high risk of being replaced by narrow AI applications. There are real questions as to the risks posed by AI to citizens' digital security, physical security (e.g. autonomous weapons), political security, as well as to the ability of society's communication infrastructure to maintain the integrity of information in the face of 'deepfakes' (Brundage et al., 2018). Realization of these risks would in turn undermine the perceived integrity of ('trust' in) society's key institutions (Schneier, 2012). Shannon Vallor (2016) summarizes that in the 21st Century, AI will be (i) a significant *magnifier* of economic,[1] political, and physical risk, as well as (ii) an important tool for *managing & mitigating* these same forms of risk.

In the face of these risks there have been calls to take precautionary measures immediately, rather than wait for full scientific certainty, by which time the window of opportunity for avoiding those bad outcomes will have closed (Brundage et al., 2018; Kuziemski, 2018; Del Castillo, 2020).

Responding to these calls for application of the Precautionary Principle (PP), the Information Technology & Innovation Foundation (ITIF) in the United States has countered by arguing that AI should fall under the Innovation Principle (IP) instead (Castro and McLaughlin, 2019). Meanwhile in the European Union (EU), the European Commission embedded the IP into its procedures for evaluating AI innovation (EC, 2018),[2] and it served as a guiding principle throughout the development of the *EU Artificial Intelligence Act* (2024).

Do these principles offer wholly incompatible and contradictory guidance for the governance and regulation of AI? Does one *necessarily* negate the other? Some advocates appear to think so. How can government reconcile these seemingly contradictory demands?

In this paper I argue that among the multitude of formulations — there are more than 20 definitions of PP alone — the *weak* forms of PP and IP are compatible. This is due to the weak

---

[1] Accentuating inequality in the real economy, but also creating new systemic risks and amplifying existing systemic risks in the financial economy. Danielsson et al. (2020) note that AI "has the potential to destabilise the financial system, creating new tail risks and amplifying existing ones due to procyclicality, endogenous complexity, optimisation against the system and the need to trust the AI engine."

[2] "For any new regulatory proposals that shall be needed to address emerging issues resulting from AI and related technologies, the Commission applies the Innovation Principle, a set of tools and guidelines that was developed to ensure that all Commission initiatives are innovation friendly: [EC (2016) https://data.europa.eu/doi/10.2872/626511]" (EC, 2018, p 15, fn 56).

forms' focus on cost structure, where these costs are 'comparable' in a decision-theoretic sense, whereas the stronger forms of PP and IP feature 'protected values' which are non-comparable. Formally, the latter violate the continuity axiom, whereby it is not possible to strike trade-offs between them and other values. For this reason, the stronger forms of PP and IP offer conflicting, incompatible guidance for AI governance.

An important consequence of the focus on type-I and type-II error cost structure is that it also necessitates pinning down explicit candidate interventions — because the costs of mistaken intervention depend on the specific characteristics of the intervention that is implemented. This specificity resolves the open problem of what types of actions can be considered precautionary, and what specific interventions are to be implemented in different PP (or IP) applications (Bodansky, 2004). Context and implementability considerations determine the set of potential interventions, each of which has its own type-I and type-II error cost structure. If this set of candidate interventions is non-empty, Benefit-Cost Analysis may be applied across the candidates to select the socially 'best' intervention.

The key novel insight underpinning this paper is that one of the two corner solutions in Signal Detection Theory (SDT) corresponds to non-postponement of weak-PP red-light intervention, while the remaining corner solution corresponds to non-postponement of weak-IP green-lighting. The corner solutions occur when the expected error cost structure is highly asymmetric. This paper addresses the question of just how asymmetric the cost structure must be in order to trigger weak-PP red-light (or weak-IP green-light) intervention. The answer depends on the distinguishing power of current AI-risk science, captured in the discriminability parameter d', as well as an inferential-surprise-tolerance parameter ϵ.[3] Superficial resemblance aside, the latter parameter should not be confused with a *de minismis* threshold.

Careful consideration of SDT mathematical structure yields further insights into the consequences of interactions between base rates, expected error cost structure, and discriminability d'. For instance when public trust in science is eroded, from the public's standpoint the relevant discriminability parameter is zero. If one sets $d' = 0$ in SDT, it follows that for all less-than-unity (< 1) expected error cost ratios, weak-PP red-light intervention is optimal, whereas for all greater-than-unity (> 1) expected error cost ratios, weak-IP green-lighting is optimal. And if the erosion of public trust extends to the work of risk professionals and regulatory institutions, then each respective interest group will prioritize their own perception of the expected error cost ratio and the associated red-light or green-light determination — i.e. the result is societal polarization by interest group, even without the effect of non-comparable protected values.

The SDT-based model can also be used to explain how 'regulatory sandboxes' structure the

---

[3]For details, see Section 4, p 18.

risk-regulation problem to allow weak-PP red-lighting and weak-IP green-lighting as possible outcomes within a single coherent framework. As an institutional innovation, sandboxing has been growing in popularity among regulators around the world. This regulatory instrument allows AI testing and experimentation to take place within a structured environment of limited duration and societal scale — crucially, constraining the expected cost ratio — so that the innovative firm begins its sandbox journey within the SDT-model's amber-light 'wait-and-monitor' range. And throughout, as the regulator closely monitors testing and experimentation, it is well-placed to make a weak-PP red-light determination or a weak-IP green-light determination. By the end of the sandbox journey, regulators and innovating firms learn more about the expected cost ratio, and what respective adaptations — of regulation, of technical solution, of business model, or combination thereof, if any — are needed to keep the ratio out of the weak-PP red-light zone. Regulatory sandboxes are real-world examples of institutions in which the *weak forms* of PP and IP coexist without contradiction, consistent with the SDT-based model's predictions. But the SDT-based model also provides guidance on the limits to implementing weak-form PP and IP with regulatory sandboxes — which is particularly germane to initiatives that aim to create regulatory sandboxes for general-purpose AI systems, also known as 'AI foundation models'.

In the sequel, Section 2 reviews the definitions and interpretations of PP and IP. Section 3 introduces SDT-based optimal inferential thresholds for policy decisions. Section 4 develops necessary conditions for non-postponement of weak-PP red-lighting and weak-IP green-lighting determinations within SDT. Section 5 clarifies the nature of comprehensive cost estimation required for implementing weak-PP and weak-IP determinations within SDT, as well as the method for solving the associated intervention-selection problem. Section 6 contains the discussion, and Section 7 concludes.

2   PRINCIPLES

In a strict sense it is misleading to refer to *the* PP and *the* IP, as there are many different variants and interpretations of each.

2.1   Precautionary Principles

Of the twenty-plus definitions of the PP in existence (Sandin, 1999; Löfstedt et al., 2002; Sunstein, 2005), I focus here on three key spinal points in an ascending scale of stringency: weak PP, strong PP, and super-strong PP.

The PP emerged in the former West Germany during the 1970s[4] as part of the country's social-democratic-planning response to large-scale environmental problems including acid rain, pollution of the North Sea, and climate change (Boehmer-Christiansen, 1994; O'Riordan and Jordan, 1995; deFur and Kaszuba, 2002). 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. (I.L.M., 1988)

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. (I.L.M., 1992)

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 the weak variant of the PP, there is no mention of which party bears the burden of proof.

Remark 2.1. Under the weak PP, preventive intervention is undertaken even though "full scientific certainty" is lacking. Although Knightian uncertainty satisfies this condition, it is not synonymous with "lack of full scientific certainty". Knightian-uncertainty-based formalizations are consistent with the weak PP, but not exclusively so. The weak PP does not exclude the formalization of scientific uncertainty as a unique, if possibly high-dispersion probability distribution. Hence the weak PP is a candidate for being extended to the domain of risk.

Remark 2.2. The combination of threats of serious or irreversible damage — i.e. non-trivial (mis-)classification costs — and scientific uncertainty entails that there is no critical experiment currently available to resolve the societal-level intervention/nonintervention question. The weak PP asserts that in these circumstances cost-effective intervention measures shall not be postponed. In this sense the weak PP requires *non-postponement* of preventive intervention.

---

[4]Some authors trace the PP's emergence further back, to the late 1960s, in West Germany, Sweden and Switzerland; by the late 1970s the PP was already being cited in U.S. law as well as in international agreements (Löfstedt et al., 2002; Wiener et al., 2011).

Remark 2.3. The weak PP explicitly restricts preventive intervention to "cost-effective measures". Under one interpretation, this refers to positive net-benefit measures established through Benefit-Cost Analysis (BCA). Under a second interpretation, this refers to an intervention selected using Cost-Effectiveness Analysis (CEA). The latter CEA method emerged in response to some of the methodological problems as well as the perceived ethical problems involved in assigning monetary valuations to non-market benefits. However CEA implicitly invokes a separate set of assumptions, including in particular (i) that Willingness To Pay (WTP) per effectiveness unit is constant for all levels of effectivity and (ii) that this WTP is identical for all members of the population (Johannesson, 1995; Garber and Phelps, 1997; Bleichrodt and Quiggin, 1999; Dolan and Edlin, 2002). For these reasons, I henceforth interpret references to 'cost-effective measures' through the lens of BCA.

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.[5]

Unlike the weak PP, the strong PP (i) does not mention costs, (ii) does not acknowledge that different nation-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 tradeoffs 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

---

[5]http://www.sehn.org/wing.html

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:

> 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.* [emphasis added] (Blackwelder, 2002)

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 exceeds even the highest standard of proof employed in US Common Law, which is 'beyond a reasonable doubt', corresponding to a Bayesian posterior probability greater than 99% but less than 100% (Weiss, 2003a,b, 2006). Neither is it the 'full conviction of the judge' (90%, 95%, or 99.8%) standard of proof employed in continental European Civil Law (Schweizer, 2015). A literal reading of the super-strong PP requires a 100% standard of proof to be achieved before preventive intervention may be withdrawn.

Henceforth, this paper focuses on the weak-form PP.

2.2   Innovation Principles

Since 2013, representatives of the European Regulation & Innovation Forum (ERIF)[6] grouping of industrial concerns in the European Union (EU) have advocated a new risk principle, dubbed the *Innovation Principle*, to complement the PP:

> ...whenever the EU's institutions consider policy or regulation proposals, the impact on innovation should be fully assessed and addressed (Hudig, 2015).

Some have viewed the creation and promotion of the IP as a vehicle though which corporate lobbyists are attempting to balance, if not neutralize, application of the PP in the EU (Garnett et al., 2018; Del Castillo, 2020; Ducuing, 2022). However, a conservative reading of the wording suggests no more than comprehensive, non-myopic cost estimation requiring a full inventory of

---

[6]https://www.eriforum.eu/

direct and opportunity costs. In particular, the IP requires full accounting of the false-positive cost components associated with forgone diffusion of the innovation through the economy. Upon this reading, the IP sits naturally within the weak-PP framework.

In contrast, the final exhortation to *fully address* the impact on innovation could have expansive implications under a less conservative reading. Conceivably, one set of expectations as to how impacts on innovation would be 'fully addressed' involve the prevention of all such negative impacts outright. This in turn suggests that underlying valuations of false-positive misclassification costs are so extreme so as to invoke 'protected values' and associated infinite costs.

Thus with the IP, as with the PP, some interpretations spill over from the domain of normative rationality, characterized by comparability and the continuity axiom (*weak IP*) into the domain of protected values, characterized by non-comparability and violation of the continuity axiom (*super-strong IP*).

An example of the latter super-strong IP is contained in the Information Technology & Innovation Foundation's (ITIF's) 2019 publication, "Ten ways the Precautionary Principle undermines progress in Artificial Intelligence" (Castro and McLaughlin, 2019). The authors urge policymakers to embrace the 'hope-based' IP, and to thoroughly reject the fear-, doubt-, and anxiety-based PP. Although lacking a concise formal definition of the IP, they argue that "speculative concerns [about potential harms] should not hold back concrete benefits" of AI innovation. Moreover, the authors draw from the work of Adam Thierer, a popularizer of *permissionless innovation*:

> ...the notion that experimentation with new technologies and business models should generally be permitted by default. Unless a compelling case can be made that a new invention will bring serious harm to society, innovation should be allowed to continue unabated and problems, if any develop, can be addressed later. (Thierer, 2016)

In Castro and McLaughlin's (2019) ITIF formulation, the IP is a mirror image of super-strong PP, designed to supplant the PP for the governance of AI innovation.

In the European and World Trade Organization context, Portuese and Pillot (2018) advance economic and legal arguments for the promotion of an IP to counterbalance the prevailing PP and to shift the burden of proof. This too therefore is also a super-strong IP variant, but Portuese and Pillot (2018) envision a continual judicial balancing exercise between the IP and the PP, in the interest of serving the ideal of justice.

Even within the context of European Union (EU) legislation and regulation, there is no commonly agreed definition of the IP (Renda and Simonelli, 2019; Renda and Pelkmans, 2023).

Pēteris Zilgalvis, a Judge at the General Court of the European Union, characterizes the IP in regulatory impact assessment as "prioritizing regulatory approaches that serve to promote innovation while also addressing other regulatory aims" (2025). Within EU regulatory impact assessment then, the IP is conceived in terms that are not intrinsically incompatible with other regulatory principles such as the PP. Meanwhile within the EU's research & innovation laws and regulations, the IP is defined not in terms of specific criteria for intervention or non-intervention, but in terms of 'innovation-friendliness' of legislation and regulation preparation, review, and revision:

> The Innovation Principle is a tool to help achieve EU policy objectives by ensuring that legislation is designed in a way that creates the best possible conditions for innovation to flourish.

> The principle means that in future when the Commission develops new initiatives it will take into account the effect on innovation.

> This will ensure that all new EU policy or regulations support innovation and that the regulatory framework in Europe is innovation-friendly. (EC, 2019)

The EU's commitment to innovation-friendliness is currently implemented through (i) Research and Innovation Tool #21 in the Better Regulation Toolbox, (ii) innovation deals, which reveal obstacles to innovation, making it possible for future revision of regulation to address these obstacles, (iii) foresight and horizon scanning to anticipate future trends and enable anticipatory policymaking, (iv) impact-assessment support, and (v) the introduction of regulatory sandboxes in Member States (EC, 2024) to allow a responsive, experimental approach to regulation for e.g. FinTech, blockchain/distributed ledger technologies, artificial intelligence, and the internet of things. In AI-specific policy documents the European Commission (EC) reiterates this innovation-principle derived innovation-friendliness (EC, 2018). This approach is also being adopted in the United Kingdom, where the Department for Science, Innovation and Technology (DSIT) has produced guidance for AI regulation under the "pro-innovation principle" banner (DSIT, 2023, 2024).

Although a diverse range of activity types are being carried out under the IP rubric, henceforth this paper focuses on the *weak-form IP*, which we conceive in terms that parallel those of weak-form PP: (a) weak-IP green-lighting is undertaken even though full scientific certainty is lacking over eventual benefits or harms of the innovation; (b) no critical experiment is currently available to resolve the societal-level intervention/nonintervention question; (c) cost-effectiveness is a constraint for both the weak IP as well as the regulators implementing the weak IP, which we interpret here through the lens of BCA; and (d) silence on which party bears the burden of

proof.

2.2.1 Regulatory sandboxes

The progenitor of present-day regulatory sandboxes, Project Catalyst, was launched in 2012 by the U.S. Consumer Financial Protection Bureau (CFPB) with the purpose of promoting innovative consumer-friendly financial solutions (CFPB, 2012). The term was popularised by the U.K. Financial Conduct Authority's (FCA's) Project Innovate, which launched a "regulatory sandbox" in 2016 to provide a supportive environment for fintech innovation. A 2020 World Bank survey found 73 fintech-focussed sandboxes in operation in 57 countries (World Bank, 2020). Four years later the number of active sandboxes worldwide stood at 142 (Markellos et al., 2024).[7] Around the world, regulatory sandboxes cater not only to financial technology (fintech) innovation, but also to distributed ledger innovation, the internet of things innovation, digital authentication and verification innovation, communications innovation, energy innovation, public health innovation, transport innovation, and insurance technology (insurtech) innovation. We are not aware of any active regulatory sandboxes catering to pure AI innovation.[8] The flow of new sandbox initiatives is ongoing. For instance Israel launched a new regulatory sandbox program in January 2025 with the stated purpose of boosting innovation, and sandboxing is included as a key component in the country's National AI Program.[9]

According to the World Bank study, a *regulatory sandbox*

> ...is typically a virtual environment that enables live testing of new products or services in a controlled and time-bound manner [usually 6 months duration]. Controlled experimentation in a live environment provides a structured approach to promoting innovation and guiding interactions with firms while allowing regulators good oversight of emerging financial products. Regulatory sandboxes are open to innovative business models, products, and processes, whether regulated, unregulated, or slated for possible future regulation. ...[and] are usually classified into four types, based on their objectives: (i) policy-focused; (ii) product- or innovation-focused; (iii) thematic; and (iv) cross-border. (World Bank, 2020)

Most regulatory sandboxes are created to satisfy multiple objectives. Firm-sandbox relationships are collaborative, offering the firms informal 'steers' and 'regulatory comfort' as part of an indi-

---

[7]The Centre for Competition Policy (CCP) Regulatory Sandboxes Database may be filtered by sandbox Status: Announced, Closed, Considering, Open, Suspended, and Unknown. The figure of 142 is obtained by filtering by 'Open' status. https://competitionpolicy.ac.uk/research-projects/portal-on-regulatory-sandboxes/

[8]The CCP Regulatory Sandbox Database includes one entry for a 2021 announcement of the possibility of an EU-wide AI Sandbox. As of March 2025, this AI-dedicated sandbox had not been created.

[9]https://aiisrael.org.il/, accessed 4 March 2025.

vidually tailored plan whereby temporary derogations can be granted without altering existing rules outside the sandbox. Regulator-operated sandboxes offer the opportunity for regulatory experimentation — to learn what regulatory solutions are suitable for new technologies and novel business models, while still satisfying the regulator's societal and legal obligations. Once sufficient regulatory experimentation and associated experience is gained, operation of the sandbox may be contracted out to a service provider. For example the U.K. FCA's permanent digital sandbox is currently operated under contract by fintech company NayaOne Ltd.[10] Alongside the digital sandbox, NayaOne offers complementary services, including synthetic data to enable secure testing and experimentation without risking data privacy & protection or commercial confidentiality conflicts.

The OECD's 2023 survey of AI regulatory sandboxes found that the number of regulatory sandbox initiatives around the world had grown to 100, with AI featuring in the technological propositions of many sandbox startups (OECD, 2023). According to the Centre for Competition Policy's Regulatory Sandboxes Database, the total number of sandbox initiatives around the world (past, current, planned) numbers 199, spanning 92 countries (Markellos et al., 2024).

Among the new entries, the U.K. Information Commissioner's Office (ICO) developed a regulatory sandbox in 2021 to support firms launching products and services reliant upon the safe and legal use of personal data, with particular reference to compliance with the U.K. General Data Protection Regulation (UK GDPR) and the U.K. Data Protection Act 2018 (DPA 18). AI is a key technological enabler of these firms, though the areas of focus are diverse: Central Bank Digital Currencies (CBDCs), commercial use of drones, consumer health technology (healthtech), decentralised finance (DEFI), genomics, immersive technology and virtual worlds, neurotechnologies, next-generation Internet of Things (IoT), next-generation search, personalized AI, and quantum computing.

And in May 2024, the U.K. Medicines and Healthcare products Regulatory Agency (MHRA) launched 'AI Airlock', a new regulatory sandbox to help the Agency to identify and address the challenges of regulating standalone AI as Medical Devices (AIaMDs). The initial cohort of projects are virtual or real-world projects, and simulation exercises within the sandbox allow MHRA to test a range of regulatory issues these devices face when deployed within the U.K. National Health Service (NHS). AI Airlock therefore is narrowly focused on AIaMDs intended for the U.K. NHS.

Explicit in-scoping of AI 'foundation models'[11] (also known as 'frontier AI systems' or

---

[10]https://nayaone.com/

[11]The U.S. draft AI Foundation Model Transparency Act (2023) defines a foundation model as "an artificial intelligence model trained on broad data, generally uses self-supervision, generally contains at least 1,000,000,000 parameters, is applicable across a wide range of contexts, and exhibits, or could be easily modified to exhibit,

'general-purpose AI systems') is notably absent from the OECD (2023) global survey of sandboxes, the Centre for Competition Policy Regulatory Sandbox Database (Markellos et al., 2024),[12] as well as from the mission statements of the U.K.-based sandboxes.

The same observation also applies to the European Union's major legislative effort, the EU AI Act (EC, 2024).[13] Article 57 of the EU AI Act requires each Member State (possibly in co-operation with the competent authorities of other Member States) to establish at least one "AI regulatory sandbox" at the national level, to be in operation by 2 August 2026. Within EU legislation, the introduction of regulatory sandboxes is understood as one of the key means of implementing the IP.

> These sandboxes are controlled environments where AI systems can be developed, tested, and validated before being released to the market. The goal is to foster innovation while identifying and mitigating any risks, particularly those related to fundamental rights, health, and safety. The sandboxes will also provide guidance on regulatory expectations and requirements. If an AI system is successfully tested in a sandbox, the provider can use this as proof of compliance with regulations. The sandboxes are also intended to facilitate cross-border cooperation and share best practices." (EC, 2024, summary of Article 57)

Nevertheless, much of the fast-moving, high-stakes, large-scale investment in general-purpose AI foundation model innovation is taking place in the U.S. and in China, where there currently are no AI-focussed regulatory sandboxes.[14] The *general-purpose* nature of AI foundation models such as LLMs poses a specific challenge for weak-form PP and IP, which we discuss in Section 6 below.

A number of legal scholars have produced academic analysis and commentary upon evolving regulatory sandbox policy initiatives in connection with the IP and other aspects of regulation (Renda and Simonelli, 2019; Ducuing, 2022; Truby et al., 2022a,b; McCarthy, 2023; Yordanova

---

high levels of performance at tasks that could pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters."

Generative language models are the most prominent category of foundation models at the time of this writing. Google launched its Bidirectional Encoder Representations from Transformers (BERT) Large Language Model (LLM) in October 2018. BERT has 340 million parameters and a 16 GB, 3.3 billion word training dataset. Five years later, OpenAI's mixture-of-experts architecture GPT-4 features 17.6 trillion parameters trained on 13 trillion tokens ($\approx$ 9.75 trillion words). In January 2025 China's DeepSeek released their DeepSeek-R1 model, which has a sparse mixture-of-experts architecture with 671 billion parameters (of which 37 billion active per token), trained with extensive use of large-scale reinforcement learning on a dataset of 14.8 trillion tokens. 1 million tokens $\approx$ 750,000 words; therefore 14.8 trillion tokens $\approx$ 11.1 trillion words.

[12]https://competitionpolicy.ac.uk/research-projects/portal-on-regulatory-sandboxes/

[13]The EU AI Act (2024) includes separate chapters (Artiles 53 and 54) on general-purpose AI systems. See https://artificialintelligenceact.eu/article/53/ and https://artificialintelligenceact.eu/article/54/.

[14]Eleven countries are listed in OECD (2023) as having AI-related regulatory sandboxes: Canada, Colombia, Estonia, France, Germany, Korea, Lithuania, Malta, Norway, Singapore, and U.K. (Annex B: AI-related sandboxes, pp. 30–32.)

and Bertels, 2024). The present paper aims to complement this legal scholarship with a decision-theoretic model from which both weak-PP and weak-IP arise naturally as particular solutions, and which furthermore identifies the essential features of regulatory sandbox instruments that enable them to implement weak-form PP and IP for *specific applications* of AI foundation models, but not for the AI foundation models themselves.

3   Optimal inferential thresholds for policy decisions

Some of the problems inherent in Null Hypothesis Significance Testing (NHST) as currently practiced[15] 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 Theory (SDT) (Egan, 1975; Green and Swets, 1966; Macmillan and Creelman, 1991). 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.

I 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: $TPR = TP/(TP + FN)$, $FNR = FN/(TP + FN)$, $FPR = FP/(FP + TN)$, $TNR = TN/(FP + TN)$. 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 immediate (i.e. generation-specific) as well as intertemporal (i.e. inter-generational) externalities.

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 (Kaivanto, 2014), applications of SDT are couched in terms of *mini-*

---

[15]See Appendix A

Table 1: Classification matrices.

(a) Confusion matrix (counts).				(b) Misclassification cost matrix.			
		Inference under x'				Inference	
		$\neg H_0$	H_0			$\neg H_0$	H_0
Actual	H_1	TP	FN	Actual	H_1	C_{TP}	C_{FN}
	H_0	FP	TN		H_0	C_{FP}	C_{TN}

mizing expected misclassification cost. The central results of classical SDT are all derived under this 'expected misclassification cost' objective function. For present purposes the parsimony and tractability of this objective function serve well. Adopting an expected utility objective function introduces additional structure within the square brackets of (3.5), but the rest of the qualitative structure remains intact.

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

$$\alpha = P(X > x' \,|\, \theta_0) = \int_{x'}^{+\infty} f(x|\theta_0)\, \mathrm{d}x \tag{3.1}$$

$$1 - \beta = P(X > x' \,|\, \theta_1) = \int_{x'}^{+\infty} f(x|\theta_1)\, \mathrm{d}x \ . \tag{3.2}$$

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

$$\left(\frac{\mathrm{d}P(X > x' \,|\, \theta_1)}{\mathrm{d}P(X > x' \,|\, \theta_0)}\right)_{x'} = \frac{-f(x'|\theta_1)}{-f(x'|\theta_0)} = l(x') \ , \tag{3.3}$$

which is the likelihood ratio at x'. I assume $G' > 0$ and $G'' < 0$, ensuring that the monotone-likelihood ratio condition holds.[16]

Solving the constrained-minimization problem

$$\min_{x'} E(C) \qquad \text{s.t.} \qquad 1 - \beta = G(\alpha) \tag{3.4}$$

gives the optimality condition

$$l(x^*) = \frac{P(\theta_0)}{P(\theta_1)} \left[\frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}}\right] = \left(\frac{\mathrm{d}(1 - \beta)}{\mathrm{d}\alpha}\right)_{\bar{C}^*} \ , \tag{3.5}$$

---

[16]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 (Hills and Berbaum, 2011).

Figure 1: ROC curve, assuming $\sigma_0 = \sigma_1 = 1$ and four different discriminability parameters.
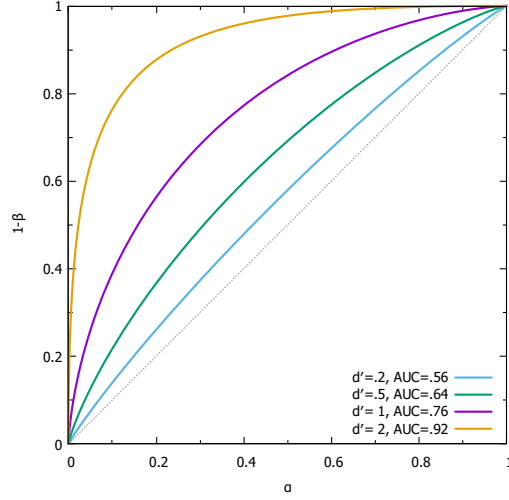
which states that the slope of the cost-minimizing iso-expect-cost line at the optimal operating point is given by the ratio of the expected opportunity cost of misclassifying a Negative to the expected opportunity cost of misclassifying a Positive. From (3.3) and (3.5) it is also clear that the optimality condition defines the critical likelihood ratio $l(x^*)$, and that (3.5) is a tangency condition between the least-cost iso-expected-cost line and the ROC curve. From (3.1) and (3.2) we have that

$$\alpha^* = \int_{x^*}^{+\infty} f(x|\theta_0)\,\mathrm{d}x \tag{3.6}$$

$$1 - \beta^* = \int_{x^*}^{+\infty} f(x|\theta_1)\,\mathrm{d}x \ . \tag{3.7}$$

When the cutoff threshold is optimally determined by (3.5), 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:

$$x^* = \frac{1}{\theta_1}\left(\ln(C_{FP} - C_{TN}) - \ln(C_{FN} - C_{TP}) + \ln(P(\theta_0)) - \ln(P(\theta_1)) + \frac{\theta_1^2}{2}\right) \ . \tag{3.8}$$

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

15

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. $1 - \beta^{\text{NP}} = G(0.05)$.

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.

I 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* (emphasis added, Cox, 1982).

4   Non-postponement

The weak PP admonishes *against waiting* for a critical experiment before introducing protective intervention when there are threats of serious or irreversible damage and science currently does not offer a definitive, certain answer (see Remark 2.2 on p. 5). When the current scientific state of the art lacks conclusive resolving power and is accompanied by credible threats of serious damage, scientific uncertainty in itself does not constitute sufficient grounds to suspend or postpone preventive intervention. In SDT, prevailing scientific uncertainty may be operationalized through sampling distributions under the null and alternative hypotheses. Let $H_0 : \theta_0 = \theta$ be the status-quo level of the critical index variable, and define $H_1 : \theta_1 = \theta$ ($\theta_1 > \theta_0$) to be the (irreversible) higher value of the critical index-variable induced by a commercial innovation in the state of the world where it is harmful. Maintaining the assumption of equal-variance sampling distributions,[17] the resolving power of state-of-the-art scientific experiments may be

---

[17]This assumption is done purely for mathematical convenience. The existing SDT literature, now well established, explores the multitude of deviations from this assumption, involving different combinations of underlying distribution family and unequal variance.

summarized succinctly with the *discriminability* index:

$$d' = \frac{\theta_1 - \theta_0}{\sigma} \ . \tag{4.1}$$

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[18] — 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 over mere chance (see Figure 1).

In this model, *non-postponement* of the red-light/green-light regulatory response arises when either

$$x^* \to -\infty \qquad \text{and thus} \qquad \alpha^* = 1 \ , \ 1 - \beta^* = 1 \tag{4.2}$$

or

$$x^* \to +\infty \qquad \text{and thus} \qquad \alpha^* = 0 \ , \ 1 - \beta^* = 0 \tag{4.3}$$

Under (4.2), any finite x draw falls to the right of x^*, and H_0 is rejected in favor of H_1, implying weak-PP red-light intervention. Hence, current experiments will generate scores satisfying $x^* < x$ with probability $p(x^* < x) = 1$. Similarly, under (4.3), any finite x draw falls to the left of x^*, and H_0 is not rejected, implying weak-IP green-lighting. Hence, current experiments will generate scores satisfying $x < x^*$ with probability $p(x < x^*) = 1$.

In ROC space, non-postponement occurs when either the (1,1) top-right or (0,0) bottom-left corner solution obtains. Depending on the combination of discriminability d', prior probability $P(\theta_0), P(\theta_1)$, and misclassification costs $C_{FP} - C_{TN}$, $C_{FN} - C_{TP}$, it is possible for the optimal solution to be a top-right corner solution (non-postponement of weak-PP red-light preventive intervention despite prevailing scientific uncertainty) or a bottom-left corner solution (non-postponement of weak-IP green-lighting, allowing commercial activity to continue despite scientific uncertainty).

In order for (4.2) to hold for an arbitrary $d' > 0$, $AUC > 0.5$ ROC curve, *either* the numerator of the optimality condition expression (3.5) is equal to zero, *or* the denominator is equal to infinity, *or* both. In other words the iso-expected cost line must be horizontal. In turn for (4.3) to hold for an arbitrary $d' > 0$, $AUC > 0.5$ ROC curve, the numerator of (3.5) must

---

[18]fully developed treatment of which is deferred to future work

be infinite so that the iso-expected cost line is vertical.

Moral imperatives and trade-off-resistant 'protected values'[19] — such as those implicit in the strong- and super-strong-form PPs — can cause individuals to impute infinite values to either C_{FP} or C_{FN}. However in order to remain consistent with normative decision theory, misclassification costs must remain *finite*. For this reason, the slope of the iso-expected-cost line (3.5) can be neither infinite nor zero, contrary to the possibilities identified in the preceding paragraph.

With finite misclassification costs it is nevertheless possible for the optimal solution to be arbitrarily close to the corner solution. This can be formalized by setting an arbitrarily small ϵ_1 ($\epsilon_1 \in \mathbb{R}_{++}$) such that $x^* < x$ with probability tolerance $1 - p(x^* < x) < \epsilon_1$ for weak-PP intervention in the δ_1-neighborhood of the top-right corner of the ROC space. Similarly, $x < x^*$ with probability tolerance $1 - p(x < x^*) < \epsilon_0 \in \mathbb{R}_{++}$ for non-intervention in the δ_0-neighborhood of the bottom-left corner of the ROC space, where ϵ_0 is arbitrarily small.

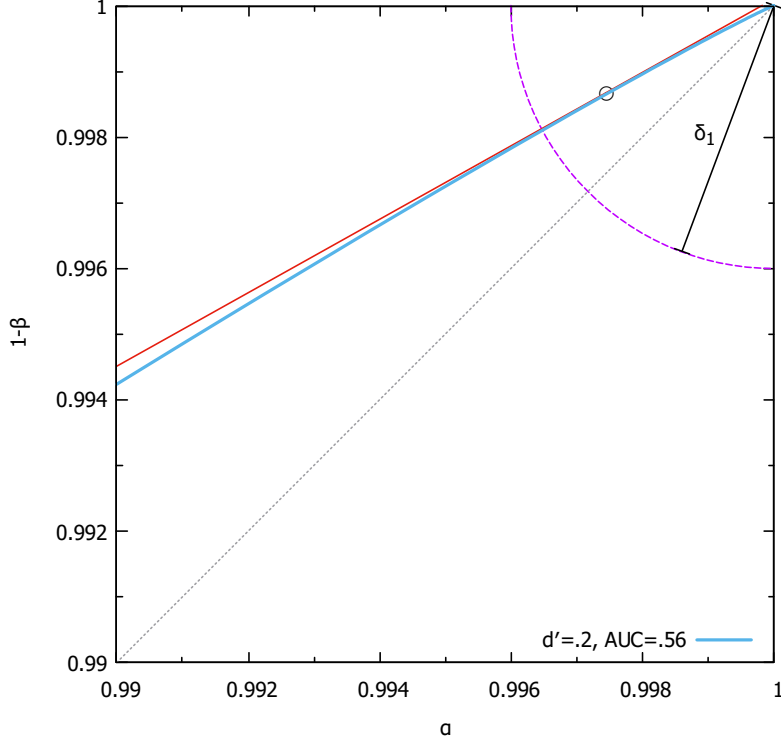*Condition* 4.1. (ϵ_1-tolerance non-postponement)

$$(\alpha^*, 1 - \beta^*) \in B_{\delta_1}(1,1) \qquad \Rightarrow \qquad 1 - p(x^* < x) < \epsilon_1 \ , \qquad \text{where} \qquad \delta_1 = \sqrt{2} \cdot \epsilon_1$$

This states that when the optimal operating point $(\alpha^*, 1 - \beta^*)$ falls within the δ_1-neighborhood of the top-right $(1,1)$ corner solution, the probability of any current experiment generating a score value to the right of x^* is within ϵ_1 tolerance of 1.

Condition 4.1 is illustrated in Figure 2. In this figure we let $\delta_1 = 0.004$ be the proximity threshold representing 'arbitrary closeness' to the corner solution. The expected incremental costs of type-II error (the denominator in (3.5)) are sufficiently large relative to the expected incremental costs of type-I error (the numerator in (3.5)), that the iso-expected cost line's point of tangency with the ROC curve falls within the δ_1 neighborhood of the $(1,1)$ corner solution.

---

[19]Protected Values (PVs) are characterized by an 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. To illustrate: 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. (Baron and Spranca, 1997)

Figure 2: Iso-expected cost line (in red) that is tangent with the ROC curve (in cyan) within the $\delta_1 = 0.004$ neighborhood of (1,1) (in magenta).

Condition 4.2. (ϵ_0-tolerance non-postponement)

$$(\alpha^*, 1 - \beta^*) \in B_{\delta_0}(0,0) \qquad \Rightarrow \qquad 1 - p(x < x^*) < \epsilon_0 \ , \qquad \text{where} \qquad \delta_0 = \sqrt{2} \cdot \epsilon_0$$

This states that when the optimal operating point $(\alpha^*, 1-\beta^*)$ falls within the δ_0-neighborhood of the bottom-left $(0,0)$ corner solution, the probability of any current experiment generating a score value to the left of x^* is within ϵ_0 tolerance of 1.

Condition 4.2 is illustrated in Figure 3. Here $\delta_0 = 0.004$ is the proximity threshold representing 'arbitrary closeness' to the corner solution. The expected incremental costs of type-II error (the denominator in (3.5)) are sufficiently small relative to the expected incremental costs of type-I error (the numerator in (3.5)), that the iso-expected cost line's point of tangency with the ROC curve falls within the δ_0 neighborhood of the $(0,0)$ corner solution.

Figure 3: Iso-expected cost line (in red) that is tangent with the ROC curve (in cyan) within the $\delta_1 = 0.004$ neighborhood of (0,0) (in magenta).
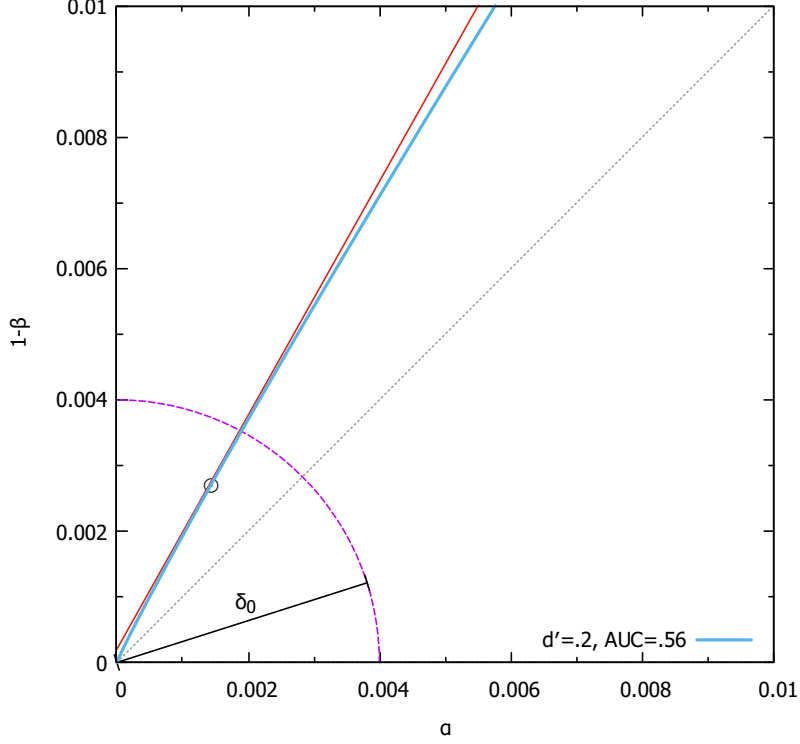
Condition 4.1 formalizes a necessary condition for ϵ_1-tolerance weak-PP red-light preventive intervention. In turn, Condition 4.2 formalizes a necessary condition for ϵ_0-tolerance weak-IP green-lighting, despite prevailing scientific uncertainty over whether the commercial activity is harmful or not. This second possibility, which has not been modeled decision-theoretically before, arises naturally from the structure of SDT.

Derivations supporting Conditions 4.1 and 4.2 are contained in the Appendices C and D.

5   Comprehensive cost estimation and intervention selection

The misclassification-cost matrix was succinctly introduced in Section 3. Yet the validity of any particular application of the SDT model hinges critically upon the care and attention with which entries in the misclassification-cost matrix are constructed and estimated. Comprehensive, non-myopic cost estimation requires full inventory of direct and opportunity costs. These costs are specific to the intervention that is to be implemented in the event that the null hypothesis is rejected. Hence the very first step is to fix a specific intervention to be implemented in the event that the null hypothesis is rejected. As I show below in Section 5.3, this does not preclude evaluating multiple candidate interventions or indeed combinations of interventions, and avoids

the simplifying assumptions involved in positing stylized continuous-function representations.[20]

Sections 5.1–5.3 below analyze weak-PP red-lighting interventions s_j where $j \in (1, 2, ..., J)$ and whether the associated solutions fall within the δ_1-neighborhood of the (1,1) corner solution. Section 5.4 then presents the Benefit-Cost Analysis needed to select an optimal intervention from among those satisfying the δ_1-neighborhood condition. The same logic, adapted to the δ_0-neighborhood of the (0,0) corner solution, can be applied to analyzing weak-IP green-lighting interventions s_k where $k \in (1, 2, ..., K)$.

5.1 The scope of 'comprehensive'

Misclassification costs may be inventoried at different entity levels: for different classes of individuals, for different organizations or regions, for different state-level and supra-state-level political entities, or, for society as a whole. The higher entity levels aggregate across lower-level heterogeneity. Successive levels thereby resolve heterogeneity, culminating in societal-level analysis, which for present purposes is assumed to yield a well-defined, unique misclassification-cost matrix.

Although even comprehensive cost inventory cannot anticipate unforeseen side effects, it must nevertheless capture direct costs as well as opportunity costs, many of which are not fully deterministic and therefore require separate estimation. These include not only the ancillary impacts, but also the 'countervailing' or 'competing' risks associated with the intervention (Lave, 1981; Graham and Wiener, 1995; Cross, 1996; Viscusi, 1998; Graham, 2001). The internationally agreed phaseout of chlorofluorocarbons (CFCs) to prevent further depletion of the ozone layer provides an interesting illustration of such countervailing risks.[21] This phaseout was accomplished by substituting CFCs with hydrochlorofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs), which pose respectively one-tenth and zero ozone-depletion potency compared to CFCs. Yet these compounds pose in turn other problems: (i) some have been found to be toxic in laboratory-animal testing, (ii) they can be many times more expensive than CFCs, slowing the diffusion of refrigeration and its health benefits in poorer countries, and (iii) they may substitute global-warming risk for ozone-depletion risk. From the standpoint of global warming, the effect of CFCs is roughly neutral: their global-warming effect via direct radiative forcing is roughly offset by their global-cooling effect via depletion of lower-stratosphere ozone. In turn the direct global-warming potential (GWP) of HCFCs and HFCs is approximately one-fifth that of CFCs — but without an offsetting global-cooling effect.

---

[20]which for tractability reasons must abstract from representing the full complexity of nonlinear and interaction effects

[21]This example draws on Graham and Wiener (1995).

In the example above, the GWP of CFCs must be reflected in C_{TN}, i.e. the state of nature in which the ozone layer is not being depleted by CFCs and CFCs are correctly not phased out, as well as in C_{FN}, the state of nature in which the ozone layer is being depleted by CFCs and CFCs are erroneously not phased out. In turn the GWP of HCFCs and HFCs must be reflected in the costs C_{TP} and C_{FP} where CFCs are correctly, and respectively incorrectly, phased out. Thus the SDT-based model implements an incremental-risk calculation within each state of nature (within the numerator and within the denominator in square brackets of (3.5)) as well as a risk-risk tradeoff between states of nature (through the ratio between the numerator and the denominator in square brackets of (3.5)).

5.2   Option value

Option value forms an important class of opportunity cost that is an essential component of comprehensive misclassification-cost estimation. Note that it is possible for the SDT-based model to support the weak-PP corner solution even in regions of the parameter space where option value is zero — for instance where no improvements in the resolving power of scientific experiments are expected in the near future. Hence non-zero option value is not a necessary prerequisite for normatively underpinned weak-PP red-light intervention. Yet where option value is strictly positive, its inclusion into the misclassification-cost matrix is necessary for normative validity of the SDT-based model.

Already Arrow and Fisher (1974) examined the problem of when to allow the irreversible development of wilderness land when society learns over time, resolving uncertainty about the benefits of preservation. This combination of irreversibility, uncertainty resolution (learning), and sequential decision-making structure gives rise to quasi-option value that raises the opportunity cost of development, reflecting the value of information. Optimal decisions early in the timeline favor flexibility preservation, as later on it will be possible to undertake the irreversible development if the benefits of preservation are revealed to be low.

In turn Gollier et al. (2000) examine the broader sequential problem in which today's actions can irreversibly increase society's future risk, and propose an economic interpretation of the PP rooted in the standard Bayesian framework. They show that in this broader framework, irreversibility and learning combine to yield a quasi-option, as a consequence of which more scientific uncertainty optimally induces society to favor more conservative measures (e.g. tighter restrictions on emissions) early in the timeline to allow more flexibility later on. (Gollier et al., 2000) also study additional effects that occur in the irreversible-accumulation model of harmful substances associated with consumption. They find that restrictions on the shape of the utility function — absolute prudence must be at least twice the absolute risk aversion — ensure that

the optimal response to more scientific uncertainty is to restrict current consumption associated with the accumulation of harmful substances. Alteratively, restrictions on the set of distributions of risk ensure the same policy implication.

Whereas the SDT-based model supports the weak PP interpreted as a (1,1) corner solution within a one-date model, it is also possible to introduce value-of-information quasi-option value into the misclassification cost matrix by explicitly considering the possibility of scientific progress occurring across two or more dates. This does not require abandoning the risk-neutrality assumption. Such quasi-option value enters as an opportunity-cost component in C_{FN}. It is a measure of the value of information: the benefit obtained by making the decision at time $t = 1$ with discriminability d_1', rather than at time $t = 0$ with discriminability d_0', where $d_1' \geq d_0'$ due to improved resolving power of scientific experiments. This value of information is foregone, i.e. the quasi-option is extinguished, when H_0 is falsely accepted, and the harmful commercial activity is erroneously permitted.

5.3 Multiple potential policy interventions

In this section the preventive intervention under consideration has been 'cessation of the harmful commercial activity'. However this represents only one extreme of a range of possible policy interventions. More generally, the SDT-based model of weak PP can be used to partition a collection of policy interventions into those that satisfy the δ_1-neighborhood corner-solution condition $(\alpha^*, 1 - \beta^*) \in B_{\delta_1}(1, 1)$ and those that do not. To fix notation, consider a particular policy intervention s_j from among a collection of discrete policy interventions $S = \{s_1, s_2, ..., s_J\}$, where each policy intervention s_j is assigned an index j according to its misclassification-cost ratio:

$$\left[\frac{C_{\mathrm{FP}}^j - C_{\mathrm{TN}}^j}{C_{\mathrm{FN}}^j - C_{\mathrm{TP}}^j} \right] \geq \left[\frac{C_{\mathrm{FP}}^{j+1} - C_{\mathrm{TN}}^{j+1}}{C_{\mathrm{FN}}^{j+1} - C_{\mathrm{TP}}^{j+1}} \right] \quad . \tag{5.1}$$

Thus larger values of the index j are associated with smaller misclassification-cost ratios, ceteris paribus. The collection S may include multiple candidate 'maximum-permissible emissions limits', for example. However, due to non-linear effects and complex interaction effects with rest of the society and the environment, it is not possible to simply assume that the misclassification-cost ratio $\left[\frac{C_{\mathrm{FP}} - C_{\mathrm{TN}}}{C_{\mathrm{FN}} - C_{\mathrm{TP}}} \right]$ is monotonic in the permitted emission limit. For AI, the equivalent to an emission limit would be the extent of permitted diffusion across jurisdictions and applications.

With elements of S indexed according to (5.1), any policy intervention s_j that is found to satisfy the δ_1-neighborhood corner-solution condition $(\alpha_{s_j}^*, 1 - \beta_{s_j}^*) \in B_{\delta_1}(1, 1)$ implies that those policy interventions with equally small or smaller misclassification-cost ratios — i.e. $\{s_{j+1}, s_{j+2}, ..., s_J\}$ — also satisfy the δ_1-neighborhood corner-solution condition, and equally

are feasible candidates for weak-PP red-lighting. Of course it is also possible to identify the critical policy intervention $s_{j'}$

$$j' \ = \ \arg \max_{j \in (1,2,...,J)} \sqrt{(1-\alpha^*_{s_j})^2 + (\beta^*_{s_j})^2} \qquad \text{s.t.} \qquad (\alpha^*_{s_j}, 1 - \beta^*_{s_j}) \ \in \ B_{\delta_1}(1,1) \qquad (5.2)$$

to the right of which all remaining policy interventions are feasible candidates for weak-PP red-lighting.

However if policy-intervention *combinations* are contemplated, the possibility that the misclassification costs of a combination are not merely additive but non-linearly impounded (e.g. through interaction effects) necessitates (i) re-estimation of combination-specific misclassification costs, requiring (ii) re-evaluation through (3.5) whether the *combination* satisfies the δ_1-neighborhood corner-solution condition. In general, policy makers may wish to consider the power set (excluding the null set) of simple policy interventions $\mathcal{S} = 2^S \setminus \{\varnothing\}$, comprising the union of simple S and combination policy interventions $\mathcal{S} \setminus S$. The elements of this set $\mathcal{S}$ must be re-indexed according to (5.1), yielding $\mathcal{S} = \{s_1, s_2, ..., s_{|\mathcal{S}|}\}$, among the elements of which the critical policy intervention $s_{j'} \in \mathcal{S}$ may be found by applying (5.2) over the expanded index list $j \in (1, 2, ..., |\mathcal{S}|)$.

5.4 Intervention selection

Expansion of the consideration set beyond the singleton introduces a further selection problem: which policy measure, or indeed which policy-measure combination, should be implemented? Recall that the weak PP states that *lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation* (I.L.M., 1992). Recall also that the direct cost of implementing a policy intervention does not enter into the optimality condition (3.5) which determines the optimal cutoff (3.8). Consequently it is natural to maximize net social benefit across the candidate PP policy interventions $\mathcal{S}_{j \geq j'}$ subject to a non-negativity constraint. In other words, Benefit-Cost Analysis (BCA) is applied to each candidate PP policy intervention $s_j \in \mathcal{S}_{j \geq j'}$, and society selects for implementation that policy (combination) with the highest strictly positive net social benefit.

Indeed if $\mathcal{S}_{j \geq j'}$ is a singleton, then in order for it to be implemented under the present normative conception of the weak PP, then it also must pass the net-social-benefit test. The necessity of this BCA test is encoded in the verbal statement of the weak PP (see above). Moreover, it is also necessitated by the method of *inframarginal analysis*, which has been developed for problems involving corner solutions, in which the final step is to evaluate candidate corner and interior solutions by BCA (Cheng and Yang, 2004).

Sections 5.3 and 5.4 show that the SDT-based normative model of the weak PP supports multiple candidate policy interventions, different degrees of stringency in permissible exposures or emissions limits, as well as policy-intervention combinations. Furthermore, the SDT-based normative model naturally culminates in a policy-selection stage that examines the full portfolio of costs and impacts associated with each candidate policy intervention (and all combinations thereof). Much of the PP literature neglects this final stage, which is necessary for completing a normative analysis of preventive intervention.

6 DISCUSSION

6.1 Implications for the PP

Within the present framework it is not meaningful to separate the question, 'When is PP-based preventive intervention warranted?', from 'What is the preventive intervention being contemplated?'. The former can only be answered for specific candidates of the latter. Different interventions will have different impacts, both direct and ancillary, which result in different misclassification costs and different benefit-cost ratios. Much of the PP literature has remained silent on what types of actions can be considered precautionary and what specific action(s) are to be deployed under specific applications of the PP (Bodansky, 2004). This paper's contribution hinges on recognizing that by focusing on one specific intervention at a time, misclassification costs can be pinned down to allow evaluation of whether the δ_1-neighborhood $(1,1)$ corner solution holds, the δ_0-neighborhood $(0,0)$ corner solution holds, or neither holds. The normative analysis culminates with total benefit-cost comparison among those specific policy interventions — and combinations thereof — satisfying the relevant corner-solution condition. Hence the present SDT-based framework provides an integrated formalization not only of the conditions supporting the existence of the weak PP (weak IP), but also of the associated procedure for identifying the societally optimal policy intervention. In this sense, it offers a solution for the open problem highlighted by Bodansky (2004), among others.

Graham (2001) observes that existing PP-definition variants leave unanswered how a public decision maker should resolve the dilemma that arises when "a precautionary action might prevent one hazard but induce[s] another hazard." The present SDT-based operationalization illustrates how such countervailing risks associated with a particular action s_j may be incorporated into the misclassification-cost matrix (see Section 5.1). The SDT-based model of the weak PP offers the following answer to John Graham's question: the hazard introduced by the precautionary action s_j is incorporated into C_{TP} and C_{FP}. Since increasing C_{TP} decreases the denominator of the term in square brackets appearing in expression (3.5), while increasing C_{FP}

increases the numerator, the hazard induced by the precautionary action s_j unambiguously reduces the likelihood of a δ_1-neighborhood corner solution occurring, ceteris paribus. Whether the $(\alpha^*, 1 - \beta^*) \in B_{\delta_1}(1,1)$ condition holds following incorporation of the countervailing risk depends on (i) the initial (pre-incorporation) proximity of the optimal operating point to the boundary of the δ_1-radius neighborhood of (1,1), (ii) the local curvature of the ROC curve, and (iii) the magnitude of the countervailing risk induced by s_j. Comprehensive cost estimation, inclusive of countervailing risks, diminishes potential bias against innovation. With balanced and comprehensive cost estimation, the SDT-formalization of weak PP-IP reconciles promotion and control of technological innovation, which is a characterization of PP advocated by Todt and Luján (2014).

Just because an industrial innovation poses a credible risk of serious damage to the environment, or an AI innovation poses a credible risk to current technical and economic structures, it does not necessarily follow that the optimal operating point will fall within the δ_1-neighborhood of (1,1). It is entirely possible for there to be serious risks of damage, yet when expected incremental misclassification costs are computed according to equation (3.5), the resulting optimal operating point remains within the amber-light 'wait-and-monitor' interior of the ROC space. For smaller discriminabilities and AUCs, however, the likelihood of the optimal operating point reaching the neighbourhood of a corner solution increases, ceteris paribus. In the limit, as $d' \to 0$ and AUC$\to 0.5$, even minute asymmetry between the expected costs of false positives and false negatives will lead to a corner solution. Indeed when societal trust in the ability of science to deliver credible, politically even-handed positive discriminability $d' > 0$ is eroded, then even under the SDT-based model one expects to see polarization. Part of society will have steeper expected misclassification cost ratios, a (0,0) corner solution, and will argue for the removal of all restrictions on AI innovation. Another part of society will have flatter expected misclassification cost ratios, a (1,1) corner solution, and will argue for the shuttering of AI innovation.

As for Boyer-Kassem's (2019) question on the general scope of the PP and its relation to other decision rules, the present work's analytical model offers rich potential for conceptual contextualization of the weak PP. Consider the *Argument from Inductive Risk* (AIR), long studied in the philosophy of science (Rudner, 1953; Douglas, 2000, 2009).[22] The AIR's relevance to the PP is well understood among philosophers (Steel, 2014). But whereas the AIR literature has focused on whether scientists should (or should not) adjust their inferential thresholds in light of non-epistemic values, only recently has this been operationalized using SDT to show precisely how such adjustments may be made in light of not only misclassification costs, but also

---

[22]According to the AIR, the standard of evidence required to accept or reject a hypothesis should be responsive to the *non-epistemic costs* associated with inferential error.

prior probabilities (Kaivanto and Steel, 2019). The present paper proceeds from the observation that in widespread scientific practice, inferential thresholds are not in fact adjusted to reflect non-epistemic values. Instead, the inferential thresholds used in scientific practice — for many disciplines, $\alpha = 0.05$ — are a long-standing convention reflecting *epistemic values*. This leaves a gap — in suitability and relevance — between the inferences generated by science, which largely ignore non-epistemic values, and the needs of policy making, which are driven by the need to strike trade-offs between non-epistemic values. Hence, one can characterize the PP as a *patch*[23] for the mismatch between (i) the scientific community's fixed-threshold inferential practices and (ii) policy makers' need to strike trade-offs among non-epistemic values.

This *PP-as-a-patch* characterization also yields insight into an even deeper reason for the existence of PPs, as well as insight into PPs' relations to other decision frameworks. The inferential conventions and practices of science currently reflect frequentist NHST. If instead all scientists and policy makers were full-fledged Bayesians, there would be no need for PPs. Since each experiment and empirical study would yield a posterior probability distribution rather than a dichotomous significant/non-significant result, there would be no role for a patch. Both scientists and policy makers would use the same posterior probability distribution, but each community would combine it with a different loss function, capturing epistemic and non-epistemic values respectively, in implementing Bayesian Decision Theory.

It has become widely accepted that reasonable formulations of the PP must be compatible with the *de minimis* risk principle (Carter and Peterson, 2015), according to which sufficiently improbable risks, falling under some threshold, should be ignored. Yet when examining the *de minimis* risk principle, several fundamental problems have been identified (Peterson, 2002; Carter and Peterson, 2015; Lundgren and Stefánsson, 2020). The structure of the SDT-based model of weak PP shows that, strictly speaking, it is erroneous to examine base rates in isolation and to apply a threshold to these probabilities. Instead, in SDT these probabilities must be combined multiplicatively with their respective incremental misclassification costs and related to the trade-off between α and $(1 - \beta)$ encoded in the ROC curve. However, as the discussion of equation (3.8) in Section 3 points out, base-rate probabilities are unusually powerful 'levers' acting upon the location of the *optimal cutoff threshold* x^*. If either base rate is sufficiently small, its effect will dominate those of finite, non-extreme misclassification costs, and the optimal cutoff threshold will move far into the tail that accommodates the small-probability hypothesis. Using the SDT-based model one can derive a *de minimis-like* base-rate-risk threshold that corresponds to the regulator's ϵ_1 parameter, which captures its tolerance to inferential surprise

---

[23]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. Previous versions of this paper emphasized this aspect of the PP.

in issuing a weak-PP red-light determination. However, this threshold is not unconditional; it also depends upon error costs, the discriminability d' of current risk science, as well as the regulator's above-mentioned inferential-surprise tolerance ϵ_1. As such, this would have to be viewed as a *conditional de minimis* risk principle.

6.2   Implications for the IP

The IP, being much younger than the PP, is sometimes viewed as lacking the deeper foundations and legitimacy claimed for the PP. There has been less time — and perhaps, for varied reasons, less interest among scholars — to develop deeper foundations for the IP. This paper argues that within the SDT-based model, the weak IP is theoretically no less well-founded than the weak PP. Indeed weak-form PP and IP emerge naturally and simultaneously from the SDT framework. Delineation of decision-theoretic (SDT-based) foundations for weak-form IP and its compatibility with weak-form PP is a central contribution of this paper.

This SDT-based model underpinning weak-form IP is novel within the context of existing IP scholarship, which adduces a variety of motives and arguments for the IP. Given the diffuse nature of those motives, some policy makers interpret the IP equally diffusely as a 'pro-innovation principle' which requires regulation to be 'innovation friendly'. The value of the SDT-based model is not so much in providing exact, parametric estimates or a specific formula that regulators can apply slavishly, but in providing a structure within which to reason through questions pertaining to risk regulation and regulator intervention. The first distinction that the SDT-based model highlights is between strong-form PP and IP, which are expressions of protected, non-comparable values, and weak-form PP and IP in which costs are finite and comparable. Thus, to advocate the weak IP does not entail adopting the position that all AI innovation should be allowed to proceed without hindrance. Instead, the SDT model identifies the essential considerations in making a weak-form PP or IP determination: expected misclassification-error costs, the discriminability that captures the distinguishing power of current AI-risk science, and the regulator's tolerance to inferential surprise.

A case in point is how the SDT model's structure aids in reasoning through the challenges faced in AI innovation governance. Should AI innovation governance be guided by the PP, or by the IP? Is it coherent for AI innovation governance to be guided by both? Furthermore, what types of policies and instruments does this imply for AI innovation governance?

Answers to the first two questions are clear by now: *weak-form* PP and IP are compatible, implying no incoherence in a regulator guided by both principles. As for the third question, policies and instruments can target (i) discriminability, (ii) the misclassification cost ratio, and (iii) the qualitative (green-light, amber-light, red-light) properties of the regulatory solution.

Policy makers in most countries now recognize the importance of innovation in increasing national productivity, which is the most important driver of welfare in the long run.[24] Accordingly policy makers are not agnostic to the qualitative properties of the regulatory solution for AI innovation governance: they prefer to green-light where possible, else to wait-and-monitor, and only to red-light when it is risk-optimal to do so. Moreover, optimal classification into green-light and red-light categories is improved (expected misclassification costs are smaller) the larger the discriminability d' — the increase of which can be enhanced by learning more about the technological underpinnings and business models of the new AI-based products and the firms bringing them to market. The same learning process also allows regulators to estimate more precisely the type-I and type-II error cost structures of those technologies and business models. Furthermore, by placing temporary bounds on the societal scale at which the AI innovation is deployed — e.g. virtual deployment, or beta testing on a limited scale, or limited in-sandbox duration, or conditional, time-limited post-sandbox regulatory approval — the extremity of the expected misclassification-cost ratio is limited, keeping the optimal solution in the wait-and-monitor range of the ROC curve. The aforementioned elements all come together in *regulatory sandbox* instruments. In this sense the SDT model provides a decision-theoretic foundation for regulatory sandbox instruments.[25]

As noted in Section 2.2.1, IP-inspired innovation-friendly 'AI regulatory sandboxes' are beginning to propagate across countries. Where the sandbox operator can identify the application or set of applications that the AI-based innovation will be employed in, the techniques discussed in Section 5 can be brought to bear to identify the set of possible type-I and type-II error costs with which to solve the SDT model for red-light, amber-light, or green-light classification by the regulator. Indeed the collaborative sandbox process allows both the regulator and firms to discover these potential costs, and how these costs vary with different measures that the regulator (or sandbox operator) can take, as well as with different design, implementation, and business-operation measures that the firm can take.

The regulator and the firm have less control over *how* the innovation will be used in the market. At least three different categories of users must be considered: naïve, well-behaved, and nefarious. Naïve users are liable to mis-use the product through oversight unless hard guardrails are built into the product. Nefarious third parties may compromise naïve users by exploiting their poor cyber hygiene or susceptibility to social engineering. Well-behaved users conform with

---

[24]In Paul Krugman's oft-repeated words, "Productivity isn't everything, but, in the long run, it is almost everything. A country's ability to improve its standard of living over time depends almost entirely on its ability to raise its output per worker." (Krugman, 1994)

[25]Even though in practice regulatory sandbox implementations often bundle additional objectives beyond (i)–(iii), such as facilitating early-stage access to finance (Cornelli et al., 2024).

product developers' instructions, while maintaining good cyber hygiene and rebuffing social-engineering attacks. Finally, nefarious users may 'game' the product itself, steal data revealed by the product, or use the product in ways not intended by the developer.

AI-based innovations in particular pose new security, regulatory, and societal risks, insofar as they offer so much more scope for unintended use by naïve users, nefarious users, and naïve users' accounts hijacked by nefarious third parties (Anderljung et al., 2023). Difficult regulatory questions also arise over the extent to which firms and regulators must anticipate possible *combinations* of AI tools to achieve nefarious effects. These questions bring forth profound ethical and legal debates underpinning regulation and its operational implementation.

Finally, the SDT model also offers guidance on the limits of regulatory sandbox instruments. The defining characteristic of 'AI foundation models' is that they are general-purpose AI systems capable of generating deep-fake text, audio, static images, and video. These capabilities are now being harnessed to create AI assistants that can perform numerous conditionally linked operations, including making and responding to telephone calls, as well as searching and executing tasks on the internet — with the potential of reaching any internet-connected system. Moreover, AI foundation models can be run either online or locally — so AI foundation models propagate easily across national borders. Currently, regulatory sandbox initiatives remain silent on whether AI foundation models fall within their scope or not (see Section 2.2.1). To the extent that regulatory sandboxes are explained by the SDT model, an attempt to bring AI foundation models within the scope of regulatory sandboxing would constitute over-reach. Given the general-purpose nature of AI foundation models, the set of possible applications and the range of legitimate and nefarious uses are vast and open-ended. It is therefore not possible to explicitly identify all possible interventions and the associated misclassification costs required in the SDT-based approach. Hence if regulatory sandboxing is viable for AI foundation models, it must be based on a different theoretical underpinning than the SDT model developed here.

6.3   Contestability

Finally, whereas the present formalization of rationality-preserving weak-form PP and IP has theoretical appeal, in practical application it is vulnerable to contestability. Can there be widespread agreement within society about the appropriate values for C_{FP}, C_{TN}, C_{FN}, and C_{TP}? Can the same be said about the parameters of the sampling distributions, and the operative value for ϵ? In practical application, these parameters are heavily contested by opposing interest groups. The advantage of relaxing normative rationality to admit valuations based on moral imperatives or protected values is that the framework then delivers clear regulatory-policy recommendations *within* each like-minded interest group. But at that point the PP and IP cease

to be compatible guides for AI innovation governance.

6.4 winner takes all

Underpinning the surge of investment in the development of AI foundational models is the view that the market for services powered by generative AI will be highly concentrated, dominated by the firm that can drive its technology ahead of the field and thereafter stay ahead. In other words, it has been viewed as a winner-takes-all market. The consequences of this arms-race are explored in numerous formal models. Armstrong et al. (2016) study how teams in such technology-development races have incentives to cut corners and under-invest in safety precautions to increase the probability of winning the race. Naudé and Dimitri (2020) develop an all-pay contest model for deriving public-policy measures to avoid the emergence of a poor-quality (possibly dangerous to humans) Artificial General Intelligence (AGI) out of the winner-takes-all race.[26] The paper argues for globally co-ordinated taxation and regulation of AI.

To date, AI regulatory initiatives do not show clear evidence global co-ordination. In keeping with SDT-motivated limitations of regulatory sandboxes,[27] there are no public calls by regulators requiring frontier AI system development to take place within regulatory sandboxes.

Competition authorities are also concerned about the possibility of a 'winner-takes-most' outcome among AI foundation models (CMA, 2023a). At the international level, competition authorities have argued that competitive markets for AI services are needed to ensure that their benefits will be widely felt, and that "if there are 'winner-takes-all/most' characteristics, intervention risks being ineffective if too late" (OECD, 2024).

In many IT product markets, firms gain a dominant position through network effects and leveraging those network effects into neighboring markets — i.e. through diffusion of their product through the economy and through their associated market conduct. In such IT product markets, capturing dominant positions and exploiting them takes place in the market-conduct space overseen by competition authorities. However the race for frontier AI-system dominance takes place in technology space, navigated by R&D investment — which is not conventionally the purview of competition authorities. Although it may seem that the IP and the related 'pro-innovation principle' (DSIT, 2023, 2024) are ranged against national-level competition policy concerns, they in fact operate in different spaces — the former in the technological space, the latter in the space of market conduct.

---

[26]They recommend the taxing AI and using public procurement to reduce the probability of the emergence of an 'unfriendly' AGI. The underlying logic being that these measures would reduce teams' payoffs, increase the total amount of R&D investment needed, and incentivize co-operation and co-ordination.

[27]See discussion of the limitations of regulatory sandboxes on p 30.

7 CONCLUSION

The presently developed SDT-based model operationalizes the weak PP, and this informs fundamental questions pertaining to the general scope of the PP in relation to other decision rules. This model allows integrated consideration of factors and dimensions that previously have been considered in isolation.

The first of these, which can be interpreted as the reason for the existence of the PP, is the bridging of the disparate inferential conventions and requirements of science (NHST) and policy decision making. 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 tradeoffs. The weak PP, understood as a verbal formulation of an SDT corner solution, serves as a *patch* for this disparity between the inferential conventions of science and the needs of practical policy decision making.

The second is the non-separability of the 'applicability of the PP' and the 'nature of the intervention' questions. The third is the possibility to incorporate quasi-option value. The fourth is the explicit incorporation of countervailing risks caused by intervention measures themselves. The fifth is the weak PP's relation to strong PP and super-strong PP, and the sixth is the natural complementarity between weak PP and weak IP as the two possible SDT corner solutions. Although there are other formalizations of the PP,[28] the present work is the first to formalize (weak) IP and to do so in a single common framework together with (weak) PP.

Furthermore, the SDT-based model provides a structure within which to analyze a variety of questions and issues arising in AI innovation governance. This paper focuses on sandboxing and how this regulatory instrument — which is increasingly popular globally — can be explained through the lens of SDT. Each key element of SDT — discriminability, the expected misclassification cost ratio, and the qualitative properties of the optimal inferential threshold — is reflected in the makeup and operation of regulatory sandboxes. In short, the SDT model provides decision-theoretic underpinning for regulatory sandboxing. This model suggests that AI foundation models are ill-suited for regulatory sandboxing, as their general-purpose nature entails that the set of applications by naïve, well-behaved, and nefarious users is open-ended, with the consequence that the relevant misclassification costs cannot be uniquely identified.

---

[28]Recent examples include Steel and Bartha (2022) and Boyer-Kassem and Duchêne (2024).

References

Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, Avin S, Brundage M, Bullock J, Cass-Beggs D, Chang B, Collins T, First T, Hadfield G, Hayes A, Ho L, Hooker S, Horvitz E, Kolt N, Schuett J, Shavit Y, Siddarth D, Trager R, Wolf K. Frontier AI regulation: Managing emerging risks to public safety. 2023. https://arxiv.org/abs/2307.03718

Armstrong S, Bostrom N, Shulman C. Racing to the precipice: A model of artificial intelligence development. AI & Society, 2016; 31:201–206.

Arrow KJ, Fisher AC. (1974). Environmental preservation, uncertainty and irreversibility. Quarterly Journal of Economics, 1974; 88:312–319.

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

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

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

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, https://www.iatp.org/sites/default/files/Testimony-By_Dr_Brent_BlackwelderBefore_the_Se.htm

Bleichrodt H, Quiggin J. Life-cycle preferences over consumption and health: When is cost-effectiveness analysis equivalent to cost-benefit analysis? Journal of Health Economics, 1999; 18:681–708.

Bodansky D. Deconstructing the Precautionary Principle. In Caron T, Scheiber HN (eds) Bringing New Law to Ocean Waters. Leiden, The Netherlands: Brill Academic Publishers. 2004:381–391.

Boehmer-Christiansen S. The Precautionary Principle in Germany: Enabling government. In: O'Riordan T, Cameron J (eds) Interpreting the Precautionary Principle. London, UK: Cameron May, 1994:31–60.

Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford University Press: Oxford, 2014.

Boyer-Kassem T. Philosophy and the precautionary principle: Science, evidence, and environmental policy. Ethics, Policy & Environment, 2019; 22:103–105.

Boyer-Kassem T, Duchêne S. Reasonable precaution or unjust discrimination? Applying a lexical utility model of the Precautionary Principle to moral choices. Journal of Business Ethics, 2024: 1–14. https://doi.org/10.1007/s10551-024-05833-6

Brundage M, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. 2018. https://doi.org/10.17863/CAM.22520

Carter JA, Peterson M. On the epistemology of the precautionary principle. Erkennitis, 2015; 80:1–13.

Del Castillo AP. Labour in the age of AI: Why regulation is needed to protect workers. European Trade Union Institute (ETUI) Foresight Brief, 2020; #8:1–16. https://www.etui.org/publications/foresight-briefs/labour-in-the-age-of-ai-why-regulation-is-needed-to-protect-workers

Castro D, McLaughlin M. Ten Ways the Precautionary Principle Undermines Progress in Artificial Intelligence. The Information Technology & Innovation Foundation, Washington, DC, 2019. https://itif.org/sites/default/files/2019EuropeanTradeUnionInstitute(ETUI)-precautionary-principle.pdf

Consumer Financial Protection Bureau. CFPB Launches Project Catalyst to Spur Consumer-Friendly Innovation. 14 Nov 2012. https://www.consumerfinance.gov/about-us/newsroom/consumer-financial-protection-bureau-launches-project-catalyst-to-spur-consumer-friendly-innovation/

Cheng W, Yang X. Inframarginal analysis of division of labor: A survey. Journal of Economic Behavior and Organization, 2004; 55:137–174.

Chéve M, Congar R. Managing environmental risks under scientific uncertainty and controversy. In: van Ierland EC, Weikard HP, Wessler J (eds), Proceedings of the International Conference on Risk and Uncertainty in Environmental and Resource Economics. Wageningen, The Netherlands: Wageningen University. 2002:5–7.

Competition and Markets Authority. AI foundation models: Initial report. 18 Sept 2023. https://www.gov.uk/government/publications/ai-foundation-models-initial-report

Cornelli G, Doerr S, Gambacorta L, Merrouche O. Regulatory sandboxes and fintech funding: Evidence from the UK. Review of Finance, 28(1): 203–233.

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

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

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

Danielsson J, Macrae R, Uthemann A. Artificial Intelligence and Systemic Risk. LSE Systemic Risk Centre working paper, 2020. http://dx.doi.org/10.2139/ssrn.3410948

Dolan P, Edlin R. Is it really possible to build a bridge between cost-benefit analysis and cost-effectiveness analysis? Journal of Health Economics, 2002; 21:827–843.

Douglas H. Inductive Risk and Values in Science. Philosophy of Science, 2000; 67:559–579.

Douglas H. 2009. Science, Policy, and the Value-Free Ideal. Pittsburgh, PA: University of Pittsbrugh Press. 2009.

Department of Science, Innovation and Technology. A pro-innovation approach to AI regulation. 2023. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

Department of Science, Innovation and Technology. Implementing the UK's AI regulatory principles: Initial guidance for regulators. 2024. https://www.gov.uk/government/publications/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators

Ducuing C. A legal principle of innovation? Need for an assessment against the principle of democracy. Law, Innovation and Technology, 2022; 14(1):237–266.

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

European Commission. Towards an innovation principle endorsed by better regulation. European Political Strategy Centre, Issue 14, 30 June 2016, Publications Office, 2016, https://data.europa.eu/doi/10.2872/626511

European Commission. Artificial Intelligence for Europe. COM/2018/237, EUR-Lex-52018DC0237. European Commission, Brussels, BE, 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237

European Commission. Ensuring EU legislation supports innovation. 2019. https://research-and-innovation.ec.europa.eu/law-and-regulations/ensuring-eu-legislation-supports-innovation_en

European Commmission. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L series, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

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

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

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

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

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

Garber AM, Phelps CE. Economic foundations of cost-effectiveness analysis. Journal of Health Economics, 1997; 16:1–31.

Garnett K, Van Calster G, Reins L. Towards an innovation principle: An industry trump or shortening the odds on environmental protection? Law, Innovation and Technology, 2018; 1–14.

Gollier C, Jullien B, Treich N. (2000). Scientific progress and irreversibility: An economic interpretation of the 'Precautionary Principle'. Journal of Public Economics, 2000; 75:229–253.

Gollier C, Treich N. (2003). Decision-making under scientific uncertainty: The economics of the precautionary principle. Journal of Risk and Uncertainty, 2003; 27:77–103.

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

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

Grant S, Quiggin J. Bounded awareness, heuristics and the Precautionary Principle. Journal of Economic Behavior & Organization, 2013; 93(C):17–31.

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

Henry C, Henry M. Formalization and applications of the Precautionary Principle. Unpublished manuscript, Department of Economics, Columbia University, New York, NY, 2002.

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.

Hudig D. Fostering innovation – Better management of risk. Brussels: European Risk Forum. 2015. http://www.riskforum.eu/uploads/2/5/7/1/25710097/monograph_innovation_principle.pdf

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.

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.

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

Johannesson M. The relationship between cost-effectiveness analysis and cost-benefit analysis. Social Science and Medicine, 1995; 41:483–489.

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

Kaivanto K, Steel D. Adjusting inferential thresholds to reflect non-epistemic values. Philosophy of Science, 2019; 86:255–285.

Krugman, P. The age of diminshed expectations: U.S. economic policy in the 1990s. MIT Press, 1994.

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

Kuziemski M. A Precautionary Approach to Artificial Intelligence. Project Syndicate, 2018. https://t.co/JyChqIDN1f

Lave L. The Strategy of Social Regulation: Decision Frameworks for Policy. Washington, DC: Brookings. 1981.

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

Lehmann EL, Romano JP. Testing Statistical Hypotheses (3rd ed). New York, NY: Springer. 2005.

Lundgren B, Stefánsson HO. Against the *de minimis* principle. Risk Analysis, 2020; 20:908–914.

Löfstedt R E, Fischhoff B, Fischhoff IR. Precautionary Principles: General definitions and applications to genetically modified organisms. Journal of Policy Analysis and Management 2002; 21:381–407.

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

Markellos, R, Ennis S, Enstone B, Manos A, Pazaitis D, Psychoyios D. Worldwide adoption of regulatory sandboxes: Drivers, constraints and Policies. Centre for Competition Policy working paper 24-02. University of East Anglia, Norwich, UK. https://competitionpolicy.ac.uk/publications/worldwide-adoption-of-regulatory-sandboxes-drivers-constraints-and-policies/

McCarthy J. From childish things: The evolving sandbox approach in the EU's regulation of financial technology. Law, Innovation and Technology, 2023; 15(1):1–24.

Naudé W, Dimitri N. The race for an artificial general intelligence: Implications for public policy. AI & Society, 2020; 35:367–379.

Neyman J. Discussion of Fisher, R. A. (1935). The logic of inductive inference. Journal of the Royal Statistical Society, 1935; 98:74–75.

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

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.

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

NOAA. Environmental Data Management at NOAA: Archiving, Stewardship, and Access. Washington, DC: The National Academies Press. 2007.

OECD. Regulatory sandboxes in artificial intelligence. OECD Digital Economy Papers, No. 356, OECD Publishing, Paris, 2023. https://doi.org/10.1787/8f80a0e6-en

OECD. Artificial intelligence, data and competition. OECD Artificial Intelligence Papers, No. 18, OECD Publishing, Paris, 2024. https://doi.org/10.1787/e7e88884-en

O'Riordan T, Jordan A. The Precautionary Principle in contemporary environmental politics. Environmental Values, 1995; 4:191–212.

Peterson M. What is a *de minimis* risk? Risk Management: An International Journal, 2002; 4:47–55.

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

Portuese A, Pillot J. The case for an innovation principle: A comparative law & economics analysis. Manchester Journal of International Economic Law, 2018; 15(2):214–237.

Renda A, Pelkmans J. EU regulation: Hindering or stimulating innovation? In: Pontus Braunerhjelm, Martin Andersson, Knut Blind, Johan E. Eklund (eds.) "Handbook of Innovation and Regulation." Edward Elgar. 2023, pp. 263–293.

Renda A, Simonelli F. Study supporting the interim evaluation of the innovation principle: Final report. European Commission: Directorate-General for Research and Innovation, Publications Office, 2019. https://data.europa.eu/doi/10.2777/620609

Rudner R. The Scientist *Qua* Scientist Makes Value Judgments. Philosophy of Science, 1953; 20:1–6.

Sandin P. Dimensions of the Precautionary Principle. Human and Ecological Risk Assessment: An International Journal, 1999; 5:889–907.

Schneier B. Liars & Outliers: Enabling the Trust that Society Needs to Thrive. John Wiley: Indiana, IN, 2012.

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.

Steel D. Philosophy and the Precautionary Principle: Science, Evidence, and Environmental Policy. Cambridge: Cambridge University Press. 2014.

Steel D, Bartha P. Trade-offs and the precautionary principle: A lexicographic utility approach. Risk Analysis, 2022; 43(2):260–268.

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

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

Thierer A. Permissionless innovation: The continuing case for comprehensive technological freedom. Published by the Mercatus Center at George Mason University, Arlington, VA. 2016.

Todt O, Luján JL. Analyzing precautionary regulation: Do precaution, science, and innovation go together? Risk Analysis, 2014; 34:2163–2173.

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

Truby J, Brown RD, Ibrahim IA, Parellada OC. A sandbox approach to regulating high-risk artificial intelligence applications. European Journal of Risk Regulation, 2022; 13:270–294.

Truby J, Dahdal A, Ibrahim IA. Sandboxes in the desert: Is a cross-border 'gulf box' feasible? Law, Innovation and Technology, 2022; 14(2):447–473.

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

Vallor S. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, Oxford, 2016.

Viscusi WK. Rational Risk Policy: The 1996 Arne Ryde Memorial Lectures. Oxford, UK: Oxford University Press. 1998.

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

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

Weiss C. Expressing scientific uncertainty. Law, Probability and Risk, 2003; 2:25–46.

Weiss C. Scientific uncertainty and science-based precaution. International Environmental Agreements: Politics, Law and Economics, 2003; 3:137–166.

Weiss C. Can there be science-based precaution? Environmental Research Letters, 2006; 1:1–7.

Wiener JB, Rogers MD, Hammitt JK, Sand PH. The Reality of Precaution: Comparing Risk Regulation in the United States and Europe. Washington, DC: RFF Press. 2011.

World Bank. Global experiences from regulatory sandboxes. Fintech Note No. 8, Washington, DC, 2020. http://documents.worldbank.org/curated/en/912001605241080935

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

Yordanova K, Bertels N. Regulating AI: Challenges and the way forward through regulatory sandboxes. In: Henrique Sousa Antunes, Pedro Miguel Freitas, Arlindo L. Oliveira, Clara Martins Pereira, Elsa Vaz de Sequeira, Luís Barreto Xavier (eds). "Multidisciplinary Perspectives on Artificial Intelligence and the Law." Springer. 2023. pp. 441–456.

Zilgalvis P. The need for an innovation principle in regulatory impact assessment. *Network Law Review*, summer 2025. https://www.networklawreview.org/innovation-principle/

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.

APPENDICES

A   Scientific inferential convention: NHST

Null Hypothesis Significance Testing (NHST) is the workhorse method of statistical inference in modern science. It combines the Neyman-Pearson concept of a critical rejection region (Neyman and Pearson, 1933) with Fisher's formulation of p-values (Fisher, 1959). Although there are basic, pointed philosophical differences between the developers of these two concepts,[29] in modern usage these differences have been glossed over or subsumed within a unified framework (Lehmann, 1993; Berger, 2003).

That NHST has become a central preoccupation within empirical science was critically noted already by Yates (Yates, 1951). Since Yates, criticism of this preoccupation and of NHST per se has been repeated and expanded (Nickerson, 2000; Sterne and Smith, 2001; Ziliak and McCloskey, 2008). The widely cited paper entitled 'Why most published research findings are false' by John Ioannidis represents one culmination of this stream of criticism Ioannidis (2005). Some of the strongest and most persistent critics of NHST are advocates of Bayesian statistical methodology (Kruschke, 2010). Nevertheless NHST remains the prevailing convention — thus far, in all but one journal.[30]

A.1   Observations

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

First, note that the $\alpha = 0.05$ level is, ostensibly, arbitrary (Sterne and Smith, 2001; Lehmann and Romano, 2005). Appendix B 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 (Cox, 1982). 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-

---

[29]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." (Neyman, 1935; Lehmann, 1993)

[30]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* (Trafimow and Marks, 2015).

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' (Cox, 1982). 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 minuscule, contributing little to overall scientific understanding or to the understanding of effective policy levers for decision making.

However, as is shown in Section 3, these three detractions lose force when a fixed α is abandoned in favor of a contextually optimal inferential threshold α^*.

B   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 $T = t(X)$,[31] 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.

> Fisher (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 significant or not. ... ...We shall not often be astray if we draw a conventional line at 0.05 ... .
>
> Fisher (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.
>
> Fisher (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... .

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

---

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

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 (Lehmann, 1993).

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" (Berger, 2003). 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 B.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. I 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, 2005), which may also be consulted for the associated proof.

Theorem B.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*

$$E_{\theta_0} \phi(X) = \alpha \tag{B.1}$$

and

$$\phi(x) = \begin{cases} 1 & \text{if} \quad \frac{f(x|\theta_1)}{f(x|\theta_0)} > k \\ 0 & \text{if} \quad \frac{f(x|\theta_1)}{f(x|\theta_0)} < k \end{cases} \tag{B.2}$$

(ii) *Sufficient condition for a most powerful test. If* ϕ *satisfies* (B.1) *and* (B.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* (B.2) *for some* k, *and it also satisfies* (B.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. (Neyman and Pearson, 1933)

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.(Neyman, 1950)

From these beginnings, inertia took hold (Cowles and Davis, 1982). 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. (Lehmann and Romano, 2005)

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 *tradeoffs* between type-I and type-II errors.

C Derivations supporting Condition 4.1

Let ϵ be the maximum probability of inferential surprise that the regulator is willing to tolerate in issuing a weak-PP red-light determination or a weak-IP green-light determination. In other words, given the current discriminating ability of scientific experiment and analysis d', the regulator is willing to issue a weak-PP red-light determination if the optimal threshold x' is such that the associated probability of surprise is bounded by ϵ:

$$1 - p(x' < x) \leq \epsilon \tag{C.1}$$

$$1 - [P(\theta_0)\alpha' + (1 - P(\theta_0))(1 - \beta')] \leq \epsilon \tag{C.2}$$

$$P(\theta_0)(1 - \alpha') + (1 - P(\theta_0))\beta' \leq \epsilon \tag{C.3}$$

The ROC curve does not fall below the principal diagonal for all $d' \geq 0$, and thus

$$1 - \beta \geq \alpha \qquad \Leftrightarrow \qquad 1 - \alpha \geq \beta \ . \tag{C.4}$$

Using this, notice that inequality (C.3) continues to be satisfied if we substitute $1 - \alpha'$ with β'. But instead substituting β' with $1 - \alpha'$ we define the upper bound of inferential surprise ϵ_1 ($\epsilon \leq \epsilon_1$) as

$$P(\theta_0)(1 - \alpha') + (1 - P(\theta_0))(1 - \alpha') \leq \epsilon_1 \tag{C.5}$$

$$(1 - \alpha') \leq \epsilon_1 \tag{C.6}$$

Let α'_{ϵ_1} be the value that solves (C.6) as an equality:

$$(1 - \alpha'_{\epsilon_1}) = \epsilon_1 \ . \tag{C.7}$$

Turning to the δ-neighborhood of (1,1)

$$\sqrt{(1 - \alpha')^2 + (1 - (1 - \beta'))^2} \leq \delta \tag{C.8}$$

$$\sqrt{(1 - \alpha')^2 + (\beta')^2} \leq \delta \tag{C.9}$$

Making use of (C.4) again and defining δ_1 ($\delta \leq \delta_1$) to be the upper bound of the neighborhood-defining radius

$$\sqrt{(1 - \alpha')^2 + (1 - \alpha')^2} \leq \delta_1 \tag{C.10}$$

$$\sqrt{2(1 - \alpha')^2} \leq \delta_1 \tag{C.11}$$

$$\sqrt{2}(1 - \alpha') \leq \delta_1 \tag{C.12}$$

Solving for upper bound δ_1 that corresponds with ϵ_1 in (C.7) gives

$$\sqrt{2}(1 - \alpha'_{\epsilon_1}) = \delta_1 \tag{C.13}$$

$$\sqrt{2} \cdot \epsilon_1 = \delta_1 \tag{C.14}$$

D Derivations supporting Condition 4.2

Let ϵ be the maximum probability of inferential surprise that the regulator is willing to tolerate in issuing a weak-PP red-light determination or a weak-IP green-light determination. In other

words, given the current discriminating ability of scientific experiment and analysis d', the regulator is willing to issue a weak-IP green-light determination if the optimal threshold x' is such that the associated probability of surprise is bounded by ϵ

$$1 - p(x < x') \leq \epsilon \tag{D.1}$$

$$1 - [P(\theta_0)(1 - \alpha') + (1 - P(\theta_0))\beta'] \leq \epsilon \tag{D.2}$$

$$P(\theta_0)\alpha' + (1 - P(\theta_0))(1 - \beta') \leq \epsilon \tag{D.3}$$

Using (C.4), notice that inequality (D.3) continues to be satisfied if we substitute $1 - \beta'$ with α'. But instead substituting α' with $1 - \beta'$ we define the upper bound of inferential surprise ϵ_0 ($\epsilon \leq \epsilon_0$) as

$$P(\theta_0)(1 - \beta') + (1 - P(\theta_0))(1 - \beta') \leq \epsilon_0 \tag{D.4}$$

$$(1 - \beta') \leq \epsilon_0 \tag{D.5}$$

Let β'_{ϵ_0} be the value that solves (D.5) as an equality:

$$(1 - \beta'_{\epsilon_0}) = \epsilon_0 \ . \tag{D.6}$$

Let the δ-neighborhood of (0,0) be

$$\sqrt{(\alpha')^2 + (1 - \beta')^2} \leq \delta \tag{D.7}$$

Making use of (C.4) again and defining δ_0 ($\delta \leq \delta_0$) to be the upper bound of the neighborhood-defining radius, then

$$\sqrt{(1 - \beta')^2 + (1 - \beta')^2} \leq \delta_0 \tag{D.8}$$

$$\sqrt{2(1 - \beta')^2} \leq \delta_0 \tag{D.9}$$

$$\sqrt{2}(1 - \beta') \leq \delta_0 \tag{D.10}$$

Solving for the upper bound δ_0 that corresponds with ϵ_0 in (D.6) gives

$$\sqrt{2}(1 - \beta'_{\epsilon_0}) = \delta_0 \tag{D.11}$$

$$\sqrt{2} \cdot \epsilon_0 = \delta_0 \tag{D.12}$$