Position Paper: Values and Argumentation in Environmental Risk Regulation

Peter McBurney
Department of Computer Science
University of Liverpool
Liverpool L69 7ZF U. K.
p.j.mcburney@csc.liv.ac.uk

1. INTRODUCTION

When a new potential environmental or health risk arises, regulatory policy makers typically seek to identify the consequences, costs and benefits of different regulatory options before determining on a course of action [9]. In doing so, they usually rely on scientific or epidemiological studies of the effects of the chemical or activity which poses the potential risk to health or the environment. These studies normally involve the collection and analysis of sample data, analyzed using statistical hypothesis testing procedures. Although such testing procedures are widespread in contemporary western science, and often required by academic journals and peer-reviewers, they are not objective; indeed, they encode subjective relative valuations of the consequences of different alternatives [1, 22]. While this problem has long been recognized, an absence of formal reasoning machinery has precluded any solution to the problem. In this note, I show how recent advances in argumentation theory may be applied to this problem.

2. STATISTICAL HYPOTHESIS TESTS

Statistical inference is not deductively valid: the truth of a statement made about a sample (for example, that the mean of the sample lies within a certain range) provides us with no guarantees of the truth of the same statement when made about the population from which the sample was drawn. This is the case even when we know that the sample was selected randomly from the population. A major achievement of mathematical statistics in the twentieth century was to place bounds on the possibility of error when we infer from sample to population. We still cannot say that statements about the population are true; however, under certain assumptions about the distribution of the variables of interest in the population and about the sampling procedures used, we can say that such statements, when made repeatedly, will only be false at most an estimated percentage of times.

Thus, in the terminology of Jerzy Neyman and Egon Pearson [16], the probability of a Type I error, that of wrongly rejecting an hypothesis of no effect, can be guaranteed (under suitable assumptions) to be less than some pre-determined level α, while that of a Type II error, that of wrongly accepting an hypothesis of no effect, can be guaranteed to be less than another pre-determined level β. Thus, α is the proportion of "false positive" results, and β the proportion of "false negative" results. These two values are called the *critical levels* of the test. The challenge is that for any given sample size, the values of α and β are inversely-related: we cannot reduce both values simultaneously without an increase in the sample size, n. Moreover, the type II error typically depends on which alternative hypothesis is in fact true: the value of β will be a function of the difference between the true alternative and the hypothesis of no effect, typically declining as this difference increases.[1]

So, at what levels should we set α and β? A rational determination of these two error bounds would take into account the consequences of each type of error, relative to the costs of undertaking samples of different sizes. Indeed, Neyman and Pearson in their original paper [15] refer explicitly to determining the error bounds based on consideration of error consequences. This idea was taken up most prominently in the statistical decision theory of Abraham Wald [24], and applied to industrial quality control applications, where quantification of the consequences of inference errors is usually straightforward. However, the primary application considered by Neyman and Pearson was not industrial quality control, but scientific experiments, and here the approach they adopted was an informal consideration of the consequences of inference errors: If the null hypothesis is the hypothesis of no scientific effect, then it is more important (they argued) not to reject it wrongly than to accept it falsely, i.e. better to err on the side of knowledge-revision-caution than to wrongly assert evidence for the presence of scientific causal mechanisms where there are none. Such an approach leads to the setting of α at low levels (typically 5% or 1%), and, for a given sample size, choosing an hypothesis-testing procedure which minimizes β. This can result in β being much greater than α. Due its dominance across the sciences in the 75 years since then, we might call this the *standard approach* to determining the error bounds, and the resulting levels of α and β the *standard levels*.

The main application of statistical hypothesis testing in the 1920s and 1930s was for agricultural experiments testing new crop varieties following the post-Great War famines [11], and for these applications, Neyman and Pearson's informal reasoning seems applicable. Indeed, one can view the error bounds from an information-theoretic perspective as acting to control the extent of noise in a scientific communications network [5]: the level of α is an upper bound on the proportion of falsely positive reports circulated by scientists to each other across the network. From this perspective, the standard levels of α and β are set appropriately. Although many scientists now present their work with p-values [18], most biomedical scientists still view the values of 5% and 1% as

decision-thresholds, both for publication decisions and for the revision of the corpus of scientific knowledge. Irwin Bross [5] presents a compelling case why such decision-thresholds are desirable for a scientific communications network, by describing practices in pharmacology before the widespread use of standard hypothesis testing procedures in clinical trials. Moreover, even if scientists use p-values and so avoid themselves making a decision, anyone else using the research results, such as an environmental policy-maker, still needs to assume specific critical levels in order to make a decision.

3. ENVIRONMENTAL REGULATION

Although the standard decision thresholds may be appropriate for tests of scientific hypotheses when there are only scientific consequences, they are not necessarily rational when other consequences are involved. For example, in hypothesis tests undertaken as part of a determination of environmental regulatory policy process, there are usually many more legal, social and economic consequences for each error. Moreover, as Talbot Page argued almost three decades ago [17], in assessing the impacts of chemicals on human health or the environment the consequences of the two types of error may be markedly asymmetric: consequences may differ in their nature, incidence, location, extent, timings, duration, impact and intensity. Indeed, not all these consequences may be negative for those people impacted, as for example when an existing chemical is banned and the manufacturers of substitute products enjoy an increase in demand. Even if the consequences were to be symmetric and equal, those people affected by each may differ greatly in their relative political, economic or social power; society may therefore, or for other reasons, place different value on the impacts falling on the different groups. Using the standard values uniformly across all cases ignores such case-specific detail.

Indeed, dissatisfaction with the use of the standard levels in risk regulation decision-making may motivate much of the recent interest in the Precautionary Principle [4, 19]: it is precisely because scientists and risk regulators have *not* adequately considered the consequences of falsely negative results, proponents argue, that we have suffered serious health and environmental effects from new chemicals and substances. Some (e.g., [12, 20]) have even argued that the consequences of regulation based on false positive results (e.g., imposition of a regulatory burden on an industry when none was required) are invariably far less serious than the consequences of regulation based on false negative results (e.g., illness or death due to use of a chemical wrongly thought to be safe). Such a view argues for a direct reversal of the standard approach, namely for setting β first and at a low level, while accepting possibly much greater levels of α. Proponents of an extreme version of the Precautionary Principle would ban all new technologies unless and until proven safe, thus setting β theoretically at zero.

Both this approach and the standard approach, however, are mistaken in believing that one determination of the critical levels is appropriate for all risk decisions. As Frank Cross [6], among others, has argued, even regulations outlawing chemicals or technologies so as to protect public health may have adverse public health impacts. A rational approach — rational in the sense of seeking to maximize society's overall welfare — would decide the critical levels, and hence the decision thresholds, for risk regulation decisions on a case-by-case basis, taking into account the consequences of the two different errors, and society's respective valuations of these consequences at the time of the decision.

In fact, such a case-by-case determination of hypothesis-testing decision thresholds is in fact what good statistical practice recommends, as described for example in [23]. But such deliberation, if it occurs at all, occurs only informally, and uses no explicit representation of consequences or their valuations. In earlier work [14], the use of formal argumentation was proposed for tackling this problem without identifying any specific formalism. In this position paper, I now identify a specific argumentation framework appropriate to this domain.

4. PROPOSAL

In previous work [10, 2], my colleagues and I have developed a general framework for arguments over proposals for action, in which the values promoted or demoted by the consequences of actions are made explicit. In this framework, a proposal for some action A intended to take the world from the present state R to a specified future state S, where some goal G will be true, the achievement of which will promote a value v, is represented as follows:

$$R \xrightarrow{A} S \models G \uparrow v$$

The goal G is some well-formed formulae which is true in the state S; we do not identify it with S as it may not include all the propositions which are true in S. Our reason for separating G from the value v is to distinguish those elements of the consequences of the action A which are objectively true and therefore can, at least in principle, be verified (namely, goal G), from those elements which involve subjective evaluation of the state brought about by the action A (namely, value v). At present, this framework does not incorporate uncertainty regarding the consequences of actions, or regarding their evaluation.

I propose to use the same representation for comparison and evaluation of the consequences of errors in hypothesis tests. Here we begin with two alternative actions, corresponding to Neyman and Pearson's classical formulation:

- A_1: Rejecting an hypothesis of no effect.
- A_2: Accepting an hypothesis of no effect.

To assess the consequences of each of these actions, we need to recognize that the outcomes of the actions differ according to the true state R of the world when the action is taken. We therefore have four alternative actions whose four separate consequential states need to be assessed:

- $R_1 \xrightarrow{A_1} S_{1,1}$ Rejecting an hypothesis of no effect, when it is in fact false.
- $R_2 \xrightarrow{A_1} S_{2,1}$ Rejecting an hypothesis of no effect, when it is in fact true. (Type I Error)
- $R_1 \xrightarrow{A_2} S_{1,2}$ Accepting an hypothesis of no effect, when it is in fact false. (Type II Error)
- $R_2 \xrightarrow{A_2} S_{2,2}$ Accepting an hypothesis of no effect, when it is in fact true.

Assessment would involve defining the possible relevant consequences (desirable and undesirable) in each state, defining the beneficiaries or victims of these consequences, and then seeking to quantify the magnitude of their benefits or losses. Even without quantification, a qualitative valuation would be possible, to determine which social values are enhanced or demoted by the realization of each consequence. Deciding between the alternative actions would then involve attempting to impose a preference ordering over the different values impacted, as in the value-based argumentation framework of [3].

5. EXAMPLE

I now present an example to illustrate these ideas. In the 1970s and 1980s, the Australian Government faced pressure from Australian veterans for the Second Indochinese War to provide compensation for the effects of exposure to the herbicide "Agent Orange", a mixture of 2,4,5-trichlorophenoxyacetic acid and 2,4--dichlorophenoxyacetic acid. Veterans claimed a number of ill-effects, particularly an increase in the proportion of their offspring born with birth defects. To assess these claims, statistical studies were proposed to compare the proportions of birth defects in the population of veterans with those in the population as a whole, the control population.[2] Here, the hypothesis of no effect is:

H_0: There is no difference in the proportion of offspring born with birth defects in the two populations.

Let us consider each hypothesis-testing action in turn. The first action (A_1) involves rejecting the hypothesis of no effect. What would be the consequences of such an action? One could imagine many consequences, ranging from the banning of the chemical, compensation being paid to the veterans, and efforts expended to minimize or mitigate the health effects believed present. The beneficiaries of these efforts would include the veterans, their offspring, the health and welfare industry, and manufacturers of alternative herbicides. In contrast, the manufacturers and distributors of Agent Orange would presumably suffer from its banning. All of these statements would be true in states $S_{1,1}$ and $S_{2,1}$. But suppose this action were taken wrongly, i.e, suppose we are in state $S_{2,1}$. Then, these beneficiaries would all receive compensation and benefits when they should not.

Similarly, the second action (A_2) involves accepting the hypothesis of no effect. What would be the consequences of this action? We could imagine that the chemical would remain in use, and that veterans and their offspring would receive no compensation or benefits. These statements would be true in states $S_{1,2}$ and $S_{2,2}$. But if this action were taken in error (i.e., we are actually in State $S_{1,2}$), then the real ill-effects of the chemical would not be recognized, and those who suffer these effects would not be compensated.

Which error is to be more preferred, or which more avoided? The answer depends upon one's relative valuations of these consequences. The standard critical levels would put α at say 0.05, with β dependent upon which alternative hypothesis is in fact true. In the case of birth defects in western countries, the background population rate is in fact very small (of the order of 2% of births aggregated across all defects, with much smaller rates for each specific defect), and so the probability of incorrectly accepting the null hypothesis may be high.[3] Let us suppose that $\beta = 0.4$, and that all the outcomes above are realized. We therefore have a 5% chance of veterans receiving compensation for the ill-effects of exposure to Agent Orange when they should not, and a 40% chance of them not receiving compensation when they should. I believe that most citizens in a western democracy would find these relative proportions unjust and unacceptable. Consequently, any democratic determination of the critical levels would likely not use the standard levels.

6. CONCLUSIONS

I have presented a proposal to represent the evaluated consequences of errors in hypothesis testing in environmental risk regulation. This is intended as the first step in the development of a formal argumentation procedure for comparing such evaluated consequences in order to determine appropriate values for the critical levels used in the tests of scientific hypotheses which underpin contemporary risk regulation. By explicit representation of the consequences of these errors and their valuation, such a procedure would assist regulatory decision-makers in balancing competing valuations, and ensure full evidential support for their trade-off decisions. Such procedures are consistent with recent proposals for deliberative democracy in environmental decision-making, e.g., [7, 25].

There are, of course, major challenges involved in implementing such a framework in determination of regulatory policy. One challenge is that of identifying all the consequential outcomes of different errors. Clinical trials were conducted, for example, on both human and animals subjects prior to the commercial release of Thalidomide, but none of these trials involved pregnant subjects [21], presumably because no one thought of the possibility that there may be adverse effects specific to such subjects. The challenge of identifying all possible consequences of proposed actions has received some attention in the Artificial Intelligence community, under the names of *possibilistic risk assessment* [13], although this work is still very preliminary.

The second challenge to case-by-case determination of critical levels is quantification: assessing the likelihoods of different outcomes, assessing their positive and negative impacts, and assessing the valuations (or utilities) that those affected and society would place on these impacts. For most new substances and activities, evidence to support an objective assignment of quantitative values to these variables is scarce or non-existent. Subjective quantification (e.g. assignment of subjective probabilities) is always possible, but that simply magnifies the third challenge, that of reaching agreement between the different parties involved. Different participants are likely to have very different values, and different preference-orderings over values. This will mean that making regulatory decisions using such a formal framework will require policy-makers to strike a balance between the different interests involved, with the result being that any decision is ultimately a political one. However, this is already the case and understood as such in the environmental regulation domain [17, 22].

7. REFERENCES

[1] L. Atkins and D. Jarrett. The significance of "significance tests". In J. Irvine, I. Miles, and J. Evans, editors, *Demystifying Social Statistics*. Pluto Press, London, UK, 1979.

[2] K. Atkinson, T. Bench-Capon, and P. McBurney. Computational representation of practical argument. *Synthese: Knowledge, Rationality and Action*, 2005. *In press*.

[3] T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[4] D. Bodansky. Scientific uncertainty and the Precautionary Principle. *Environment*, 33(7):4–5, 43–44, 1991.

[5] I. Bross. Critical levels, statistical language, and scientific inference. In V. Godambe and D. Sprott, editors, *Foundations of Statistical Inference*, pages 500–513. Holt, Rinehart and Winston, Toronto, Canada, 1971.

[6] F. B. Cross. Paradoxical perils of the Precautionary Principle. *Washington and Lee Law Review*, 53(3):851–925, 1996.

[7] D. J. Fiorino. Environmental risk and democratic process: a critical review. *Columbia Journal of Environmental Law*, 14:501–547, 1989.

[8] J. A. Freiman, T. C. Chalmers, H. Smith, and R. R. Kuebler. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized control trial.

Survey of 71 'negative' trials. *New England Journal of Medicine*, 299(13):690–694, 1978.

[9] J. D. Graham and L. Rhomberg. How risks are identified and assessed. *Annals of the American Academy of Political and Social Science*, 545:15–24, 1996.

[10] K. Greenwood, T. Bench-Capon, and P. McBurney. Towards a computational account of persuasion in law. In G. Sartor, editor, *Proceedings of the Ninth International Conference on AI and Law (ICAIL-03)*, pages 22–31, New York, NY, USA, 2003. ACM Press.

[11] L. Hogben. *Statistical Theory*. W. W. Norton, 1957.

[12] D. T. Hornstein. Reclaiming environmental law: a normative critique of comparative risk analysis. *Columbia Law Review*, 92:562–633, 1992.

[13] P. Krause, J. Fox, P. Judson, and M. Patel. Qualitative risk assessment fulfils a need. In A. Hunter and S. Parsons, editors, *Applications of Uncertainty Formalisms*, Lecture Notes in Artificial Intelligence 1455, pages 138–156. Springer, Berlin, Germany, 1998.

[14] P. McBurney and S. Parsons. Determining error bounds for hypothesis tests in risk assessment: a research agenda. *Law, Probability and Risk*, 1(1):17–36, 2002.

[15] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference, Part I. *Biometrika*, 20A:175–240, 1928. Pages 1–66 of [16].

[16] J. Neyman and E. S. Pearson. *Joint Statistical Papers*. Cambridge University Press, Cambridge, UK, 1967.

[17] T. Page. A generic view of toxic chemicals and similar risks. *Ecology Law Quarterly*, 7 (2):207–244, 1978.

[18] K. J. Rothman and S. Greenland. *Modern Epidemiology*. Lippincott-Raven, Philadelphia, PA, USA, second edition, 1998.

[19] P. Sandin. Dimensions of the Precautionary Principle. *Human and Ecological Risk Assessment*, 5(5):889–907, 1999.

[20] S. Shapiro. Keeping the baby and throwing out the bathwater: Justice Breyer's critique of regulation. *Administrative Law Journal*, 8:721–, 1995.

[21] H. Teff and C. R. Munro. *Thalidomide: The Legal Aftermath*. Saxon House, Westmead, Farnborough, Hampshire, UK, 1976.

[22] J. E. Toll. Elements of environmental problem-solving. *Human and Ecological Risk Assessment*, 5(2):275–280, 1999.

[23] R. Wakeford, K. Binks, and D. Wilkie. Childhood leukaemia and nuclear installations. *Journal of the Royal Statistical Society, Series A*, 152(1):61–86, 1989.

[24] A. Wald. *Statistical Decision Functions*. Wiley, New York, NY, USA, 1950.

[25] T. Webler, S. Tuler, and R. Krueger. What is a good public participation process? Five perspectives from the public. *Environmental Management*, 27(3):435–450, 2001.

Notes

[1] We would expect the probability of wrongly accepting the null hypothesis if it is in fact false to be small if the true alternative is very different from the null hypothesis.

[2] Note that an argument could be made that the appropriate control population should be people of similar gender, age and educational background to the veteran population.

[3] This is especially so if the sample sizes are small. A study of 71 clinical trials of new medical treatments estimated that 50% of the trials had $\beta > 0.74$ when the true difference was a 25% improvement in the efficacy of the treatment [8]. In other words, half these trials had at least a 74% probability of not detecting a 25% health improvement arising from the medical treatment.