Climatic Change (2011) 108:693–700
DOI 10.1007/s10584-011-0190-x

Differentiating theory from evidence in determining confidence in an assessment finding

Kristie L. Ebi

Received: 15 April 2011 / Accepted: 1 July 2011 / Published online: 9 August 2011

Abstract The *Guidance Notes for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties* describes a process for consistently evaluating and communicating levels of certainty in findings. The process begins with an assessment of the scientific evidence and agreement supporting a finding, where evidence is defined as including mechanistic understanding, theory, data, models, and expert judgment. The appropriateness of categorizing theory as one line of evidence varies by scientific discipline; for the natural and social sciences, developing theory and collecting data are different steps in the scientific method. Further, decision-makers often find it valuable for scientists to differentiate situations where a theory is generally agreed but for which supporting data are limited, from situations where empirical data lack an explanatory theory. The paper describes the approach used by the International Agency for Research on Cancer (IARC) for assessing the relative robustness of a theory separately from the strength and quality of its supporting evidence, and then developing consensus statements of whether an agent is a human carcinogenic. Although the IARC and IPCC processes are very similar, the IARC process also differs by combining theory, evidence, and agreement as equal partners in a limited set of standardized categories of confidence. Incorporating aspects of the IARC approach into the IPCC guidance could improve the evaluation and communication of theory, evidence, and agreement in future versions of the uncertainty guidance.

Keywords Uncertainty guidance · Theory · Evidence

1 Introduction

Decision-makers want to understand not only the key findings from scientific assessments, but also the confidence the author team and broader scientific community have in those findings based on the underlying theory, evidence, and agreement. The *Guidance Notes for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties* (Mastrandrea et al. 2010) describes a process for consistently evaluating and communicating levels of certainty in the findings of the IPCC 5th Assessment Report

K. L. Ebi (✉)
Carnegie Institution for Science, 260 Panama Street, Stanford, CA 94305, USA
e-mail: krisebi@stanford.edu

(AR5). Decision-makers use the key findings and their associated level of confidence as a basis for prioritizing appropriate strategies, policies, and measures to iteratively manage the risks of climate change.

The process proposed for the AR5 assessment assesses the validity of a finding "based on the type, amount, quality, and consistency of evidence (e.g., mechanistic understanding, *theory,* data, models, expert judgment) and the degree of agreement" (Mastrandrea et al. 2010). However, the appropriateness of categorizing theory as one line of evidence varies by scientific discipline. Theories are explanations of observations and experimental results, along with the rules that describe relationships within and among observations and results. Theories provide an explanatory framework for evaluation of evidence. Empirical data can be used to support, modify, or reject theories and hypotheses. Clearly, there is greater certainty in findings, and the policy responses based on findings, when there is a robust theory, multiple lines of independent evidence, and high agreement in the scientific community. Challenges include cases where there is a theory with limited evidence (e.g. climate change makes a significant contribution to migration), and where evidence lacks an explanatory theory or has multiple, competing theories (e.g. causes of observed changes in the geographic range of a pest or pathogen). Climate change poses a particular challenge because confidence is evaluated for projected future events where there are, by definition, no empirical data; although there are results supported by data, theory, and agreement that link climate variables to changes in impacts of interest.

In the physical sciences, extensive observations and accumulated knowledge yield very high confidence that understanding of system structure, processes, and basic relationships will not change in the future. Therefore, theories based on these relationships could be considered as one line of evidence when evaluating confidence in projections. However, for the natural and social sciences, categorizing theory as a type of evidence conflates different steps in the scientific method. There are many examples in which theories draw on basic principles (such as more developed societies will be better prepared for future climate change) but evidence is limited or will not be available in the short term; and there are also many examples in which evidence is available (such as community-based adaptation increases resilience to climate variability) but is supported by limited theory (in this case, on how to scale up community-based adaptation beyond the local scale). Differentiating situations where a theory is generally agreed but for which supporting data are limited (suggesting that collecting additional evidence may reduce uncertainty), from situations where empirical data lack an explanatory theory (suggesting that increasing understanding of the underlying processes and relationships may be needed to explain observations) is valuable for decision-makers and, therefore, desirable for authors of assessments that inform their decisions. The complexity of human, natural, and social systems makes it much more difficult to identify clear and consistent evidence of trends that can be attributed to climate change (or another driver of the outcome).

The well-established process used by the International Agency for Research on Cancer (IARC) to develop consensus statements of whether an agent is a human carcinogenic is an approach for separately assessing theory and evidence, and then assigning confidence (IARC 2006). IARC differentiates between empirical data from human studies and theories developed from cellular and animal studies when reaching on overall conclusion about human carcinogenicity. Cellular and animal studies cannot provide sufficient evidence to conclude that an agent is carcinogenic (to humans) because humans may respond significantly differently when exposed to the agent. Routes of exposure, absorption, mechanism, and metabolism can all influence the response to varying degrees. There is a large literature base on the similarities and differences between how animals and humans metabolize specific compounds, and on whether animal results are relevant to assessing human carcinogenicity (e.g. Ennever et al. 1987). Successfully predicting human carcinogenicity requires cellular and animal studies to be both sensitive (few false negatives) and specific (few false positives). Therefore, cellular and animal studies are used to develop theories of how a specific agent could act in humans; these theories then need to be tested through human observational and experimental studies.

2 International Agency for Research on Cancer

IARC was founded in 1965 to provide government authorities with expert, independent, scientific opinion on the causes of human cancer (IARC 2006). The process followed is similar to that used for an IPCC assessment. International working groups of experts are convened to examine all relevant information on the evidence of carcinogenicity of a specific agent in order "to assess the strength of the available evidence that an agent could alter the age-specific incidence of cancer in humans." The results are generally published in Monographs that evaluate cancer hazards, where hazard is defined as an agent that is capable of causing cancer under some circumstance, and cancer risk is an estimate of the carcinogenic effects expected from exposure to a cancer hazard. Similar to IPCC assessments, IARC Monographs are used by national and international authorities in risk assessments, to formulate decisions concerning preventive measures, to provide effective cancer control programs, and to decide among alternative options for public health decisions. The evaluations of IARC working groups are scientific judgments of the evidence for or against carcinogenicity based on the available data. Just as in IPCC assessments, no recommendations are made with regard to regulation or legislation (i.e. the evaluations are not policy prescriptive); these are the responsibility of individual governments or other international organizations (they are policy relevant)

The working groups review all pertinent epidemiological studies, cancer bioassays in experimental animals, and mechanistic and other relevant data. Studies reviewed have to be published or accepted for publication in the "openly available scientific literature." Data from government agency reports that are publicly available also are considered.

The members of a working group for a particular agent are selected on the basis of knowledge and experience (experts generally have published significant research related to the carcinogenicity of the agents being reviewed), and an absence of real or apparent conflicts of interest. Consideration is given to demographic diversity and balance of scientific findings and views. Each participant serves as an individual scientist and not as a representative of any organization, government, or industry.

The IARC process differs from the IPCC in that major data are collected and working papers are prepared on key topics (e.g. chemical and physical properties, analysis, production and use, and occurrence) under a separate contracts and sent to the members of a working group 6 months before a meeting. During that time, the members of the working group may prepare additional working papers. The working group then meets for 7 to 8 days to discuss and finalize the text and formulate the evaluations. Consensus on the text and summary evaluations are finalized in plenary. Therefore, the entire Monograph (and not just individual chapters) is the joint product of the entire working group. The working groups strive to achieve a consensus evaluation that reflects broad agreement among the members, but unanimity is not necessary (although dissenting opinions are rare).

IARC developed four standard terms for evaluations of the strength of evidence for carcinogenicity arising from human and experimental animal data, and for the strength of mechanistic evidence.[1]

- There is *sufficient evidence of carcinogenicity* in humans when a causal relationship has been established between exposure to the agent and human cancer; chance, bias, and confounding can be ruled out with reasonable confidence.
- There is *limited evidence of carcinogenicity* when a positive association has been observed between exposure to the agent and cancer for which a causal interpretation is considered to be credible, but chance, bias or confounding could not be ruled out with reasonable confidence.
- There is *inadequate evidence of carcinogenicity* when available studies are of insufficient quality, consistency, or statistical power to permit a conclusion regarding the presence or absence of a causal association between exposure and cancer, or no data on cancer in humans are available.
- There is *evidence suggesting lack of carcinogenicity* when there are several adequate studies covering the full range of levels of exposure that humans are known to encounter that are consistent in not showing a positive association between exposure to the agent and any studied cancer at any observed level of exposure. The results from these studies alone or combined should have narrow confidence intervals, and bias and confounding should be ruled out with reasonable confidence.

Evidence of carcinogenicity in experimental animals is classified into similar categories. The terms 'weak', 'moderate' or 'strong', and an assessment of whether a particular mechanism is likely to be operative in humans, are used to evaluate the strength of evidence that any carcinogenic effect observed is due to a particular mechanism. Data on humans or biological specimens obtained from exposed humans provide the strongest indications that a particular mechanism operates in humans.

The IARC process then combines theory, evidence, and degree of agreement into a summary evaluation. The reasoning for an evaluation integrates the major findings from studies of cancer in humans, studies of cancer in experimental animals, and mechanistic and other relevant data. The summary includes concise statements of the principal line(s) of argument that emerged, the conclusions of the working group on the strength of the evidence for each group of studies, citations to indicate which studies were pivotal to these conclusions, and the reasons for any differential weighting of data. On the rare occasions when there are significant differences of scientific interpretation, a brief summary of the alternative interpretations is provided, together with their scientific rationale and an indication of the relative degree of support for each alternative.

The summary IARC categories are[2]:

Group 1: The agent is *carcinogenic to humans*.

This category is used when there is *sufficient evidence of carcinogenicity* in humans. Exceptionally, an agent may be placed in this category when evidence of carcinogenicity in humans is less than *sufficient* but there is *sufficient evidence of carcinogenicity* in experimental animals and strong evidence in exposed humans that the agent acts through a relevant mechanism of carcinogenicity.

[1] Text taken verbatim from IARC 2006.

[2] Text taken verbatim from IARC 2006.

Group 2 This category includes agents for which, at one extreme, the degree of evidence of carcinogenicity in humans is almost *sufficient*, as well as those for which, at the other extreme, there are no human data but for which there is evidence of carcinogenicity in experimental animals. Agents are assigned to either Group 2A (*probably carcinogenic to humans*) or Group 2B (*possibly carcinogenic to humans*) on the basis of epidemiological and experimental evidence of carcinogenicity and mechanistic and other relevant data. The terms *probably carcinogenic* and *possibly carcinogenic* have no quantitative significance and are used simply as descriptors of different levels of evidence of human carcinogenicity, with *probably carcinogenic* signifying a higher level of evidence than *possibly carcinogenic.*

Group 2A: The agent is *probably carcinogenic to humans.*

This category is used when there is *limited evidence of carcinogenicity* in humans and *sufficient evidence of carcinogenicity* in experimental animals. In some cases, an agent may be classified in this category when there is *inadequate evidence of carcinogenicity* in humans and *sufficient evidence of carcinogenicity* in experimental animals and strong evidence that the carcinogenesis is mediated by a mechanism that also operates in humans. Exceptionally, an agent may be classified in this category solely on the basis of *limited evidence of carcinogenicity* in humans. An agent may be assigned to this category if it clearly belongs, based on mechanistic considerations, to a class of agents for which one or more members have been classified in Group 1 or Group 2A.

Group 2B: The agent is *possibly carcinogenic to humans.*

This category is used for agents for which there is *limited evidence of carcinogenicity* in humans and less than *sufficient evidence of carcinogenicity* in experimental animals. It may also be used when there is *inadequate evidence of carcinogenicity* in humans but there is *sufficient evidence of carcinogenicity* in experimental animals. In some instances, an agent for which there is *inadequate evidence of carcinogenicity* in humans and less than *sufficient evidence of carcinogenicity* in experimental animals together with supporting evidence from mechanistic and other relevant data may be placed in this group. An agent may be classified in this category solely on the basis of strong evidence from mechanistic and other relevant data.

Group 3: The agent is *not classifiable as to its carcinogenicity to humans.*

This category is used most commonly for agents for which the evidence of carcinogenicity is *inadequate* in humans and *inadequate* or *limited* in experimental animals. Exceptionally, agents for which the evidence of carcinogenicity is *inadequate* in humans but *sufficient* in experimental animals may be placed in this category when there is strong evidence that the mechanism of carcinogenicity in experimental animals does not operate in humans. Agents that do not fall into any other group are also placed in this category.

Group 4: The agent is *probably not carcinogenic to humans.*

This category is used for an agent for which there is *evidence suggesting lack of carcinogenicity* in humans and in experimental animals. In some instances, agents for which there is *inadequate evidence of carcinogenicity* in humans but *evidence suggesting lack of carcinogenicity* in experimental animals, consistently and strongly supported by a broad range of mechanistic and other relevant data, may be classified in this group.

3 Relevance of the IARC process for handling uncertainties in IPCC reports

IARC and IPCC are similarly mandated to conduct comprehensive evaluations of scientific research to identify key findings and to summarize confidence in those findings. The processes differ in several ways, including that the IARC process combines theory, evidence, agreement, and confidence as equal criteria when determining certainty in key findings; the evaluations are summarized into a limited set of standardized categories. This is possible in part because the IARC process focuses only on the question of cancer causality and not on the wide range of causalities and confounding factors that can influence many IPCC conclusions. The wide diversity of issues covered in an IPCC assessment makes it unlikely that such a limited set of categories could be agreed for all confidence statements. However, there are situations, such as findings on detecting and attributing a trend to anthropogenic climate change, where it may be possible for the scientific community to agree on a limited set of well-defined categories for confidence in key findings (such as established, probable, possible, and uncertain, with definitions comparable to those used by IARC), or categories similar to the four qualitative state of knowledge descriptors used for the IPCC Third Assessment Report (Table 1; Moss and Schneider 2000). Consistent application of agreed categories, along with accompanying explanations of the principal lines of evidence, would be a useful step in helping decision makers understand the degree and sources of certainty.

The IARC process also offers an approach to assessing theory, evidence, and agreement that, if applied in the next version of the IPCC uncertainty guidance, would provide greater clarity on the sources of certainty in findings. There are advantages within the natural and social sciences to communicate not just evidence and agreement, but also the extent to which there is a robust theory underlying the evidence. Adding theory as a third dimension to evidence and agreement would indicate the research needed to reduce major sources of uncertainty and enhance understanding of the strength of scientific support for policy actions.

One challenge is that evidence and agreement in the AR5 guidance are each categorized on a three-point scale (for nine combinations; see Fig. 1 in Mastrandrea et al. (2010). Adding theory as a separate consideration (on a three-point scale) would result in twenty-seven summary combinations; too many to clearly differentiate when subsequently assigning specific confidence levels (when doing so is warranted). Indeed, the AR5 uncertainty guidance does not recommend assigning confidence to all categories of evidence and agreement; it recommends that a level of confidence or quantified measure of

Table 1 Qualitative state of knowledge descriptors used in the uncertainty guidance in the IPCC Third Assessment Report

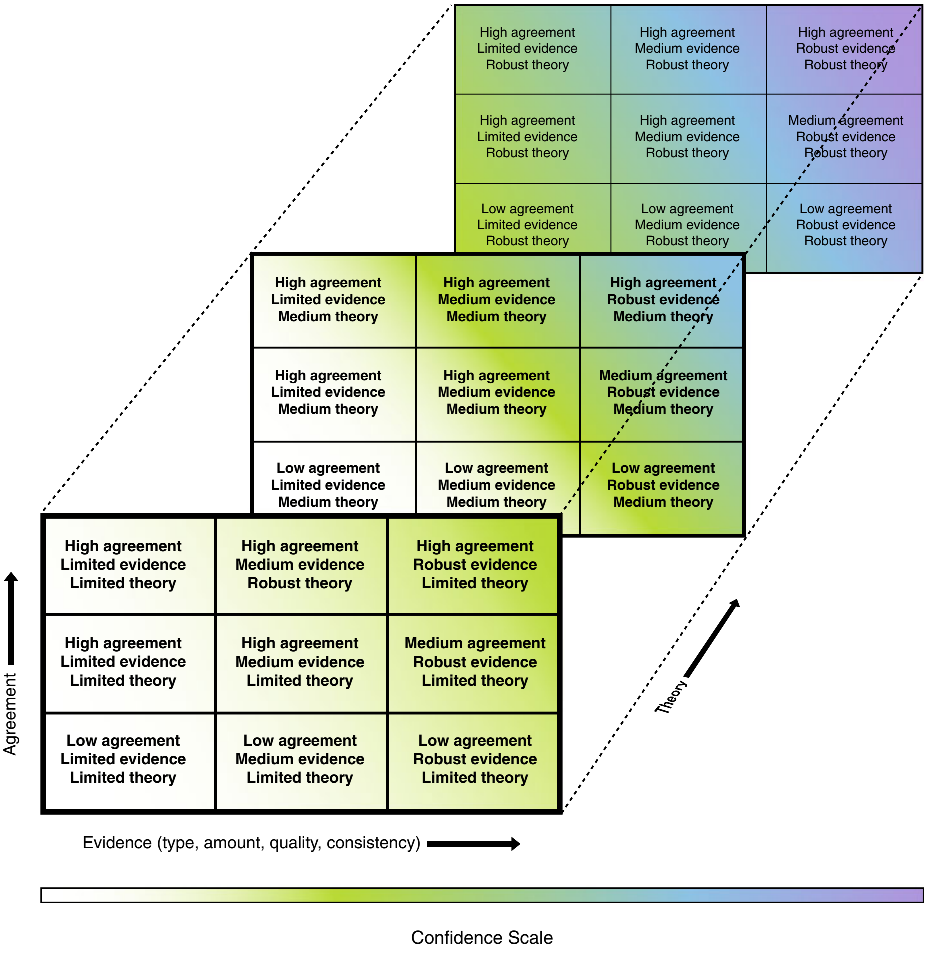
Well-established: models incorporate known processes; observations largely consistent with models for important variables; or multiple lines of evidence support the finding

Established but Incomplete: models incorporate most known processes, although some parameterizations may not be well tested; observations are somewhat consistent with theoretical or model results but incomplete; current empirical estimates are well founded, but the possibility of changes in governing processes over time is considerable; or only one or a few lines of evidence support the finding

Competing Explanations: different model representations account for different aspects of observations or evidence, or incorporate different aspects of key processes, leading to competing explanations

Speculative: conceptually plausible ideas that haven't received much attention in the literature or that are laced with difficult to reduce uncertainties or have few available observational tests

Source: Moss and Schneider 2000



High agreement Limited evidence Robust theory	High agreement Medium evidence Robust theory	High agreement Robust evidence Robust theory
High agreement Limited evidence Robust theory	High agreement Medium evidence Robust theory	Medium agreement Robust evidence Robust theory
Low agreement Limited evidence Robust theory	Low agreement Medium evidence Robust theory	Low agreement Robust evidence Robust theory

High agreement Limited evidence Medium theory	High agreement Medium evidence Medium theory	High agreement Robust evidence Medium theory
High agreement Limited evidence Medium theory	High agreement Medium evidence Medium theory	Medium agreement Robust evidence Medium theory
Low agreement Limited evidence Medium theory	Low agreement Medium evidence Medium theory	Low agreement Robust evidence Medium theory

High agreement Limited evidence Limited theory	High agreement Medium evidence Robust theory	High agreement Robust evidence Limited theory
High agreement Limited evidence Limited theory	High agreement Medium evidence Limited theory	Medium agreement Robust evidence Limited theory
Low agreement Limited evidence Limited theory	Low agreement Medium evidence Limited theory	Low agreement Robust evidence Limited theory

Fig. 1 A depiction of evidence, agreement, and theory and their relationship to confidence. Confidence increases with increasing strength of shading

uncertainty be assigned for findings with robust evidence and high agreement, and for findings with either robust evidence or high agreement (Mastrandrea et al. 2010). That is, assigning confidence is recommended for only three of the nine categories of evidence and agreement. The author team decides whether it is appropriate to assign confidence when there is medium evidence and agreement. Similarly, several of the twenty-seven summary combinations of theory, evidence, and agreement could lead to similar levels of confidence. It would be possible to create a smaller set of combinations to consider when assigning confidence. One possible combination, assuming the five-point confidence scale currently used, is (Figure):

- Robust theory, robust evidence, high agreement. Findings would generally have very high confidence.
- Among theory, evidence, and agreement, at least one is in the highest category and the other(s) in the medium category. Depending on the specifics, findings would generally have high or very high confidence.

- Among theory, evidence, and agreement, at least one is in the highest category and the other(s) in the medium or low category. Depending on the specifics, findings would generally have low or medium confidence.
- Medium theory, evidence, and agreement. Depending on the specifics, findings would generally have medium confidence.
- Other combinations would generally have low or very low confidence.

Whether this or another approach is used in the next IPCC uncertainty guidance, separately assessing theory, evidence, and agreement, and then combing these into an evaluation of confidence, would improve communication of the sources of that confidence and would highlight where additional information could reduce uncertainties to provide a firmer foundation for informed decision-making.

Acknowlegements The author would like to thank Terry Root, Joel Smith, Thomas Wilbanks, Gary Yohe and the reviewers for their very helpful comments. The author also would like to thank Yuka Ostuki Estrada for developing the figure.

References

Ennever FK, Noonan TJ, Rosenkranz HS (1987) The predictivity of animal bioassays and short-term genotoxicity tests for carcinogenicity and non-carcinogenicity in humans. Life Sci Med Res 2:73–78

International Agency for Research on Cancer (IARC) (2006) Preamble. IARC monographs on the evaluation of carcinogenic risks to humans. World Health Organization, IARC, Lyon, France. Available at http://monographs.iarc.fr/ENG/Preamble/index.php, accessed 1 March 2010

Mastrandrea MD, Field CB, Stocker TF, Edenhofer O, Ebi KL, Frame DJ, Held H, Kriegler E, Mach KJ, Matschoss PR, Plattner G-K, Yohe GW, Zwiers FW (2010) Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change (IPCC). Available at http://www.ipcc.ch, accessed 1 March 2010

Moss RH, Schneider SH (2000) Uncertainties in the IPCC TAR: recommendations to lead authors for more consistent assessment and reporting. In: Pachauri R, Taniguchi T, Tanaka K (eds) Guidance papers on the cross cutting issues of the third assessment report of the IPCC world meteorological organization. Geneva, pp 33–51