Synthese (2019) 196:3251–3262
https://doi.org/10.1007/s11229-017-1573-y

S.I. : EVIDENCE AMALGAMATION IN THE SCIENCES

Extrapolation and the Russo–Williamson thesis

Michael Wilde[1] · **Veli-Pekka Parkkinen**[1]

Received: 15 February 2017 / Accepted: 16 September 2017 / Published online: 29 November 2017

Abstract A particular tradition in medicine claims that a variety of evidence is helpful in determining whether an observed correlation is causal. In line with this tradition, it has been claimed that establishing a causal claim in medicine requires both probabilistic and mechanistic evidence. This claim has been put forward by Federica Russo and Jon Williamson. As a result, it is sometimes called the Russo–Williamson thesis. In support of this thesis, Russo and Williamson appeal to the practice of the International Agency for Research on Cancer (IARC). However, this practice presents some problematic cases for the Russo–Williamson thesis. One response to such cases is to argue in favour of reforming these practices. In this paper, we propose an alternative response according to which such cases are in fact consistent with the Russo–Williamson thesis. This response requires maintaining that there is a role for mechanism-based extrapolation in the practice of the IARC. However, the response works only if this mechanism-based extrapolation is reliable, and some have argued against the reliability of mechanism-based extrapolation. Against this, we provide some reasons for believing that reliable mechanism-based extrapolation is going on in the practice of the IARC. The reasons are provided by appealing to the role of robustness analysis.

Keywords Causality · Extrapolation · IARC · Robustness · Russo–Williamson thesis

✉ Michael Wilde
m.e.wilde@kent.ac.uk

✉ Veli-Pekka Parkkinen
v.k.parkkinen@kent.ac.uk

[1] Department of Philosophy, University of Kent, Canterbury, UK

1 The Russo–Williamson thesis

A particular tradition in medicine claims that a variety of evidence is helpful in determining whether an observed correlation is causal: A variety of different types of evidence is included in the indicators of causality put forward by Hill (1965). Some of these indicators concern probabilistic evidence, such as the strength of an observed correlation between a putative cause and effect. Other indicators concern mechanistic evidence, for example, the biological plausibility of a causal interpretation of the observed correlation. Hill believed that none of the indicators is necessary or sufficient for establishing causality (1965, p. 299). However, the general approach still makes clear the importance of considering a variety of evidence in determining causality. Recently, this way of thinking has been extended by Russo and Williamson (2007). They claim that 'two different *types of evidence*—probabilistic and mechanistic—are at stake when deciding whether or not to accept a causal claim' (2007, p. 163). In other words: 'To establish causal claims, scientists need the mutual support of mechanisms and dependencies' (2007, p. 159). This has been called the Russo–Williamson thesis (Illari 2011; Gillies 2011).

Illari (2011) has argued that the commitments of the thesis are not immediately clear, because it is not clear exactly what counts as probabilistic evidence or mechanistic evidence. She argues that there is no useful distinction between probabilistic and mechanistic evidence-gathering methods. An epidemiological study may provide not only evidence for the existence of an appropriate correlation but also evidence for the existence of a mechanism. Similarly, basic science research may provide not only evidence for the existence of a mechanism but also evidence for the existence of a correlation. Instead, the requirement for the two types of evidence is plausible only if probabilistic evidence is taken to mean evidence of the existence of a correlation between the putative cause and effect, and mechanistic evidence is taken to mean evidence of the existence of a mechanism that can account for the extent of this correlation (2011, pp. 141–148). However, even granting this disambiguation, it remains unclear just how much probabilistic and mechanistic evidence is required in order to establish the relevant causal claim. Here different interpretations are possible. At one extreme, all that is required to establish a causal claim in medicine is to have some evidence of the existence of a correlation and some evidence of the existence of an appropriate mechanism. This is the weak interpretation. At the other extreme, it is necessary to have evidence that establishes both the existence of a correlation and the existence of an appropriate mechanism, where establishing a claim about the existence of a correlation or mechanism requires the claim to meet a high enough epistemic standard that it may be rightly employed as evidence for other claims. This is the strong interpretation.

In this section, we argue that the strong interpretation is clinched by the theoretical motivation for the thesis provided by Russo and Williamson. In particular, Russo and Williamson argue that even if the probabilistic evidence establishes the existence of a correlation between the putative cause and effect, this is not by itself enough to establish the corresponding causal claim, since the correlation may have some non-causal explanation, such as being the result confounding, bias, or chance. They argue that it is mechanistic evidence that rules out alternative non-causal explanations of the established correlation (2007, pp. 162, 163). They say that 'if there is no plausible

mechanism…, then any correlation is likely to be spurious' (2007, p. 159). However, notice that a causal claim is established only if all non-causal explanations of the established correlation have been ruled out, and this requires more than simply some evidence of a mechanism. In fact, it requires having *established* the existence of a mechanism, which may not be straightforward.

Russo and Williamson also argue that even if mechanistic evidence establishes the existence of an appropriate mechanism, this is also not by itself enough to establish the corresponding causal claim. This is because there may also exist further mechanisms by which the putative cause also cancels out any difference made by way of the established mechanism. In their own words: 'it is uncontroversial that mechanistic evidence on its own can not warrant a causal claim, as it may be the case that the purported cause, although prior to the effect and mechanistically connected to it, actually makes little or no difference to it' (2007, p. 162). This is sometimes called the problem of masking: an established mechanism is not enough to establish a causal claim, since there may also exist further counteracting mechanisms by which the putative cause also cancels out any difference made by way of the established mechanism (Illari 2011). Russo and Williamson argue that it is probabilistic evidence that helps to overcome the problem of masking by providing evidence of an overall correlation between the putative cause and effect (2007, p. 162) (see also Illari 2011, pp. 146, 147). However, some evidence of correlation is not enough to overcome the problem of masking, since this is consistent with the existence of some counteracting mechanisms. Instead, in order to overcome the problem of masking, it is necessary to *establish* the existence of an appropriate correlation.

All this suggests that the strong interpretation of the Russo–Williamson thesis is correct: in order to establish a causal claim in medicine, it is necessary to have evidence that establishes both the existence of a correlation and the existence of an appropriate mechanism. It should be noted that the strong interpretation is confirmed by a later and more careful formulation of the thesis:

> In order to establish that A is a cause of B in medicine one normally needs to establish two things. First, that A and B are suitably correlated—typically that A and B are probabilistically dependent, conditional on B's other known causes. Second, that there is some underlying mechanism linking A and B that can account for the difference that A makes to B (Clarke et al. 2014, p. 343).

2 The International Agency for Research on Cancer

In support of the thesis, Russo and Williamson appeal to the practice of the International Agency for Research on Cancer. They claim that '[t]he crucial and equal importance of probabilistic and mechanistic considerations is recognized by the International Agency for Research on Cancer' (2007, p. 161). Unfortunately, it is difficult to evaluate this claim, because no details about this practice are provided. However, the relevant details are provided in the *Preamble* to the *IARC Monographs* (IARC 2015).

IARC regularly brings together a working group of experts in order to determine the strength of the available evidence concerning whether a particular exposure is

carcinogenic to humans. The working group is not concerned with determining the extent of the cancer risk of the exposure, but only with determining the strength of the evidence that the exposure increases the risk of cancer, in other words, that the exposure is a cancer hazard. Accordingly, they aim to classify the exposure into one of five groups: (1) The exposure is carcinogenic to humans; (2A) The exposure is probably carcinogenic to humans; (2B) The exposure is possibly carcinogenic to humans; (3) The exposure is not classifiable as to its carcinogenicity to humans; (4) The exposure is probably not carcinogenic to humans. The overall evaluation and rationale is published as one of the *IARC Monographs*.

The overall evaluation is informed by a variety of evidence, including epidemiological studies, cancer bioassays, and mechanistic and other relevant data. In particular, one subgroup reviews only studies of cancer in humans, which are typically epidemiological studies. A distinct subgroup reviews only studies of cancer in experimental animals. Both subgroups then provide an evaluation of the strength of the respective evidence in terms of four categories: (i) *Sufficient evidence of carcinogenicity*; (ii) *Limited evidence of carcinogenicity*; (iii) *Inadequate evidence of carcinogenicity*; (iv) *Evidence suggesting lack of carcinogenicity*. A further subgroup reviews the mechanistic and other relevant data. This group provides an evaluation of the strength of the mechanistic data in terms of whether there is weak, moderate or strong evidence of relevant mechanisms underlying any increased cancer risk. It then attempts to determine whether any identified mechanisms also operate in exposed humans. The working group as a whole then attempts to reach consensus on an overall evaluation of the carcinogenicity of the exposure, on the basis of the individual evaluations of each of the subgroups: 'The overall evaluation is a matter of scientific judgement, reflecting the combined weight of the evidence' (Cogliano et al. 2008, p. 102).

A natural account of this practice does seem to provide some support for the Russo–Williamson thesis. According to this account, the subgroup looking at the studies of cancer in humans is concerned only with establishing whether there exists a correlation between the exposure and cancer in humans. And the subgroup looking at the mechanistic and other relevant data is concerned only with establishing whether there exists a mechanism that can explain the extent of this correlation. As long as this practice is the right way of doing things, it looks like the natural account provides some support for the Russo–Williamson thesis. Indeed, Russo and Williamson claim that the practice of IARC shows that an '[a]ssessment of causality depends on the presence of a plausible mechanism and on probabilistic evidence' (Russo and Williamson 2007, p. 161).

However, there are some problematic cases for this natural account. In particular, there are cases in which an exposure was classified as carcinogenic to humans, even though the mechanistic and other relevant data fall short of establishing a mechanism in humans. In a recent case, the consumption of processed meat was classified as carcinogenic to humans on the basis of sufficient evidence of carcinogenicity in the studies of cancer in humans alone (Bouvard et al. 2015, p. 1599). According to the natural account, this is a case in which an exposure is classified as carcinogenic to humans only on the basis of probabilistic evidence. As a result, this looks like a counterexample to the Russo–Williamson thesis.

In response to such cases, a proponent of the thesis may claim that some reform is required in the practice of IARC. This response is suggested by Leuridan and Weber (2011). They argue that in establishing carcinogenicity to humans, it is not enough to rely only on the studies of cancer in humans but that '[m]echanistic evidence should also be used to better exclude the possibility of confounding' (2011, p. 99).

Alternatively, a proponent of the thesis may give up one of the commitments of the natural account, namely, that the studies of cancer in humans are concerned only with establishing whether there exists a correlation between the exposure and cancer in humans. Instead, they might maintain that in some cases the studies of cancer in humans are of sufficient quality that they rule out the possibility of confounding, bias, and chance (cf. Clarke et al. 2014, p. 343). In this case, the studies of cancer in humans may establish not only the existence of a correlation in humans, but also the existence of an appropriate mechanism. Indeed, this seems to be the account taken by IARC: 'On the basis of the large amount of data and the consistent associations of colorectal cancer with consumption of processed meat across studies in different populations, which make chance, bias, and confounding unlikely as explanations, the majority of the Working Group concluded that there is sufficient evidence in human beings for the carcinogenicity of the consumption of processed meat' (Bouvard et al. 2015, p. 1599).

In this paper, we focus on a different problematic case for the natural account. In particular, there is the case of benzo[a]pyrene, a polycyclic aromatic hydrocarbon resulting from the incomplete combustion of organic material which is found in certain foods, coal tar, tobacco smoke, among other things (IARC 2009, pp. 111, 112). Benzo[a]pyrene was evaluated as carcinogenic to humans in the absence of sufficient evidence of carcinogenicity from the studies of cancer in humans (IARC 2009) (cf. Baan et al. 2009, p. 1143). This case is not an anomaly because it is consistent with the procedures employed by IARC (2015, p. 22). In this case, the evidence from the studies in humans was not only inadequate to establish the relevant causal claim, it was also inadequate to establish even a correlation between benzo[a]pyrene and cancer in humans. According to the natural account, this is a case in which an exposure is classified as carcinogenic to humans in the absence of probabilistic evidence. As a result, it looks like a counterexample to the Russo–Williamson thesis.

Again, it may be responded that exposure to benzo[a]pyrene has not been appropriately evaluated. Alternatively, the proponent of the thesis may make an analogous change to the natural account: in the benzo[a]pyrene case, it may be that the strength of the mechanistic and other relevant data was such that it established the existence of the appropriate mechanism linking benzo[a]pyrene and cancer and at the same time ruled out the possibility of masking. In this case, the mechanistic and other relevant data are claimed to be sufficient to establish not only that there exists an appropriate mechanism linking benzo[a]pyrene and cancer, but also that they are appropriately correlated. However, the problem with this response is that biomedical mechanisms are so complex that it is unlikely that it is possible to establish an appropriate correlation between a putative cause and effect even if it has been established that they are linked by an appropriate mechanism (Howick 2011, pp. 140–146). It is unlikely that the mechanistic and other relevant data may alone establish that an exposure causes cancer in humans, because it alone cannot establish an appropriate correlation between

the exposure and cancer in humans. As a result, there seems to be a case in which a causal claim about the carcinogenicity of benzo[a]pyrene exposure to humans is established on the basis of the strong mechanistic data, but where these mechanistic data fail to establish that there exists an appropriate correlation in humans, namely, the benzo[a]pyrene case. It would be good for the proponent of the Russo–Williamson thesis to have some other response.

3 Mechanisms and extrapolation

An alternative response may begin by pointing out that although there was no evidence from studies of cancer in humans, the overall evaluation was based upon more than simply the strong mechanistic and other relevant data that had established an appropriate mechanism linking benzo[a]pyrene exposure and cancer in humans. In particular, there was also evidence from the studies of cancer in experimental animals, which was deemed sufficient to establish a correlation between benzo[a]pyrene and cancer in experimental animals (IARC 2009, pp. 112–131). The alternative response maintains that this further evidence is also sufficient to establish an overall correlation between benzo[a]pyrene and cancer in humans, at least when taken with the mechanistic and other relevant data. As a result, the benzo[a]pyrene case would not present a counterexample to the Russo–Williamson thesis, because it is not a case in which a causal claim in humans is established without the appropriate correlation established in humans.

How exactly is the evidence from the studies of cancer in experimental animals supposed to help establish the appropriate correlation in humans? The proposal here is that the established correlation between benzo[a]pyrene and cancer in experimental animals may be carried over to humans by means of extrapolation. An influential suggestion is the mechanism-based approach to extrapolation: 'the mechanisms approach to extrapolation suggests that knowledge of mechanisms and factors capable of interfering with them can provide a basis for extrapolation' (Steel 2008, p. 85). According to this approach, a claim about an experimental animal model may be extrapolated to humans only if there is some knowledge that the mechanisms which support the causal relation of interest in the animal models are sufficiently similar to those in humans, where the mechanisms are sufficiently similar only if there are no differences in the mechanisms that would result in a difference in the outcome of interest produced by the mechanisms. Notice that this means that the requirements for mechanistic similarity are more stringent the more specific the outcome that is being extrapolated (Steel 2008, pp. 93, 94). In other words, extrapolating a precise quantitative claim about the extent to which an exposure increases the risk of cancer requires closer mechanistic similarity than extrapolating the qualitative claim that the exposure is a cancer hazard.

Leuridan and Weber (2011, p. 97) claim that 'it is clear that the IARC procedures do take into account the role that information about mechanisms can play in extrapolating results from animal experiments to humans'. Indeed, this claim seems to be supported by the explanation of the role of mechanistic data that is provided by IARC:

> In an evaluation of cancer bioassays, mechanistic studies can provide data to address questions about animal-to-human similarities and differences. This

> implies sufficient data to identify the mechanisms contributing to tumor induction in experimental animals and to determine whether those mechanisms can also operate in humans. …Since 1991, IARC has also allowed an agent to be classified as carcinogenic to humans (Group 1) when there is less than sufficient evidence in humans …but there is sufficient evidence in experimental animals and "strong evidence in exposed humans that the agent acts through a relevant mechanism of carcinogenicity" (Cogliano et al. 2008, pp. 101–103).

In addition, the mechanisms approach to extrapolation also seems to be endorsed in the IARC *Preamble*: 'Mechanistic and other relevant data may provide evidence of carcinogenicity and also help in assessing the relevance and importance of findings of cancer in animals and in humans' (IARC 2015, p. 15). Given this, there is reason to believe that mechanism-based extrapolation sometimes plays a role in the overall evaluation of exposures by IARC.

Importantly, there is also reason to believe that extrapolation played a role in the overall evaluation of benzo[a]pyrene. Here is the rationale provided in support of this evaluation:

> In making the overall evaluation, the Working Group took the following into consideration: The strong and extensive experimental evidence for the carcinogenicity of benzo[a]pyrene in many animal species, supported by the consistent and coherent mechanistic evidence from experimental and human studies provide biological plausibility to support the overall classification of benzo[a]pyrene as a human carcinogen (Group 1) (IARC 2009, p. 138).

Given all this, a case may be made in favour of claim that extrapolation took place in the benzo[a]pyrene case, and that a correlation between benzo[a]pyrene and cancer was established in humans on this basis. Indeed, elsewhere the following more detailed rationale of the benzo[a]pyrene case is provided:

> The most widely investigated polycyclic aromatic hydrocarbon is benzo[a]pyrene, which induces tumours in mice, rats, guinea pigs, hamsters, rabbits, monkeys, newts, and ducks. In mice, strong evidence shows that benzo[a]pyrene causes lung tumours through the diolepoxide mechanism, and skin tumours through the diolepoxide and radical-cation mechanisms. Important steps of these mechanisms, including the complete activation pathways, have been reported in individuals exposed to polycyclic aromatic hydrocarbons. G to T transversions in the *Kras* proto-oncogene identified in lung tumours from mice treated with benzo[a]pyrene are causally associated with formation of DNA adducts derived from the diolepoxide. Human beings exposed to benzo[a]pyrene activate this molecule metabolically to diolepoxides that form DNA adducts. One of these adducts has been measured in chimney sweepers and coke-oven workers, who are frequently exposed to mixtures of polycyclic aromatic hydrocarbons that contain benzo[a]pyrene. Similar mutations in *KRAS* were found in lung tumours from non-smokers exposed to coal combustion products rich in polycyclic aromatic hydrocarbons containing benzo[a]pyrene (Straif et al. 2005, p. 931).

In other words, two types of mechanisms are identified in the experimental animal models: the diolepoxide mechanism, and the radical-cation mechanism. In the former, intermediate metabolites of benzo[a]pyrene react with DNA to form DNA adducts associated with tumorigenesis. In the latter mechanism, benzo[a]pyrene is oxidized resulting in a free radical that similarly forms DNA adducts. And the mechanistic data from humans provide evidence that similar mechanisms are activated by benzo[a]pyrene in humans (2009, pp. 131–137). This led to the conclusion that benzo[a]pyrene exposure is also correlated with cancer in humans. As a result, the benzo[a]pyrene case would not present a counterexample to the Russo–Williamson thesis: it is not a case in which a causal claim in humans is established without an established correlation in humans.

The problem with this response is that it works only if the correlation between benzo[a]pyrene and cancer is *established* in humans by extrapolation on the basis of the evidence provided by the studies of cancer in experimental animals and the knowledge of shared mechanisms. This turns on the question of whether the extrapolation from animal models to humans is reliable. Some have argued that there are always salient differences in the relevant mechanisms, such that no conclusion about humans can be established in this way (LaFollette and Shanks 1995; Howick et al. 2013). It may be that the established carcinogenicity of benzo[a]pyrene to experimental animals, together with evidence of relevant mechanistic similarities in the animal models and humans, is still not sufficient to establish the carcinogenicity of benzo[a]pyrene to humans. A precondition of mechanism-based extrapolation is that the mechanisms in question are relatively insensitive to their causal context, as well as the mechanism exhibiting a certain degree of modularity such that local changes in component properties do not ramify intractably to other parts of the mechanism (Steel 2008; Andersen 2012). If these conditions are not met, even slight differences in parts of a mechanism or its context might lead to significant differences in the mechanisms' output, thus defeating extrapolation. These are problems related to the problem of extrapolation in heterogeneous populations (Steel 2008). In addition, it has been argued that there are further distinct quality problems in experimental animal research—the sample sizes are often small, many of the studies are not randomized nor blinded, results are not adequately reported, and the genetic homogeneity of experimental animals fails to reflect the heterogeneity of natural human populations (Hackam 2007).

If it is not possible to reliably extrapolate the correlation between benzo[a]pyrene exposure and cancer from experimental animal models to humans, then the appeal to the role of extrapolation does little to defend the Russo–Williamson thesis. *Is there any reason to believe that the extrapolation in the benzo[a]pyrene case is sufficiently reliable that the correlation is established in humans?*

4 Robustness analysis

In this section, we provide some reasons to believe that the conclusion that there exists a correlation between benzo[a]pyrene exposure and cancer in humans was in fact established by mechanism-based extrapolation. The reasons involve an appeal to the robustness of the experimental animal evidence.

Robustness analysis is a method for studying the reliability of various types of scientific inferences, and robustness itself has been defined in relation to robustness analysis in the following way:

> To analyse a variety of independent derivation, identification, or measurement processes. 2. To look for and analyse things that are invariant over or identical in the conclusions or results of these processes. 3. To determine the scope of the processes across which they are invariant and the conditions on which their invariance depends. 4. To analyse and explain any relevant failures of invariance. I call things that are invariant under this analysis robust (Wimsatt 2007, p. 44).

Here, the term "process" is an umbrella term that may refer to experimental procedures, models, or theoretical descriptions of phenomena (see Wimsatt 2007, pp. 45–46). The epistemic function often attributed to robustness analysis is that of identifying results that are independent of particular modelling assumptions or detection methods and experimental set-ups (Weisberg 2006; Kuorikoski and Marchionni 2016). For example, mathematical models inevitably involve some unrealistic assumptions made solely for the tractability of the model. A method is required for distinguishing the results that depend on empirically verifiable core assumptions of the model from results that crucially depend on the unrealistic assumptions. One way to do this is to systematically vary the model specification and search for results that persist against such modifications. It is these robust results that are most likely to be applicable to the model's real world target, as they are not "infected" by the unrealistic assumptions to the same degree as more fragile results.

The interest here is in empirical evidence that comes from heterogeneous experimental sources. In believing a conclusion drawn from empirical evidence, it is assumed that the evidence is a reliable symptom of the underlying target phenomenon in the following sense: (1) The evidence is caused by the target phenomenon through its interaction with the detection devices and other causal features of the experimental set-up; (2) It is possible to correctly distinguish the contribution of the underlying phenomenon from the contribution of the causal background of the experimental set-up. But this is not a trivial assumption, since experiments may be unreliable for a number of reasons: it may be that the causal context of a laboratory or the methods used to prepare the experimental system interfere with the phenomenon of interest, or detection techniques pick up background noise instead of tracking the target phenomenon. To determine that the evidence tracks the phenomenon of interest, the evidence-generating processes and their error characteristics may be studied in order to come up with a detailed explanation of the experimental results. If the true explanation of the outcome of an experiment is that causal factors idiosyncratic to the experimental set-up drown out the contribution of the target phenomenon, then there is reason to believe that conclusions drawn from those results about the underlying phenomenon could be false.

The problem is that an extensive empirical analysis of the evidence-generating process is not always possible in practice. Typically, even in controlled experiments, there is less than perfect knowledge of how exactly some result was produced. As a result, there is room for competing explanations of the result, some of which implicate that the evidence fails to track the phenomenon of interest. This is where robustness analysis is purported to help: By analysing the robustness of evidence, the plausibility of

these alternative explanations may be lowered, in this way increasing the plausibility that inferences from the evidence to claims about the underlying phenomenon are in fact reliable.

The processes subject to robustness analysis here are the concrete causal processes by which items of empirical evidence are produced. Robust evidence consists of those results that are determined to be invariant in the output of these processes. The crucial precondition for attributing robustness is that these processes are sufficiently independent of each other. Consider a phenomenon that can be elicited and detected by many causally dissimilar experimental set-ups and detection methods. If each of the individual experiments consists of a causal mechanism for producing a result, and each of these mechanisms is sufficiently dissimilar to the others, then it is plausible that the contribution of the experimental context and detection methods themselves is different in each experiment. In other words, each experiment will likely produce artefacts that are characteristic to that experiment alone. Given that the artefacts caused by the different experimental set-ups are likely to vary, whatever remains stable in the results of the many experiments is likely to reflect something that is a common causal component to each experimental set-up, namely, the underlying phenomenon. Robust results are less likely to be explained by the idiosyncrasies of individual experiments. By eliminating or discriminating against hypothetical alternative explanations of the robust result, robustness analysis may be able to determine that a particular inference process is in fact reliable (cf. Schupbach 2016).

How might robustness reasoning help to determine the reliability of a mechanism-based extrapolation? In the benzo[a]pyrene case, it seems that the basic components for running a robustness argument are in place. Firstly, there is a phenomenon in the world, namely, the correlation between benzo[a]pyrene exposure and cancer. Secondly, there are multiple lines of evidence pointing to the same result, generated using varying methods in causally dissimilar experimental systems, namely, the experimental animals. The further requirement that the sources of evidence be independent of each other requires some elaboration. Experimental animals are clearly not causally independent of each other in one sense—all animals share a common ancestor somewhere in the phylogenetic tree, and their traits have evolved from the traits of this ancestor species. But this is exactly what one wants: it is this evolutionary connectedness that explains the plausibility that humans share the mechanisms of interest with the experimental animals. In addition, it is equally clear that divergent evolution has created significant physiological differences between species. A result from an experimental animal counts as robust only if it is reproducible in sufficiently divergent groups of species of experimental animals, animals that differ from each other in significant parts of physiology, as well as being dissimilar to humans.

Here, the main concern in extrapolating from animal models is that given evolved differences in physiology, a result in an experimental animal might be due to some mechanism idiosyncratic to that experimental animal. The function of robustness analysis here is to identify results that, in virtue of their robustness across various experimental animals, are less likely to be explained in terms of species-specific mechanisms. The more stable a result is across varying experimental animals, the less likely it is that it is due to a mechanism that is an evolutionary novelty of a particular experimental animal. By discriminating against explanations of the experimental animal results in terms

of species-specific physiology, and given the plausible assumption that some mechanisms are shared between many species, the robustness of a result in experimental animals can boost the plausibility that the result is established on the basis of mechanism-based extrapolation. In other words, robustness of results across different experimental animals can make it more likely that a mechanism-based extrapolation is reliable, even though these experimental animals are in ways disanalogous to humans. As a result, this helps to address the problem of extrapolation in heterogeneous populations.

In the benzo[a]pyrene case, a correlation between benzo[a]pyrene exposure and cancer had been demonstrated in a variety of experimental animals. Indeed, the relevant IARC *Monograph* reports carcinogenicity of benzo[a]pyrene in eight species of non-human model animals, in many cases including several different strains of single species (IARC 2009, pp. 112–131). It is this robustness which suggests that a correlation in a particular experimental animal is not the result of some species-specific mechanism. As a result, this boosts the plausibility that a correlation between benzo[a]pyrene exposure and cancer has been established in humans by mechanism-based extrapolation. But it is not that the robustness of the result in experimental animals is alone sufficient to establish the causal claim in humans. Establishing the causal claim also required a reliable mechanism-based extrapolation. Instead, the robustness analysis provides some reason to believe that this mechanism-based extrapolation is in fact reliable.

5 Conclusion

The Russo–Williamson thesis maintains that establishing a causal claim in medicine requires both probabilistic and mechanistic evidence. In motivating their thesis, Russo and Williamson appeal to the practice of IARC. A closer inspection of this practice suggests some problems for this motivation. In particular, we have pointed out the benzo[a]pyrene case, in which it looks like the carcinogenicity of an exposure in humans is established without probabilistic evidence in humans. In response, we have argued that the benzo[a]pyrene case is in fact consistent with the Russo–Williamson thesis. In the benzo[a]pyrene case, rather than considering probabilistic evidence gleaned from studies on humans, the correlation between benzo[a]pyrene exposure and cancer in humans was established by mechanism-based extrapolation from experimental animal models. However, this response works only if the correlation between benzo[a]pyrene exposure and cancer is in fact established in humans by mechanism-based extrapolation. As a result, the question then turns into one about the reliability of this extrapolation. In this paper, we have argued that there are reasons to believe that this extrapolation is reliable. The reasons are provided by appealing to robustness analysis, where robustness refers to the stability of an experimental result across heterogeneous detection methods and experimental systems. In the benzo[a]pyrene case, the robustness of the correlation in experimental animals, together with the mechanistic data, makes plausible the reliability of this mechanism-based extrapolation. In conclusion, the practice in the benzo[a]pyrene case is consistent with the Russo–Williamson thesis.

Acknowledgements Thanks to Holly Andersen, John Campbell, Brendan Clarke, Leen De Vreese, Donald Gillies, Phyllis Illari, Federica Russo, Kurt Straif, Erik Weber, and Jon Williamson. Thanks also to audiences

at Aarhus University, the University of Belgrade, the University of Kent, and the University of Lisbon. This work was funded by the Arts and Humanities Research Council and the Leverhulme Trust.

References

Andersen, H. (2012). Mechanisms: What are they evidence for in evidence-based medicine? *Journal of Evaluation in Clinical Practice*, *18*(5), 992–999.

Baan, R., Grosse, Y., Straif, K., Secretan, B., Ghissassi, F. E., Bouvard, V., et al. (2009). A review of human carcinogens—Part F: Chemical agents and related occupations. *The Lancet Oncology*, *10*(12), 1143–1144.

Bouvard, V., Loomis, D., Guyton, K., Grosse, Y., Ghissassi, F. E., Benbrahim-Tallaa, L., et al. (2015). Carcinogenicity of consumption of red and processed meat. *The Lancet Oncology*, *16*(16), 1599–1600.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, *33*, 339–360.

Cogliano, V. J., Baan, R. A., Straif, K., Grosse, Y., Secretan, B., & Ghissassi, F. E. (2008). Use of mechanistic data in IARC evaluations. *Environmental and Molecular Mutagenesis*, *49*, 100–109.

Gillies, D. (2011). The Russo–Williamson thesis and the question of whether smoking causes heart disease. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 110–124). Oxford: Oxford University Press.

Hackam, D. G. (2007). Translating animal research into clinical benefit. *British Medical Journal*, *7586*, 163.

Hill, A. B. (1965). The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine*, *58*, 295–300.

Howick, J. (2011). *The philosophy of evidence-based medicine*. London: BMJ Books.

Howick, J., Glasziou, P., & Aronson, J. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics*, *34*(4), 275–291.

IARC. (2009). *IARC monographs on the evaluation of carcinogenic risks to humans: A review of human carcinogens chemical agents and related occupations* (Vol. 100F). Lyon: International Agency for Research on Cancer.

IARC. (2015). *IARC monographs on the evaluation of carcinogenic risks to humans: Preamble*. Lyon: World Health Organization International Agency for Research on Cancer.

Illari, P. (2011). Mechanistic evidence: Disambiguating the Russo–Williamson thesis. *International Studies in the Philosophy of Science*, *25*, 139–157.

Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science*, *83*(2), 227–247.

LaFollette, H., & Shanks, N. (1995). Two models of models in biomedical research. *The Philosophical Quarterly*, *45*(179), 141–160.

Leuridan, B., & Weber, E. (2011). The IARC and mechanistic evidence. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 91–109). Oxford: Oxford University Press.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, *21*, 157–170.

Schupbach, J. N. (2016). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axw008.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

Straif, K., Baan, R., Grosse, Y., Secretan, B., Ghissassi, F. E., & Cogliano, V. (2005). Carcinogenicity of polycyclic aromatic hydrocarbons. *The Lancet Oncology*, *6*, 931–932.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730–742.

Wimsatt, W. (2007). *Re-engineering philosophy for limited beings. Piecewise approximations to reality*. Cambridge: Harvard University Press.