Determining Error Bounds for Hypothesis Tests in Risk Assessment: A Research Agenda

Peter McBurney and Simon Parsons
Department of Computer Science
University of Liverpool
Liverpool L69 7ZF U.K.
{p.j.mcburney,s.d.parsons}@csc.liv.ac.uk

December 11, 2001

Abstract

We argue for renewed attention to the problem of the selection of Type I and Type II error bounds in statistical tests undertaken as part of environmental risk decision-making. Because of the challenges involved in quantifying likelihoods and valuations (or utilities) for the consequences of these errors, default error bounds are typically used. However, while these may be appropriate for scientific domains, their uniform use for risk regulation is not necessarily rational. Recent work in Artificial Intelligence, particularly Computational Dialectics and qualitative decision theory, may provide a way to formalize deliberations regarding the appropriate levels for the error bounds on a case-by-case basis. This would make explicit the anticipated consequences of errors and the trade-offs involved in decisions, thus assisting regulatory decision-makers. We outline a research agenda to develop such a formalization and report on progress to date towards its achievement.

KEYWORDS: Argumentation, Computational Dialectics, Hypothesis Tests, Precautionary Principle, Risk Assessment.

1 Introduction: The Problem

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. An achievement — perhaps the supreme 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 [54], 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. 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.

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 [53] refer explicitly to determining the error bounds based on formal consideration of error consequences. This idea was taken up most prominently in the statistical decision theory of Abraham Wald [75], 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 we might term 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 70 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 [38], 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 [12]: 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 and many scientific publications require this [64], our experience is that most biomedical scientists still think of 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 [12] presents a compelling case why such decision-thresholds are desirable for a scientific communications network, by describing practice in pharmacology before the widespread use of standard hypothesis testing procedures in clinical trials.

However, these decision thresholds are not necessarily appropriate for other decisions, such as deciding regulatory policies, because they ignore other consequences. As Talbot Page [55] argued a quarter-century ago, in assessing the impacts of chemicals on human health or the environment the consequences of the two types of error may be markedly asymmetric: they may differ in their nature, incidence, location, extent, timings, duration, impact and intensity. Moreover, not all the error consequences may be negative for those people impacted, as for example when a chemical is banned and the manufacturers of substitute products enjoy an increase in demand. Even were the consequences to be symmetric and equal, those people affected by each may differ greatly in their relative political, economic or social power and 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 be seen as motivating much of the recent debate on the Precautionary Principle [10, 66]: 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., [39, 67]) 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 values is appropriate for all risk decisions. As Frank Cross [16], among others, has argued, even regulations outlawing chemicals or technologies so as to protect public health may have adverse public health impacts. The mistaken belief that one set of decision thresholds is appropriate to all circumstances might be viewed analogously to *kurdaitcha*,[1] the traditional Australian Aboriginal practice of "pointing the bone" at someone as part of a spell to make them ill or die. By using a certain set of error bounds (this belief implies), we eliminate the problem of the consequences of inference errors by a uniform set of thresholds, in the same way perhaps that pointing the bone solves all inter-personal problems.

A rational approach — rational in the sense of seeking to maximize society's overall welfare — would decide the critical values, and hence the decision thresholds, for risk regulation decisions on a case-by-case basis. Good statistical practice may involve, prior to each hypothesis test, a deliberation over the error bounds and the judicious balancing of levels of α against levels of β, as described for example in [74], but such deliberation, if it occurs at all, rarely takes into account all the consequences of the errors. There are several reasons for this. One is the challenge of identifying all the consequential outcomes. 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 [70], 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 assess-*

[1] from the Aranda word $g^w erdaje$.

ment [21, 42] and *chance discovery* [46], although this work is still very preliminary. The second challenge to case-by-case determination of error bounds 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. Finally, making regulatory decisions means striking a balance between the different interests involved and thus any decision is ultimately a political one [71].

Recent work in Artificial Intelligence in developing qualitative decision theory [24, 56] may provide techniques to meet these challenges. It may be easier, for example, for different stakeholders to reach agreement when likelihoods are represented by labels from a qualitative dictionary such as *"Very Likely," "Likely,"* etc, than when they are represented as probabilities. Although this work is also only preliminary, in the next Section we outline an approach we believe could from the basis for a structure in which formal consideration of consequences and determination of error bounds could be undertaken. It is important to note that we are not proposing that different error bounds be used by scientists, where Bross's arguments [12] about the information-theoretic role of the standard values are persuasive. In addition, as Sven Hansson has argued [33], the web of science is by now such a thickly-woven tapestry that pulling at the thread of the error bounds in one area of science may have complex implications far beyond that area and so unravel science's many interconnected parts. Rather, we are arguing that possibly different error bounds be used for decisions in risk regulation, and that these be decided on a case-by-case basis. As will be seen in the next section, we may view the scientific question (*What to Believe?*) as being distinct from the regulatory question (*What to Do?*).

2 Towards a Solution?

2.1 A list of requirements

In this Section, we present a list of requirements for a structure which would permit formal deliberation over the levels of α and β appropriate for risk regulation decisions in the domain of environmental health; we refer to this structure as a *deliberation structure.* Only an outline is presented here, because the work is still ongoing; we are presenting it now so as to raise awareness in the risk regulation community of the potential of these developments. Our deliberation structure builds on recent work in Artificial Intelligence (AI), Philosophy and Linguistics, in particular an emerging discipline known as Computational Dialectics [26, 60]. Underlying this work is the theory of argumentation, the formal study of argument [18], which has a history in Philosophy dating back at least to Aristotle [6]. Argumentation theories have been applied successfully for some time in Artificial Intelligence, for example in the design of expert systems for medical diagnosis and for personal health risk assessment [15, 41], in legal expert systems [7], and in the design of systems of autonomous software agents [58]. A recent review of such applications is given in [13].

What would be required of a structure for formal deliberation of error bounds in the risk domain? We believe there are a number of components. Firstly, an understanding of the precise nature of the debate(s) being undertaken. In an influential typology, philosophers Doug Walton and Erik Krabbe [76] identified several primary types of dialogue, distinguished by their initial situations, the goals of each of their participants, and the goals of the dialogue itself (which may differ from the individual goals of the participants). The dialogue types were: *Information-seeking dialogues*, in which one participant seeks the answer to some question from another participant; *Inquiries*, in which all participants collaborate to answer some question to which none has the answer; *Persuasion dialogues*, in which one participant seeks to convince others of the truth of some proposition; *Negotiations*, in which participants jointly attempt to divide a scarce resource; *Deliberations*, in which participants collaborate to decide what actions to take in some situation; and *Eristic* (strife-ridden) dialogues, in which participants quarrel verbally as a substitute to physical fighting. While this typology is quite rich, Walton and Krabbe do not claim it is comprehensive, and there are certainly other types of dialogue.

In an idealized sense, one may view scientific dialogues as Inquiry dialogues, where participants collaborate to prove or disprove some hypothesis of interest. In this sense, a scientific dialogue concerns the question, *What to Believe?* However, this assumes an hypothesis has already been explicitly stated, and prior work — involving data collection, data analysis, theory development and much thinking and discussion, especially of a counterfactual nature — may be needed to induce or form an hypothesis. All these activities may be undertaken or supported through dialogue, of types which are not necessarily Inquiries. Moreover, once a scientist adopts a position on an open issue, the debate which then ensues is best described as a multi-way Persuasion dialogue, where both those in favour of a proposition and those against it seek to convince others to accept or reject the proposition at issue. These exchanges can be quite emotionally charged, to the point where they may resemble Eristic dialogues. We may consider these other dialogues as sub-dialogues embedded in the main Inquiry dialogue. Likewise, a similar analysis may be undertaken for regulatory decisions. Treated ideally, these are Deliberations, i.e. deciding what course of action to take in specific circumstances, and so the question under consideration is, *What to Do?* However, in reality there are also many embedded, sequential or parallel sub-dialogues: Information-seeking, Persuasion and Negotiation dialogues and (again perhaps) Eristic dialogues. Recent work in Artificial Intelligence has sought to model the embedding of one type of dialogue in another, and in the next section we discuss some of this research.

A second requirement for the error bounds deliberation structure are formal models of the appropriate dialogue-types. Starting with the work of philosophers Charles Hamblin [32] and Jim MacKenzie [44], formal dialogue-game models have been used as a means of analysis of philosophical and logical questions, such as fallacies in argument. These models treat each dialogue as a game, with the permitted speech acts, or locutions, of the players being the legal moves of the game. The rules of the game determine which combinations of locutions are possible and the circumstances under which the dialogue ends. Walton and Krabbe [76], for example, developed formal dialogue game models for Persuasion dialogues. Researchers in AI have recently proposed formal dialogue games as protocols for interaction between between autonomous software entities, known as *agents*, as a means to enable automated dialogues. This research has led to dialogue game models for various forms of Information-seeking dialogues [40], Inquiries [51], Persuasion dialogues [2, 17], Negotiations [4, 40, 65] and Deliberation dialogues [37].[2] To apply dialogue game models to the error bounds selection problem, we will first require formal models of the relevant scientific and regulatory discourses, and then dialogue game models of each of the component discourses. In the next section, we present some of our own recent work in dialogue game models for Inquiry and Deliberation dialogues, which will be relevant for such an undertaking.

The third component of the deliberation structure is the qualitative representation of uncertainty, for example of consequential outcomes of errors, of their likelihoods, and of the values (utilities) assigned to these outcomes. Here again, our approach is to use a form of argumentation, where the cases for and against various propositions can be articulated, contrasted and possibly combined. This use of argumentation to model uncertainty has also received attention in Artificial Intelligence [23, 31, 51]. How might arguments be combined? A common approach is to define various attack and defeat relationships between arguments, and to assess which arguments survive all attempts to defeat them. This approach is analogous to the conduct of legal proceedings, where claims are accepted as "true" if and only if they survive attempts to defeat them in a validly-constituted and appropriately-conducted court.[3] In the next section, we give an example of qualitative uncertainty labels defined in terms of argument relationships in the modeling of scientific inquiry dialogues.

The next component of the deliberation structure is a decision calculus for formal consideration of qualitative information. One approach to this could be to view it as operating analogously to traditional quantitative decision theory [43]. For example, quantitative decision theory assigns likelihoods to possible outcomes on the basis of numeric probabilities, i.e. elements of the set $[0, 1]$. A qualitative decision theory could assign likelihoods from a qualitative dictionary, such as the linguistic set, {*Open, Supported, Plausible, Probable, Accepted*}. Similarly, we can also imagine utilities being assigned to outcomes from a qualitative dictionary of utility-labels, such as the set $\{-, 0, +\}$. Here, assignment of the label "$+$" to an outcome may be understood to mean that the outcome is perceived to have a net positive (beneficial) utility, the label "$-$" a net negative utility and the label "0" that there is neither positive nor negative utility. Clearly other qualitative dictionaries for both likelihoods and utilities are possible. A qualitative decision theory could seek to combine qualitative likelihoods with qualitative utilities, in a manner similar to that in which quantitative decision theory combines probabilities and utilities to obtain expected values. As mentioned earlier, qualitative calculi for decision-making have been a focus of recent research in Artificial Intelligence [24, 56, 57, 77].

However, even with these modifications, approaches based on classical decision theory will still assume that utilities of different outcomes are comparable and may

[2] Dialogue games have also been applied to problems in legal reasoning [8, 59], in software specification [20] and in automated software design [68].

[3] These are instances of a Game-Theoretic semantics in the sense of Jaako Hintikka [35].

be partially ordered [25]. As our discussion in the section above made clear, these assumptions are not necessarily appropriate in the domain of risk regulation. Moreover, classical decision theory, arising as it did within economics, has assumed that participants to a decision commence with their preferences fully-formed and known, at least to themselves. However, as deliberative democracy theorists in political science have noted, e.g., [14, 30], preferences may be altered or even formed in the very process of making a decision. Indeed, the decision process may enable participants to acquire a *group* or social perspective on a problem, for example awareness of the wider social consequences of individual actions, which may otherwise elude them [61]. In the light of these considerations, it is clear that classical decision-theoretic models are not adequate for the complexity of decisions in the risk regulation domain. This does not mean, however, that adequate decision models cannot be formulated. Henry Richardson, for example [63], has argued that it possible for different people to argue rationally about final ends and values, and not merely about the most effective means for achieving those ends. While much work would be needed to turn these ideas into computational models of decision-making, we do not believe the task is impossible. Given the interactive nature of stakeholder involvement in risk assessment and regulation decision-making, we would anticipate that any adequate model of decision-making would draw on argumentation theory, as we have argued in [49].

Finally, the deliberation structure needs to coherently combine all the components mentioned above. To achieve this, a formal model which specifies the different components and their inter-relationships will be required. Having such a model would enable the exploration of its properties in a systematic and rigorous way, in a manner which is relevant for regulatory decision-making. For example, formalisation may better enable determination of the salient differences between two decision-cases which otherwise appear very similar, or, conversely, may reveal two different cases to be essentially equivalent. One element of such a model may well be a theory of *scenarios*, under which alternative future courses of action could be identified and rigorously compared with one another. In recent work [50], we have begun the development of a computational theory of scenarios in contexts of decision-making under uncertainty. In the domain of risk regulation, the final decisions will always be made by human beings, taking account of all the scientific, political, cultural, social, economic and other factors relevant to the decision. Because of the stakes involved, no formal calculus or intelligent computer system will ever replace the final human decision-makers. However, we see a role for such formal structures and systems in assisting the human decision-makers in contemplating, making, communicating, recording and evaluating those decisions. Elsewhere [62], we have referred to systems providing this type of assistance to human decision-making as serving an *orrery* function, on the analogy of mechanical models of the solar system. In addition to development of coherent, overall models of the decision process and of associated support systems, research attention will also be required for assessment criteria for such models and systems, an issue which has received relatively little attention within AI thus far [28].

2.2 Progress to date

In this section, we present the progress to date in specifying systems which meet the requirements just listed for the error bounds deliberation structure. As mentioned earlier, this work draws on recent developments in Artificial Intelligence and Argumentation Theory in an original way.

First, we have recently defined a logic-based formalism for representing complex dialogues of multiple types [48, 52]. This formalism is hierarchical, with three levels of representation, and is computational. At the highest level is a control dialogue in which participants discuss whether or not to engage in particular types of dialogues over specified topics of discussion. At the second level are dialogue-types themselves; here, the representation is modular and thus permits incorporation not only of the various types articulated by Walton and Krabbe [76], but also other types of dialogues, including types not yet defined. At the third and lowest level are the dialogue-game rules of each specific dialogue-type. Drawing on a species of modal logic known as propositional dynamic logic [34] developed to model formally the operations of computer programs, this three-level representation permits the combination of dialogue-types in complex ways; for example, dialogues may be repeated; they may be undertaken in sequence or in parallel; they may be embedded within one another; and they may be interrupted at any time. In this way, the formalism enables representation of complex human dialogues in a single, unifying framework. We have also been shown [52] that the formalism is potentially generative: that is, it may be used to generate arguments automatically when used for dialogues between suitably-programmed software entities. This is important for applications seeking automatic identification and resolution of any differences between participants, and has not previously been a feature of hierarchical dialogue models in Artificial Intelligence.

The second requirement for the error bounds deliberation structure listed in the previous section involved formal models of the appropriate dialogue-types, namely Deliberation and Inquiry dialogues. Once completed, such models can be incorporated readily at the third level of the hierarchical structure just described. Such formal models are currently under development, building on various principles for the conduct of these dialogues due to philosophers Robert Alexy [1] and David Hitchcock [36]. For instance, we have defined [51] a dialogue game for scientific discourses with locutions allowing a debate participant $\mathcal{P}_i$ to assert a claim θ with a degree of support d_θ, as follows:

$$assert(\mathcal{P}_i : (\theta, d_\theta)).$$

The degree of support label d_θ is taken from a dictionary of labels agreed between the participants, and may be quantitative or qualitative. Once asserted, the claim θ may be questioned by another participant, which move then obliges the first participant to present an argument for the assertion; our locution syntax enables such arguments to be presented to the forum. As in a real debate, participants may also query or contest assertions, premises, arguments, rules of inference and degrees of support, via specific locutions. The dialogue game definitions specify the preconditions required for each locution to be validly executed, for example, a participant may not assert a claim θ and then assert the contradictory claim $\neg\theta$, without in the interim retracting the first assertion. Likewise, post-conditions are specified, as when a dialogue move imposes a burden of proof on a participant.

Together with David Hitchcock, we have also recently developed the first formal model for Deliberation dialogues [37], building on work in the philosophy of argumentation undertaken by Harald Wohlrapp [78]. This model of a deliberation comprises eight elements:

Open: Opening of the deliberation dialogue.

Inform: Discussion of the goal(s) of the deliberation dialogue (i.e. for what purpose is action being considered), of any constraints on the possible actions which may be considered, of the perspectives by which proposals may be evaluated, and of any premises (facts) relevant to this evaluation.

Propose: Suggesting of possible action-options to achieve the agreed goal.

Consider: Commenting on proposals from various perspectives.

Revise: Revising discussion goal, constraints and/or action-options in the light of the comments presented.

Recommend: Recommending an option for action, and acceptance or non-acceptance of this recommendation by the participants.

Confirm: Confirming acceptance.

Close: Closing of the deliberation dialogue.

These eight elements may be undertaken in any order, subject only to a small number of constraints (e.g. that the **Confirm** stage always follows the **Recommend** stage). As with the scientific dialogue example, appropriate locutions for participants in each of the eight elements, along with rules for their use, have been articulated. These specifications are quite general; to be applied to debates over error bounds, will require domain-specific locutions and dialogue rules, an issue we have given some attention [47, 49]. For instance, drawing on the generic theory of Communicative Action of philosopher Jürgen Habermas [29], we proposed the following types of locutions as appropriate for debates in the domain of environmental and health risk assessment [49]:

Factual Statements: These are statements which seek to represent the state of the external world, such as claims about scientific reality, and the scientific, economic or social consequences of particular actions. Contesting such a statement means denying that it is a true description of objective, external reality.

Value Statements: These are statements which seek to represent the state of the internal world of the speaker, i.e. they reveal publicly the speaker's subjective preferences or value assignments. Such statements may only be contested by doubting the sincerity of the speaker.

Connection Statements: These are statements which assert some ethical, social or legal relationship between different parties, in the common world of the debate participants. Contesting these statements means denying the existence, relevance or importance of such relationships.

Inferential Statements: These are statements which refer to the content of earlier statements made in a debate, drawing inferences from them or noting implications. Once a scientific theory has been proposed, a specific risk assessment model and the ensuing calculations based on the model fall into this category. Contestation of such statements can take the form of questioning the appropriateness or the validity of the inferences made.

Procedural Statements: These are statements about the activity of speaking itself, such as the rules for participation and debate. In many real-life discourses, these often become the focus of debate, overtaking issues of substance.[4]

Obligation Statements: These are statements which assert some obligation on the participants, for example, that they must limit the commercial sale of a new substance. Only the authorized regulator has the power to make such assertions, and once made, they cannot be contested within the debating forum. In real-life, they may of course be contested in the courts, and often are.

As can be seen, this typology is still very coarse, and more work will be required to produce a finely-grained set of locutions, along with a full specification of the pre- and post-conditions appropriate for each.

The third requirement of the error bounds deliberation structure listed was the qualitative representation of uncertainty. Recent work in AI has made explicit use of argumentation to represent knowledge uncertainty, arising for example from inconsistent, contested or missing evidence [42, 51]. In this approach, a claim θ is assigned an uncertainty label from the qualitative linguistic dictionary $\{$*Open, Supported, Plausible, Probable, Accepted*$\}$ according to rules such as the following:

- If θ is a claim for which no argument has yet been provided by a participant, then θ is assigned the value *Open*.
- If θ is a claim for which at least one argument has been provided by a participant, then θ is assigned the value *Supported*.
- If θ is a claim for which a consistent argument has been provided by a participant, then θ is assigned the value *Plausible*.
- If θ is a claim for which a consistent argument has been provided by a participant, and for which neither rebutting nor undercutting arguments have been provided, then θ is assigned the value *Probable*. Rebutting arguments (*rebuttals*)

[4]For instance, in the scientific debate over Genetically-Modified Organisms in Britain during 1999, an argument between the medical journal *The Lancet* and The Royal Society ensued over whether the latter was entitled to comment on a paper submitted to the journal before it had been accepted or rejected for publication [5].

are arguments for the negation of θ, and undercutting arguments (*undercutters*) are arguments for the negation of a premise of an argument for θ. Rebuttals and undercutters are together called attacking arguments.

- If θ is a claim for which a consistent argument has been provided by a participant, and that argument is well-defended, then θ is assigned the value *Accepted*. A well-defended argument is one for which counter-attacking arguments have been presented for each attacking argument.

The labels are listed in order of increasing strength, and they are a property of the debate as a whole, not of any one participating individual. Thus, the labels assigned by these rules may differ from any labels assigned by the individual debate participants.

The uncertainty label applied to a claim may change as new arguments are presented, so this assignment of uncertainty labels is defeasible. If it is assumed that all relevant arguments are eventually presented by one participant or another, then after a finite (but possibly very long) time, the value of the uncertainty label should be stable. We may refer to this stable value as the *uncertainty label at infinity*. Suppose we commence a debate and then, after some finite time t, take a snap-shot of the uncertainty label at that time, and learn that it is *"Accepted"*. Will the uncertainty label at infinity also be *"Accepted?"* Under some reasonable assumptions about the timing of the snapshot, it can be shown [51] that if the probability of new information arising after the snapshot is taken is less than ϵ, for some real number $0 < \epsilon < 1$, then the probability that the uncertainty label at infinity is also *"Accepted"* is at least $1 - \epsilon$. This result demonstrates that assignment of qualitative uncertainty labels based on the arguments presented for and against claims is well-behaved: As with statistical inference, we cannot guarantee that inference from a finite snapshot of the uncertainty label to its value at infinity is always valid, but we can place a probabilistic bound on the likelihood of error in making this inference. This and the other formal properties proved in [51] provide confidence in the use of dialogue and argumentation systems to represent uncertainty in domains where the absence of scientific information precludes quantification of uncertainties, or where agreement over such quantification is not achievable. These features are typical of the environmental risk assessment domain.[5]

Work on the next element of the error bounds deliberation structure, namely an appropriate qualitative decision calculus, is still too preliminary to report at this time. Likewise with the development of a formal model which combines coherently all the listed components. There is likely to be more than one way to combine these components rigorously, and different combinations may result in different decision support systems or even different decision outcomes. In that case, research will be needed to assess the most appropriate formal model for deliberation over statistical error bounds.

How does this research on computational dialectics and qualitative decision-making relate to the problems of determination of error-bounds in hypothesis testing for risk assessment? Firstly, the work we have outlined on computational dialectics should, when completed, enable the formal modeling and explicit representation of debates over the consequences of inference errors, in all their glorious complexity. Arguments for and

[5] For an example, see the recent report on the issue of Genetically-Modified foodstuffs prepared for the U.K. Economic and Social Research Council Global Environmental Change programme [72].

against various error consequences, rebuttals and challenges to these arguments, along with their various implications, valuations and degrees of confidence, will all be representable in a suitable dialectical argumentation formalism. That such an application to the error bounds problem is potentially feasible is shown by the applications of argumentation theory already to be found in deployed computer systems [13, 22, 27]. Secondly, the actual decision-making task involved in selecting non-default values of error bounds will also be amenable, we believe, to formal modeling and representation. Here the decisions involve making trade-offs between different sets of consequences, in a context of competing views of their relative importance and value. As we suggest above, formalization of such decision-making may be possible in an appropriate qualitative decision calculus; like argumentation, qualitative approaches to reasoning under uncertainty developed in Artificial Intelligence have already proven themselves in real-world applications [22, 56, 77].

2.3 Example

We now present a hypothetical example of a simple Inquiry dialogue regarding an uncertain proposition. This example, adapted from [51], concerns a debate over the possible carcinogenicity of a chemical $\mathcal{X}$, for which evidence is conflicting. The purpose of this example is to show how the relationships between the various arguments uttered in a dialogue for and against a proposition can be used to generate qualitative uncertainty labels for the proposition. In a real debate, participants would be free to introduce supporting evidence and modes of inference at any time. However, to aid understanding, in this example we first list the assumptions and modes of inference to be used in assertions and proposals. The various assumptions — statements which can be used as grounds for arguments — are numbered K1 through K4.

K1: $\mathcal{X}$ is produced by the human body naturally (i.e. it is endogenous).

K2: $\mathcal{X}$ is endogenous in rats.

K3: If $\mathcal{X}$ is an endogenous chemical then it is not carcinogenic.

K4: Bioassay experiments applying $\mathcal{X}$ to rats result in significant carcinogenic effects.

The modes of inference used by participants in the dialogue are labeled R1 through R3:

R1 (And Introduction): Given a statement ϕ and a statement θ, we may infer the statement $(\phi \wedge \theta)$.

R2 (Modus Ponens): Given a statement ϕ and the statement $(\phi \rightarrow \theta)$, we may infer the statement θ.

R3: If a chemical is found to be carcinogenic in an animal species, then we may infer it to be carcinogenic in humans.

We now give an example of a dialogue concerning the statement: *$\mathcal{X}$ is carcinogenic to humans*, which we denote by ϕ. The dialogue utterances are numbered M1, M2,. . ., in sequence. The locutions used in this dialogue are those for the scientific Inquiry dialogue protocol of [51], which also contains details of the locution syntax. We assume that participants agree to use the following qualitative dictionary of labels to express their degree of support for statements: {*Certain, Confirmed, Probable, Plausible, Supported, Open*}. Moreover, they agree to assign qualitative uncertainty labels to statements on the basis of arguments presented in the debate according to the dictionary defined in Section 2.2, namely: {*Accepted, Probable, Plausible, Supported, Open*}. To assist understanding of the example, each utterance is followed by an annotation.

At the commencement of the dialogue, no arguments have been advanced for either ϕ or $\neg\phi$, and so both statements are assigned the uncertainty label *Open.*

M1: *assert(*$\mathcal{P}_1 : (\phi, \mathit{Confirmed})$*).*

Participant $\mathcal{P}_1$ asserts the claim ϕ, that $\mathcal{X}$ is carcinogenic to humans, which she believes has strength *Confirmed.* The uncertainty label assigned to ϕ remains as *Open*, because no argument has yet been presented for ϕ.

M2: *query(*$\mathcal{P}_2 : assert(\mathcal{P}_1 : (\phi, \mathit{Confirmed}))$*).*

Participant $\mathcal{P}_2$ asks $\mathcal{P}_1$ for her argument for ϕ.

M3: *show_arg(*$\mathcal{P}_1 : (K4, R3, \phi, (\mathit{Confirmed}, \mathit{Valid}, \mathit{Confirmed}))$*).*

Participant $\mathcal{P}_1$ presents her argument for ϕ, which rests on grounds that bioassay experiments of $\mathcal{X}$ have been shown to produce carcinogenic effects in rats (Assumption K4), and that one can infer from these results to humans, by means of Inference Rule R3. Participant $\mathcal{P}_1$ assigns this rule a modality of *Valid.* Utterance **M3** has presented an argument for ϕ, and so a new uncertainty label for this statement is required. This label is *Probable*, because the argument presented is consistent, and also because no rebuttals or undercutters have yet been presented against ϕ.

M4: *contest(*$\mathcal{P}_2$ *:assert(*$\mathcal{P}_1 : (\phi, \mathit{Confirmed})$*)).*

Participant $\mathcal{P}_2$ contests the assertion of ϕ with modality *Confirmed* by $\mathcal{P}_1$.

M5: *query(*$\mathcal{P}_3$ *:contest(*$\mathcal{P}_2$ *:assert(*$\mathcal{P}_1 : (\phi, \mathit{Confirmed})$*)))*

Participant $\mathcal{P}_3$ asks $\mathcal{P}_2$ for her reasons for the contestation in Utterance **M4.**

M6: *propose(*$\mathcal{P}_2 : (\neg\phi, Plausible)$*).*

Participant $\mathcal{P}_2$ proposes the claim $\neg\phi$, i.e. that $\mathcal{X}$ is not carcinogenic to humans, and says she believes this is *Plausible.*

M7: *query(*$\mathcal{P}_1$ *:propose(*$\mathcal{P}_2 : (\neg\phi, Plausible)$*)).*

Participant $\mathcal{P}_1$ asks $\mathcal{P}_2$ for her argument for $\neg\phi$.

M8: *show_arg(*$\mathcal{P}_2 : ((K1, K3), R2, \neg\phi, (\mathit{Confirmed}, Probable, \mathit{Valid}, Plausible))$*).*

Participant $\mathcal{P}_2$ presents her argument for $\neg\phi$. This argument starts from the premises that $\mathcal{X}$ is endogenous (K1) and that endogenous chemicals are not carcinogenic (K3), and then uses Modus Ponens (Rule R2) to conclude that $\mathcal{X}$ is not carcinogenic to humans. This utterance means that an argument against ϕ has been presented in the dialogue, and so the uncertainty label applied to ϕ changes from *Probable* to *Plausible*. Similarly, the uncertainty label for $\neg\phi$ statement is now *Plausible*, since the argument of ϕ in Utterance **M3** forms a rebuttal of $\neg\phi$.

M9: *contest_ground(*$\mathcal{P}_4$: *show_arg(*$\mathcal{P}_2$: $((K1, K3), R2, \neg\phi, ($ *Confirmed*, *Probable*, *Valid*, *Plausible* $)$*:* $(K3,$ *Probable*$)))$.

Participant $\mathcal{P}_4$ contests a grounds of the argument presented by $\mathcal{P}_2$ in Utterance **M8**, namely the premise K3, that an endogenous chemical is carcinogenic.

M10: *show_arg(*$\mathcal{P}_4$: $((K2, K4), R1, \neg K3, ($*Confirmed*, *Confirmed*, *Valid*, *Confirmed*$)))$

Participant $\mathcal{P}_4$ immediately follows the contestation with a presentation of her own argument for the negation of K3, i.e. an argument for the claim that it is not the case that an endogenous chemical is carcinogenic. This argument uses *And Introduction* (Rule R1) on the premises K2, that $\mathcal{X}$ is endogenous in rats, and K4, that $\mathcal{X}$ has been shown to cause cancers in rats. This argument is an undercutter for the argument for $\neg\phi$, presented by $\mathcal{P}_2$ in Utterance **M8**, since it attacks an assumption of that argument. The uncertainty label assigned to $\neg\phi$ does not change, however, as a rebuttal already had been presented. However, the undercutter of **M10** attacks a rebuttal (that in **M8**) of ϕ, and so is a counter-attacking argument. This rebuttal is the only rebuttal or undercutter presented which attacks ϕ. Thus, the argument for ϕ presented in **M3** is now well-defended. Hence, the uncertainty label assigned to ϕ changes from *Plausible* to *Accepted*.

This example, although hypothetical and simplified, shows how the arguments articulated for and against a proposition may be used to generate qualitative uncertainty labels for that proposition. Thus a formalism for dialogue over an uncertain proposition may incorporate representations of the arguments for and against it, and these arguments may be resolved — and the debate summarized at any time — on the basis of their formal relationships with one another. In this example, we have treated all arguments equally, and not given some more weight than others. However, in most real-world domains some arguments will be weighted more strongly than others, as, for example, in the 1986 Guidelines for Carcinogen Risk Assessment of the USA Environmental Protection Agency, where human evidence for carcinogenicity is afforded priority over animal evidence [73, p. 34000]. Recent work in AI has developed formal approaches to resolving arguments based on conflicting preferences, e.g., [3], and can accommodate differential weighting of arguments.

3 Discussion and Conclusion

In this paper, we have restated arguments long known to statisticians that the selection of Type I and Type II error bounds should be decided on a case-by-case basis, taking account of the consequences of each type of inference error. Indeed, as we mentioned, informal considerations of this nature motivated the values now standard in the sciences for these error bounds. We believe this is especially important in the regulation of environmental and health risks, where such consequences may differ greatly in their nature, dimensions and incidence. Such an approach is contrary to the uniform use either of the values for α and β standardly used by scientists or of the diametrically-contrary values implicit in applications of the Precautionary Principle.

A key challenge in deciding error bounds case-by-case is doing so in a rigorous and formal manner, given the diversity of considerations appropriate to the decisions, the diversity of interests, values and preferences of the stakeholders, and the difficulties of quantification of many of the variables at issue. Recent work in Artificial Intelligence in modeling dialogues in computational systems, in developing qualitative representations of uncertainty and in developing non-standard decision theories provide a basis, we believe, for approaches to deal with this challenge. We have thus outlined a set of requirements for a structure for deliberation over error bounds, and presented the current status of research work to develop the techniques needed to meet these requirements. When fully developed, this structure should provide a rigorous and coherent formal framework to assist debate and decision over these bounds on a case-by-case basis. Such a framework could be seen as an example of a meta-statistical tool, in the sense of Deborah Mayo [45].

We believe a number of benefits would arise from adopting such a formal deliberation structure. Firstly, it would make all assumptions, inferences and conclusions in the decision-process explicit and transparent. It would also reveal the explicit trade-offs necessary to the making of these decisions. Transparency in both areas is especially important in the domain of risk regulation, where matters of wide public importance, including life and death, may be involved. Secondly, the use of argumentation formalisms enables the reasons for and against conclusions to be represented alongside those conclusions. Moreover, using a formalized argumentation and decision structure enables all stakeholders (whether participants in the deliberation process or onlookers) to judge the arguments and procedures used in any case against the formal structure, independent of the particular case. In both respects, formalization acts to improve the decision-making process and to increase its transparency. Thirdly, the use of a computer-based deliberation structure could enable greater public participation in these decisions and thus potentially give effect, as we and others have argued [19, 49], to ideas of deliberative democracy current in political theory [9, 11]. At present, risk decision thresholds are those based on the standard values used in the scientific community, and which may be appropriate there. One could argue, as we have done above, that they are inappropriate to the domain of public policy decision-making over potential environmental and health risks. Whatever the merits of that case, however, their deployment in this domain has certainly not followed any public debate over their use; indeed, the debate over the Precautionary Principle is the closest that western society has come to a public discussion over error bounds in hypothesis test procedures, and to date error bounds issues have been implicit rather than explicit in that debate.

Finally, we believe it important to stress that there are no "kurdaitcha bones," error-bounds and thus decision-thresholds appropriate for every situation; instead, rational decision-making — at least within the limits of resource constraints on the decision-making process itself — requires that the error bounds for each case be decided taking account the consequences of inference errors in that particular case. The scientific, economic and social issues involved in any debate over error bounds will usually be quite complex, and the values placed on the same outcomes by different participants often very discordant; this can be seen from the public debate over almost any potential major environmental or health risk. There is no guarantee of agreed resolution of such differences, as has been found in applying multi-criteria scoring techniques to the issue of Genetically-Modified foods [69]. However, even without a guarantee of resolution, representation of a debate within such a formalism would force greater clarity in the statements articulated, and thus facilitate attempts at reaching trade-offs between different regulatory alternatives. At the very least, explicit representation would make everyone involved aware of the case-specific consequences of inference errors, awareness which the unthinking use of uniform error bounds inhibits.

Acknowledgments

This work was partly funded by the British Engineering and Physical Sciences Research Council (EPSRC) under grant GR/L84117 and a PhD studentship, and this support is gratefully acknowledged. We also thank Chip Heathcote for discussions and the anonymous referees for their comments on earlier versions of this paper. However, we alone are responsible for the views expressed here.

References

[1] R. Alexy. A theory of practical discourse. In S. Benhabib and F. Dallmayr, editors, *The Communicative Ethics Controversy*, Studies in Contemporary German Social Thought, pages 151–190. MIT Press, Cambridge, MA, USA, 1990. Translation by D. Frisby of: *Eine Theorie des praktischen Diskurses.* In: W. Oelmüller, Ed., *Normenbegründung-Normendurchsetzung.* Paderborn, Germany: Schöningh, 1978.

[2] L. Amgoud, N. Maudet, and S. Parsons. Modelling dialogues using argumentation. In E. Durfee, editor, *Proceedings of the Fourth International Conference on Multi-Agent Systems (ICMAS 2000)*, pages 31–38, Boston, MA, USA, 2000. IEEE Press.

[3] L. Amgoud and S. Parsons. Agent dialogues with conflicting preferences. In J-J. Meyer and M. Tambe, editors, *Pre-Proceedings of the Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL 2001)*, pages 1–14, Seattle, WA, USA, 2001.

[4] L. Amgoud, S. Parsons, and N. Maudet. Arguments, dialogue, and negotiation. In W. Horn, editor, *Proceedings of the Fourteenth European Conference on Artificial Intelligence (ECAI 2000)*, pages 338–342, Berlin, Germany, 2000. IOS Press.

[5] Anon. Editorial: Health risks of genetically modified foods. *The Lancet*, 353(9167), 29 May 1999.

[6] Aristotle. *Topics*. Clarendon Press, Oxford, UK, 1928. (W. D. Ross, Editor).

[7] T. J. M. Bench-Capon, F. P. Coenen, and P. Orton. Argument-based explanation of the British Nationality Act as a logic program. *Computers, Law and AI*, 2(1):53–66, 1993.

[8] T. J. M. Bench-Capon, T. Geldard, and P. H. Leng. A method for the computational modelling of dialectical argument with dialogue games. *Artificial Intelligence and Law*, 8:233–254, 2000.

[9] J. Bessette. Deliberative Democracy: The majority principle in republican government. In R. A. Goldwin and W. A. Schambra, editors, *How Democratic is the Constitution*, pages 102–116. American Enterprise Institute, Washington, DC, USA, 1980.

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

[11] J. Bohman and W. Rehg, editors. *Deliberative Democracy: Essays on Reason and Politics*. MIT Press, Cambridge, MA, USA, 1997.

[12] 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.

[13] D. V. Carbogim, D. S. Robertson, and J. R. Lee. Argument-based applications to knowledge engineering. *Knowledge Engineering Review*, 15(2):119–149, 2000.

[14] T. Christiano. The significance of public deliberation. In J. Bohman and W. Rehg, editors, *Deliberative Democracy: Essays on Reason and Politics*, pages 243–277. MIT Press, Cambridge, MA, USA, 1997.

[15] A. S. Coulson, D. W. Glasspool, J. Fox, and J. Emery. RAGs: a novel approach to computerized genetic risk assessment and decision support from pedigrees. *Methods of Information in Medicine*, 40:315–322, 2001.

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

[17] F. Dignum, B. Dunin-Kęplicz, and R. Verbrugge. Creating collective intention through dialogue. *Logic Journal of the IGPL*, 9(2):305–319, 2001.

[18] F. H. van Eemeren, R. Grootendorst, F. S. Henkemans, J. A. Blair, R. H. Johnson, E. C. W. Krabbe, C. Plantin, D. N. Walton, C. A. Willard, J. Woods, and D. Zarefsky. *Fundamentals of Argumentation Theory*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.

[19] C. Ess. The political computer: Democracy, CMC, and Habermas. In C. Ess, editor, *Philosophical Perspectives on Computer-Mediated Communication*, pages 197–230. State University of New York Press, Albany, NY, USA, 1996.

[20] A. Finkelstein and H. Fuks. Multi-party specification. In *Proceedings of the Fifth International Workshop on Software Specification and Design*, Pittsburgh, PA, USA, 1989. ACM Sigsoft Engineering Notes.

[21] J. Fox. Will it happen? Can it happen? A new approach to formal risk analysis. *Risk, Decision and Policy*, 4 (2):117–128, 1999.

[22] J. Fox and S. Das. *Safe and Sound: Artificial Intelligence in Hazardous Applications*. MIT Press, Cambridge, USA, 2000.

[23] J. Fox, P. Krause, and S. Ambler. Arguments, contradictions, and practical reasoning. *Proceedings of the Tenth European Conference on Artificial Intelligence (ECAI-92), Vienna, Austria*, pages 623–626, 1992.

[24] J. Fox and S. Parsons. Arguing about beliefs and actions. In A. Hunter and S. Parsons, editors, *Applications of Uncertainty Formalisms*, volume 1455 of *Lecture Notes in Artificial Intelligence*, pages 266–302. Springer, Berlin, Germany, 1998.

[25] S. J. Gollop. Paradoxes of the Black Box: The Allais paradox, instransitive preferences and Orthodox Decision Theory. M.A. Thesis in Philosophy, University of Auckland, Auckland, New Zealand, 2000.

[26] T. F. Gordon. Computational dialectics. In P. Hoschka, editor, *Computers as Assistants: A New Generation of Support Systems*, pages 186–203. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.

[27] T. F. Gordon and N. Karacapilidis. The Zeno argumentation framework. In *Proceedings of the Sixth International Conference on AI and Law*, pages 10–18. ACM Press, 1997.

[28] M. M. Groothuis and J. S. Svensson. Expert system support and juridical quality. In J. Breuker, editor, *Proceedings of the Thirteenth International Conference on Legal Knowledge-Based Systems (JURIX-2000)*, pages 1–10, Amsterdam, The Netherlands, 2000. IOS Press.

[29] J. Habermas. *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society*. Heinemann, London, UK, 1984. Translation by T. McCarthy of: *Theorie des Kommunikativen Handelns, Band I, Handlungsrationalitat und gesellschaftliche Rationalisierung.* Suhrkamp, Frankfurt, Germany. 1981.

[30] J. Habermas. *The Inclusion of the Other: Studies in Political Theory*. MIT Press, Cambridge, MA, USA, 1998. Edited by C. Cronin and P. De Greiff.

[31] M. Haggith. *A Meta-level Argumentation Framework for Representing and Reasoning about Disagreement*. PhD thesis, University of Edinburgh, UK, 1996.

[32] C. L. Hamblin. *Fallacies*. Methuen, London, UK, 1970.

[33] S. O. Hansson. Can we reverse the burden of proof? *Toxicology Letters*, 90:223–228, 1997.

[34] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. Foundations of Computing Series. MIT Press, Cambridge, MA, USA, 2000.

[35] J. Hintikka. *The Game of Language: Studies in Game-Theoretical Semantics and Its Applications*, volume 22 of *Synthese Language Library*. D. Reidel, Dordrecht, The Netherlands, 1983.

[36] D. Hitchcock. Some principles of rational mutual inquiry. In F. van Eemeren et al., editor, *Proceedings of the Second International Conference on Argumentation*, pages 236–243, Amsterdam, The Netherlands, 1991. SICSAT.

[37] D. Hitchcock, P. McBurney, and S. Parsons. A framework for deliberation dialogues. In H. V. Hansen, C. W. Tindale, J. A. Blair, and R. H. Johnson, editors, *Proceedings of the Fourth Biennial Conference of the Ontario Society for the Study of Argumentation (OSSA-2001)*, Windsor, Ontario, Canada, 2001.

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

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

[40] J. Hulstijn. *Dialogue Models for Inquiry and Transaction*. PhD thesis, Universiteit Twente, Enschede, The Netherlands, 2000.

[41] P. Krause, J. Fox, and P. Judson. An argumentation based approach to risk assessment. *IMA Journal of Mathematics Applied in Business and Industry*, 5:249–263, 1994.

[42] 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*, volume 1455 of *Lecture Notes in Artificial Intelligence*, pages 138–156. Springer, Berlin, Germany, 1998.

[43] D. V. Lindley. *Making Decisions*. John Wiley and Sons, London, UK, second edition, 1985.

[44] J. D. MacKenzie. Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8:117–133, 1979.

[45] D. G. Mayo. Increasing public participation in controversies involving hazards: the value of metastatistical rules. *Science, Technology and Human Values*, 10:55–68, 1985.

[46] P. McBurney. First international workshop on chance discovery. *Knowledge Engineering Review*, 16(2):215–218, 2001.

[47] P. McBurney and S. Parsons. Risk Agoras: using dialectical argumentation to debate risk. *Risk Management*, 2(2):17–27, 2000.

[48] P. McBurney and S. Parsons. Agent ludens: games for agent dialogues. In S. Parsons and P. Gmytrasiewicz, editors, *Game-Theoretic and Decision-Theoretic Agents: Proceedings of the 2001 AAAI Spring Symposium*, pages 70–77, Menlo Park, CA, USA, 2001. AAAI Press. Technical Report SS-01-03.

[49] P. McBurney and S. Parsons. Intelligent systems to support deliberative democracy in environmental regulation. *Information and Communications Technology Law*, 10(1):33–43, 2001.

[50] P. McBurney and S. Parsons. Reasoning across scenarios in planning under uncertainty. In C. Gomes and T. Walsh, editors, *Using Uncertainty within Computation: Papers from the 2001 AAAI Fall Symposium*, pages 85–92, Menlo Park, CA, USA, 2001. AAAI Press. Technical Report FS-01-04.

[51] P. McBurney and S. Parsons. Representing epistemic uncertainty by means of dialectical argumentation. *Annals of Mathematics and Artificial Intelligence*, 32(1–4):125–169, 2001.

[52] P. McBurney and S. Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, (In press), 2002.

[53] 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 [54].

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

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

[56] S. Parsons. *Qualitative Methods for Reasoning Under Uncertainty*. MIT Press, Cambridge, MA, USA, 2001.

[57] S. Parsons and S. Green. Argumentation and qualitative decision making. In A. Hunter and S. Parsons, editors, *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, volume 1638 of *Lecture Notes in Artificial Intelligence*, pages 328–339. Springer, Berlin, Germany, 1999.

[58] S. Parsons, C. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261—292, 1998.

[59] H. Prakken and G. Sartor. Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law*, 6:231–287, 1998.

[60] C. Reed and T. Norman, editors. *Bonskeid Symposium on Argument and Computation*. In press, 2002.

[61] W. Rehg. The argumentation theorist in deliberative democracy. *Controversia*, 2002 (*to appear*). (Keynote Address at the Second Conference of the International Debate Education Association, Prague, Czech Republic, 13 October 2001).

[62] W. Rehg, P. McBurney, and S. Parsons. Computer decision-support systems for public argumentation: Criteria for assessment. In H. V. Hansen, C. W. Tindale, J. A. Blair, and R. H. Johnson, editors, *Proceedings of the Fourth Biennial Conference of the Ontario Society for the Study of Argumentation (OSSA 2001)*, Windsor, Ontario, Canada, 2001.

[63] H. S. Richardson. *Practical Reasoning about Final Ends*. Cambridge University Press, Cambridge, UK, 1994.

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

[65] F. Sadri, F. Toni, and P. Torroni. Logic agents, dialogues and negotiation: an abductive approach. In M. Schroeder and K. Stathis, editors, *Proceedings of the Symposium on Information Agents for E-Commerce, Artificial Intelligence and the Simulation of Behaviour Conference (AISB-2001)*, York, UK, 2001. AISB.

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

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

[68] K. Stathis. A game-based architecture for developing interactive components in computational logic. *Electronic Journal of Functional and Logic Programming*, 2000(5), March 2000.

[69] A. Stirling and S. Mayer. *Rethinking Risk: A Pilot Multi-Criteria Mapping of a Genetically Modified Crop in Agricultural Systems in the UK*. Report, SPRU, University of Sussex, Brighton, UK, 1999.

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

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

[72] U. K. Economic and Social Research Council (ESRC) Global Environmental Change (GEC) Programme. *The Politics of GM Food: Risk, Science and Public Trust*. Special Briefing 5, University of Sussex, Brighton, UK, 1999.

[73] U. S. A. Environmental Protection Agency. Guidelines for carcinogen risk assessment. *U.S. Federal Register*, 51:33991–34003, 24 September 1986.

[74] 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.

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

[76] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Series in Logic and Language. State University of New York Press, Albany, NY, USA, 1995.

[77] M. P. Wellman. *Formulation of Tradeoffs in Planning under Uncertainty*. Pitman, London, UK, 1990.

[78] H. Wohlrapp. A new light on non-deductive argumentation schemes. *Argumentation*, 12:341–350, 1998.