World Trade Review (2012), 11: 2, 171–221
 doi:10.1017/S1474745611000516

Commentary on the Appellate Body Report in *Australia–Apples* (DS367): judicial review in the face of uncertainty

SIMON A.B. SCHROPP*
Senior Economist, Sidley Austin LLP

Abstract: To justify certain sanitary and phytosanitary measures *vis-à-vis* New Zealand apple fruit, Australian authorities performed an import risk assessment that was riddled with scientific uncertainty. The quality and reliability of this assessment was the central bone of contention between the parties in the *Australia–Apples* dispute. In consequence, it featured prominently in the Appellate Body's discussion. In commenting on the Appellate Body Report, we are concerned with the standard of review of Members' decisions that are taken in situations of factual uncertainty. We discuss the elements in a Member's risk assessment that, in our view, should be considered 'off limits' for review by Panels if such assessment was performed in the presence of uncertain scientific facts, and how high degrees of uncertainty affect a Panel's ability to assess alleged less-trade-restrictive alternatives. Relying on findings from the disciplines of economics, decision analysis, and risk management, we introduce a basic framework for decisionmaking in conditions of uncertainty. We submit that certain types of decisions taken by the risk assessor cannot be subject to external review, since they *necessarily* contain subjective trade-offs on the part of the risk assessor. Applying our theoretical insights to the *Apples* case, we offer a series of recommendations concerning Panels' discretion to review Members' risk assessments and suggest concrete modifications to the standard of review currently applied by the Appellate Body.

1. Introduction and summary: the standard of review under factual uncertainty

In dealing with the Appellate Body Report in *Australia–Apples* (DS367),[1] we discuss what can be called 'second generation' issues of the *Agreement on the Application of Sanitary and Phytosanitary Measures* (henceforth the '*SPS Agreement*'). We will be less, or only indirectly, concerned with the material

* Email: sschropp@sidley.com
The author would like to thank all the participants of the 2011 ALI workshop, and especially Joost Pauwelyn, Petros Mavroidis, and Joseph Weiler, for valuable input. Generous support by his employer Sidley Austin LLP is gratefully acknowledged. All opinions are the author's and should neither be attributed to the institution he is affiliated with, nor with its clients. Any errors, flaws, or lapses remain the author's sole responsibility.

1 Appellate Body Report, *Australia – Measures Affecting the Importation of Apples from New Zealand*, WT/DS367/AB/R, adopted 17 December 2010.

obligations spelled out in the *SPS Agreement*. Instead, our focus will be on the contestation of an SPS measure taken by a country, and, in particular, the process of judicial review by a dispute Panel or the Appellate Body (AB). More specifically, we are interested in the proper standard of review (SoR) of SPS measures, as it relates to the responding party's SPS risk assessment.[2]

In addressing issues of the SoR, we discuss which aspects of the responding party's risk assessment should or should not be reviewed by a Panel or the AB in situations characterized by high levels of factual uncertainty, that is in the absence of complete and comprehensive scientific knowledge concerning the sanitary and phytosanitary risks of importation.

Every risk assessment is subject to some degree of uncertainty.[3] Scientific uncertainty notwithstanding, the domestic institution charged with performing a risk assessment (the risk assessor) has to find ways of dealing with such uncertainty and ultimately must come up with results. The *Apples* dispute seems a good object of study here, because Australia was faced with substantial factual uncertainty in the creation of its risk assessment. The issue that we wish to address in this article is to what extent the presence of higher levels of factual uncertainty should influence the level of scrutiny with which Panels may review Members' risk assessments under Articles 2.2, 5.1, and 5.2 of the *SPS Agreement*. Are dispute Panels and their appointed outside experts in a position to judge whether the risk assessor properly dealt with situations where scientific evidence is scant? Which aspects of Members' risk assessments, if any, should be considered 'off limits' for Panel review? We also discuss how high degrees of factual uncertainty affect a Panel's ability to assess less-trade-restrictive alternatives under Article 5.6 of the *SPS Agreement*, and what role should Panel-appointed scientific experts play in the SoR of this provision.

As we will argue, the AB in *Australia–Apples* has not provided Panels with sufficient guidance as to how they should review Members' attempts at overcoming factual uncertainty in connection with their risk assessments. In search of answers on what constitutes the proper SoR, we turn to work conducted in the disciplines of economics, decision analysis, and risk management. Literature in these fields has dealt extensively with the issue of assessing risks in the presence of massive factual

2 The *SPS Agreement* requires WTO Members to perform risk assessments whenever they introduce sanitary or phytosanitary (SPS) measures (Articles 2.2, 5.1, and 5.2). This risk assessment must be based on scientific evidence. Alternatively – and exceptionally – Members may be able to enact provisional measures without having conducted a risk assessment in cases where relevant scientific evidence is insufficient (Article 5.7). However, for the purposes of this paper, we will focus mainly on instances where enacting Members decide that the available scientific evidence is sufficient to conduct a risk assessment, despite being faced with considerable factual uncertainty while performing their assessments.

3 The presence of factual uncertainty over probabilities of occurrence and/or economic, social, and biological consequences is probably a result of the extreme context-sensitivity of the task at hand. Even in situations, where a number of scientific studies have been conducted on, say, the health effects in connection with the import of hormone-treated beef, a risk assessment still must be adapted to the particular circumstances at issue: a certain type of hormone, a certain type of beef, and other contextual idiosyncrasies.

uncertainty. We introduce a basic framework for decisionmaking under uncertainty. Based on this framework, we distinguish different types of uncertainty that exacerbate the assessment of risk. We find that, for theoretical and practical reasons, risk assessments *necessarily* involve judgment calls and subjective trade-offs on the part of the risk assessor. Based on the theoretical insights, we develop a series of recommendations as to what constitutes a proper SoR. We then discuss these theoretical findings in the context of the *Australia–Apples* AB Report and suggest concrete modifications to the SoR currently applied by Panels and the AB.

This paper is structured as follows. Section 2 provides a summary of the *Apples* case and the AB's findings. With a view to the later discussion of factual uncertainty, we place considerable emphasis on a detailed account of Australia's risk assessment, and how Australia dealt with overcoming areas of scientific uncertainty. Section 3 gives a nontechnical summary of the concept of SoR, and reviews the approach to SoR adopted by Panels and the AB in the five SPS disputes leading up to *Australia–Apples*.[4] Reflecting on the theoretical literature on risk management, Section 4 discusses how the presence of factual uncertainty should influence Panels' SoR. Based on this discussion, we examine the SoR as pronounced by the AB in *Australia–Apples* and propose some modifications. Section 5 concludes.

2. Summary of the case and Appellate Body findings

The *Australia–Apples* dispute concerned Australia's sanitary and phytosanitary measures against three types of contagious diseases affecting apple fruit. The consequences of infection can be substantial for fruit production and the entire native flora. Australia has in place very strict sanitary and phytosanitary rules to protect its territories from an infestation of fruit and vegetable diseases.[5]

2.1 Factual background

The dispute between Australia and New Zealand over the importation of apples has a long history. New Zealand is an important producer and exporter of apples; so is the Australian state of Tasmania. Ever since the entry and outbreak of fire blight[6]

4 Readers acquainted with the details of the *Apples* dispute and the current SoR in SPS cases may wish to skip Sections 2 and 3.

5 Australia's strict rules not only concern the import of fresh and processed fruit. Even within Australia there are tight interstate quarantine restrictions on the transport of fruit and vegetables. Travelers within Australia are prohibited from importing certain fruit and vegetables when crossing state borders. See http://www.quarantinedomestic.gov.au/index.php (last visited 13 December 2011) for Australia's interstate quarantine instructions.

6 Fire blight is a plant disease caused by the bacterium *Erwinia amylovora*. It affects mainly apple and pear trees by infecting flowers, young leaves, stems, and fruits. The disease development may be severe enough to result in plant death. Springtime is the primary time for spread of fire blight. The pest may spread within host plants, infecting blossoms, fruits, twigs, branches, and leaves. Depending on environmental and orchard conditions, fire blight may spread from tree to tree. On rare occasions, an infested flower can develop into a mature apple carrying the bacteria. However, in orchards with fire-blight symptoms,

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

in Auckland, New Zealand, in the early 1920s the Australian market has remained closed to New Zealand apples. In 1986, 1989, and 1995, New Zealand officially, yet unsuccessfully, requested permission to resume exportation to the Australian market. New Zealand's fourth request in February 1999 resulted in the initiation of an import risk analysis by the Australian Quarantine and Inspection Service (AQIS), certainly one of the most experienced and skilled authorities in the field of sanitary and phytosanitary investigation worldwide. Biosecurity Australia (back then a division of AQIS and now a separate agency) needed three attempts in the form of 'draft' risk assessments (in 2000, 2004, and 2005), to finally publish its *Final Import Risk Analysis Report for Apples from New Zealand* (IRA) in 2006 (Biosecurity Australia, 2006).

In its risk assessment, Biosecurity Australia determined the risk of entry, establishment, and spread of three apple-related diseases from imports of mature apple fruit from New Zealand: (i) fire blight;[7] (ii) apple leafcurling midge (ALCM), an insect whose larvae uniquely feed on apple leaves;[8] and European canker, a fungus that affects apple orchards. It is noted that the Australian continent, that is the Australian mainland and proximate islands including Tasmania, is free from any of these pests, while all three pests exist in New Zealand.

Biosecurity Australia characterized its approach to risk assessment as 'semi-quantitative'.[9] For each pest, it combined a *quantitative* assessment of the likelihood of entry, establishment, and spread with a *qualitative* assessment of the likely associated potential biological and economic consequences. The combination of these assessments then yielded an overall determination of 'unrestricted risk', that is the risk of establishment and spread of the three pests associated with the importation of apples from New Zealand in the *absence of any risk-management measures*. Whenever the 'unrestricted risk' associated with a specific pest was deemed to exceed Australia's appropriate level of protection (ALOP[10]), the agency evaluated and ultimately recommended the adoption of certain risk-management measures to mitigate the risk.

After having conducted the baseline case of 'unrestricted risk' and comparing it against Australia's ALOP, the IRA recommended a number of risk-management measures to the Director of Animal and Plant Quarantine. The Director

fire-blight bacteria can be found on the surface of mature apples, but such external populations typically do not multiply and diminish over time. See AB Report, *Australia–Apples*, para. 134.

7 See footnote 6 above.

8 ALCM is a small short-lived fly, usually found in cool to temperate apple-producing regions. Its larvae develop by feeding on the opening leaves of apple trees, thus preventing the leaves from unfurling normally. This may result in reduced shoot and tree growth. The larvae usually pupate after having dropped to the ground. Some larvae, however, may lodge in stalk ends of the fruit, thus spreading the pest to other areas. See AB Report, *Australia–Apples*, para. 136.

9 Ibid., para. 145.

10 According to Annex A(5) to the *SPS Agreement*, the term 'ALOP' is equivalent to 'acceptable level of risk'.

subsequently determined that compliance with these 'requirements' was a precondition for the importation of apples from New Zealand. Sixteen of these requirements relating to the importation of apple fruit from New Zealand constituted the 'measures at issue' in this dispute.[11]

These 16 measures were so strict, intrusive, and expensive for New Zealand apple farmers that the New Zealand Government did not consent to the required standard operating procedures needed to meet these requirements. This resulted in an effective import ban on New Zealand apples into Australia.

2.2 *Panel proceedings and measures on appeal*

In January 2008, a Panel was established. New Zealand claimed that the 16 measures at issue, both individually and as a whole, were inconsistent with Articles 2.2, 2.3, 5.1, 5.2, 5.5, 5.6, and 8, as well as Annex C(1)(a) to the *SPS Agreement*.[12] In particular, New Zealand alleged that the Australian measures: (i) are maintained without scientific evidence; (ii) are not based on a proper risk assessment; (iii) subject imported fruit from other countries to measures substantially less restrictive than those imposed on imports of New Zealand apples; and (iv) are more trade-restrictive than necessary to achieve Australia's appropriate level of protection.

Both parties agreed that Australia's risk assessment was *the* central element of the *Apples* dispute. The quality and results of Australia's IRA, as it relates to the three pests, was seen as the fundamental bone of contention by both litigating parties.[13] New Zealand claimed that the IRA ignores available scientific evidence; Australian border-inspection processes; relevant apple-production processes in New Zealand; relevant diseases or pests in New Zealand; and relevant climatic conditions in both New Zealand and Australia in violation of Article 5.2. Furthermore, New Zealand alleged that the delay of almost eight years between its initial request for the admission of New Zealand apples into Australia and the completion of Australia's approval procedures was 'undue' in the sense of Annex C(1)(a). Hardly surprising, Australia contested these claims.

During the proceedings, the Panel decided to seek expert advice, and consequently appointed seven experts: two in the area of fire blight; two for European canker; one for ALCM; and two for pest risk assessment in general, including the use of semi-quantitative methodologies.[14]

11 It is not necessary for the purpose of this article to discuss the 16 measures in detail. They are explained at length in para. 125 of the AB Report, *Australia–Apples*.

12 See AB Report, *Australia–Apples*, para. 4.

13 In its Panel request, New Zealand linked each of the measures at issue to a specific part of Australia's IRA. Australia, too, referred to 'the reasonable measures recommended in the Final IRA Report', and argued that '[t]he Final IRA Report provides the basis for Australia's measures'. See Panel Report, *Australia – Measures Affecting the Importation of Apples from New Zealand*, WT/DS367/R, adopted 17 December 2010, as modified by Appellate Body Report WT/DS367/AB/R, para. 2.97.

14 See AB Report, *Australia–Apples*, para. 6.

After a number of quite substantial delays and a preliminary ruling on consistency of New Zealand's Panel request with Article 6.2 of the *Understanding on Rules and Procedures Governing the Settlement of Disputes* (*DSU*), the Panel Report was circulated to Members in August of 2010. The Panel found that:

- Australia's measures at issue regarding fire blight, European canker, and ALCM, as well as certain 'general measures',[15] were inconsistent with Articles 5.1 and 5.2 of the *SPS Agreement*.[16] By implication, these requirements were also inconsistent with Article 2.2 of the *SPS Agreement*.[17]
- New Zealand failed to demonstrate that the measures at issue were inconsistent with Article 5.5 and, consequentially, also failed to demonstrate that these measures are inconsistent with Article 2.3.
- Australia's measures regarding three pests were inconsistent with Article 5.6.[18]
- New Zealand's claim under Annex C(1)(a) and its consequential claim under Article 8 were outside of the Panel's terms of reference in this dispute.[19]

15 New Zealand challenged three 'general' measures, which were linked to all three pests in the present dispute. These general measures were part of the 16 measures at issue. See AB Report, *Australia–Apples*, paras. 3, 125.

16 Article 5, paras. 1 and 2, of the *SPS Agreement* read as follows:

> ***Article 5: Assessment of Risk and Determination of the Appropriate Level of Sanitary or Phytosanitary Protection***
>
> 1. Members shall ensure that their sanitary or phytosanitary measures are based on an assessment, as appropriate to the circumstances, of the risks to human, animal or plant life or health, taking into account risk assessment techniques developed by the relevant international organizations.
> 2. In the assessment of risks, Members shall take into account available scientific evidence; relevant processes and production methods; relevant inspection, sampling and testing methods; prevalence of specific diseases or pests; existence of pest- or disease-free areas; relevant ecological and environmental conditions; and quarantine or other treatment.

17 In Appellate Body Report, *EC – Measures Concerning Meat and Meat Products (Hormones)*, WT/DS26/AB/R, WT/DS48/AB/R, adopted 13 February 1998, DSR 1998:I, 135, para. 180, the AB stated that 'Articles 2.2 and 5.1 should constantly be read together' and that Article 2.2 informs and imparts meaning to Article 5.1. The same type of relationship was found to exist between Articles 2.2 and 5.2 and between Articles 2.2 and 5.6. See AB Report, *Australia–Apples*, para. 339. A violation of Articles 5.1, 5.2, or 5.6 can thus be presumed to imply a violation of Article 2.2 (but not *vice versa*). See Appellate Body Report, *Australia – Measures Affecting Importation of Salmon*, WT/DS18/AB/R, adopted 6 November 1998, DSR 1998:VIII, 3327, para. 138.

18 Article 5, paragraph 6 of the *SPS Agreement* reads as follows:

> ***Article 5: Assessment of Risk and Determination of the Appropriate Level of Sanitary or Phytosanitary Protection***
>
> 6. Without prejudice to paragraph 2 of Article 3, when establishing or maintaining sanitary or phytosanitary measures to achieve the appropriate level of sanitary or phytosanitary protection, Members shall ensure that such measures are not more trade-restrictive than required to achieve their appropriate level of sanitary or phytosanitary protection, taking into account technical and economic feasibility.

19 See AB Report, *Australia–Apples*, para. 7.

On appeal, Australia accepted the Panel's findings with respect to European canker, but challenged practically any Panel finding adverse to it with respect to fire blight and ALCM. Relevant for this article, on appeal Australia raised that, in finding that the measures regarding fire blight and ALCM are inconsistent with Articles 5.1, 5.2, and, consequently, 2.2 of the *SPS Agreement*, the Panel misinterpreted and misapplied these provisions.[20] More specifically, Australia sought clarification from the AB on the following issues:

- whether, in evaluating Australia's IRA and the consistency of Australia's SPS measures with these provisions, the Panel applied an improper SoR;
- whether, in reviewing Australia's IRA and its use of expert judgment at several intermediate steps, the Panel required too high a standard of transparency and documentation and, thereby, erred in its assessment of the objectivity and coherence of the reasoning of the risk assessor; and
- whether the Panel erred in failing to assess the materiality of the faults it found with Australia's IRA, and in failing to determine whether any alleged flaws were so serious as to call into question the risk assessment as a whole.

With respect to Article 5.6, Australia asked the AB to find that the Panel inappropriately relied on its findings under Articles 5.1, 5.2, and 2.2 of the *SPS Agreement* in concluding that New Zealand's proposed alternative measures would achieve Australia's appropriate level of protection.

In its Report, the Appellate Body upheld the Panel's findings that Australia's measures regarding fire blight and ALCM, as well as the general measures relating to these pests, are inconsistent with Articles 5.1 and 5.2 of the *SPS Agreement*, and that, by implication, these measures are also inconsistent with Article 2.2. However, the AB reversed the Panel's reasoning underlying the finding that Australia's measures at issue are inconsistent with Article 5.6. Although Australia had not asked it to do so, the AB attempted to complete the legal analysis of New Zealand's claim under that provision. Eventually, the AB declared itself unable to complete the legal analysis.[21]

Ultimately, the AB recommended that the Dispute Settlement Body request Australia to bring its WTO-inconsistent measures, as determined by the AB and in the Original Panel Report (as modified), into conformity with its obligations under the *SPS Agreement*.

The fact that it crushingly lost on appeal probably did not come as a great surprise for Australia. Whereas the first IRA had taken more than six years and several revisions to be completed, Australia's next attempt was completed in less

20 Ibid., para. 124.
21 Ibid., para. 444.

than six months after the publication of the AB Report.[22] In its latest version of the draft report, AQIS concludes that:

> when the New Zealand apple industry's standard commercial practices for production of export grade fruit are taken into account, the unrestricted risk for all three pests assessed [fire blight, ALCM, European canker] achieves Australia's appropriate level of protection (ALOP). Therefore, no additional quarantine measures are recommended, though New Zealand will need to ensure that the standard commercial practices detailed in this review are met for export consignments.[23]

Given that Australia's new import requirements are much less rigid than the 16 measures at issue in the dispute, and partially reflect ideas that New Zealand itself had suggested as less-trade-restrictive alternatives,[24] it can be expected that the New Zealand Government will accede to Australia's novel requirements. Although Australia and New Zealand are still negotiating at the time of writing, commentators are confident that the parties will settle their dispute soon.[25]

2.3 *Australia's import risk assessment of unrestricted import of New Zealand apples*

In this subsection, we review the IRA's structure, methodology, and conclusions, as well as the Panel's findings relating to the IRA.[26] The reason for introducing the methodology of the IRA in considerable detail is to demonstrate the substantial complexity of the task at hand, how Biosecurity Australia addressed the various aspects of factual uncertainty it faced,[27] and what the Panel had to say about the IRA's handling of such uncertainty upon its review of the IRA. This review will

22 The report (AQIS, 2011), entitled 'Draft Report for the *Non-Regulated* Analysis of Existing Policy for Apples from New Zealand', is currently undergoing comments from stakeholders. See http://www.daff.gov.au/ba/reviews/final-plant/non-regulated_analysis_apples_from_new_zealand/baa_2011-06_release_on_non-regulated_analysis_nz_apples (last visited 13 December 2011).

23 See AQIS (2011: xv), emphasis added. The standard commercial practices required of New Zealand farmers include: application of the integrated fruit-production system to manage pests and diseases in orchards; testing to ensure that only mature fruit is exported to Australia; maintenance of sanitary conditions in dump-tank water; high-pressure water washing and brushing of fruit in the packing house; and that a minimum 600-fruit sample from each lot of fruit packed be inspected and found free of quarantine pests for Australia. See ibid.

24 A requirement of inspection of a 600-fruit sample of each import lot was suggested by New Zealand as a less-trade-restrictive alternative to Australia's catalogue of measures. See AB Report, *Australia–Apples*, para. 124.

25 See Kolsky Lewis (2011).

26 Readers acquainted with the details of the dispute may wish to skip this part. For the summary of Australia's IRA, see AB Report, *Australia–Apples*, paras. 131–164 and 186–189.

27 Such instances of factual uncertainty were manifold, because Australia, in its assessment, went into such detail that it oftentimes exceeded the limits of what had previously been discussed in the relevant literature.

provide essential background for our discussion of Panels' standard of review in situations of factual uncertainty later on (Section 4).

Australia's approach to assessing the risks from the two pests at issue was quite complex. It consisted of two parts: the Pest Risk Assessment and the Pest Risk Management.

The Pest Risk Assessment discussed the risks of entry, establishment, and spread of fire blight and ALCM against an unmitigated benchmark scenario ('unrestricted risk').

The IRA's Pest Risk Assessment consisted of three major parts:

(i) a *quantitative* estimation of the probability of entry, establishment, and spread of mature infected apples imported from New Zealand;
(ii) a *qualitative* estimation of the biological, economic, social, and environmental effects of the entry, establishment, and spread of infected apples; and
(iii) a '*semi-quantitative*'[28] combination of the estimated probability of entry, establishment, and spread with the estimate of consequences.

The Pest Management Assessment succeeded the Pest Risk Assessment. It is described as the process of identifying, evaluating, and implementing several (combinations of) measures aimed at mitigating the risks of the pest, so as to achieve Australia's ALOP, whenever the 'unrestricted risk' exceeds Australia's ALOP. The IRA defined Australia's ALOP as 'providing a high level of sanitary and phytosanitary protection aimed at reducing risk to a very low level, but not to zero'.[29] The Pest Risk Assessment followed the same methodological approach, just replacing the baseline situation of 'unrestricted risk' for (combinations of) risk-mitigation measures.[30]

(i) The probability of entry, establishment, and spread. The IRA's assessment of the probability of entry, establishment, and spread consisted of three subparts: an estimation of (a) the probability of entry, that is the probability of the importation of *infected* apples; (b) the probability of establishment, that is the likelihood of propagation of the pest once an infected apple is imported; and (c) the probability of spread, that is the likelihood of dispersion of the pest in susceptible hosts. As a baseline scenario, the IRA limited itself to the importation of mature, symptomless apple fruit from New Zealand within 12 months.[31] The volume of imports was

28 AB Report, *Australia–Apples*, para. 145.

29 Ibid., footnote 9, and para. 148.

30 As explained by the AB, there are slight differences in the IRA methodologies for pathogens (for example, bacteria and viruses, such as fire blight), and those used for arthropod pests (such as ALCM). See ibid., para. 136. These differences notwithstanding, we will introduce Australia's IRA as it relates to fire blight, and limit ourselves to the comment that the assessment for ALCM is no less complex.

31 Australia may have learned a lesson from the *Japan–Apples* case, where Japan unsuccessfully tried to assess the risk from imports other than mature, symptomless apples. See Appellate Body Report,

estimated to be between 50 and 400 million apples per year, with a most likely volume of 150 million apples.[32]

The *probability of entry* was further divided into three partial probabilities: (1) the probability of importation (P_{imp}), that is the probability of contamination of New Zealand apples ('vectors' or 'carriers'); (2) the probability of proximity (P_{prox}), that is the probability of contaminated apples getting sufficiently close, and thus exposed, to susceptible hosts ('receivers') in Australia; and (3) the probability of exposure (P_{exp}), that is the probability of transfer or contamination of susceptible Australian fauna with fire blight.

To assess P_{imp}, the IRA combined different importation steps with certain importation scenarios: it defined eight ways of how a New Zealand apple could be contaminated (infected or infested) during the harvesting and exportation process, and assigned a probability to each alternative.[33] These eight contamination probabilities were then recombined into ten scenarios, or 'biological pathways', in which an infested/infected New Zealand apple would reach Australia's shores after the quarantine.[34] Note that the estimation of P_{imp} concerned contamination rates within New Zealand and thus contained data proprietary to New Zealand. The IRA used expert judgment to generate the probabilities for each importation scenario.[35] To derive P_{imp}, the IRA summed all partial probabilities generated from the ten importation scenarios. The IRA estimated the overall probability of importation of apples infested with fire blight to be 3.9% of the total number of apples that would be imported from New Zealand.[36]

Things get slightly more complicated with respect to P_{prox}, the likelihood that a contaminated carrier (a mature New Zealand apple) comes sufficiently close to a susceptible host plant in Australia. The IRA defined five so-called 'utility points' (vectors), that is handlers and users of (contaminated) New Zealand apples,[37] and four 'exposure groups' (receivers), that is susceptible hosts that potentially could

Japan – Measures Affecting the Importation of Apples, WT/DS245/AB/R, adopted 10 December 2003, DSR 2003:IX, 4391, paras. 150–160.

32 A volume of 150 million apples constitutes roughly 7.5% of Australia's annual production of apples.

33 For example, one importation step is simply the proportion of contaminated fruit harvested. Other, more complicated importation steps are also included, such as the proportion of previously clean fruit becoming contaminated by the pest during processing in the packing house or during palletization, containerization, and transportation to Australia. See AB Report, *Australia–Apples*, footnote 181.

34 A straightforward example for such biological pathway is the case of an infested apple that remains infested all through its way to Australia, despite routine inspections and quarantine. Biological pathways are more complex processes for a 'clean' apple from a 'clean' orchard that gets infected in the packing house and survives routine inspection and quarantine until reaching Australia. It is not reported how the IRA effectuated this 'combination' (rather, permutation) of importation steps into importation scenarios.

35 The IRA assigned an importation step value of 1 (100%) to two importation steps, and assigned importation step values based on probability intervals and a triangular probability distribution function to the other six importation steps.

36 See AB Report, *Australia–Apples*, para. 149.

37 These utility points were: orchard packing houses/wholesalers, urban packing houses/wholesalers, retailers, food-service industries, and consumers. See ibid., para. 139.

become infected.[38] Recombining utility points and exposure groups, the IRA assigned proximity probabilities for each of the 20 combinations. The IRA determined that for one of the 20 combinations P_{prox} is 100% (i.e., certainty), while it defined probability intervals with maximum and minimum values for all the other 19 combinations.[39] It can be safely assumed that there was no pre-existing literature available on the probability of proximity between the unique 'utility points' and 'exposure groups' as described in the IRA. The IRA filled this scientific gap with expert judgment.

Finally, for the determination of P_{exp}, the probability of actual contamination of Australian fauna with fire blight, the IRA came up with what it described as a 'complex variable dependent on a number of factors, such as viability of the pest, survival mechanism of the pest, transfer mechanism(s) of the pest, host receptivity, and environmental factors'.[40] The IRA noted that a sequence of events needed to be completed for successful contamination of host plants, and analyzed key steps in such a sequence of events. Ultimately, the IRA, based on expert opinion, assigned to each of the 20 combinations an exposure value for an individual apple within a range of zero and one-millionth, using uniform distribution.[41]

The *probability of establishment*, or likelihood of propagation of the pest, was derived by the IRA for each of the five 'exposure groups' from a comparative assessment of various factors, such as the availability of suitable hosts; alternate hosts and vectors in the pest risk-analysis area; suitability of the environment; cultural practices and control measures; and other characteristics of the pest relevant to the probability of establishment.[42] How exactly each criterion factored into the calculation of probability is not conveyed in the AB Report, but it cannot have been a straightforward task, given the extreme case-sensitivity of the single pathways determined in the IRA.

Next, the *probability of spread*, or likelihood of dispersion in susceptible hosts, was derived for each of the five exposure groups from a comparative assessment of various factors, including the suitability of the natural and/or managed environment for natural spread of the pest; presence of natural barriers; the potential for movement with commodities or conveyances; intended use of the commodity; potential vectors of the pest in the pest risk-analysis area; and potential natural enemies of the pest in the pest risk-analysis area.[43] Again, details of how the

38 These exposure groups were susceptible commercial fruit crops, nursery plants, household and garden plants, and wild and amenity species.

39 AB Report, *Australia–Apples*, para. 150. It is noted that not all 20 vector-host combinations were deemed equally likely by the IRA. Some weighting of the combination scenarios seems to have occurred, although this weighting scheme seems rather random. Ibid.

40 Ibid., para. 140.

41 Ibid., para. 152.

42 Ibid., footnote 193.

43 Ibid., footnote 194.

IRA got to the final probability for each exposure group is unclear. The IRA just remarked that it was following international guidelines for risk analysis and that each partial probability was described as a probability interval with uniform distribution.[44]

To determine the *overall* estimate of the probability of entry, establishment, and spread, the IRA combined the partial probabilities of establishment and spread for each exposure group with the overall probability of importation, along with the estimated annual importation volume of 150 million apples in a computer program that performed a Monte Carlo simulation (a simulation model with random sampling of each subprobability, taking into account the probability range and form of distribution of each of these subprobabilities in the various steps). Assuming importation of 150 million apples per year, the IRA's estimation for the overall probability of entry, establishment, and spread was between 0.0087 and 0.18, with a median of 0.044, corresponding to a contamination risk with the qualitative descriptor of 'very low'.[45]

(ii) The assessment of consequences. Regarding consequences of establishment and spread of fire blight in Australia, the IRA discussed biological, economic, social, and environmental effects in two dimensions – direct and indirect effects.[46] The impact (magnitude) of the pest was considered at four levels: local, district, regional, and national effects. The IRA described the impact in *qualitative* terms, then used a correspondence table to convert those qualitative terms into 'impact scores' on an eight-step scale, ranging from 'A' (unlikely to be discernible) to 'G'

44 Ibid., paras. 141 and 152.

45 See ibid., footnote 232. We have numerous concerns about the way the IRA estimated the overall probability of occurrence of the pests at issue. We shall limit ourselves to the following observations: *First*, to us it is not clear whether the final probability estimated by the IRA is one of entry, establishment, *or* spread (as Annex A(4) of the *SPS Agreement* mandates), or one of entry, establishment, *and* spread. The difference is that the probability of entry, establishment, *and* spread seems to add the subprobabilities, thereby inflating the overall assessment, since only the probability of *spread* should be causing consequences in Australia. *Second*, the estimation of P_{imp}, as part of the assessment of the probability of entry, seems highly arbitrary, and its treatment cursory and under-researched. The derivation of P_{imp} includes issues of contamination, transfer, and propagation occurring in New Zealand during the process of exportation. These are complex processes, as conceded by the IRA in its later assessment of contamination, transfer, and propagation as it occurs in Australia. To our mind, there must be other, more tractable, ways of assessing the probability of importing an infested or infected apple from New Zealand. *Third*, there seems to be significant overlap between the probability of exposure (as part of the probability of entry) and the probability of establishment: both describe the likelihood of contamination of Australian apples. Such overlap could lead to a substantial overestimation of the overall probability. *Fourth*, the process by which the IRA arrived at the probabilities of establishment and spread is highly opaque. Even though it is claimed that the IRA is based on 'international guidelines', this is no replacement for explaining in detail how the IRA took into consideration the unique circumstances and conditions in Australia.

46 'Direct' effects relate to plant and life or health and to other environmental effects. 'Indirect' effects relate to control and eradication costs, repercussions on domestic trade and international trade, environment, and communities. See AB Report, *Australia–Apples*, para. 144.

(highly significant).[47] These individual impact scores were evaluated and combined, based on certain 'decision rules' set forth in the IRA, to arrive at an overall conclusion on the potential biological and economic consequences. The overall conclusion was expressed in qualitative terms on a six-step scale, ranging from 'negligible' to 'high'.[48] In applying these decision rules, the IRA arrived at the outcome that the overall potential biological and economic consequences of fire blight were 'high'.[49]

(iii) Combining probabilities and consequences. Usually, risk is measured by the mathematical product of total effects (measured along a single dimension such as monetary costs or mortality) and overall probability of occurrence. In the case of New Zealand apples, the IRA instead chose a 'semi-quantitative' approach,[50] which expressed the risk in qualitative terms. To combine the IRA's *quantitative* estimate of the probability of entry, establishment, and spread of fire blight with its *qualitative* estimate of consequences, the IRA first devised a way of converting the numerical probability results into qualitative ratings.[51] Next, the IRA used a self-defined 'risk estimation matrix' to yield an overall qualitative determination of 'unrestricted risk' associated with fire blight if apples from New Zealand were imported to Australia for 12 months without any phytosanitary measures. This matrix is reproduced on the next page as Figure 1.

The matrix depicts two dimensions: the *likelihood* of entry, establishment, and spread of a pest, and the *consequences* that the entry, establishment, and spread entails. Different combinations of likelihoods/probabilities and consequences get assigned six grades of risk: negligible, very low, low, moderate, high, and extreme risk. Each dimension is broken down in qualitative terms on a six-step scale,

47 The IRA assigned the following impact scores in its assessment of the potential consequences of the entry, establishment, and spread of fire blight in Australia: (i) for the direct consequences, the IRA assigned an 'F' to the effects on plant life or health, and an 'A' to the effects on human life or health and on any other aspects of the environment; and (ii) for the indirect consequences, the IRA assigned an 'E' to control and eradication and to impact on the domestic industry, a 'D' to the effects on international trade and on communities, and an 'A' to effects on the environment.

48 See AB Report, *Australia–Apples*, para. 144. The mechanics of the 'decision rules' along which the economic and biological effects are weighted and compared are opaque. For example, it is unclear, how 'medium' consequences on human life are paired with 'significant effects' on domestic trade. Also, it is not evident, how effects along the four levels – local, district, regional, and national – are kept distinct, so as to avoid double-counting of effects.

49 Ibid. In other words, the IRA converted a *nominal scale* of consequences (a scale of different, and *a priori* incommensurable, *qualitative* dimensions) into an *ordinal scale* (a rank-ordered scale), using rather opaque 'decision rules'.

50 Ibid., para. 145.

51 To this end, the IRA established a correspondence between the *quantitative* probability values and *qualitative* likelihood ratings according to a self-defined 'nomenclature' table. This table is reproduced in para. 147 of the AB Report, *Australia–Apples*. In other words, the IRA converted (rather: downgraded) a *cardinal* scale of probabilities (a scale of measurable quantitative attributes) into an *ordinal scale* (a rank-ordered scale), using a rather subjective correspondence table.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

Figure 1. Risk estimation matrix used by IRA, and assessment result for fire blight

Likelihood of entry, establishment and spread	Negligible	Very low	Low	Moderate	High	Extreme
High	Negligible risk	Very low risk	Low risk	Moderate risk	High risk	Extreme risk
Moderate	Negligible risk	Very low risk	Low risk	Moderate risk	High risk	Extreme risk
Low	Negligible risk	Negligible risk	Very low risk	Low risk	Moderate risk	High risk
Very low	Negligible risk	Negligible risk	Negligible risk	Very low risk	Low risk	Moderate risk
Extremely low	Negligible risk	Negligible risk	Negligible risk	Negligible risk	Very low risk	Low risk
Negligible	Negligible risk	Negligible risk	Negligible risk	Negligible risk	Negligible risk	Very low risk

Consequences of entry, establishment and spread

Source: AB Report, *Australia–Apples*, para. 147.

ranging from 'negligible' to 'high'. Since the overall probability of entry, establishment, and spread was considered to fall within the category of 'very low', and the impact score of the overall potential biological and economic consequences of fire blight was considered 'high', the IRA concluded, in accordance with the matrix, that the unrestricted risk of fire blight was 'low' (see boxed cell in Figure 1).[52]

The outcome of 'low' risk exceeded Australia's ALOP.[53] This required Pest Risk Management action to mitigate the risk resulting from the importation of mature apples from New Zealand.

52 Let us recall how this result came about. The IRA converted (rather, downgraded) a *cardinal scale* of probabilities into an *ordinal scale*, using a rather arbitrary correspondence table (see footnote 51 above). For the assessment of consequences, it converted (rather, upgraded) a series of qualitative outcome dimensions into one single *ordinal scale*, using a set of unexplained decision rules (see footnote 48 above). These two ordinal scales were then combined into one matrix, which included yet another creative act: the denomination (and thus evaluation) of the matrix cells as 'extreme', 'high', 'moderate', etc. risk. In sum, Australia's 'semi-quantitative' approach may look scientific, but is fundamentally dependent on a number of judgment calls by the IRA.

53 Recall that Australia's ALOP was defined as 'providing a high level of sanitary and phytosanitary protection aimed at reducing risk to a *very low* level, but not to zero'. See AB Report, *Australia–Apples*, footnote 9, and para. 148.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

(iv) Pest Risk Management. The IRA used the same risk-assessment methodology as in its Pest Risk Assessment of unrestricted risk to compare different risk-mitigation alternatives, and to assess the effects of, and risk associated with, potential risk-management measures.[54] Since none of the individual options would sufficiently reduce risk, combinations of measures were examined and found to reduce the risk to 'very low' on the 'risk estimation matrix', and therewith *within* Australia's ALOP.[55] These combinations of measures formed the basis of AQIS's risk-management recommendations to the Director of Animal and Plant Quarantine, who then formulated certain 'requirements' for the importation of apples from New Zealand which ultimately constituted the 'measures at issue' in this dispute.

2.4 *Factual uncertainty in Australia's risk assessment*

Apart from facing a number of methodological issues, such as how to combine various types of consequences into a single result and how to combine qualitative probabilities with qualitative consequences,[56] Australia, in performing its risk assessment, was faced with a number of issues that have not been addressed previously in such detail in the literature.[57] Due to the unique circumstances of the task at hand (and owing to the particular way it tackled its task), the IRA had to deal with a series of 'knowledge gaps', or areas of factual uncertainty, that needed to be 'bridged'.

The IRA bridged such areas of uncertainty by allowing for flexibility in the form of introducing confidence bands and event scenarios, and by involving expert opinion.[58] When assigning quantitative values to the likelihoods associated with the importation steps and the factors relating to proximity, exposure, establishment, and spread, the IRA rarely used point values (single numbers). More frequently, and whenever the IRA saw limited evidence or considered the underlying biological process to be naturally highly variable, it used probability *intervals*, or ranges.[59]

54 Ibid., para. 148.

55 Ibid., para. 155. The IRA's assessment on the other pest, namely ALCM, led to the following results: the overall probability of entry, establishment, and spread was seen as 'moderate', while the estimate of consequences was judged 'low'. Based on these ratings, and according to the risk-estimation matrix depicted in Figure 1, the IRA concluded that the overall unrestricted risk of ALCM was 'low', and thus above Australia's ALOP. The resulting Pest Risk Management recommended a series of actions to mitigate that risk and bring it below Australia's ALOP threshold. See ibid., paras. 156–164.

56 See footnotes 48, 49, and 52 above.

57 Take for example the IRA's estimation of proximity, where it paired five 'utility points' (vectors) with four 'exposure groups' (hosts) and assessed the likelihood that proponents of these two groups came sufficiently close. There is probably not much scientific literature on the likelihood of an apple wholesaler crossing paths with susceptible nursery plants.

58 Expert judgment was exercised collectively by the six members of the IRA team, allegedly taking into consideration relevant stakeholder comments. As to the extent of experts' input, Australia admitted that, while the IRA 'did not rely on 100% expert judgment', 'it is true that certain steps in the pathways assessed were better supported by evidence than others' and that '[i]n those latter cases, expert judgment was employed'. See AB Report, *Australia–Apples*, footnotes 360 and 373.

59 Ibid., footnote 175, and para. 232.

According to the IRA, these ranges were based on scientific evidence and resources, as well as on the IRA expert judgment.[60] In most cases, the IRA, supported by its experts, selected the *numerical endpoints* of the intervals from a set of predefined *quantitative ranges* which it then converted into numerical values.[61] As to the probability distributions in connection with these probability intervals, the IRA selected one of three distribution types: uniform, triangular, and pert.[62] The IRA also considered two alternative scenarios regarding P_{prox}, the probability of proximity, to allow for different weights regarding apple outlets in Australia.[63]

2.5 *Findings of the Panel regarding Australia's risk assessment*

The Panel was not convinced by the quality of the IRA.[64] For the risk of fire blight from New Zealand apples, the Panel found that:[65]

- With respect to the IRA's estimation of the *probability of importation* (P_{imp}), for four out of eight 'importation steps', there was no sufficient support in the scientific evidence relied upon.[66]
- With respect to the *overall probability of exposure, establishment, and spread*, a number of the assumptions and qualifications expressed in the IRA were not

60 Ibid., para. 136.

61 This conversion of expert judgment into numerical terms casts doubt on the nature of the 'quantitative' approach towards probabilities of entry, establishment, and spread. It was also subject to disagreement between the parties. For example, the fact that the IRA matched the qualitative category 'negligible' with a probability interval of [0; $1*10^{-6}$] with a midpoint of 0.5 in a million was seen as excessive by the Panel. The IRA's own qualitative descriptor for events occurring with a 'negligible' likelihood was that '[t]he event would almost certainly not occur', which included 'explicit acknowledgment that in some circumstances the chances … would be zero'. The Panel ruled that an event that almost certainly will not occur cannot objectively be matched with a positive mean of $0.5*10^{-6}$, which, after all, corresponds to 75 incidents a year under the assumption of an import volume of 150 million apples. The Panel thus ruled that the IRA's conversion impermissibly inflated the estimate of the unrestricted overall risk. See ibid., paras. 319–326.

62 The probability density function describes the relative likelihood of occurrence within a certain interval. All three distribution types utilized by the IRA have a maximum and a minimum value. Triangular and pert distributions also have a third parameter: the most likely value. In a uniform distribution, in contrast, each value in the continuous range between these minimum and maximum values occurs with equal probability. The IRA stated that uniform distribution was used where there was insufficient information to determine a most likely value. Ibid., para. 146.

63 Ibid., para. 150.

64 We have expressed our own concerns about the methodology and approach taken by the IRA in its risk assessment. See footnotes 45, 48, and 52 above. Yet, our opinion is secondary for the purposes of this article.

65 AB Report, *Australia–Apples*, paras. 190–195 and 252–257.

66 See ibid., para. 191, citing Panel Report, *Australia–Apples*, para. 7.447 ('[T]he Panel concluded that: "… the IRA's estimation that [fire-blight bacteria] will be always present in the source orchards in [N]ew Zealand (importation step 1); that fruit coming from an infected or infested orchard is infected or infested with [fire-blight bacteria] (importation step 2); that clean fruit from infected or infested orchards is contaminated with [fire-blight bacteria] during picking and transport to the packing house (importation step 3); and that clean fruit is contaminated by [fire-blight bacteria] during processing in the packing house (importation step 5); do not find sufficient support in the scientific evidence relied upon and, accordingly, are not coherent and objective"').

convincing, and the IRA had 'not properly considered a number of factors that could have [had] a major impact on the assessment of this particular risk'.[67] This led to reasonable doubts about the evaluation made by the risk assessor.

- Concerning *IRA expert involvement*, the IRA resorted to expert judgment to estimate the quantitative probability of certain events, even where scientific evidence was available, without offering reasoned and adequate explanations for having done so.[68] In addition, the IRA failed to explain and document how it arrived at the expert judgment.[69]
- Regarding the potential *biological and economic consequences* associated with fire blight, the IRA had a tendency to overstate the severity of the consequences of fire blight in certain aspects, particularly in respect of criteria concerning plant life or health, and domestic trade or industry.[70] In addition, the IRA did not adequately consider the existence of certain mitigating factors when assessing the effects of establishment and spread of the pests in Australia.[71]
- The IRA contained certain *technical flaws* that magnified the risk assessed and, for that reason, too, did not constitute a proper risk assessment within the meaning of Article 5.1.[72]

In sum, the Panel found that the reasoning articulated in the IRA did not rely on adequate scientific evidence and, accordingly, was not coherent and objective. Thus, the Panel concluded that the IRA did *not* constitute a proper risk assessment within the meaning of Articles 5.1 and Annex A(4) to the *SPS Agreement*, and consequently was inconsistent with Article 2.2.[73] The Panel also held that the flaws it had found in the IRA constituted a failure, under

67 Such factors concerned climatic conditions, mode of trade, pest viability, and others. See AB Report, *Australia–Apples*, paras. 253, 256.

68 Ibid., para. 241.

69 Ibid., para. 248.

70 Ibid., paras. 254, 257.

71 For example, the Panel held that the IRA had adequately considered the existence of the climatic conditions necessary for the establishment and spread of ALCM in Australia. See ibid., paras. 197, 300, and 312. We do not find this convincing. It seems that the Panel effectively mixed up the dimensions of 'probability' and 'consequences'. Negligence of climatic conditions may reduce the probability of occurrence of fire-blight spread, but not the existence of negative effects, or their magnitudes.

72 One technical flaw the Panel found concerns the IRA's choice of probability interval for events with a 'negligible' likelihood of occurring (see footnote 61 above). The second flaw, according to the Panel, was that the IRA's use of *uniform* distribution (see footnote 62 above) to model the likelihood of negligible events would tend to generate less realistic samples, and thus 'to result in an additional overestimation of the likelihood of such "negligible events"'. See AB Report, *Australia–Apples*, footnote 459. The form of a probability function is often as important as its range, mean value, and variance. Where differing irregular or asymmetrical probability density functions overlap, this can have significant nonlinear implications on the outcomes. See Stirling (2001: 55).

73 Annex A(4) defines pest risk assessment as:

> The evaluation of the likelihood of entry, establishment or spread of a pest or disease within the territory of an importing Member according to the sanitary or phytosanitary measures which might be applied, and of the associated potential biological and economic consequences.

Article 5.2, to take adequately into account factors such as the available scientific evidence; the relevant processes and production methods in New Zealand and Australia; and the actual prevalence of fire blight and viable ALCM.[74]

3. The standard of review in SPS disputes pre *Australia–Apples*

In this section, we briefly address the standard of review ('SoR') adopted by Panels and the AB in the five SPS disputes leading up to *Australia–Apples*.[75] The SoR is usually understood as the 'degree of deference or discretion that the court accords to legislator or regulator' or the 'degree of intrusiveness or invasiveness into the legislator's or regulator's decision-making process'.[76] In the WTO context, the SoR is a mechanism for allocating the power between national governments and the international reviewing body (Panel or AB).[77]

No WTO Agreement bar one (the *Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994*, more commonly known as the 'Antidumping Agreement') contains a detailed provision on the applicable SoR. In consequence, it fell on the AB to identify the applicable standard for the *SPS Agreement*. In *EC–Hormones*, the AB established that Article 11 of the *DSU* 'articulates with great succinctness but with sufficient clarity the appropriate standard of review for Panels in respect of both the ascertainment of facts and the legal characterization of such facts under the relevant agreements'.[78]

74 For the wording of Articles 5.1 and 5.2, see footnote 16 above. With respect to ALCM, the Panel reached a similar conclusion. See AB Report, *Australia–Apples*, paras. 196–198. Since the two 'general' (cross-cutting) measures were inextricably linked to those for fire blight and ALCM, the Panel found a violation of Articles 5.1, 5.2, and 2.2 of the *SPS Agreement*, too. See AB Report, *Australia–Apples*, para. 199.

75 The relevant SPS cases leading up to *Australia–Apples* were: *EC–Hormones; Australia–Salmon; Japan – Measures Affecting Agricultural Products*, WT/DS76/AB/R, adopted 19 March 1999, DSR 1999:I, 277; *Japan – Measures Affecting the Importation of Apples*, WT/DS245/AB/R, adopted 10 December 2003, DSR 2003:IX, 4391; *EC – Measures Affecting the Approval and Marketing of Biotech Products*, WT/DS291/R, WT/DS292/R, WT/DS293/R, Add.1 to Add.9, and Corr.1, adopted 21 November 2006, DSR 2006:III-VIII, 847; and *US – Continued Suspension of Obligations in the EC – Hormones Dispute*, WT/DS320/AB/R, adopted 14 November 2008, DSR 2008:X, 3507 and *Canada – Continued Suspension of Obligations in the EC – Hormones Dispute*, WT/DS321/AB/R, adopted 14 November 2008.

76 Bohanes and Lockhart (2009: 379). From a methodological point of view, one has to distinguish between the SoR applied to factual and legal determinations (see, e.g., the difference between Articles 17.6 (i) and 17.6(ii) of the *Antidumping Agreement*). In this article, we are only concerned in the first type of SoR, namely *factual* determinations by a national authority.

77 Or in the words of the AB: 'the standard of review . . . reflect[s] the balance established [in WTO law] between the jurisdictional competences conceded by the Members to the WTO and the jurisdictional competences retained by the Members for themselves'. AB Report, *EC–Hormones*, para. 115.

78 AB Report, *EC–Hormones*, para. 116. Article 11 of the *DSU* reads in pertinent part:

> [A] panel should make an objective assessment of the matter before it, including an objective assessment of the facts of the case and the applicability of and conformity with the relevant covered agreements . . .

Having determined Article 11 *DSU* applicable, the AB characterized the SoR as 'neither a *de novo* review as such, nor "total deference", but rather the "objective assessment of the facts"'.[79]

This overly broad definition has left Panels with nothing much but an insipid middle ground between the two extremes of *de novo* review and total deference.[80] As a result, Panels and the AB have adopted strikingly different types of review over the years. In fact, the cases leading up to *Australia–Apples* arguably showed distinctive seesaw swings in the SoR between considerable deference and substantial intrusion by reviewing WTO bodies.[81]

In the very first SPS case, *EC–Hormones*, the AB afforded a considerable degree of deference to WTO Members. Although it rejected a totally deferential SoR, such as that of Article 17.6(i) of the *Antidumping Agreement*, the AB recognized that Members were entitled to base their SPS measures on minority scientific opinions.[82] The AB also declared that Members are free to choose and adopt their own ALOP, and that this choice is not subject to any review. Moreover, the AB held that the connection between scientific evidence, risk assessment, and the SPS measure is merely a *reasonable* relationship which may be influenced by the character of risk.[83]

While *Hormones* may be seen to have granted considerable deference to national authorities, the subsequent case law seems to have swung towards the other extreme, at least when it comes to the assessment of scientific data provided by the respondent Member to justify its SPS measures. In *Japan–Agricultural Products II*, the AB interpreted the term 'sufficiency' (used in Article 2.2 in the context 'maintained without sufficient scientific evidence') to require a 'rational or objective relationship' between the SPS measure and the scientific evidence.[84] In *Japan–Apples*, the Panel arguably substituted factual determinations of national authorities with its own, coming quite close to *de novo* review, when it determined that 'as a matter of fact', and based on 'reliable' evidence, it was unlikely that mature apple fruit would serve as a pathway for the entry, establishment, and spread of fire blight in Japan.[85] The AB upheld these findings and opined that Panels were not obliged to favor the respondent's approach to risk assessment and scientific evidence over the views of independent experts consulted by the Panel.[86]

79 AB Report, *EC–Hormones*, para. 117. Note that the requirement of 'objective assessment of facts' can hardly be seen as a helpful addition, because any assessment of the facts, whether highly deferential or highly intrusive, should be expected to be 'objective' in nature. See Bohanes and Lockhart (2009: 389).

80 Under a *de novo* SoR, the reviewing body is able to evaluate all determinations of the national authorities, and substitute them with its own judgment. A fully deferential standard restricts the reviewing powers to checks of procedural compliance, without interfering into the substantive determinations made by national authorities.

81 Gruszczynski (2010a: Section 3).

82 AB Report, *EC–Hormones*, para. 194.

83 Ibid.

84 AB Report, *Japan–Agricultural Products II*, paras. 73–85.

85 Panel Report, *Japan–Apples*, para. 8.176. See also Prevost (2005: 11), Gruszczynski (2010a: 9).

86 AB Report, *Japan–Apples*, para. 165.

The compliance Panel in *Japan–Apples* (Article 21.5) arguably came close to a *de novo* review, when it rejected studies conducted by Japan, the subsequent modified risk assessment, and ultimately Japan's revised SPS measure. The Panel ruled that Japan's new measure was not based on 'sufficient scientific evidence', since the scientific studies reflected a result that had little practical basis under natural – as opposed to laboratory – conditions.[87] While the scientific value of the studies relied on by Japan may have been *unscientific* in nature, the Panel failed to draw an ascertainable line between a 'valid minority opinion' and a 'flawed scientific study' that lacks affirmation by scientific evidence.[88]

The Panel Reports in *EC–Biotech* and *US/Canada–Continued Suspension* maintained the trend of Panels assessing the quality, persuasive force, and correctness of the national scientific determinations. The Panel in *EC–Biotech* conducted a detailed survey of scientific evidence put before it, and chose between competing scientific claims articulated by its appointed experts, preferring some opinions over others.[89] The Panel in *US/Canada–Continued Suspension*, in its review of factual evidence, opined that Panels enjoyed a wide discretion when making factual findings. The Panel stated that it was free to value some expert opinions over others, choosing those opinions that were 'most specific in relation to the question at issue and ... best supported by arguments and evidence'.[90]

The *Biotech* and *Continued Suspension* Panels may have taken their discretion too far for the AB's taste. The AB report in the *Continued Suspension* case thus may be seen as a response by the AB to a too dramatic swing of the pendulum in the direction of intrusion into respondents' policymaking. In its report, the AB overturned the Panel on many important points. First and foremost, the AB recalled that it is the task of the particular WTO Member to *perform* the risk assessment while a Panel is only entitled to *review* it. Consequently, a Panel is only entitled to review the risk assessment, not to substitute its own scientific judgment for that of the risk assessor.[91] According to the AB, it is not for Panels to determine whether a risk assessment is correct but only 'whether that risk assessment is supported by coherent reasoning and respectable scientific evidence and is, in this sense,

87 Panel Report, *Japan – Measures Affecting the Importation of Apples – Recourse to Article 21.5 of the DSU by the US*, WT/DS245/RW, adopted 20 July 2005, DSR 2005:XVI, 7911, paras. 8.45–8.72.

88 See Mercurio and Shao (2010: 215), Gruszczynski (2010b: Chapter 4).

89 See Gruszczynski (2010b: 141 and 191–197) for a detailed discussion.

90 Panel Report, *US/Canada–Continued Suspension*, para. 7.420. The AB summarized the Panel's view as follows:

> [T]he panel seems to have conducted a survey of the advice presented by the scientific experts and based its decisions on whether the majority of the experts, or the opinion that was most thoroughly reasoned or specific to the question at issue, agreed with the conclusion drawn in the EC's risk assessment [rather than a discussion of the evidence relied upon in the European Communities' risk assessment].

See AB Report, *US/Canada–Continued Suspension*, para. 598.

91 See ibid., para. 590.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

objectively justifiable'.[92] More specifically, at paragraph 591 of its report, the AB developed a three-step test for Panels to assess respondents' risk assessments.

1 A Panel must identify the *scientific basis* underlying an SPS measure and verify that the scientific basis comes from a respected and qualified source. Although the scientific basis need not represent the majority view within the relevant scientific community, its views must nevertheless be considered *legitimate science* according to the standards of the relevant community.
2 A Panel then should assess whether the *reasoning* articulated on the basis of the scientific evidence is *objective and coherent*. In other words, a Panel should review whether the particular conclusions drawn by the Member assessing the risk find sufficient support in the scientific evidence relied upon.
3 Finally, a Panel should determine whether the results of the risk assessment *sufficiently warrant* the challenged SPS measures.[93]

Regarding the involvement of Panel-appointed experts, the AB in *Continued Suspension* ensured that Panels first had to consider and respond to the risk assessment (and the underlying scientific evidence) presented by the responding Member before consulting outside experts.[94] Panel-appointed experts should be consulted to assess whether the scientific basis relied upon in the risk assessment came from a respected and qualified source and whether the reasoning articulated by the respondent was objective and coherent. The AB warned that the consultations with outside experts could not seek to test whether they would have performed the risk assessment in a different way as the WTO Member, but only to review such assessment as to its scientific value.[95]

The approach taken by the AB in *Continued Suspension* was hailed by some commentators as reversion towards a less intrusive SoR, since the AB explicitly prohibited Panels from reviewing the correctness of evidence. Instead, it instructed Panels to concentrate on methodological issues in order to assess the epistemic value and coherence of scientific findings. Some authors interpreted the AB Report as a move to a fully deferential approach under which the main focus would be on the *process* of the risk assessment, rather than a *substantive* analysis of outcomes.[96] Other commentators did not go quite as far, but argued that this standard leaves WTO Members with a greater degree of discretion as to how to assess scientific data and what kind of inferences to draw on that basis.[97] Some authors remained

92 See ibid.

93 See ibid., para. 591.

94 Ibid., para. 598. This is in contrast to Panels' approaches in *Japan–Agricultural Products II, Japan–Apples*, and *US/Canada–Continued Suspension*, where Panels first identified 'best science' and then asked whether the measures at issue conformed to this. See footnote 90 above.

95 Ibid., para. 592.

96 Robert Howse on the International Economic Law and Policy Blog (cited in Gruszczynski, 2010a: 13).

97 Gruszczynski (2010a: 12–14).

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

skeptical whether the AB really resolved all the uncertainties as to the scale of deference to be granted to Members' risk assessments.[98]

4. The standard of review in the presence of factual uncertainty

This section deals with issues arising from factual uncertainty, or situations where complete scientific evidence is unavailable. We are interested in the proper SoR and the role that Panel-appointed scientific experts (should) play in this process. To that end, we assess how factual uncertainty over probabilities of occurrence and/or economic, social, and biological consequences can, or should, affect the review of Members' risk assessments by WTO Panels and the AB. In particular, we raise and discuss two key questions:

- From a perspective of economic theory, should the SoR for risk assessments under Articles 5.1, 5.2, and 2.2 of the *SPS Agreement* be more deferential towards the enacting Member in situations that feature a relatively high degree of factual uncertainty? Which elements in a Member's risk assessment (if any) should be considered 'off limits' for review by Panels or the Appellate Body?
- From a perspective of economic theory, how do higher degrees of factual uncertainty affect a Panel's ability to assess less-trade-restrictive alternatives (LTRA) under Article 5.6 of the *SPS Agreement*, and what role should Panel-appointed scientific experts play in the SoR of this provision?

To address these questions, we proceed as follows. At the outset, we provide some theoretical background and introduce a basic framework for decisionmaking in situations of increased factual uncertainty, that is situations where scientific evidence is scant and 'knowledge gaps' abound. This is followed by a discussion of important practical and theoretical problems that risk assessors face when dealing with high levels of factual uncertainty. Depending on the nature and degree of the uncertainty, overcoming these obstacles not only requires technical skill, but also oftentimes necessitates subjective trade-offs and value judgments on the part of the risk assessor. Reflecting on the necessary subjectivity involved in conducting risk assessments in the presence of scientific uncertainty, we formulate a series of recommendations as to the proper SoR in such situations of knowledge gaps. We then discuss these theoretical findings in the context of the *Australia–Apples* AB Report and propose a modified SoR that better deals with Members' risk assessments conducted in situations of low scientific certainty.

4.1 Risk appraisal in the presence of 'incertitude' – a framework of analysis

The term 'risk' features prominently in the *SPS Agreement*, in the Panel Report, and in the AB Report. It thus seems worthwhile taking a closer look at the concepts of

98 Mercurio and Shao (2010: 218).

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

risk, risk assessment, and risk management. A useful point of departure is to see what the disciplines of economics, decision analysis, and risk management have contributed in this area.

The concept of risk in the broad sense is conventionally regarded to be comprised of two basic elements: (i) the *probability of occurrence* of a contingency, and (ii) the *consequences* this entails.[99] A well-established definition of 'risk' is a 'condition under which it is possible both to define a comprehensive set of all possible outcomes (or effects), and to resolve a discrete set of probabilities (or a [probability] density function) across this array of outcomes'.[100]

It is important to appreciate the limits of the concept of 'risk' (in the narrow sense), because neither the probability of occurrence of a contingency,[101] nor its consequences can always be predicted with complete certainty. Factoring in the existence and presence of factual *un*certainty[102] requires us to include other types of risk (in the broad sense), or 'incertitude', as we prefer to call it.[103] To that end, we introduce a simple framework for structuring the different types of 'incertitude'.

4.1.1 *A framework for analyzing different concepts of 'incertitude'*

As mentioned, risk in the general sense is correctly defined along the dimensions of (i) *consequences* of a contingency, and (ii) of *probability* (or *likelihood*) *of occurrence* of that contingency. The dimension *consequences* is better understood as a

99 Indeed, this corresponds to the definition of 'risk' in Annex A of the *SPS Agreement* ('*likelihood* of entry, establishment or spread of a pest or disease within the territory of an importing Member according to the sanitary or phytosanitary measures which might be applied, and of the associated potential biological and economic *consequences*'; emphasis added).

100 ESTO (1999: 16). Mathematically, to yield an expression of the total risk, the probability of occurrence of a contingency is multiplied with the sum over the set of consequences (if measured along a unique metric). This enables the risk assessor to compare different contingencies or events.

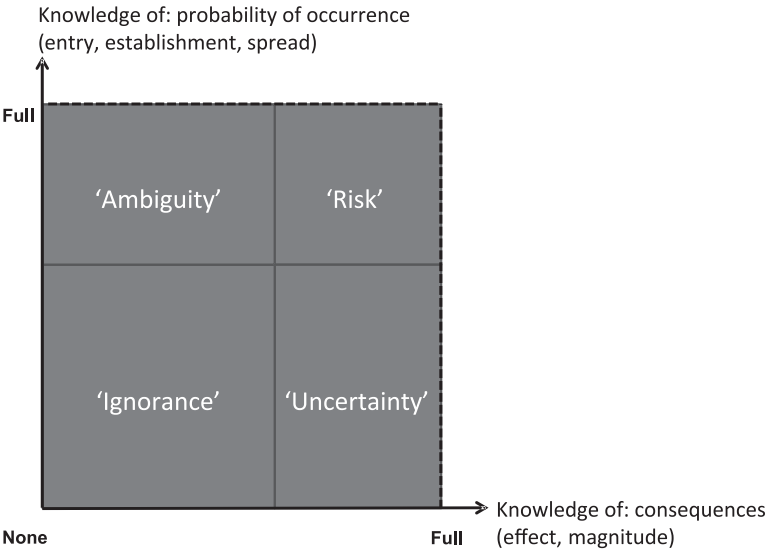
101 A contingency here is understood as either a consequence of human activity, or a state of nature (volcanic eruption, earthquake, etc.).

102 We acknowledge that (un)certainty is not a uniform concept and can be interpreted broadly to include concepts such as disagreement of views, incomplete knowledge, contradictory information, conceptual imprecision, divergent frames of reference, and indeterminacy of data. See Klinke and Renn (2002). The term 'uncertainty' can also connote what the AB has termed 'theoretical' uncertainty (the uncertainty that is 'inherent in the scientific method and which stems from the intrinsic limits of experiments, methodologies, or instruments deployed by scientists to describe a given phenomenon' – see AB Report, *Japan–Apples*, para. 241). In this paper, however, we are only interested in uncertainty in an objective sense. Hence, we limit the notion of factual certainty and uncertainty to indicate the level of *scientific confidence* or *ability* to make reliable predictions over the occurrence of certain outcomes. This includes situations of insufficient knowledge, observation, and data; indeterminacy and variability of scientific observations and data; practical immeasurability; and issues of systematic and random measurement errors. We believe that our definition of factual (un)certainty is largely congruent with the AB's interpretation of '(in)sufficiency of scientific evidence', spelled out in paragraphs 2.2 and 5.7 of the *SPS Agreement*. See Gruszczynski (2010b: 187–191) for a discussion on the jurisprudence.

103 The term was coined by Stirling (2001: 52).

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

Figure 2. Concepts of 'incertitude' in risk management



Source: Based on ESTO (1999: Figure 4).

function of the *nature* (form or type) of the outcome/effect,[104] and its *magnitude* (severity or impact). Following ESTO (1999), Figure 2 above spans a matrix over these two dimensions, and so allows for variable levels of *factual* certainty.

On the *x*-axis (the horizontal dimension), we plot knowledge of, or certainty about, *consequences*. The *x*-axis informs how well a comprehensive set of consequences can be circumscribed, that is how well the *nature* of outcomes, or effects, can be predicted, how precisely *causal mechanisms* can be defined, and how accurately *magnitudes*, or impacts, can be predicted. On the *y*-axis (vertical dimension), we depict knowledge of *probability of occurrence*, which in the SPS context usually means knowledge of probability of entry, establishment, and spread of pests, diseases, toxins, food stuffs, or contaminants.[105] This dimension effectively asks whether the likelihood with which a contingency is occurring can be characterized. The axes of Figure 2 have two conceptual endpoints, namely 'None' (no knowledge/total uncertainty) and 'Full' (full knowledge/complete certainty). In between these extremes lie various gradations of knowledge or certainty. We distinguish four archetypes of 'incertitude':

- In a situation where there exist credible grounds for the assignment of a discrete probability to each possible outcome within a well-defined set of outcomes, a risk

104 As a convention, we will use the terms 'outcomes', 'effects', 'hazards', or 'damage' interchangeably.
105 See Annex A(1) of the *SPS Agreement*.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

assessor faces the paradigm condition of '*risk*' (in the narrow sense).[106] Where it is possible to define in quantitative terms both a comprehensive set of all possible outcomes and the probabilities across this array of outcomes, unrestricted risk can easily be assessed, and various risk-management alternatives be compared.[107]

- '*Uncertainty*' (in the strict sense of the term) applies to situations, where there is confidence in the completeness of the defined set of outcomes, but where there is no or inexact theoretical or empirical basis for the assignment of probabilities to the individual outcomes.[108]
- In circumstances where there not only exists no basis for assigning the probabilities (as under 'uncertainty'), but where the definition of a complete set of outcomes, magnitudes, or causal mechanisms is also problematic, a condition termed '*ignorance*' prevails. In a state of 'ignorance', it is always possible that outcomes occur, whose magnitudes are unknown, whose causal links to the contingency are incompletely understood, or whose general existence has never been anticipated in the first place.[109]
- The final category concerns situations in which the definition of a comprehensive set of outcomes (including magnitude and causal mechanisms) is identifiable, but not calculable, yet where there is good indication of probabilities of occurrence. This category shall be termed '*ambiguity*'.[110]

In sum, while 'risk', in the technical sense, deals with what may be termed 'known knowns', there are other important concepts of 'incertitude' in the real world, namely 'uncertainty' and 'ambiguity' ('known unknowns'), and 'ignorance' ('unknown unknowns'). As will be shown below, for the risk assessor charged with evaluating unrestricted risk and comparing different risk-management alternatives,

106 In order to avoid confusion between the strict definition of the terms used in risk-management theory and the looser, colloquial usage, we utilize quotation marks, when talking of theoretical concepts, such as 'risk' or 'uncertainty'. The terms *risk assessment, risk management, risk appraisal*, and *risk assessor*, thus, apply to *all* situations of 'incertitude'.

107 See footnote 100 above. An example from the realm of SPS may help to illustrate the concept of 'risk': Imagine the risk at issue stems from lead in water pipes. Here, both adverse effects to human health, and the risk of occurrence of such morbidity effects are well known and documented, because lead poisoning has been studied for decades.

108 Examples for situations of 'uncertainty' in the real world are floods, earthquakes, or volcanic eruptions. In the realm of SPS, let us take the importation of avian-flu-infected poultry from some country *X* into country *Y*. In general, the health, economic, and social consequences of avian flu may be well known, but probabilities of importation, entry, transfer, and spread of avian flu from *X* into *Y* may be yet unknown, or at least highly uncertain.

109 Cloning of animals could be seen as an example for 'ignorance'. Given the short span and small number of instances in which cloning of animals has occurred, the scope of long-term effects on human and animal life or health, the causal pathways, and the severity of effects are very difficult to predict. The same goes for the probabilities of occurrence of these uncertain effects, which by default are highly speculative.

110 As an example, take the introduction of a new artificial sweetener. Suppose that after limited clinical studies, there is *some* knowledge of possible consequences for human life and probabilities, but no conclusive evidence yet as to the severity of effects in terms of morbidity rates.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

the presence of factual uncertainty in the two dimensions, occurrence and consequences, poses a number of significant practical and theoretical problems.

4.1.2 *Theoretical and practical problems of risk appraisal in the presence of 'incertitude'*

Whether they want it or not, risk assessors charged with examining sanitary and phytosanitary risk must find ways of dealing with all types of 'incertitude'. Below, we will describe theoretical and practical problems faced by risk assessors and describe what aspects of uncertainty they can and cannot address.

(i) Theoretical and practical problems *of risk appraisal in situations of* complete *knowledge (situations of 'risk').* We start with problems that occur in situations of complete (or sufficient) knowledge about outcomes and probabilities – the paradigm referred to above as 'risk'. Even under such circumstances of relative factual certainty, the risk assessor has to struggle with two fundamental problems, namely (i) multidimensionality and (ii) incommensurability of the manifestations of different types of risk. These two problems lead to what has been termed 'Arrovian uncertainty',[111] the impossibility to definitively aggregate preferences and to rank alternatives in a pluralistic society.[112]

The problem of *multidimensionality* becomes evident once it is acknowledged that there are different classes of outcomes, or risk factors, to be considered in situations of risk.[113] And within each broad class of effects, there are usually a multitude of relevant subgroups of hazards.[114] Moreover, effects can be assessed from the vantage point of various stakeholder groups – that of consumers, producers, using industries, pregnant women, the health-impaired, etc. – and at different levels of aggregation – local, regional, national, international.

Another aspect of multidimensionality relates to the different dimensions in which the *magnitude* of different classes/types of effects can be measured. The size of the impact of any one risk factor can be appraised in dimensions such as severity (e.g., number of deaths or injuries), immediacy (immediate or long-term effects), gravity (low probabilities of large impacts v. high probability of low impacts), reversibility (whether associated effects are reversible), controllability (whether effects can be controlled by affected parties), balance of burden (distributed v. centralized burden). These different dimensions along which magnitude can be

111 Named after economist and Nobel Laureate Kenneth Arrow. See, e.g., Rosenberg (1996: 340).

112 See Arrow (1974).

113 Take the example of genetically modified organisms ('GMO'): should one focus on effects on agriculture, or should other classes of effects, such as health, environmental hazards, biology, economy, or society be included?

114 Staying with the GMO example mentioned in the previous footnote, the broad class of 'health risks' may further be subdivided into allergenicity, toxicity, carcinogenicity, occupational safety, digestibility, etc.

measured cross-cut the different classes and types of individual effects, adding significant complexity to the appraisal of 'incertitude'.

Notwithstanding the fact that, ideally, the set of outcomes should be as complete and comprehensive as possible, any risk assessor faced with a staggering multitude of risk factors, viewpoints, and dimensions naturally must make the decision as to which elements can reasonably be included, in order to keep the assessment manageable and controllable. In the real world, constraints on budgets, staff, information, computational resources, and response time, as well as a need for tractability make a comprehensive set of effects virtually impossible. Thus, it is ultimately a subjective decision of the risk assessor which classes and dimension of risk to include, and from whose viewpoint.

However, even if it were practically feasible to consider a complete or comprehensive set of outcomes, this still would pose another fundamental issue to the risk assessor, namely that of framing and prioritizing different (classes of) effects. For one, different risk classes have different metrics: whereas human health effects can be measured in terms of mortality, rate of infection, sick time, or other measures of human morbidity, economic effects are mostly measured in monetary terms. In addition, many classes of impact cannot even be measured in quantitative (cardinal) terms in the first place; they are irreducibly qualitative (nominal, or, at best, ordinal) in nature.[115]

But even if some effort at quantification of qualitative effects is felt possible, the resulting values may nevertheless still be *incommensurable* in the sense that they cannot be reduced to a single measure of performance. Whenever types of effects cannot be converted into a single metric, it is practically impossible to objectively and definitively compare outcomes, because the relative priority attached to different types of effects is an inherently subjective value judgment.[116] Hence, for practical reasons, different outcomes are incommensurable if they cannot readily be collapsed into a single cardinal metric.

There is, however, a more fundamental problem exacerbating the assessment of risk. In a pluralistic society, it is also *theoretically impossible* to aggregate individual preferences in a rationally consistent fashion. As Kenneth Arrow demonstrated more than 40 years ago, there is no effective way to compare (i.e., aggregate in a single metric) utility across individuals or different groups in a society, even where social choices are addressed simply in ordinal (relative)

115 For example, how to measure social cohesion, aesthetic appreciation of the landscape, or attachment to the existence of unspoiled wilderness.

116 Risk assessors are regularly facing trade-offs, such as: What relative priority should be attached to different risk factors, such as allergenicity, lost revenue, biodiversity, and ecological integrity, when these are measured in different metrics? What relative weight should properly be placed on impacts to different groups such as workers, children, breastfeeding mothers, future generations, farmers, or senior citizens? If one risk-management alternative is expected to result in three deaths, five injuries, and USD 300,000 in material damage, who is to say with acceptable certainty that this alternative is better than one that causes two deaths, 20 injuries, and USD 5,000 in material damage?

terms.[117] As a result, whenever a collapse of outcomes into a single metric is not feasible, it is theoretically impossible to definitively aggregate preferences.

Even if it were possible to reduce all classes of effects in connection with a certain risk to a single metric (say, in monetary terms), it is still questionable whether members of a society could agree on the relative weight and ranking of different risk dimensions in which magnitude is measured, such as gravity, severity, immediacy, or fairness.[118]

In sum, the problems of multidimensionality and incommensurability unavoidably give rise to Arrovian uncertainty, making a definite aggregation of preferences difficult, even impossible at times. Different individuals quite reasonably draw fundamentally different conclusions about the sanitary and phytosanitary risk of different risk-management options, because they have different – yet equally legitimate – perspectives. There is no single 'objectively correct' outcome of a risk analysis.[119]

The practical implications of multidimensionality and incommensurability are significant degrees of variability in appraisal outcomes;[120] and ambiguity in ranking of alternatives.[121] According to ESTO (1999: 15), the manifest variability in results and the ambiguity in rankings expose a striking disjuncture between the *precision* with which single risk assessments report their results, and their *accuracy*. However, the fact that different studies often report vastly different appraisal outcomes need not be driven by any single factor, nor the degree of diligence with

117 This is one version of Arrow's 'impossibility theorem'. See Arrow (1974). The basic insight of Arrow's work is that there can be no single uniquely 'rational' way to resolve contradictory perspectives or conflicts of interest in the regulation of risk in a pluralistic society. There is *no analytical fix* to the problems of multidimensionality and incommensurability encountered in the social appraisal of consequences.

118 Even if all risk classes could be measured in monetary terms, should the *gravity* of outcomes be considered more or less important than their *severity*? How should the dimensions of *fairness* in terms of spatial distribution of burden be ranked relative to *intergenerational equity*? As per Arrow's impossibility theorem, it is highly unlikely that members of a society could reach a single unified stance on these issues.

119 It is important to emphasize that 'Arrovian uncertainty' is not the result of the information available, the scientific quality of the risk assessment, or the evidence relied upon by the risk assessor. It also is irrelevant how much consultation and deliberation went into the process beforehand.

120 Stirling (2001) impressively illustrates the volatility of results that may occur even in situations of 'risk', featuring complete (or sufficient) certainty about consequences and probabilities. He surveyed a large number of top-quality studies of risk and environmental impacts associated with modern coal-fired power-generating technologies. Although the individual studies he reviewed often reported their results with utmost precision, the picture changed when he compared the results across studies. The results of individual studies varied by up to a factor of 50,000. Measured in 1995 USD/kWh, the lowest estimate for external environmental cost of new coal power was 1/25th of a cent per kilowatt hour, while the highest assessment approached 20 USD/kWh. See Stirling (2001: Section 5.2).

121 Assessing the ambiguity in ranking of risk alternatives, Stirling (2001: Section 5.2), compared different risk studies on the environmental costs of different electricity options, including onshore wind, photovoltaics, biomass, hydroelectricity, nuclear fission, and fossil fuels (gas, oil, coal). Looking at the risk-management literature as a whole, and comparing dozens of top-quality studies, Stirling showed that virtually *any* conceivable ranking for these eight options was possible due to the sizable variability in results. See Stirling (2001: Figure 5), which shows that the estimated environmental costs for all eight alternatives overlapped, thus allowing any possible ranking (at least for the overlap).

which risk analyses were conducted. Rather, differences in results across studies often are an outflow of different methodological set-ups, that is the inclusion of different classes of effects, of the adoption of different – but equally scientifically valid – assumptions, and of the ranking given to the multitude of different dimensions of risk.

(ii) Ways of dealing with Arrovian uncertainty. What lessons can be learned for the risk assessor who is tasked with drafting a risk analysis and who appreciates that Arrovian uncertainty is genuine to any risk assessment, no matter how reliable available scientific evidence may be?

First, the above discussion would counsel against a purely quantitative, unidimensional, and overly analytical approach to risk assessment. Exclusively probabilistic characterizations of risk often ignore many crucial respects, because they cannot capture the multifaceted nature of risk.[122]

Second, it seems advisable not to sacrifice effects that are *qualitative* in nature, or that are difficult to compare, for those that are easily measurable, and *vice versa*.[123]

Third, risk assessors should avoid conflation of incommensurable properties in appraisal. Using multiple metrics can allow risk assessors to appraise each individual aspect of performance using whatever seems the most appropriate yardstick.

Fourth, even those risk assessments that are ultimately purely quantitative in nature should be complemented with sufficient qualitative explanation. The findings of the risk assessment should be accompanied by a systematic exploration of the choices and trade-offs taken, exposing relationships between different assumptions, and the associated relative importance of different options.[124]

(iii) Theoretical and practical problems *of risk appraisal in situations of* incomplete *knowledge of outcomes and probabilities of occurrence (situations of 'uncertainty', 'ambiguity', and 'ignorance')*. The challenges faced by a risk assessor struggling with problems in connection with Arrovian uncertainty seem formidable. However, things may get even more difficult once we factor in 'Knightian uncertainty', that is situations where knowledge of outcomes and probabilities is increasingly limited.[125]

122 See ESTO (1999: 12), Stirling (2001: 73–74). In a probabilistic risk assessment, outcomes are multiplied with probabilities of occurrence to characterize the size of the 'risk'. See footnote 100 above.

123 See ESTO (1999: 15): 'It is better to be *roughly accurate* in [the] task of mapping the social and methodological context-dependencies, than it is to be *precisely wrong* in spurious aspirations to a one-dimensional quantitative expression of . . . risk.'

124 This includes detailed explanation as to the choice of classes of effects that are included in the risk assessment, of the metric they are measured in (or converted into), of the relative weighing and ranking of classes of effects, and a description of causal chains. The same goes for the relative importance of the dimensions in which the magnitude of effects was measured.

125 See Knight (1921), Luce and Raiffa (1957).

Situations of 'uncertainty'. Let us first relax the assumption of complete (or sufficient) knowledge of *probabilities of occurrence*, leading to a situation that we termed *'uncertainty'* above. High degrees of certainty about probabilities of occurrence are commonly found in self-contained formal rule-based systems (such as games of chance), highly repetitive situations affecting a multitude of subjects in long-term stable systems (such as life insurance), or generally in situations where the past is a good predictor for future events.[126]

In contrast, knowledge of probabilities is imperfect whenever there are no similar past events under comparable circumstances to draw from situations, that is when situations are incompletely understood, open-ended, and dynamic. This is typically the case:

- when causality chains are long and complex;[127]
- when the causal mechanism between a contingency and the occurrence of a hazard is incompletely understood;[128] or
- when the probability of harmful consequences is dependent on external conditions and circumstances.[129]

The presence of a situation characterized by 'uncertainty' has two immediate implications. *First*, situations of 'uncertainty' cannot be tackled with standard probabilistic approaches. Whenever the risk assessor is unable to determine a valid theoretical or empirical basis for the assignment of probabilities, the remaining analytical toolbox is less well equipped, forcing the assessor to shift from purely *probabilistic* approaches to those that actually embrace uncertainty, such as Bayesian updating or sensitivity analysis.[130] These analytical tools are not less 'scientific', systematic, or rigorous; they are more robust in addressing the

126 See, e.g., Bernstein (1996), Salanié (1997: Chapter 7).

127 Australia's risk assessment in *Australia–Apples* is a case in point (see Section 2.2 above). The causal chain that the IRA defined in order to come up with the overall probability of entry, establishment, and spread of fire blight was rather complex in nature. Although the IRA, with dubitable success, took an attempt at calculating this probability, other risk assessors might have concluded that the causality chains cannot be expressed in quantitative terms.

128 For instance, the mechanism, by which low-toxicity herbicides used in GMO farming may disrupt reproduction in non-target crops, is still not yet fully understood. See, e.g., Krimsky (1997). Consequently, it is difficult to assign a probability to this event.

129 For example, where the occurrence of a hazard is dependent on the strict adherence to good practice by those exposed to risk, a conclusive characterization of likelihoods is difficult. See Appellate Body Report, *Japan–Apples*, paras. 138–141 (where the AB found that errors of handling or illegal actions by those involved in the import transaction were risks that could, in principle, legitimately be considered in the scope of the risk analysis).

130 Bayesian updating is an extension of the probabilistic paradigm. The risk assessor expresses the relative likelihood of eventualities, given the best available information, which may be updated in light of new information that transpires. See Gibbons (1992: Chapters 3, 4). In sensitivity analyses, point values are replaced by probability *intervals* and probability density functions. Sensitivity analyses can be used to vary one parameter at a time within a predefined range of values, or to explore all permutations of parameter variations. The presentation of appraisal results is given in the form of systematic 'maps', rather than discrete values. See Stirling (2001: 78).

underlying uncertainty. However, this robustness comes at the cost of less precision in formulating results.

Second, whereas different risk-mitigation options under appraisal may still be broadly characterized,[131] it is difficult to rank them. An ordering of risk-mitigation alternatives, even in relative terms, is impossible without some knowledge of the relative likelihoods at which the different effects under consideration occur.

Situations of 'ambiguity'. In situations of *'ambiguity'*, the set of outcomes cannot be conclusively defined, although there is (sufficient) knowledge concerning the probabilities of occurrence. This is typically the case in situations where the scope of possible *effects* is known, but their *magnitude* is uncertain or unknown. As with 'uncertainty', the analytical toolkit of *probabilistic* risk assessment is not effective in situations of 'ambivalence'. The risk assessor, however, may apply fuzzy logic[132] or scenario analysis.[133] As with 'uncertainty', a ranking of outcomes is a challenging task in situations of 'ambiguity'. Any attempt at ranking options requires some knowledge of the relative magnitude of the different effects under consideration. Where such ranking of magnitudes is not possible, a ranking of options cannot be completed, either.

Situations of 'ignorance'. A state in which there exist neither grounds for the assignment of probabilities, nor a solid basis for the definition of a comprehensive set of outcomes is denoted as *'ignorance'*. It may occur in the form of 'surprise contingencies', that is in the form of effects whose very existence were previously *unforeseen*.[134] The analytical toolkit for risk appraisal in situations of unforeseen contingencies is virtually empty: no risk assessor can qualitatively or quantitatively assess hazards whose existence it previously was not privy to. This means that

131 Note that in situations of 'risk', full characterizations and orderings of different options – in theory – are possible. The result, however, is subject to the condition that all effects can be compressed into one single common metric.

132 Fuzzy logic is an appraisal technique that can be applied where magnitudes of effects are unknown, but can be expressed in terms of ranges and density functions. It can also be used where classes of outcomes are conceived in terms of set theory. In such cases, members of outcome sets ('probabilities of eventuation') can be treated like probabilities of occurrence and expressed as numerical weightings normalized to sum unity. See Stirling (2001: 56).

133 Scenario analysis is a systematic examination of different possible outcomes and contingencies which bear on a decision. It can take a quantitative or qualitative format as a flexible tool for exploring alternative risk-mitigation options. See ESTO (1999: 29).

134 In the case of *unforeseeable* contingencies, knowledge of outcomes is by definition zero, which necessarily implies complete ignorance of probabilities of occurrence. There are numerous real-world examples of unforeseen outcomes. The failure to anticipate the potential depletion of the ozone layer by halogenated hydrocarbons was not a matter of neglecting low-probability 'risk' or a mis-assignment of probabilities ('uncertainty'), but rather an utter failure to identify the very possibility of this outcome, and thus a manifestation of 'ignorance'.

different risk-mitigation strategies cannot be meaningfully compared. All that risk assessors can do is to resort to measures of precaution.[135]

Situations of 'ignorance', however, may also arise in circumstances where outcomes *are* identifiable, but inexplicable. Here, all effect classes are correctly predicted (there are no 'surprise' events), but there is no precise knowledge that these classes are consequences of the occurrence in question. There is uncertainty over the precise *causal mechanism* and thus the magnitude of such effects and their probability of occurrence.[136] Equally, where a risk assessor decides to take into consideration relevant additive, cumulative, and synergistic relationships between different effects, precise statements of magnitudes and likelihoods of occurrence are unlikely, again resulting in a state of 'ignorance'.[137] Finally, 'ignorance' prevails whenever a risk assessor attempts to evaluate a variety of incommensurable effects and each type of effects defines a scale of possible states. In such instances, it is daunting, if not impossible, to characterize outcomes and likelihoods with any certainty.[138]

Many sophisticated techniques of risk assessment fail – both in principle and in practice – to successfully tackle the condition of 'ignorance'. To address situations of 'ignorance', risk assessors occasionally use matrix analyses and multicriteria techniques, although qualitative approaches oftentimes seem more apt to deal with the inherent uncertainties. As a matter of theory, a ranking of alternative risk-management options in situations of 'ignorance' is not feasible without basic

135 See, e.g., O'Riordan and Cameron (1994) and O'Riordan and Jordan (1995). The *SPS Agreement* does not permit Members to adopt precautionary measures to safeguard against *unforeseeable* contingencies. Article 5.7 specifically requires that 'a Member may provisionally adopt sanitary or phytosanitary measures on the basis of *available pertinent information*…' (emphasis added). This has been interpreted by the AB to mean that 'where there is *some evidentiary basis* indicating the possible existence of a risk, but not enough to permit the performance of a risk assessment … Article 5.7 provides a "temporary safety valve"'. See Appellate Body Report, *US/Canada–Continued Suspension*, para. 678 (emphasis added). Yet, owing to the very nature of unforeseen consequences, a Member cannot base its measure on any evidentiary basis. Thus, the *SPS Agreement* is inapplicable to situations of unforeseen contingencies.

136 Take the potential allergenic property of genetically modified soybeans as an example. Allergenic properties were so far entirely absent from soya products. The incorporation of genetic material from Brazil nuts into soybeans may have changed this. Yet, according to the British Royal Society, the mechanisms of allergenicity are generally poorly understood, thus rendering possible allergenic effect of GM soybeans very difficult to predict. See Royal Society (2002).

137 Predicting the effects arising from interactions of different agents of harm associated with a particular risk (or with a particular risk-management option) is problematic. As argued by Rosenberg (1996: 340), whenever 'uncertainty exists along more than one dimension, and the decisionmaker does not have information about the joint distribution of all random variables, there is little reason to believe that a "rational" decision is possible or that there will be a well-defined "optimal" … strategy'.

138 For example, climate change can be measured solely in terms of temperature change. Measured in one metric, the problem may be one of 'uncertainty'. However, when further classes of effects are considered, such as to ecological, agricultural, economic, and social hazards and costs, the permutations of outcome parameters rises geometrically with the associated possible manifestations, which quickly renders any assessment untraceable.

knowledge of the relative likelihoods and magnitudes of the different effects under consideration.

(iv) Ways of dealing with Knightian uncertainty. What lessons can be drawn from this review of the literature on high degrees of factual uncertainty over outcomes and probabilities of occurrence?

First, regulators and policymakers are far more often confronted with situations of 'ambivalence', 'uncertainty', and 'ignorance' than one might expect. Indeed, in the realm of SPS the imponderables associated with GMO, cloning, endocrine disruptors, or BSE may be argued to render 'ignorance' and 'uncertainty' dominant conditions in the management of sanitary and phytosanitary risk and regulation.[139]

Second, it is worth reemphasizing that situations of Arrovian uncertainty (as explained, uncertainty resulting from the problems of multidimensionality and incommensurability) are independent from, and additional to, those of Knightian uncertainty (uncertainty over outcomes and probabilities). This may require the risk assessor to address both types of uncertainty in the analysis of incertitude and different risk-management options.

Third, risk assessors should acknowledge that 'incertitude' is a much more multifaceted and complex phenomenon than that of 'risk' in the narrow sense. Importantly, risk assessors should refrain from shoving the square peg of a probabilistic risk assessment into the round hole of appraisal of 'incertitude'. Given the manifest inapplicability of standard probabilistic techniques in the context of 'uncertainty', 'ambiguity', and 'ignorance', risk assessors should resist the intuitive appeal of the elegance and facility of probabilistic calculus which may produce precise, but ultimately worthless results.[140] Reductionism does not make the analysis more correct or more scientific. Instead, *fourth*, risk assessors ought to embrace the complexity of the situation at hand by acknowledging, characterizing, and articulating the kind of incertitude they are facing. They should then adapt their appraisal to the complexity of the situation at hand: analytical methods such as scenario analyses, sensitivity analyses, fuzzy logic, multivariate approaches, or qualitative appraisals are no less 'scientific' than utilizing probabilistic instruments, but potentially yield more robust results. Risk assessors also should acknowledge and address the subjectivity and variability ingrained in their choices by giving detailed account of the way they chose, framed, and prioritized their appraisal

139 See Stirling (2001: 58), Renn and Klinke (2001: 18).

140 For example, regulation on cloning of animals is probably inaccurately addressed with a *probabilistic* risk assessment (multiplying outcomes with probabilities of occurrence). Science does not have much experience with this subject. Not only are long-term consequences uncertain (possibly unforeseen). There is also no historical or theoretical experience to draw from for identifying discrete probabilities or probability density functions. Thus, more flexible assessment methodologies and tools should be used. See footnote 109 above.

criteria. Though this places very demanding requirements on the process of appraisal, it ultimately makes the outcomes more tractable and transparent.

Fifth, risk assessors should invest in gathering more and higher-quality scientific evidence to help them overcome uncertainties and to further substantiate their assessments. However, recognizing that there are limits to what can be quantified and compared, risk assessors should also invest in new methodologies and research tools that help them better address inevitable factual uncertainties.

Finally, in some situations, however, factual uncertainty reaches a level, at which the risk assessor can no longer rely on a formal risk assessment.[141] When such a threshold point is reached and different risk-mitigation options no longer can be sensibly compared by means of qualitative or quantitative analysis, the risk assessor may resort to the adoption of precautionary measures.[142]

Figure 3, on the next page, summarizes the different approaches to be considered by risk assessors when appraising different situations of 'incertitude'. Evidently, the borders between the four archetypes are blurry and largely subjective. So are the analytical risk-assessment instruments, applied by each risk assessor:[143] depending on the contextual circumstances, its access to scientific information, level of sophistication, and attitude to risk, a risk assessor may choose what it deems to be the most appropriate risk-appraisal instrument. Alternatively, it may opt to take precautionary measures.

4.1.3 Implications from the theoretical literature on 'incertitude' for reviewing bodies: recommendations regarding the proper SoR

As was shown in the previous subsection, the creation of a risk assessment can be a risky business. Our systematic analysis of these 'risks of

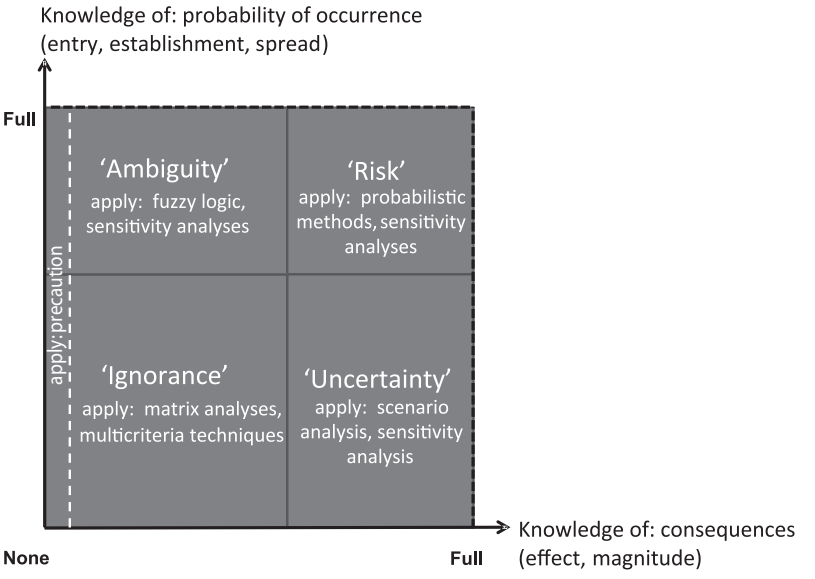
141 Formal risk-appraisal methods necessarily fail in the case of *unforeseeability* of consequences, as discussed above (see footnote 134). The same goes for situations in which there is insufficient scientific evidence available to enable the risk assessor to engage in a structured assessment of the 'incertitude' at issue. Further, a threshold may be reached when calculations for such analysis get so complex such that they exceed computing capacity. The same can be said for instances where the results of the assessment display extreme levels of volatility, that is when small changes in input parameters cause dramatic swings in output.

142 The ultimate decision whether to initiate a formal assessment of incertitude, or to resort to precaution is probably best left to the risk assessor. Not only is such decision highly context-sensitive, it also depends on the risk-aversion of the risk assessor, which may be a function of cultural, societal, political, or historic factors. An important element in the risk assessor's choice may be attitude towards *errors* that may occur during the analysis. A risk assessor may be biased towards *Type-II error* (false negatives), which may result in over-regulation, and against *Type-I errors* (false positives), which may cost human lives.

143 It should be evident by now that a risk assessment is not just a tool to address situations of 'risk' under the 'probabilistic paradigm'. Rather, the term encompasses all structured attempts and analytical tools used in appraising 'incertitude'. The AB certainly has recognized that the definition of risk assessment in Annex A(4) of the *SPS Agreement* (see footnote 73 above) is broad enough to include all sorts of analytical approaches, by explicitly allowing Members to utilize quantitative and qualitative tools. See AB Report, *EC–Hormones*, para. 187.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

Figure 3. Ways of assessing different types of incertitude



risk appraisals' now enables us to formulate a series of recommendations for those external entities charged with reviewing the results of risk appraisals.[144]

Again, we start with a discussion of Arrovian uncertainty, which may even occur in situations of complete knowledge of consequences and probabilities. This uncertainty, and the problems of multidimensionality and incommensurability of outcomes it entails, may lead to variable risk-appraisal results and to ambiguous ranking of risk-mitigation options. As explained, such variability often is an outflow of the inevitable subjectivity in the trade-offs that the risk assessor has to make in order to keep the analysis manageable. And it is this subjectivity that may end up driving the results. For practical and theoretical reasons, there is no one 'right' or 'objectively best' outcome of any risk appraisal. Hence, any reviewing body should exercise caution when assessing the results of a risk assessment. Instead of calling into question those elements of a risk assessment that are necessarily subjective, reviewing bodies should rather call on aspects of the risk assessment that are objectively verifiable.

144 For the purposes of this theoretical discussion, we use 'reviewing body' as a generic term and thus do not distinguish between a Panel and Panel-appointed experts.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

Based on the economic theory reviewed above, an external reviewing body should *refrain from challenging* the risk assessor's choice concerning:

- the basic *methodological set-up* (e.g., whether to conduct a quantitative, semi-quantitative, or a qualitative assessment, or whether to conduct a scenario analysis instead of a purely probabilistic risk assessment);
- which basic *classes and subcategories of effects* it wishes to include in its assessment;
- from whose *vantage point* and at what *level of aggregation* effects should be measured;
- the *level of detail* with which it depicts chains of cause and effect and chains of probability of occurrence;
- the *dimensions* in which to measure the magnitude of risk;
- whether to include *direct or indirect effects*;
- whether to consider *additive, cumulative, synergistic effects;*
- how to *prioritize and weigh* different risk dimensions/factors;
- the *choice of metric* used for measuring certain effects; and
- how the risk assessor *aggregates interpersonal utility* across different individuals and groups of society.

All these basic, and at the same time necessarily subjective, elements should be considered 'off limits' for the reviewer. The reason lies in the fundamental subjectivity inherent in *any* risk appraisal. The risk assessor's choices and trade-offs reflect certain values, cultural and societal traits, a particular attitude to risk, budgetary or computational constraints, and/or historical factors. No external reviewer can rightfully claim that its own values, choices, and trade-offs would be more scientific, objective, or rational, as compared to the risk assessor's. Hence, the risk assessor's basic choices should not be subject to challenge by an uninvolved third party.

This, however, does *not* mean that a reviewing body is obliged to take the risk assessment for granted. Yet, instead of challenging the methodological set-up and the initial choices and trade-offs the risk assessor opted for, the reviewing body should focus on (i) the quality of the execution; (ii) the internal coherence; (iii) the scientific grounding; and (iv) the presentation of the risk assessment.

The reviewer should assess the *execution and implementation* of the risk analysis. The search for calculation errors and other operational glitches falls into this category. It is also appropriate for the reviewing body to examine the *coherence* of the adopted scheme. This includes an evaluation whether there are gaps and overlaps between different classes of effects that the risk assessor recognized for purposes of the analysis, or gaps and overlaps between different probability chains.[145] It also

145 An overlap of effects or probabilities may lead to double-counting of outcomes, and thus to an overestimation of effects.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

seems within the remit of the reviewer to examine whether there are any shifts in vantage points.[146] The reviewer should determine whether outcome scenarios or sensitivity tests are applied in a coherent fashion, or rather selectively, so as to influence outcomes of the appraisal.

Next, the reviewer should assess the *scientific basis* underlying the risk assessment, both with respect to the magnitude of effects included, and to the associated probabilities of occurrence.[147] As part of that task, the reviewer should also consider 'missing factors' and 'hidden variables' in the risk assessor's causality chains that may drive the outcomes. This may be done by consulting the scientific literature on whether the risk assessment at issue omits effects, magnitudes, and partial probabilities that have been described as important. *Unless* the risk assessor explains in detail why a certain factor was excluded from consideration, such omission should be held against the risk assessor.

Whenever the risk assessor attempted to 'compress' different outcome classes into a single metric, for example mortality rates into monetary costs, the reviewing body should scrutinize this process to see whether the conversion accords to scientific conventions.[148] In cases in which inherently *qualitative* (and ordinal) effects are included, a reviewing body may wish to check how the appraisal avoided a disproportionate emphasis on the more *quantifiable* (cardinal) aspects.[149]

Fourth, a reviewing body should also pay close attention to the *presentation* of the risk assessment. As discussed above, the risk assessor's choice of effects and dimensions of risk are subjective by default. It should thus make efforts to explain in detail the rationale behind the basic choices taken, particularly with respect to the dimension of consequences, causality mechanisms, and probability chains assumed. A reviewer should feel free to scrutinize the risk assessment for transparency, tractability, and comprehensibility.

Thus far, we have drawn conclusions concerning external review in situations of Arrovian uncertainty. These conclusions are valid for *every* risk assessment, regardless of the level of factual uncertainty the risk assessor is facing. In a next step, we add the possibility of Knightian uncertainty and formulate additional recommendations for reviewing bodies.

In general, the greater the degree of factual uncertainty over probabilities and/or outcomes, the more difficult it is for a reviewing body to evaluate the 'correctness'

146 Shifts in vantage points (e.g., where the risk assessor measures effects on *consumers* on a *regional* level for alternative 1, and effects to *producers* on the *national* level for alternative 2) may imply 'cherry-picking' on the part of the risk assessor.

147 Recall that we are for the moment assuming a situation of factual certainty, that is of complete knowledge of consequences and likelihoods. Thus, in principle, a scientific basis for every factor, impact score, and probability should be available.

148 Converting deaths or major injuries into monetary terms, for example, is a delicate task and one that should be conducted with extreme care, based on scientific evidence or conventions.

149 As Holdren (1982: 38) states, there is a difference 'between things that are countable and things that count'.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

of the risk assessment. One way of reacting to high levels of uncertainty for the reviewer is to conduct a *de novo* review.[150] Another route is to shift the energies from examining the *substance* of the assessment towards examining its *process*. If the reviewer opts for the latter approach, it should not concentrate on the question whether the risk assessment is *correct*, but whether it is *correctly conducted*.

As indicated, one issue of substance that should be deemed 'off limits' for review is the risk assessor's general decision to take precautionary measures, rather than performing a risk assessment at all.[151] This, however, is not to say that the reasons for this choice, once taken, are not subject to scrutiny by an external reviewer.[152]

When faced with a risk assessment allegedly performed in the presence of high factual uncertainty, the reviewer, as an initial matter, should distinguish 'real' knowledge gaps from 'fake' gaps. Where the risk assessor claims factual uncertainty over certain consequences or probabilities, the reviewer should survey the literature to evaluate whether the alleged scientific gap really exists or maybe has been filled meanwhile.[153]

To the extent alleged knowledge gaps are 'real' and have not hitherto been addressed by science, the risk assessor is forced to 'bridge' the existing gaps in knowledge. In so doing, the risk assessor may become a genuine generator of science, not just a user.[154] This process of filling scientific gaps should be required to satisfy the minimum requirements of 'science', that is 'a process characterized by systematic, disciplined and objective enquiry and analysis, that is, a mode

150 As discussed above, a *de novo* review does not seem a winning proposition. Given the fundamental subjectivity involved in the choice of risk classes and dimensions, their ranking and weighting, etc., an external reviewer may well come up with different results. However, there is no theoretical presumption why this *de novo* risk assessment should be any better, more scientific, precise, or accurate than the original assessment.

151 See footnote 142 above. The Appellate Body seems to have consented to the view that the basic choice of whether to conduct a risk assessment in the first place ultimately resides with the WTO Member, when it stated: 'We observe that, if a Member chooses to base SPS measures on a risk assessment, it must have made the preliminary determination that the relevant scientific evidence is sufficient to perform a risk assessment. If, however, the Member considers that scientific evidence is insufficient to perform a risk assessment, it may instead choose to take provisional SPS measures based on Article 5.7 of the *SPS Agreement*.' See AB Report, *Australia–Apples*, para. 238; see also AB Report, *US/Canada–Continued Suspension*, para. 674.

152 If a Member claims that there is overwhelming factual uncertainty, or insufficient scientific evidence to perform a risk assessment, this claim should be subject to review, bearing in mind the subjective and context-sensitive nature of the term 'sufficiency'. See Gruszczynski (2010b: 192–197), Scott (2007: 117).

153 For example, whenever the risk assessor applies 'expert opinion' to determine probability ranges and densities, or intervals of magnitude citing high natural variability in the underlying data, or indeterminacy of natural or social processes, the reviewer should be able to scrutinize these claims by studying the relevant literature.

154 In the case of 'real' knowledge gaps, the risk assessor's expertise and experience is required to (i) interpret existing scientific results, (ii) apply scientific results to novel contexts, (iii) recombine available scientific results, or (iv) perform original research, for example by compiling and collecting data. All these tasks, in some way or other, may advance the frontier of science.

of studying and sorting out facts and opinions'.[155] In the presence of 'real' scientific gaps, the reviewing body should then assess whether the *process* of the risk assessment is sufficiently scientific in nature. It ought to ask whether the reasoning is coherent and adequately explained.[156] For instance, the reviewer may scrutinize the logic with which causal mechanisms and processes of probability chains are hypothesized and explained by the risk assessor. If these explanations are found to be coherent and sufficiently grounded in the scientific principles, evidence, and data, then the assessment deserves trust. On the other hand, if the reasoning is patchy, the assessment jumps to conclusions, or features 'black boxes' with opaque links between causes and effects, the credibility of the risk assessment suffers.

Whenever a reviewer is charged with evaluating the risk assessor's *ranking* of risk-mitigation alternatives, or with assessing alternatives previously not considered by the risk assessor, such alternatives can reasonably be examined only *against the original methodology applied by the risk assessor*. This is an important result. As shown above in the context of Arrovian uncertainty, different but equally scientific risk assessments may yield different rankings of risk-mitigation outcomes due to the inherent subjectivity of the exercise. In situations of high factual uncertainty, an objective ranking of alternatives is even more difficult to ascertain, since something that is not known with certainty cannot be ranked with certainty, either.[157] When charged with an assessment of the ranking of risk-mitigation options, an external reviewer should be cognizant of the inevitable subjectivity in connection with risk-management options. In practical terms, this means that the reviewer should evaluate risk-mitigation alternatives against the yardstick of the original framework intended and designed by the risk assessor.

4.2 *Australia–Apples and the standard of review under factual uncertainty*

We are now ready to return to the *Australia–Apples* case. While Australia considered the scientific evidence available to be sufficient to perform a risk assessment, it also claimed that the presence of 'scientific uncertainty' made the conclusion of the IRA particularly difficult.[158] The IRA is thus a good example of a

155 AB Report, *EC–Hormones*, para. 187 and footnote 172.

156 This includes the proper identification of scientific gaps, an explanation of the scientific principles and methodological tools applied, the scientific evidence initially relied on, and how this evidence was recombined to generate new inferences and findings.

157 Some basic knowledge of the relative likelihoods and magnitudes of the different effects under consideration are required. Where relative likelihoods/magnitudes are uncertain, they may at best be approximated. Different risk assessors may approximate in different ways, subject to their values, risk attitudes, and constraints. This, in turn, inevitably results in different rankings of alternatives by different risk assessors.

158 See, e.g., AB Report, *Australia–Apples*, paras. 189, 232, and 271. Neither New Zealand, nor the Panel, nor the AB denied this contention.

risk appraisal performed in a situation characterized by high levels of factual uncertainty. In this subsection, we compare the AB's findings in the *Australia–Apples* dispute with the insights gathered from our theoretical discussion in Subsection 4.1 above. In particular, we discuss the AB's pronouncements on the applicable SoR under Articles 2.2, 5.1, 5.2, and 5.6 of the *SPS Agreement* and on the use of Panel-appointed experts, and assess whether this SoR allows Panels to adequately deal with situations characterized by high degrees of factual uncertainty.

(i) Evaluation of the Appellate Body's approach to the standard of review under Articles 2.2, 5.1, and 5.2 of the SPS Agreement. We are concerned in this paper with the AB's methodology regarding the SoR of Members' risk assessments, not with the conclusions reached by the AB in response to the litigating parties' appeals (we note that we largely agree with the AB's ultimate findings). After our survey of the risk-management literature, the AB's line of reasoning with respect to the SoR under Articles 2.2, 5.1, and 5.2 deserves some critical attention. Specifically, two issues warrant more detailed discussion. *First*, we believe that the AB's heavy reliance on the SoR, as formulated in *US/Canada–Continued Suspension*, is misplaced, because that test is unable to appropriately address situations of factual uncertainty, as present in the *Apples* dispute. *Second*, the current *Continued Suspension* SoR lacks concrete guidance to future Panels faced with reviewing risk assessments conducted in situations of high factual uncertainty.

In its appeal of Articles 2.2, 5.1, and 5.2, Australia questioned a number of aspects of the Panel's interpretation of the SoR of Australia's risk assessment, the IRA.[159] In particular, Australia claimed that (i) the Panel was in legal error by reviewing *intermediate* findings by the IRA (instead of focusing on the final outcome); (ii) due to the presence of considerable scientific uncertainty, the phrase 'as appropriate to the circumstances' in Article 5.1 should have been interpreted such as to provide the IRA experts with substantial flexibility in the way Australia conducted the IRA; and (iii) the Panel should have examined expert opinion in the same way as it assessed scientific evidence in general: according to Australia's interpretation of the standard recognized in *US/Canada–Continued Suspension*, the Panel should have reviewed only whether expert judgment falls within a range 'considered legitimate by the standards of the scientific community'.[160]

The AB rejected all of Australia's appeals. It thereby based its argumentation on the three-step SoR it had articulated in *US/Canada–Continued Suspension*.[161] Acknowledging that 'we have not been requested to decide whether the Panel

159 Ibid., para. 201.

160 Ibid., paras. 217, 224, and 232, respectively.

161 See ibid., paras. 213–216, and 219–222. To recall, the SoR that applies to a Panel reviewing a risk assessment under Article 5.1 of the *SPS Agreement* requires it to (i) scrutinize the underlying scientific basis of the risk assessment; (ii) scrutinize the reasoning of the risk assessor based upon such underlying science;

properly assessed the underlying scientific basis that was used by the IRA to support its reasoning and conclusion'[162] (the first prong of Panels' review of a risk assessment), the AB set out to give a lengthy interpretation of the second prong of the *Continued Suspension* SoR. This second prong of review requires Panels to:

> assess whether the reasoning articulated on the basis of the scientific evidence is objective and coherent. *In other words*, a panel should review whether the particular conclusions drawn by the Member assessing the risk find sufficient support in the *scientific evidence relied upon*.[163]

Our *first concern* about the AB's reasoning relates to its excessive reliance on the SoR established in *Continued Suspension*. While it may apply satisfactorily to the risk assessment conducted in the presence of *complete* scientific certainty, the *Continued Suspension* test does not seem well suited to address circumstances of increased factual *uncertainty*. In particular, it fails in situations where the risk assessor falsely claims the existence of a knowledge gap (a 'fake' gap); and in those instances where the risk assessor creates genuine science, i.e. where expert opinion effectively substitutes scientific evidence (in a situation where there is a 'real' knowledge gap).

The context in *Continued Suspension* was one of factual certainty, because the link between the hormones at issue and the claimed effects was effectively established by the scientific studies the European Union relied upon.[164] Therefore, the first two steps of the *Continued Suspension* test were appropriate – in that very context: where there exists reputable (minority) scientific evidence to confirm the risk assessor's ultimate conclusions (first prong), the Panel must ensure that the conclusions contained in the scientific studies relied upon are not incoherent, or in conflict, with the conclusions drawn by the risk assessor (second prong).

Yet, in instances where factual *uncertainty* prevails, as was the case in *Apples*, there is by definition a disconnect between the available scientific evidence and the risk assessor's conclusions – in short, there are knowledge gaps in various parts of the analysis. To assess whether the risk assessor has properly addressed these knowledge gaps, that is whether the reasoning articulated in the risk assessment is objective and coherent, a Panel cannot be limited to review only the scientific evidence *relied upon* by the risk assessor, as the above quote prescribes. The Panel should also review the scientific evidence *not relied upon* by the risk assessor in order to evaluate whether the alleged knowledge gap is 'real' or 'fake'.

and (iii) determine whether the results of the risk assessment sufficiently warrant the challenged SPS measures. See footnote 93 above, and accompanying text.

162 AB Report, *Australia–Apples*, para. 221.

163 Ibid., para. 214, citing AB Report, *US/Canada–Continued Suspension*, para. 591, emphasis added.

164 AB Report, *US/Canada–Continued Suspension*, paras. 603–615.

No less than 12 times did the AB repeat the *Continued Suspension* language on 'whether the conclusions find sufficient support in the scientific evidence *relied upon*',[165] before it seems to have realized that its reliance on the *Continued Suspension* SoR was too confining, and ultimately an inappropriate guideline for Panels having to assess a risk assessment conducted in the presence of factual uncertainty. Hence, in paragraph 236, and again in paragraphs 240 and 241 of its Report, the AB, somewhat clandestinely, added a new facet to the SoR of Members' risk assessments, stating that:

> if a risk assessor reaches certain conclusions based on its expert judgement, having determined that there is a certain degree of scientific uncertainty, this does not preclude a panel from assessing whether those conclusions are objective and coherent and have a sufficient basis in the available scientific evidence.[166]

The discretion to review *available* scientific evidence much better suits Panels' needs in situations of factual uncertainty, since it allows them to determine 'fake' scientific gaps, that is where the responding Member falsely claimed the presence of scientific uncertainty.[167] However, the emphasis on 'available scientific evidence', directly taken from Article 5.2 of the *SPS Agreement*,[168] is clearly in conflict with the language on 'scientific evidence relied upon' used in *Continued Suspension*.

The *Continued Suspension* SoR is also unhelpful to address instances of 'real' knowledge gaps, that is where the risk assessor's expert judgment effectively helps bridge areas of true factual uncertainty by creating genuine scientific evidence. The AB stated that:

> when the exercise of expert judgement forms an integral part of the risk assessor's analysis, then it should be subject to the same type of scrutiny by the panel as all other reasoning and conclusions in the risk analysis.[169]

If by 'the same type of scrutiny as all other reasoning and conclusions' the AB means that the risk assessor's expert judgment should be a scientific process that is objective and coherent,[170] the above statement is correct. If, however, the AB takes 'the same kind of scrutiny' to mean that the expert judgment should 'find sufficient support in the scientific evidence *relied upon*', we do not find this statement helpful. Where 'real' gaps in knowledge exist, risk assessors may not necessarily 'rely upon'

165 AB Report, *Australia–Apples*, paras. 214, 215, 219, 220, 221, 222, 223, 224, and 225.

166 Ibid., para. 236, emphasis added.

167 Indeed, the Panel in *Australia–Apples* made use of its right to assess whether scientific evidence was *available* in circumstances in which the IRA claimed there was a knowledge gap. And, on a number of occasions, the Panel actually found that the IRA, without reasoned and adequate explanation, substituted expert judgment for scientific evidence that must have been readily available to the IRA. See AB Report, *Australia–Apples*, paras. 193, 233, 240, and 241.

168 See footnote 16 above, for the text of Article 5.2.

169 AB Report, *Australia–Apples*, para. 224. See also paras. 236, 241, and 244.

170 See ibid., paras. 207, 241, and 244.

scientific evidence the way that the European Union did in *Continued Suspension*. Instead, the risk assessor may perform basic research and produce authentic science, thus furthering the frontier of scientific knowledge. In such circumstances, rather than checking the results against the evidence 'relied upon', a Panel ought to ask whether this original research satisfies the established standards of a scientific process. A Member's gap-filling exercise should not be denied the status of 'science', or 'scientific evidence', simply because the Panel fails to see a connection to the concrete scientific evidence 'relied upon'. Where the risk assessor performs authentic and genuine science, its findings – if properly conducted – substitute scientific evidence, rather than *relying upon* it.[171]

Our *second concern* about the AB's SoR as it pertains to Articles 5.1 and 5.2 is that it does not seem to provide a reliable roadmap to future Panels for the evaluation of risk assessments in the presence of factual uncertainty. On the one hand, the AB in *Australia–Apples* repeated its opinion originally pronounced in its reports in *EC–Hormones* and *US/Canada–Continued Suspension* that:

> a panel's task is to review a WTO Member's risk assessment and not to substitute its own scientific judgement for that of the risk assessor. A panel should not, therefore, determine whether the risk assessment is correct, but rather 'determine whether that risk assessment is supported by coherent reasoning and respectable scientific evidence and is, in this sense, objectively justifiable'.[172]

At the same time, however, the AB seemingly rubber-stamped the Panel's complete 'unpacking' of the IRA. While stopping short of a *de novo* review, the Panel deemed it within its competence to review every relevant aspect of the IRA. The Panel's review included both substantive and procedural aspects, and comprised of an evaluation of the basic methodological set-up; fundamental assumptions and qualifications; omitted scientific evidence; probability intervals and density functions; various probability factors and impact scores; omitted classes of effects; missing justifications; documentation and transparency issues; etc.[173] The AB approved of the Panel's review of all these elements.

Hence, apart from the general prohibition for Panels to conduct their own risk assessments, we have difficulties finding any explicit language in the AB Report that

171 It could be argued that in para. 236 of the AB Report, the AB actually recognized that expert judgment may to some degree substitute for available scientific evidence, when it stated:

> We have also stressed the difference between the underlying scientific evidence, on the one hand, and the reasoning and conclusions of the risk assessor based on this scientific evidence *and, where necessary, on expert judgement*, on the other hand (emphasis added).

172 AB Report, *Australia–Apples*, para. 213, citing AB Report, *US/Canada–Continued Suspension*, para. 590, emphasis added. See also AB Report, *Australia–Apples*, para. 298: 'We observe that under Article 5.1 of the *SPS Agreement* a panel is called to review the objectivity and coherence of the risk assessment, *not whether the results of the risk assessment are correct and correspond to the results the panel itself would reach based on the advice of the appointed experts*' (emphasis added).

173 See footnote 65 above, and accompanying text.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

would curb Panels' power of review.[174] Yet, if *no* element of a Member's risk assessment is 'off limits' for review by the Panel, and *every* part of the analysis can be scrutinized, how does this *not* amount to a finding that the entire risk assessment is *incorrect* – something that the AB explicitly prohibited Panels from doing? We do not see this tension resolved in the AB Report in *Apples*. Explicit language is missing that would provide future Panels with guidelines and principles on which parts and elements of the risk assessment at issue cannot or should not be subject to Panel scrutiny.

To sum up, we are uncomfortable with the current SoR as it pertains to Members' risk assessments in the presence of increased factual uncertainty. Based on these concerns, we propose a modification of the SoR of Articles 5.1 and 5.2 that applies to contexts of factual certainty and factual uncertainty alike.

The first question a Panel should address is: Does the responding Member invoke the presence of substantial factual uncertainty? If this is *not* the case, then we are in a situation of relative factual certainty. The three-prong SoR pronounced by the AB in its Report on *Continued Suspension* generally seems appropriate in such a case, with an important modification, however. In light of the discussion of the inevitable Arrovian uncertainty, Panels should be reminded to exercise caution reviewing the basic parameters of the risk analysis, that is the methodological set-up and structure, the choice of effect classes, vantage points, and risk dimensions and their relative weighing and ranking, and the choice of metric. Instead, Panels should focus on the quality of the execution, the internal coherence, the scientific grounding, and the presentation of the risk assessment.[175]

If the responding Member claims to have conducted its risk analysis in the presence of considerable factual uncertainty (as *Australia* did), the Panel, in a next step, should determine whether the alleged scientific gaps are 'real' or 'fake'. Together with its appointed experts, the Panel should evaluate whether the

174 One notable exception may be found in para. 228, where the AB stated:

> The Panel's analysis of the IRA follows the IRA's own structure and, therefore, consists of reviews of steps and factors and the methodology by which they are aggregated and combined. In so doing, the Panel *adhered to the standard of review applicable to a panel's review of a risk assessment under Article 5.1* of the *SPS Agreement*, which requires a panel to review the conclusions of a risk assessor, not to undertake its own risk assessment (emphasis added).

This quote seems to suggest that the AB is of the view that a Panel should not question the structure and basic methodological setup of the risk assessment under evaluation.

175 See the discussion in Section 4.1.2, above. In that respect, the Panel to our mind actually did a splendid job in its review of the IRA. It largely accepted the IRA's basic framework and structure, and did not challenge the IRA's choice of risk classes, vantage points, level of aggregation, and causation mechanisms. Instead, it was careful to limit itself to assessing the objectivity and coherence of the IRA's reasoning, checking for 'fake' gaps; evaluating the internal consistency of the IRA; and assessing the documentation and transparency of the process. The Panel to our mind acted amiss only when accusing the IRA of failing to consider a number of outcomes/effect classes in the context of assessing the consequences of spread of fire blight and ALCM. See AB Report, *Australia–Apples*, paras. 253, 256. As discussed in Section 4.1.2, the choice of effects is inherently subjective and should not be subject to external review.

Member, in performing its risk assessment, ignored available scientific evidence, or whether the risk assessor was correct in claiming uncertainty. If Panel and external experts together conclude that the risk assessor substituted its own judgment for available scientific evidence (a 'fake' gap), then the reasoning of the risk assessor is not objective and coherent.

Where Panel and Panel-appointed experts determine the existence of a 'real' gap in scientific knowledge, the Panel should move on to a three-step process, in which it determines whether the risk assessment: (i) is sufficiently rooted in ('based on') scientific principles; (ii) satisfies the minimum requirements of 'science'; (iii) warrants the SPS measure at issue.

Although, strictly speaking, step one is a natural element of any scientific process (step two), we separate the two for expositional purposes. The task of the Panel in step one is to assess whether there is a 'rational and objective relationship'[176] between the risk assessment and the scientific principles relied on by the risk assessor. The reference to 'scientific principles', language contained in the text of Article 2.2,[177] constitutes broader language than the reference to 'scientific evidence', as used in *Continued Suspension*. This would allow the Panel to review situations in which the risk assessor performs genuine science that is not, or not yet, regarded as 'existing scientific evidence'.

In step two, the Panel should ensure that the risk assessment satisfies all the criteria of a scientific process, despite the presence of factual uncertainty. In other words, the Panel should evaluate whether the process is systematic, disciplined, analytical, transparent, objective, and well-documented.[178] This would allow the Panel to evaluate whether the reasoning of the risk assessor is coherent and objective, as required by the AB in *US/Canada–Continued Suspension*.[179] Step three is the same as that under the current SoR.[180]

Figure 4, on the next page, summarizes our proposal for a modified SoR of a Member's risk assessment. In our opinion, this SoR is fully compatible with the wording of Articles 2.2, 5.1, and 5.2 of the *SPS Agreement*. It constitutes moderate, but substantial, development of the SoR actually applied by the AB in *Australia–Apples*.

176 This language is the AB's interpretation of the wording 'based on' in Articles 2.2 and 5.1. See AB Report, *Japan–Agricultural Products II*, para. 84; AB Report, *Japan–Apples*, paras. 162 and 163; AB Report, *Australia–Apples*, para. 210.

177 Article 2.2 of the *SPS Agreement* reads in pertinent part: 'Members shall ensure that any sanitary or phytosanitary measure . . . is based on scientific principles and is not maintained without sufficient scientific evidence.'

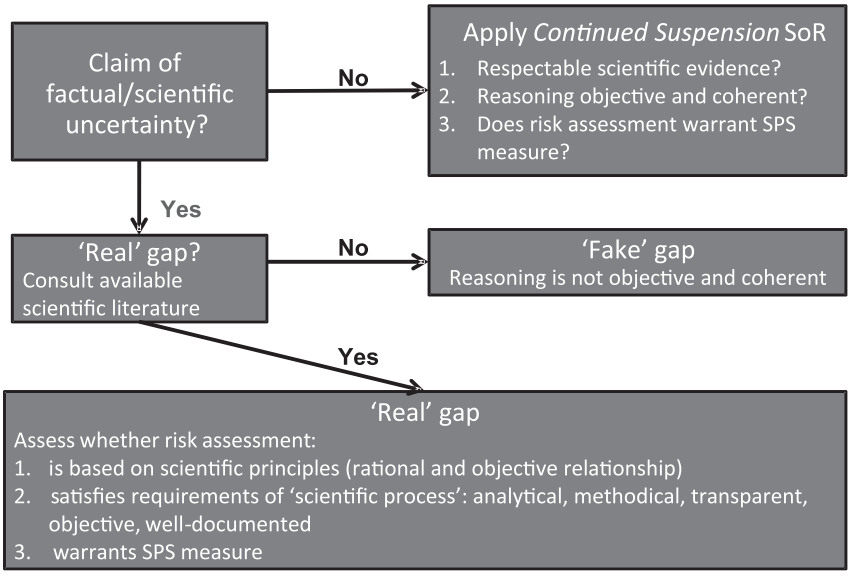
178 See AB Report, *Australia–Apples*, paras. 207, 240, and 241; and AB Report, *EC–Hormones*, para. 187.

179 AB Report, *US/Canada–Continued Suspension*, paras. 590 and 591.

180 See footnotes 93 and 161 above. This step has never been subject to AB interpretation, but could benefit from AB guidance in the future. See Gruszczynski (2010a: 12, 13).

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

Figure 4. Suggestion for a modified standard of review under Articles 5.1 and 5.2



(ii) Evaluation of the Appellate Body's approach to the standard of review under Article 5.6 of the SPS Agreement

According to Article 5.6 of the *SPS Agreement*, Members shall ensure that their SPS measures are not more trade-restrictive than required to achieve their appropriate level of sanitary or phytosanitary protection. In *Australia–Salmon*, the AB pronounced a three-prong test to determine consistency with Article 5.6.[181] The Panel in *Apples* considered various alternative measures to limit the risk of fire blight and ALCM proposed by New Zealand as candidates for less-trade-restrictive alternatives (LTRAs). For fire blight, New Zealand's proposed LTRA was identical to the 'unrestricted risk' scenario assessed by the IRA. For ALCM, the LTRA was a slight modification of one risk-mitigation option initially considered by the IRA.[182] The Panel determined that both of New Zealand's LTRAs satisfied the three-pronged test under Article 5.6.

Australia's appeal was confined to the second prong of the Article 5.6 test, namely the Panel's findings that the LTRAs put forward by New Zealand achieve

181 A violation of Article 5.6 will be established when there is an alternative measure, other than the contested measure, that (i) is reasonably available; (ii) achieves the Member's ALOP; and (iii) is significantly less trade-restrictive. These three conditions are cumulative in nature. See AB Report, *Australia – Salmon*, para. 194.

182 AB Report, *Australia–Apples*, para. 330.

https://doi.org/10.1017/S1474745611000516 Published online by Cambridge University Press

Australia's ALOP.[183] Particularly relevant here is Australia's claim that the Panel misinterpreted and misapplied Article 5.6. According to Australia, the Panel excessively relied on its findings under Article 5.1 on the inadequacy of Australia's IRA.[184]

The AB, in assessing the Panel's approach, found no basis for the Panel's approach of linking Article 5.6 to Australia's risk assessment. The AB ruled that the obligations set out in Article 5.1 and Article 5.6 are distinct and legally independent of each other. It determined that the Panel was required to undertake its own analysis of the question of whether the LTRAs proposed by New Zealand would achieve Australia's ALOP, and that this analysis was wholly independent of Australia's IRA.[185] In particular, the AB found that when considering the risk associated with the proposed LTRAs neither New Zealand nor the Panel was obliged to adopt the same methodology or structure as that employed by Australia for the IRA.[186]

The Panel had passed the question whether New Zealand's LTRAs could achieve Australia's ALOP onto the Panel-appointed experts. The AB reprimanded the Panel for having done so. The AB noted that it felt:

> certain reservations about the Panel having done so, given that this was the ultimate question that the Panel was charged with answering pursuant to Article 5.6. Experts may assist a panel in assessing … potential alternative measures, but whether or not an alternative measure's level of risk achieves a Member's appropriate level of protection is a question of legal characterization … Answering this question is not a task that can be delegated to scientific experts.[187]

While we agree with the AB's ultimate finding that the Panel's two-step process was not justified by the text of Article 5.6 of the *SPS Agreement*, we have two concerns about the AB's characterization of Panels' SoR under Article 5.6.

183 Ibid., para. 334.

184 Ibid., paras. 345–349. The Panel's analysis of whether New Zealand had established that its proposed LTRA would meet Australia's ALOP proceeded in two cumulative steps. In the first step, the Panel assessed whether the risk resulting from the importation of New Zealand apples really exceeded Australia's ALOP. Since that was the case (as per the Panel's earlier examination of Australia's risk assessment under Articles 2.2, 5.1, and 5.2), the Panel, in a second step, proceeded to assess whether New Zealand had raised a presumption that its LTRAs have a sufficient risk-reduction effect. Ibid., paras. 350–353.

185 See ibid., paras. 354–356. The AB found that 'the legal question is whether the importing Member could have adopted a less trade-restrictive measure. This requires the panel itself to objectively assess, *inter alia*, whether the alternative measure proposed by the complainant would achieve the importing Member's [ALOP]. The fact that, in the present case, the alternative measures proposed by New Zealand in the context of its claim under Article 5.6 had also been assessed in the IRA did not alter the nature of the Panel's task under Article 5.6'. Ibid., para. 356.

186 See ibid., paras. 355–357. In fact, the AB concluded that linking the assessment of New Zealand's LTRA with Australia's IRA constituted 'the fundamental flaw' in the Panel's approach and was 'an incorrect understanding of its task'. Ibid., para. 357.

187 Ibid., para. 384.

First, we find the fact that the AB mandates a Panel to 'assess, itself, whether the alternative measures … meet [the respondent's ALOP]'[188] difficult to square with its earlier statements that 'a panel is not well suited to conduct scientific research and assessment itself and should not substitute its judgment for that of the risk assessor'.[189] Comparing a risk-mitigation alternative against a country's accepted level of risk is quite a technical task, which, by the AB's own admission, is 'scientific in nature' and 'must be based on scientific principles'.[190] In fact, also by the AB's own admission, this task is not unlike conducting a risk assessment.[191] We thus do not see how a Panel could get around conducting such research – a task that it is 'poorly suited' to perform, according to the AB.[192]

This apparent contradiction is exacerbated by the above-quoted statement, in which the AB explicitly prohibits the Panel's external experts from giving answers to the ultimate 'legal' question whether a proposed LTRA meets a respondent's ALOP.[193] If Panels lack the competence to deal with complex technical questions, but expert input is curbed, this seems like an odd place for Panels to be in. In situations of increased factual uncertainty, we believe that Panels should make more, not less, use of its outside experts.

Second, we are puzzled by the AB's finding that, when it comes to assessing LTRAs, neither complainants nor dispute Panels are in any way bound by the methodological framework, structure, and overall approach taken by the original risk assessor. Proposing an LTRA is tantamount to suggesting that this alternative has a higher rank than the respondent's acceptable risk threshold. In Subsection 4.1.2, we explained that due to the problems of multidimensionality and incommensurability (Arrovian uncertainty) and those of uncertainty of outcomes and probabilities (Knightian uncertainty), a ranking of risk-mitigation options is difficult, at times even impossible. Based on findings by the risk-management literature, we suggested that any ranking of alternatives, at the very least, should be assessed within the realm of the methodological framework utilized by the risk assessor.[194]

For these theoretical reasons, we think that New Zealand was right in proposing as LTRAs two alternatives that Australia itself had previously discussed in its IRA.

188 Ibid., para. 355.

189 Ibid., para. 225. See also AB Report, *EC–Hormones*, para. 117 (Panels 'in any case [are] poorly suited to engage in a [*de novo*] review').

190 AB Report, *Australia–Apples*, para. 364.

191 Ibid., para. 364.

192 See footnote 189 above.

193 This seems a difficult position to maintain, because the higher the level of factual uncertainty, the more the ultimate 'legal' question seems to collapse with the ultimate 'factual' question. It is thus difficult to see why a Panel might be better suited than designated experts in the field to answer such question.

194 Due to the intrinsically subjective choices and trade-offs inherent in any risk assessment, two risk assessors may come up with different, yet equally correct, precise, and scientific results. Applied to the issue of LTRAs, depending on the individual methodologies applied by different risk assessors, a given risk-mitigation alternative may exceed, or fall short of, a country's ALOP.

These alternatives could easily be assessed *within the methodological framework* of the IRA. We also believe that the Panel's intuition (though, as we have explained, not its execution) was correct, when it tied its evaluation of the LTRAs closely to the original IRA.

The AB seems to think that there is one 'definitive', or 'objectively correct' ordinal ranking of risk-mitigation alternatives, with some risk-mitigation options exceeding a country's ALOP, and others falling below, and that the complainant can simply prove that its LTRA exceeds the respondent's ALOP. However, the insights of the risk-management literature reviewed above militate against such a mechanistic, and ultimately facile, view. Due to the inherent subjectivity in every risk appraisal, and the necessary trade-offs to be performed by the risk assessor, no outside third party can claim to possess superior knowledge about which risk-mitigation alternatives are objectively 'better' than others.

We have doubts whether the drafters of the *SPS Agreement* intended to give the LTRA analysis of a third party (be it the complaining party or the Panel) precedence over the risk assessment of the country that originally imposed the SPS measure at issue. The AB may wish to rethink the SoR it proclaimed under Article 5.6 and should tie the complainant's *prima facie* case and the Panel's assessment of LTRAs closer to the methodology and structure of the respondent's initial risk assessment. The AB could do so by requesting the responding Member to rerun its risk assessment using the proposed LTRAs as risk-mitigation alternatives. This 'new' risk assessment should then be assessed according to the same SoR like the responding Member's original risk analysis.

5. Conclusion

In commenting on the Appellate Body Report in *Australia–Apples*, we discussed issues surrounding the standard of review of measures taken in situations that exhibit particularly high levels of factual or scientific uncertainty. In particular, we examined whether the SoR for risk assessments under Articles 5.1, 5.2, and 2.2 of the *SPS Agreement* should be more deferential towards the enacting Member which faces a relatively high degree of factual uncertainty, and which elements in the Member's risk assessment should be considered 'off limits' for review by Panels or the Appellate Body. We also assessed how higher degrees of factual uncertainty (should) affect a Panel's ability to assess less-trade-restrictive alternatives under Article 5.6 of the *SPS Agreement*.

Relying on findings from the disciplines of economics, decision analysis, and risk management, we introduced a basic framework for decisionmaking in conditions of uncertainty. We found that risk assessments performed in situations of scientific uncertainty *necessarily* contain subjective trade-offs on the part of the risk assessor. For theoretical and practical reasons, certain types of decisions taken by the risk assessor cannot – and therefore should not – be subject to external review.

Based on these theoretical insights, we evaluated the Appellate Body's review of Australia's risk assessment in the context of the *Australia–Apples* case, and suggested concrete modifications to the SoR currently applied by Panels and the AB.

With respect to the Appellate Body's standard of review under Articles 2.2, 5.1, and 5.2, we took issue with the AB's heavy reliance on the legal test it developed in *US/Canada–Continued Suspension*, because that test was developed to address situations of factual certainty, but is unable to appropriately address situations of factual *un*certainty, as present in the *Apples* dispute. We concluded that the SoR as formulated in *Continued Suspension* lacks concrete guidance to future Panels when faced with the task of reviewing risk assessments conducted in situations of high factual uncertainty. We proposed certain modifications to the SoR of Articles 5.1 and 5.2 (and, consequently, also of Article 2.2) that are able to deal with contexts of factual certainty and factual uncertainty alike and is fully compatible with the wording of Articles 2.2, 5.1, and 5.2 of the *SPS Agreement*.

As regards the Appellate Body's review of less-trade-restrictive alternatives, we took issue with the AB's finding that neither complainants nor dispute Panels are in any way bound by the methodological framework originally chosen by the risk assessor. Our previous findings on decisionmaking in the presence of uncertainty suggested that any ranking of alternatives can only be meaningfully assessed within the realm of the methodological framework utilized by the risk assessor. Due to the inherent subjectivity in every risk appraisal, and the necessary trade-offs to be performed by the risk assessor, no outside third party can claim to possess superior knowledge about which risk-mitigation alternative is objectively 'better' than others. An external reviewer relying on its own methodology risks rejecting a ranking of alternatives that is conducted in a perfectly valid, precise, and scientific manner. Hence, we suggested that the AB may wish to rethink the SoR it proclaimed under Article 5.6 and should tie the complainant's *prima facie* case and the Panel's assessment of LTRAs closer to the methodology and structure of the respondent's initial risk assessment.

References

AQIS (2011), 'Draft Report for the *Non-Regulated* Analysis of Existing Policy for Apples from New Zealand', available at http://www.daff.gov.au/__data/assets/pdf_file/0010/1906507/Draft_report_NZ_apples_May_2011.pdf (last visited 14 December 2011).

Arrow, Kenneth J. (1963), *Social Choice and Individual Values*, 2nd edn, New Haven, CT: Yale University Press.

Arrow, Kenneth J. (1974), *The Limits of Organization*, New York, NY: W.W. Norton & Co.

Bernstein, Peter L. (1996), *Against the Gods: The Remarkable Story of Risk*, New York, NY: John Wiley & Sons.

Biosecurity Australia (2006), 'Final Import Risk Analysis Report for Apples from New Zealand', available at http://www.daff.gov.au/ba/ira/final-plant/apples-nz (last visited 15 December 2011).

Bohanes, Jan and Nicolas Lockhart (2009), 'Standard of Review in WTO Law', in Daniel Bethlehem, Donald McRae, Rodney Neufeld, and Isabelle Van Damme (eds.), *The Oxford Handbook of International Trade Law*, Oxford: Oxford University Press.

European Science and Technology Observatory (May 1999), 'On Science and Precaution in the Management of Technological Risk, Volume I', European Science and Technology Observatory, Brussels, Belgium, *available at* http://ftp.jrc.es/EURdoc/eur19056en.pdf (last visited 14 December 2011).

Gibbons, Robert (1992), *A Primer in Game Theory*, London: Prentice Hall.

Gruszczynski, Lukasz (2010), 'The Standard of Review in International SPS Trade Disputes: Some New Developments', mimeo, available at http://regulation.upf.edu/dublin-10-papers/7F4.pdf (last visited 15 December 2011).

Gruszczynski, Lukasz (2010), *Regulating Health and Environmental Risks under WTO Law: A Critical Analysis of the SPS Agreement*, Oxford: Oxford University Press.

Holdren, John P. (1982), 'Energy Hazards: What to Measure, What to Compare', *Technology Review*, **85** (3): 32–38, 74–75.

Klinke, Andreas and Renn, Ortwin (2002), 'A New Approach to Risk Evaluation and Management: Risk-Based, Precaution-Based, and Discourse-Based Strategies', *Risk Analysis*, **22** (6): 1071–1094.

Knight, Frank H. (1921), *Risk, Uncertainty and Profit*, Boston, MA: Houghton Mifflin.

Kolsky Lewis, Meredith (2011), 'Australians Get Their First Taste of New Zealand Apples in Ninety Years', *ASIL Insights*, **15**(25), available at http://www.asil.org/pdfs/insights/insight110912.pdf (last visited 14 December 2011).

Krimsky, Sheldon (1997), 'Biotechnology Safety: Enabling the Safe Use of Biotechnology: Principles and Practice and Appropriate Oversight for Plants with Inherited Traits for Resistance to Pests', *Environment*, **39**(5): 27–30.

Luce, R. Duncan and Howard Raiffa (1957), 'An Axiomatic Treatment of Utility', in R.D. Luce and H. Raiffa, *Games and Decisions: Introduction and Critical Survey*, New York, NY: John Wiley & Sons, Chapter 2.5.

Mercurio, Bryan Christopher and Dianna Shao (2010), 'A Precautionary Approach to Decision Making: The Evolving Jurisprudence on Article 5.7 of the SPS Agreement', *Trade, Law and Development*, **2** (2): 195–223.

O'Riordan, Timothy and James Cameron (eds.) (1994), *Interpreting the Precautionary Principle*, London, UK: Earthscan.

O'Riordan, Timothy and Andrew Jordan (1995), 'The Precautionary Principle, Science, Politics and Ethics', *CSERGE Working Paper* PA 95–02.

Prévost, Denise (2005), 'What Role for the Precautionary Principle in WTO Law after *Japan–Apples*?', *EcoLomic Policy and Law: Journal of Trade & Environment Studies*, **2**(4): 1–14.

Renn, Ortwin and Andreas Klinke (November 2001), 'Risk Evaluation and Risk Management for Institutional and Regulatory Policy', in *European Science and Technology Observatory 'On Science and Precaution in the Management of Technological Risk, Volume II'*, Brussels, Belgium: European Science and Technology Observatory, pp. 11–35, *available at* http://ftp.jrc.es/EURdoc/eur19056IIen.pdf (last visited 15 December 2011).

Rosenberg, Nathan (1996), 'Uncertainty and Technological Change', in Ralph Landau, Timothy Taylor, and Gavin Wright (eds.), *The Mosaic of Economic Growth*, Stanford, CA: Stanford University Press, pp. 334–356.

Royal Society (2002), 'Genetically Modified Plants for Food Use and Human Health – an Update', Royal Society, London, available at http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2002/9960.pdf (last visited 15 December 2011).

Salanié, Bernard (1997), *The Economics of Contracts: A Primer*, Cambridge, MA: MIT Press.

Scott, Joanne (2007), *The WTO Agreement on Sanitary and Phytosanitary Measures: A Commentary*, Oxford, UK: Oxford University Press.

Stirling, Andrew (November 2001), 'On "Precautionary" and "Science Based" Approaches to Risk Assessment and Environmental Appraisal', in *European Science and Technology Observatory 'On Science and Precaution in the Management of Technological Risk, Volume II'*, Brussels, Belgium: European Science and Technology Observatory, pp. 36–93, available at http://ftp.jrc.es/EURdoc/eur19056IIen.pdf (last visited 15 December 2011).