The use of expert systems for toxicology risk prediction

Simon Parsons
Department of Computer and Information Science
Brooklyn College
City University of New York
2900 Bedford Avenue, Brooklyn
NY 11210, USA
parsons@sci.brooklyn.cuny.edu

Peter McBurney
Department of Computer Science
University of Liverpool
Chadwick Building
Liverpool L69 7ZF, UK
p.j.mcburney@csc.liv.ac.uk

December 30, 2003

Abstract

One approach to predicting the toxicology of novel compounds is to apply expert knowledge. The field of artificial intelligence has identified a number of ways of doing this, and some of these approaches are briefly described in this chapter. We also examine two expert systems—DEREK, which predicts a variety of types of toxicology, and StAR, which predicts carcinogenicity—in some detail. StAR reasons about carcinogenicity using a system of argumentation. We believe that argumentation systems have great potential in this area, and so discuss them at length.

1 Introduction

One way to build a computer system to solve a problem is to replicate the way that a human would deal with the problem. For some tasks this is not a good solution. We wouldn't write a program to do arithmetic by the same symbolic manipulation that humans carry out—it just wouldn't be efficient[1]—and the

[1]Indeed, Reverse Polish Notation was invented precisely to improve the efficiency of computer manipulation of arithmetic symbols [38].

same is true of any problem for which there are clear algorithmic solutions. However, for some problems, the best we can do is to try to replicate the way that humans solve the problems. Problems such as diagnosing an illness, identifying chemical compounds from the output of a mass spectrometer, and deciding how to configure a computer are all tasks where copying what humans do seems to be the best we can do, and the same may be true of predicting the toxicology of novel compounds.

Now, all the tasks mentioned above have (at least) one thing in common. These are all tasks that humans can only complete once they have completed a significant amount of training and have gained a good deal of experience. They are tasks that can only be completed by human *experts*. When we build systems that capture the aspects of the problem-solving ability of these experts, we call them *expert systems*, and it is such systems that are the subject of this chapter.[2]

The study of expert systems is a sub-field of artificial intelligence, and rose to great prominence in the late 1970s. This was when the pioneers of the expert system field were building and trialling the earliest expert systems—MYCIN, which carried out medical diagnosis [10, 91], DENDRAL, which identified chemical compounds from mass spectrometer readings [11, 24], and R1/XCON which configured VAX computers [68]. Expert systems seemed to offer great advantages over conventional software systems and other artificial intelligence techniques. They could, it seemed, be used to replicate, and possibly replace, expensive human experts, and bring scarce expertise into the domain of mass production. This led to a huge growth in interest in the field, both academically and commercially, and the field seemed to have a bright future. However, by the late 1980s it had become clear that expert systems were not as widely applicable as some had claimed, and most of the interest in the area subsided.

This subsidence, in our view, was only to be expected. Expert systems were oversold, and it is only natural that this would become apparent in due course. However, underneath all the hype, the basic idea behind expert systems remains sound. For some problems they provide a very good solution and in that kind of role they are flourishing (as we will see, flourishing remarkably widely) and will continue to.

Our aim in this chapter is to examine the extent to which expert systems can provide a suitable solution to the problem of predicting toxicity, and to provide some pointers to those who want to try such a solution for themselves. We start with a description of two of the main approaches to building expert systems in Section 2. We then take a look, Section 3 at two particular expert systems, DEREK and StAR, that do this kind of risk prediction. The second of these systems works using a system of argumentation—that is it builds up reasons for and against predictions in order to decide which is best—and because we believe that argumentation is a particularly good approach, we describe in some detail,

[2]It should be stressed that the important feature of expert systems is this capturing of the ability of human experts. The aim of building an expert system is not to *mimic* a human expert, but to isolate an expert's problem-solving ability with the aim of using this ability to improve upon the problem-solving performance of that expert. Some expert systems do indeed manage to outperform the expert whose ability they capture, others are not so successful.

Section 4, both the general approach to argumentation used by StAR, and the directions in which we are developing the theory. Finally, Section 5, we precis the chapter.

2 Expert systems

The key idea behind expert systems is that some problems are best solved by applying *knowledge* about the problem domain, knowledge that only people very familiar with the domain are likely to have. This naturally creates a need to represent that knowledge, and *knowledge representation* is a subject that has been widely researched. The knowledge needed to solve a problem rarely includes the exact answer to particular instance of the problem. Instead, the expert system has to take the knowledge that it has and infer new information from it that bears upon the exact problem it is solving. As a result we are interested in how to perform this *reasoning* as well as how to represent knowledge. This section looks at two commonly used approaches to knowledge representation and their associated form of reasoning.

2.1 Rule-based systems

One of the earliest, and most successful approaches to knowledge representation is the use of *production rules* [18, 19] similar to:

IF battery good AND battery charging
THEN battery ok

Such rules provide a very natural means of capturing the information of a domain expert—in this case an expert in the diagnosis of problems with the electrical system of a car. These rules also provide a relatively simple means of reasoning with this information, which we can briefly illustrate with the rules in Figure 1.

If, for example, we are told that the battery is old and the alternator is broken, then we can reason as follows. "battery is old" can be used with R1 to learn that "battery is dodgy" is true, and "alternator is broken" can be used with R2 to learn that "battery is charging" is not true. Having established that these facts are true, they can then be used with R3 to learn that "battery is bad" is true, and so on. Finally we can conclude that "radio is not working" and "lights are not working" are true. This kind of reasoning is known as *forward chaining*.

We can also use the rules in *backward chaining* to show the same thing. In this form of reasoning we start from, for example, the desire to determine whether "radio is not working" and look for possible proofs of this fact. In this case there is only one possibility, that presented by R4. To use this rule also requires that "battery is bad" be true, and again there is only one way to establish this fact—the use of R3. To apply R3, it must be the case that "battery is dodgy" is true and "battery is charging" is false, and these themselves can

R1 IF battery is old
THEN battery is dodgy

R2 IF alternator is broken
THEN NOT battery is charging

R3 IF battery is dodgy AND NOT battery is charging
THEN battery is bad

R4 IF battery is bad
THEN radio is not working

R5 IF battery is bad
THEN lights are not working

Figure 1: An example rule-base

only be established by applying R1 and R2. To do this requires that "battery is old" and "alternator is broken" be true, and these, luckily, accord with what we were told to begin with.

What we have given here is a very simplified account of rule-based reasoning, but this is the essence of how it proceeds. There are additional complexities, such as how to do the necessary pattern matching—handled by algorithms like RETE [26] and TREAT [71]—but solutions have been found to these and are implemented in programming environments like OPS5 [25], CLIPS [14] and JESS [49]. These environments allow one to write rules and then invoke forward and backward chaining on them, and so make it possible to simply construct the heart of an expert system.[3]

Rules were used as the basis of many expert systems, including the early systems MYCIN [10, 91], DENDRAL [11, 24], and R1/XCON [68], as well as more recent systems, like DRACO, which helps astronomers sift through large amounts of data [69]. MYCIN, for example, makes use of both forward and backward chaining when attempting to find a diagnosis. It starts using backward chaining to determine whether there is some organism (*significant organism* in the terminology of the system) that it should be treating, and then to determine what bacteria is most likely to be the cause. The first of these tasks is achieved by rules like:

IF organism-1 comes from a sterile site
THEN organism-1 is significant

and the second is achieved by rules like:

[3] And JESS, for example, by allowing arbitrary Java function calls from within rules, makes it possible to combine rules with conventional software.

IF the identity of the organism is not known with certainty,
AND the gram stain of the organism is gramneg,
AND the morphology of the organism is rod,
AND the aerobicity of the organism is aerobic
THEN there is strongly suggestive evidence that the identity of the organism is enterobactericeae.

Once the system has determined that there is a significant organism, and has a likely identity for that organism, it then forward chains to determine what therapy should be applied (in MYCIN all therapies are courses of antibiotics). This forward chaining uses rules like:

IF the identity of the organism is bacteroides
THEN I recommend therapy chosen from among the following drugs:
clindamycin
chloramphenicol
erythromycin
tetracycline
carbenecillin

One of the reasons that rules proved so popular as a mechanism for knowledge representation is that they are very *natural*. By this we mean that it is relatively easy for anyone (including the domain expert) to understand them. It is relatively clear what rules mean, and it is relatively easy for the domain expert to learn how to write them down. However, there are problems with using rules. One such problem is that fact that rules lack a well-defined *semantics*. In other words while, because of their naturalness, it is clear roughly what rules mean, it is not clear exactly what they mean, and this lack of precision makes it hard to be sure exactly what an expert system is doing, or how it will behave. In turn that can make it hard to trust for critical applications.

Another major problem with using rules as described so far as the basis of an expert system is that they are categorical. In our example above, we are only allowed to represent the fact that "battery is old" is true, or is not true. There is no way to represent, for example, that we believe the battery to be old, but are not sure. This is a problem because so much information is not categorical. Indeed in most domains most of the information that an expert system must represent is *imperfect*, and this has prompted the development of a wide variety of mechanisms for representing and reasoning with this imperfect data [77].

Such mechanisms do not work well with rules. Although it is natural to try to associate some kind of a measure of belief with a rule,[4] producing something like:

IF battery good AND battery charging
THEN battery ok (0.8)

early attempts to do this were not very satisfactory. Indeed the best known system of attaching measures to rules, certainty factors [91, 92], turned out to have

[4]Indeed exactly this kind of approach was adopted in MYCIN.

some internal inconsistencies [40, 46]. Because of these failings, it seemed that a better solution was to look for a different kind of knowledge representation.[5]

2.2 Bayesian networks

Rather than thinking in terms of expert rules, let's consider describing a domain in terms of the important variables that it contains. For every variable X_i which captures some aspect of the current state of the domain, one way to express the imperfect nature of the information we have about X is to say that each possible value x_{i_j} of each X_i has some probability $\Pr(x_{i_j})$ of being the current value of X_i. Writing $\mathbf{x}$ for the set of all x_{i_j}, we have:

$$\Pr : x \in \mathbf{x} \mapsto [0, 1]$$

and

$$\sum_j \Pr(x_{i_j}) = 1$$

In other words, the probability $\Pr(x_{i_j})$ is a number between 0 and 1 and the sum of the probabilities of all the possible values of X_i is 1. If X_i is known to have value x_{i_j} then $\Pr(x_{i_j}) = 1$ and if it is known not to have value x_{i_j} then $\Pr(x_{i_j}) = 0$.

Given two of these variables, X_1 and X_2, then the probabilities of the various values of X_1 and X_2 may be related to one another. If they are not related, a case we distinguish be referring to X_1 and X_2 as being *independent*, then for any two values x_{1_i} and x_{2_j}, we have:

$$\Pr(x_{1_i} \wedge x_{2_j}) = \Pr(x_{1_i}) \Pr(x_{2_j})$$

If the variables are not independent, then:

$$\Pr(x_{1_i} \wedge x_{2_j}) = \Pr(x_{1_i} | x_{2_j}) \Pr(x_{2_j})$$

where $\Pr(x_{1_i} | x_{2_j})$ is the probability of X_1 having value x_{1_i} given that X_2 is known to take value x_{2_j}. Such *conditional probabilities* capture the relationship between X_1 and X_2, representing, for instance, the fact that x_{1_i} (the value "wet", say, of the variable "state of clothes") becomes much more likely when x_{2_j} (the value "raining" of the variable "weather condition") is known to be true.

If we take the set of these X_i of which the agent is aware, the set $\mathbf{X}$, then for each pair of variables in $\mathbf{X}$ we can establish whether the pair are independent

[5] Here we are broadly tracing the historical development of these techniques. Initial work on rule-based systems made use of categorical rules. Later it became apparent that mechanisms for handling imperfect knowledge were required, and techniques like certainty factors were developed. When their failings became apparent, and other efforts such as Nilsson's probabilistic logic [75] were also found to be problematic, techniques like the Bayesian networks described in the next section were invented. Subsequently much more satisfactory combinations of probability and logic have been created [20, 31], in turn overcoming limitations of Bayesian networks (which are inherently propositional).

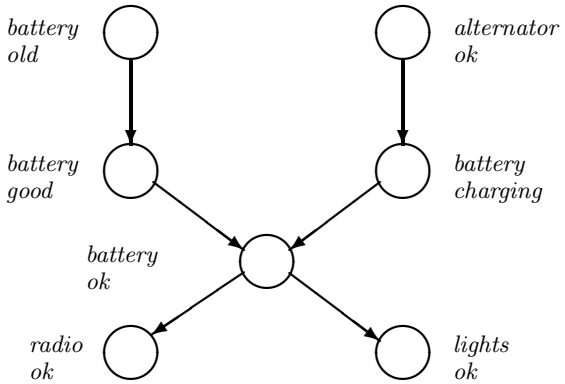

Figure 2: An example Bayesian network

or not. We can then build up a graph in which each node corresponds to a variable in $\mathbf{X}$ and an arc joins two nodes if the variables represented by those nodes are not independent of each other. The resulting graph is known as a Bayesian network[6] [79], and provides a form of knowledge representation that is explicitly tailored to representing imperfect information.

Figure 2 is an example of a fragment of a Bayesian network for diagnosing faults in cars. It represents the fact that the age of the battery (represented by the node *battery old*) has a probabilistic influence on how good the battery is, and that this in turn has an influence on whether the battery is operational (*battery ok*), the latter being affected also by whether the alternator is working and, as a result, whether the battery is recharged when the car moves. The operational state of the battery affects whether the radio and lights will work. In this network it is expected that the observations that can be carried out are those relating to the lights and the radio (and possibly the age of the battery), and that the result of these observations can be propagated through the network to establish the probability of the alternator being okay and the battery being good. In this case these latter variables are the ones that we are interested in since they relate to fixing the car.

As mentioned above, when building an expert system we are not only interested in how to represent knowledge, but also how to reason with it. It turns out that the graphical structure of a Bayesian network provides a convenient computational framework in which to calculate the probabilities of interest to the agent. In general, the expert system will have some set of variables whose values have been observed, and once these observations have been taken, we will want it to calculate the probabilities of the various values of some other set of variables. The details of how this may be achieved are rather complex (see [79] for details), but provide effective[7] algorithms that allow networks with several

[6]The notion of independence captured in the arcs of a Bayesian network is somewhat more complex than that described here, but the difference is not relevant for the purposes of this chapter. For full details, see [79].

[7]It turns out that computing values of probabilities in Bayesian networks is not computa-

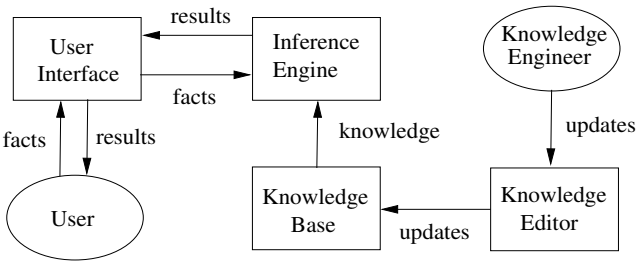

Figure 3: The general architecture of an expert system

hundred variables to be solved in only a few seconds—a speed that is sufficient for all but the most exacting real-time domains. In brief these algorithms work by passing messages between the nodes in the graph. If we observe something that changes the probability of "battery old" in Figure 2, this new value is sent to "battery good", which updates its probability, and sends a message to "battery ok". When "battery ok" updates in turn, it sends messages to "radio ok", "lights ok" and "battery charging", and so on through the network. The full range of algorithms for propagating probabilities through Bayesian networks are described in [13, 16].

We should also note that there have been many successful applications of Bayesian networks. These include PATHFINDER [41], a system for diagnosis of diseases of the lymphatic system, MIDAS [48], a system for dealing with mildew in wheat, and a system for diagnosing faults in the space shuttle [45]. These are all somewhat specialised systems, and ones that most of us will never come into contact with. However, there are expert systems based on Bayesian networks that we have all come into contact with at one time or another—these are the various systems employed by the Microsoft Windows operating systems (from Windows 95 onwards) for tasks such as troubleshooting [43].

2.3 Other aspects of expert systems

Whatever form of knowledge representation is used by an expert system—whether rules, Bayesian networks, or other mechanisms like *frames* [70] and *semantic networks* [85]—there are a number of common features of any expert system.

These common features can be best illustrated by Figure 3, which gives the general architecture of an expert system. Any such system has some form of *knowledge-base*, in which knowledge is stored in some knowledge representation. Associated with this is some form of *inference engine*, which carries out the appropriate kind of reasoning on the knowledge-base. The user of the system interacts with it through some form of graphical user interface, and this interac-

tionally efficient in general—the problem is NP-hard [15] even if the values are only computed approximately [17]—but in many practical cases, the computation can be performed in reasonable time.

tion is typically to inform the system of some things the user knows to be true. This information sparks off some reasoning by the system, which then informs the user of the results of the inference. Finally, a *knowledge engineer* maintains and updates the knowledge-base by means of some kind of *knowledge editor*.

In addition to such common features, there are also a number of problems in developing expert systems whatever the underlying knowledge representation. Perhaps the most severe of these is the *knowledge acquisition bottleneck*. This refers to the fact that the step that most limits the speed of development of an expert system is the acquisition of the knowledge it uses. The traditional approach is to interview a domain expert and obtain their knowledge, but even if this is a suitable technique,[8] it generally proves to be very slow. As a result, there has been much work on trying to automate the process, including the development of techniques for *rule induction* [84, 86], learning Bayesian networks from data [42], and *inductive logic programming* in which logical relations are inferred from data [7, 73].

Another problem that it is worth remembering is the opposition which faced the adoption of many expert systems. When many organisations came to implement expert systems, they found that their employees objected to the idea of bringing in machines as experts. Understandably, those employees saw the use of expert systems as undermining their role—either replacing them in the job that they had previously done, or removing their chance to learn the expertise now encoded in the machine. This problem led to many expert systems (of the form we have been describing), being portrayed as *expert assistants* to some human operator—the role of the assistant being, for example, to remember and point out unlikely but plausible outcomes that the human operator should consider.

3 Expert systems for risk prediction

A number of expert systems have been developed in the broad areas of toxicology and carcinogenicity risk prediction including TOPKAT [21], TOX-MATCH [56], CASETOX [57], HazardExpert [93] and MULTICASE [58]. In this section we look in some detail at DEREK, which predicts various forms of toxicity, and STAR, which predicts carcinogenicity.

3.1 DEREK

DEREK is an expert system for the prediction of toxicology risk [88, 89]. It was developed by LHASA UK, a not-for-profit company based around the Department of Chemistry at the University of Leeds.[9] DEREK is a rule-based system, with many of the features of a classic expert system, the rules being developed by

[8] There are many reasons why it might not be, including the fact that the expert in question may not wish to have their knowledge acquired for use in a computer system.

[9] An early version of DEREK was developed by Schering Agrochemicals and donated to LHASA, which is now responsible for the development of the system [52].

the expert toxicologists who work for LHASA and the various organisations that use DEREK.

A consultation with DEREK begins with the user entering the structure of the chemical in question. The system then compares the structure with rules such as [51]:

> IF a substance contains a carbamate group bearing a small N-alkyl substituents approximately 5.2 angstroms from a nitrogen, oxygen or sulfur atom, bearing a small- to medium-sized lipophilic substituent or substituents and ideally carrying a positive charge at biological pH
> AND it is not too large to fit into the enzyme cavity
> AND it has a log P = -0.5 to + 3.0
> THEN the substance is likely to be insectidal.

The structural information for such rules are written in the language PATRAN [64, 74], and additional information is recorded in the language CHMTRN.

These rules and the structural activity relationships they encode, are used to identify any structural fragments, or *toxicophores*, that are suspected of being the cause of any toxicity in the chemical. These toxicophores are then displayed, along with a description of the toxicity they are suspected of causing—the kinds of toxicity covered by the system are mutagenicity, carcinogenicity, and skin sensitization. The performance of the system is typical of expert systems. Tested on 250 chemicals from the National Toxicology Program salmonella mutagenicity database, DEREK correctly predicted the genotoxicity of 98% of the 112 Ames positive compounds, and the non-genotoxicity of 70% of the Ames negative compounds [36]. An analysis of this performance is contained in [37].

As mentioned above, the rule-base used by DEREK is under constant development, and some of the techniques used for this development have been published. A description of the development and validation of the part of the DEREK rule-base that relates to skin sensitization can be found in [3], while [51] explains how the REX system [50] can be used to automatically generate new rules for DEREK.

3.2 StAR

The StAR project, a collaboration between LHASA and the Imperial Cancer Research Fund[10] developed software for identifying the risk of carcinogenicity associated with chemical compounds [27, 61], extending the work in DEREK.

In the carcinogenicity prediction domain, environmental and epidemiological impact statistics are often unavailable, so an approach known as *argumentation* is adopted. In this approach, the expert system builds arguments, based on whatever information is available, for or against the carcinogenicity of the chemical in question, and uses the interaction between these arguments to estimate

[10]The ICRF has now become Cancer Research UK. The project also involved Logic Programming Associates and City University, London.

the gravity of the risk. Thus if there is one argument that a chemical might be carcinogenic (because it contains some functional group which is known to cause cancer in rats) then there is a risk that the chemical might cause cancer in humans. However, if there is a second argument which defeats the first (by, for instance, pointing out that the cancer-causing mechanism in rats involves an enzyme which is not present in humans) then the risk is considered to be lower. A British Government report on micro-biological risk assessment identifies StAR as a major new approach to this important problem [39].

The demonstrator system produced by the StAR project is a prototype for a computer based assistant for the prediction of the potential carcinogenic risk due to novel chemical compounds. A notion of hazard identification is taken as a preliminary stage in the assessment of risk, and the hazard identification used in StAR draws heavily on the approach taken in DEREK. As described above, DEREK is able to detect chemical sub-structures within molecules, known as structural alerts, and relate these to a rule-base linking them with likely types of toxicity. StAR builds on DEREK's ability to identify structural alerts, but uses a different set of alerts. In particular, the alerts used by StAR were taken from a U.S. FDA report identifying sub-structures associated with various forms of carcinogenic activity [23].

The user of the carcinogenicity risk adviser presents the system with the chemical structure of the compound to be assessed, together with any additional information which may be thought relevant (such as possible exposure routes, or species of animal that will be exposed to the chemical). The chemical structure may be presented using a graphical interface. The database of structural alerts is then searched for matches against the entered structure. If a match is found, a theorem prover tries to construct arguments for or against the hazard being manifest in the context under consideration. Having constructed all the relevant arguments, a report is generated on the basis of the available evidence, and the user can take appropriate action. Thus the StAR system is an expert assistant rather than a system that is intended to take action directly.

For a better understanding of how StAR works, let's look at some examples. For ease of presentation, these examples use a simplified database, and some of the following assessments may be chemically or biologically naive. The argumentation mechanism, however, is accurately described.

The first example is shown in Figure 4. Here, the user has entered a relatively simple structure based on an aromatic ring. The system has identified that it contains an alert for epoxides (the triangular structure to the top right). Whilst constructing arguments, the system has recognised that the LogP value is relevant in this case, and so queries the user for this information (loosely, the value of LogP gives a measure of how easily the substance will be absorbed into tissue). The functional group for epoxides is indicative of a direct acting carcinogen, and the value of LogP supplied by the user is supportive of the substance being readily absorbed into tissue. Hazard recognition plus supportive evidence, with no arguments countering potential carcinogenic activity, yields the classification of a "probable human carcinogen" (the result might be different for different animals). Figure 4 shows the summary report. The query box is

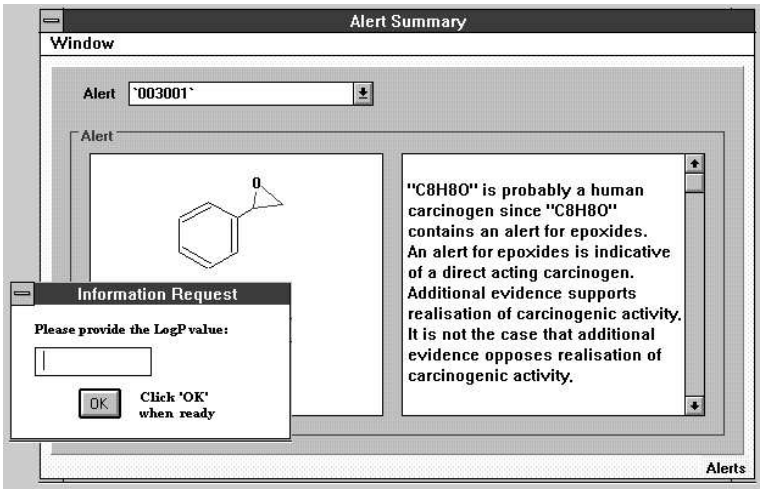

Figure 4: The StAR Demonstrator: Example 1

illustrated in this screen image, although it would normally have been closed by this stage.

The second example is shown in Figure 5. This involves a structure which contains an alert for peroxisome proliferators. The top-most screen contains a simple non-judgemental statement to this effect. The lower screen contains the summary of the argumentation stage of analysis. Here, evidence is equivocal because there is evidence both for and against the carcinogenicity of the compound.

The third example shows how StAR handles equivocal evidence in more detail. Figure 6 shows the reports generated for another compound for which there are arguments for and against carcinogenicity, and so no overall conclusion can be reached. The argument for carcinogenicity is that the structure contains the same peroxisome proliferators alert as in the previous example, and this is indicative of carcinogenic activity in rats and mice. Set against this is the argument that extrapolating from the result in rodents to carcinogenicity in humans is questionable since large doses were required to cause cancer in the test subjects. As indicated in Figure 6, the StAR system can produce further levels of explanation—in this case explanation related to the "high doses required to obtain results in rats and mice".

The use of argumentation in StAR is discussed in more detail in [53] and [60], while the representation of chemical structure—part of the specialised knowledge representation required for the toxicology domain—is described in [95], and the main reference for the system of argumentation that underpins StAR is [59]. It is also worth noting that, as part of the StAR project, experiments were carried out to test people's intuitive understanding of the terms used to

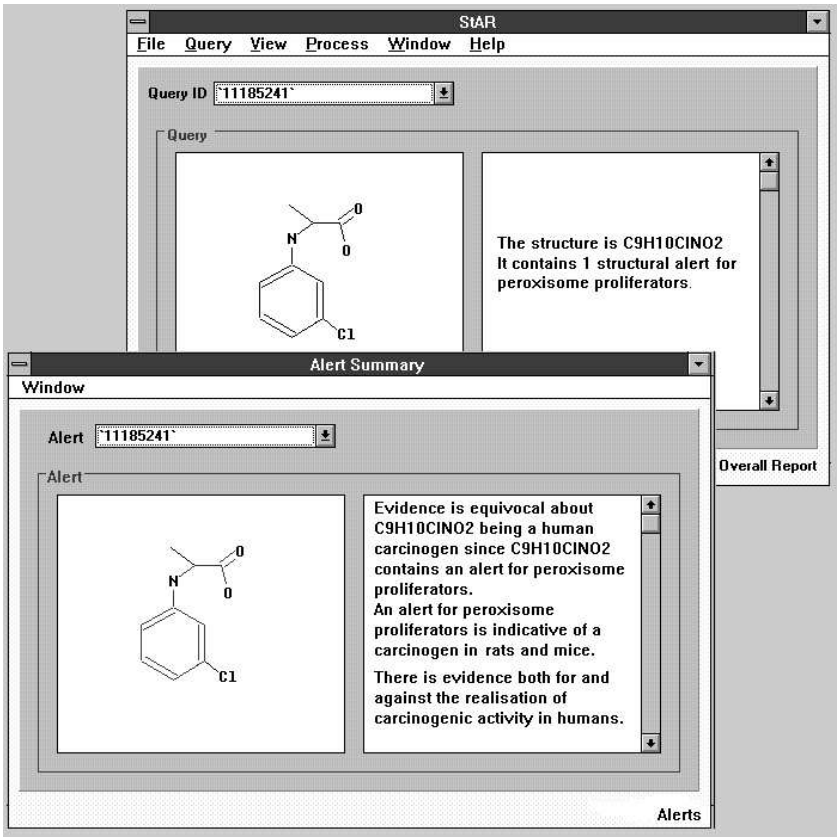

Figure 5: The StAR Demonstrator: Example 2

express the results of the argumentation process. These results are explored in [2, 28]. The experiments on how people interpret arguments are useful in the context of StAR because argumentation is being used not only as a mechanism for reaching a decision about carcinogenicity, but also as a means of explaining it. In this sense argumentation can be considered an extension of approaches like that of CASETOX [57] where properties are inferred on the basis of structural similarity. However, as we will discuss in the next section, argumentation can go far beyond this.

Before we pass on to this discussion, it is worth noting that the results from the StAR project have been incorporated in the successor to DEREK, DEREK for Windows. The model of argumentation used in DEREK for Windows is described in [55], and its use in DEREK for Windows is elaborated in [54]. Finally, [12] describes how an extended version of the argumentation system is applied in the METEOR system for predicting potential metabolic pathways for xenobiotics.

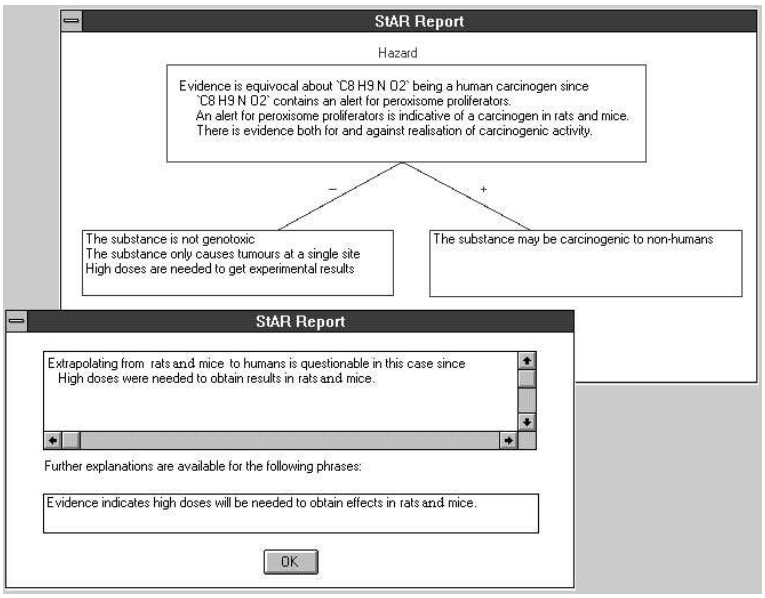

Figure 6: The STAR Demonstrator: Example 3

4 Systems of argumentation

Having given an example of the kind of system that can be built using an argumentation system, we turn to examining in more detail the kind of reasoning that can be handled using this kind of approach.

4.1 An overview of argumentation

An argument for a claim may be considered as a tentative proof for the claim. The philosopher Stephen Toulmin [96] proposed a generic framework for the structure of arguments which has been influential in the design of intelligent systems which use argumentation [29, 60, 98]. Our analysis, informed by Toulmin's structure, considers an argument to have the form of a proof, without necessarily its force.

Suppose ϕ is a statement that a certain chemical is carcinogenic at a specified level of exposure. Then an argument for ϕ is a finite, ordered sequence of inferences $G_\phi = (\phi_0, \phi_1, \phi_2, \ldots, \phi_{n-1})$. Each sub-claim ϕ_i is related to one or more preceding sub-claims ϕ_j, $j < i$, in the sequence as result of the application of an inference rule, R_i, to those sub-claims. The rules

$$\bigcup_i \{R_i\}$$

underwrite the reason why ϕ is a reasonable conclusion, and they correspond to *warrants* in Toulmin's schema and are called *step-warrants* in Verheij's legal

$r1$ $battery\_old \rightarrow battery\_dodgy$
$r2$ $alternator\_broken \rightarrow \neg battery\_charging$
$r3$ $battery\_dodgy \wedge \neg battery\_charging \rightarrow battery\_bad$
$r4$ $battery\_bad \rightarrow \neg radio\_working$
$r5$ $battery\_bad \rightarrow \neg lights\_working$

Figure 7: An example set of formulae

$$\wedge\text{-I}\frac{\vdash \varphi \quad \vdash \psi}{\vdash \varphi \wedge \psi}$$

$$\rightarrow\text{-E}\frac{\vdash \varphi \quad \vdash \varphi \rightarrow \psi}{\vdash \psi}$$

Figure 8: Two rules for natural deduction

argumentation system [98]. Note that R_i and R_j may be the same rule for different i and j.

We may present the sequence for a very simple argument graphically as follows:

$$\phi_0 \xrightarrow{R_1} \phi_1 \xrightarrow{R_2} \phi_2 \longrightarrow \ldots \longrightarrow \phi_{n-1} \xrightarrow{R_n} \phi$$

If any of these rules were rules of inference generally considered valid in deductive logic (modus ponens, say), then we would be confident that truth would be preserved by use of the rule. In other words, using a valid rule of inference at step i means that whenever ϕ_{i-1} is true, so too is ϕ_i. If all the rules of inference are valid in this sense, then the argument G_ϕ constitutes a deductive proof of ϕ. As an example, consider the logical formulae in Figure 7—a reformulation of the rules from Figure 1—and the rules of inference in Figure 8. The first of these rules says that if you can prove two things separately, then you can prove their conjunction. The second is just modus ponens. Now, if we are told that the battery is old and the alternator is broken, then we can give the following argument for $\neg lights\_working$:

$$battery\_old, battery\_dodgy, alternator\_broken,$$
$$\neg battery\_charging, battery\_bad, \neg lights\_working$$

Note that to represent even this straightforward example graphically requires a branching structure like that given in Figure 9 (which covers the first part of the argument only, but enough to show the idea).[11]

[11] A simple linear argument would be:

$$p \wedge q \wedge r \xrightarrow{\wedge-\text{E}} p \wedge q \xrightarrow{\wedge-\text{E}} p$$

where $\wedge$-E is:

$$\frac{\vdash \varphi \wedge \psi}{\varphi}$$

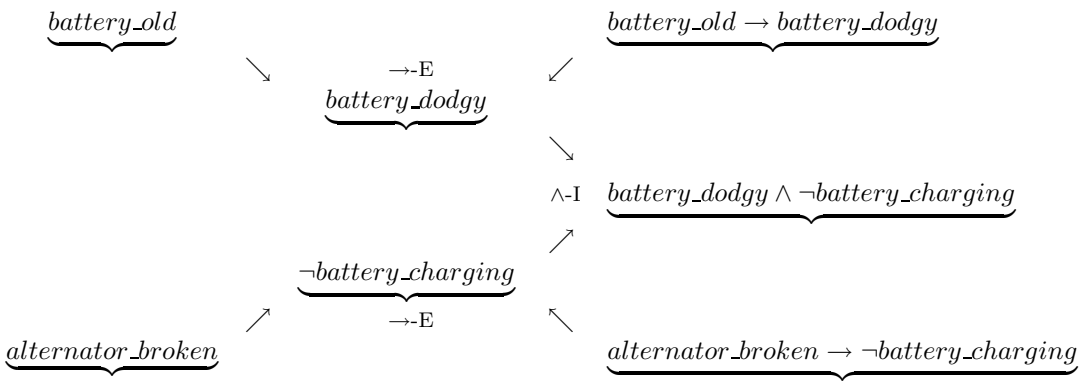

Figure 9: A branching argument structure

This machinery would be sufficient if we were only interested in inferences that were valid in the sense of preserving truth. However, the situations of interest to us in toxicology (as indeed is the case in many domains) are when some or all of the inference rules are not valid.

In pure mathematics in general, once a theorem has been proven true, further proofs do not add to its truth, nor to the extent to which we are willing to believe the theorem to be true. However, even pure mathematicians may have variable belief in an assertion depending upon the means used to prove it. For example, constructivist mathematicians (e.g. [5, 97]) do not accept inference based on proof techniques which purport to demonstrate the existence of a mathematical object without also constructing it. Typically, such proofs use a *reductio ad absurdum* argument, showing that an assumption of non-existence of the object leads to a contradiction. Thus, constructivist mathematicians will seek an alternative proof for an assertion which a non-constructivist mathematician would accept as already proven.

Although originally a contentious notion within pure mathematics, constructivist mathematics has obvious applications to computing, and has recently been proposed as a medium for the foundations of quantum physics [8]. Likewise, in another example, not all mathematicians accept the use of computers in proofs, or may do so only for some proofs. Computers have been used, for instance, to prove the Four Color Map Theorem [1] and to demonstrate the non-existence of projective planes of order 10 [62]. For an interesting deconstruction of mathematical proofs as "objectively existing real things" see Appendix D of [32].

Argumentation extends this idea of different kinds of proof being more or less convincing. In general, all alternative arguments are of great interest, and the greater the number of independent arguments that exist for a claim, the stronger is the case for it, and the stronger may be our belief in its truth. However, in arriving at a considered view as to our belief in the truth of a claim ϕ, we also need to consider the arguments against it, the arguments in favour of its negation $\neg\phi$ (which may not be the same thing), and any arguments which attack its supporting sub-claims, ϕ_i.

Given these different arguments and counter-arguments, it is possible to define a symbolic calculus, called a Logic of Argumentation, which enables the combination ("flattening") of arguments for and against a proposition [30]. Since an argument is a tentative proof of a claim, our degree of belief in the claim will likely depend upon the argument advanced for it. Thus, for each pair (ϕ, G_ϕ) consisting of a claim and an argument for it, we can associate a measure α_ϕ of our strength of belief in ϕ given G_ϕ. We represent this as a triple $(\phi, G_\phi, \alpha_\phi)$, which we call an *assessed argument*.[12] The belief-indicator may be a quantitative measure, such as a probability, or an element from a qualitative dictionary, such as {*Likely, Unlikely*}. In either case, we can define algebraic operations on the set of belief-indicators (the "denotation dictionary for belief") enabling us to generate the degree of belief in a combined argument,

[12]The use of "assessment" here is analogous to the concept of valuation in mathematical logic [83].

when we know the degrees of belief of the subsidiary arguments. In addition to belief-indicators, one can also define other labels for claim-argument pairs, such as the values of world-states and the consequences of actions arising from the claim [30].

The fact that we can attach measures to arguments based upon their relationship to other arguments (a relationship that can be based upon the contents of the grounds or the sub-claims) is very powerful. It gives us a way of defining alternative *meta-theories* for argumentation systems, and of using these meta-theories for reasoning about the arguments themselves. This is the key feature of argumentation, and one that distinguishes it from other approaches to reasoning about toxicology.

4.2 Argumentation applied to prediction of carcinogenicity

The STAR system described above is one example of how argumentation can be applied to risk prediction, and the approach has been applied by workers at Cancer Research UK to other risk prediction problems (see [78], for another example). There are other ways of using argumentation, however, and in this section we describe one direction in which we are moving [66]. We start with the question :

> On what basis do scientists claim that a chemical substance is carcinogenic?

Such claims can be based upon evidence from a number of sources (adapted from [22] and [35]):

- Using chemical theoretical reasoning, on the basis of the chemical structure of the substance and the known carcinogenicity of chemicals with congeneric structures.
- From mutagenicity tests, applying the substance to tissue-cultures in laboratory experiments.
- From experiments involving the application of the chemical to human or animal cadavars.
- From bioassays, applying the substance to animals in a laboratory experiment.
- From epidemiological studies of humans, either case-control studies (where a case group of people exposed to the substance is matched with a control group not so exposed, and their relative incidences of cancer compared), or cohort studies (where the incidence of the cancer among people exposed to the substance is compared with that in the general population, while controlling for other potential causal and interacting factors).
- From elucidation of theoretically-sound bio-medical causal pathways.[13]

[13] These are E-theories in Pera's [80, page 154] typology of scientific theories.

Now, elucidation of causal pathways is generally not undertaken until evidence of an empirical nature is observed. Hence, we focus on the other categories of evidence. There are a number of comments one can make on the relative value of these different approaches. Reasoning from chemical structure is still an imprecise and immature science for most substances; indeed, automated prediction of carcinogenicity and other properties of chemicals on the basis of their structure is an active area of Artificial Intelligence research [44, 94]. Mutagenic tests may demonstrate carcinogenicity in principle, but do not reveal what will happen in a whole, living organism (with, for instance, viral defences), nor in an environment similar to that of people exposed to the substance. Experiments with cadavars have similar difficulties. Moreover, because the incidence rates of many cancers are very small, epidemiological studies may require large sample sizes, and so can be quite expensive. Also, the time-lag between exposure to typical environmental doses and the onset of a cancer can be very long (in the order of decades), so these studies can take years to complete. For these reasons and others, the most common form of assessment of potential carcinogenicity is the bioassay.

We therefore turn our attention to animal bioassays. Because of the difficulties in inferring conclusions about humans on the basis of evidence about animal species, most cautious scientists and policy makers would not *assert* carcinogenicity to humans from a bioassay: they would, at best, only claim that there is a (perhaps high) probability of human carcinogenicity.[14] However, although it is perhaps the most contentious, the animal-to-human inference is not the only inference being deployed in concluding such a probability. It is also not the only inference deployed when quantifying the extent of risk. It therefore behooves us to examine all the modes of inference used. In doing so, we have abstracted from a number of descriptions and critiques of carcinogenic risk assessment processes [4, 6, 9, 22, 33, 34, 35, 47, 63, 72, 76, 82, 87, 90], both ideal and actual.

For the purposes of exposition, we therefore suppose an archetypal animal bioassay for a chemical substance $\mathcal{X}$ is undertaken. This will involve the administration of specific doses of $\mathcal{X}$ to selected animal subjects, usually repeatedly, in a laboratory environment. Typically, two or three non-zero dose-levels are applied to the subject animals, along with a zero-dose to the control group. The rates at which cancers of a specific nature develop is then observed in each group until a pre-determined time-point (usually the natural life-span of the animal). Those animals still alive at that time are then killed, and a statistical analysis of the hypotheses that exposure to the substance $\mathcal{X}$ results in increased incidence of cancer is then undertaken. Suppose that, based on this animal bioassay, a claim is then made that $\mathcal{X}$ is carcinogenic to humans at a specified dose. For ease of expression we will notate this claim by ϕ. In asserting ϕ from the evidence of the bioassay, a number of subsidiary inferences need to

[14] Indeed, the USA Environmental Protection Agency guidelines [22] permit one to claim probable human carcinogenicity from (sufficiently strong) animal evidence alone. Although such a claim would be classed in the second of two categories of "probable", it is still above "possible" human carcinogenicity.

be made. We have expressed these in the form of *"FROM antecedent TO consequent"*. This is short-hand for saying that an act of inference is undertaken whenever one assumes that the consequent is true (or takes a particular value) upon the antecedent being true (or, respectively, having taken a corresponding value).

The list of subsidiary inferences is as follows:

1. **FROM Administered dose TO Delivered dose.** Animal bodies defend themselves against foreign substances. Their ability to do this may be impacted by the amount of the foreign substance ingested or to which the animal is exposed. For example, chemicals applied to nasal tissues are initially repelled by defences in the tissues themselves. Larger doses may destroy this first line of defence, thereby permitting proportionately more of the chemical to enter the body's circulatory pathways than would occur for smaller doses. In other words, the dose delivered to the target tissue or organ of the body may not be proportionate to the dose administered to the animal by the experimenter.

2. **FROM A sample of animals TO A population of the same species.** Reasoning from a sample to a population from which the sample is drawn in known as statistical inference.

3. **FROM A genetically uniform animal population TO A genetically more diverse population.** Animal subjects used in laboratory experiments are often closely related genetically, both in order to control for the impact of genetic diversity on responses and because, for reasons of convenience, subjects are used from readily-available sources. Consequently, the animal subjects used in bioassays are often not as diverse genetically as would be a wild population of the same species.

4. **FROM An animal population TO The human population.** This is perhaps the most contentious inference-step in carcinogenicity claims from bioassays. Animals differ from humans in their physiology and in their body chemistry, so it is not surprising that they also differ from us in reactions to potential carcinogens. Indeed, they differ from each other. According to Graham *et al.* [34, page 18], writing more than a decade ago, "Several hundred chemicals are known to be carcinogenic to laboratory animals, but direct evidence of their human carcinogenicity is either insufficient or nonexistent." Formaldehyde, for instance, was found to cause significant nasal cancers in rats but not in mice [34], while epidemiological studies of humans whose professions exposed them to high levels of the chemical found no significant increases in such cancers. Conversely—and perversely—epidemiological studies did reveal significant increases in brain cancers and leukaemias, for which there was no biologically-plausible explanation [34].

5. **FROM A site specificity in bioassay animals TO A possibly different site specificity in humans.** Most chemicals are pre-carcinogens

which must be altered by the body's metabolic processes into an actively carcinogenic form. This happens differently in different species, because the body-chemistries are different or because the physiology or relative sizes of organs are different. Hence, a chemical may cause liver cancer in one animal species, but not in another species, or act elsewhere in another.

6. **FROM Localised exposure TO Broader exposure.** Bioassays administer a chemical to a specific site in a specific way to the subject animals, as for example, in bioassays of formaldehyde applied to nasal passages to test for nasal cancer. In contrast, humans exposed to it may receive the chemical in a variety of ways. Morticians exposed to formaldehyde may receive it via breathing and by direct application to their skin, for example.

7. **FROM Large doses TO Small doses.** At typical levels of exposure, the incidences of most individual cancers in the general population are quite small, of the orders of a few percent or much less. At equivalent dose levels, then, bioassays will require very large sample sizes to detect statistically significant increases in cancer incidence. This would be prohibitively expensive, and so most bioassays administer doses considerably greater than the equivalent doses received (allowing for the relative sizes of the animal and human species) in the environment. In order to assert carcinogenicity, then, a conversion model—a dose-response curve—is required to extrapolate back from large to small dose levels.

   While one might expect the dose-response curve to slope upwards with increasing dose levels, this is not always the case. For example, high doses of a chemical may kill cells before they can become cancerous; or a chemical may be so potent that even low doses initiate cancer in all cells able to be so initiated, and thus higher doses have no further or a lesser effect. Indeed, if the chemical is believed to be mutagenic as well as carcinogenic, then even a single single molecule of the chemical should cause an effect. The issue of whether or not a threshold level for dose exists (below which no response would be observed) is a contentious one in most cases. Fuelling controversy is the fact that claims of carcinogenicity can be very sensitive to the dose-response model used. Two theoretically-supported models for the risks associated with aflatoxin peanuts, for example, show human risk likelihood differing by a factor of 40,000 [82]. Similarly, the Chief Government Medical Officer of Great Britain recently admitted that the number of people eventually contracting CJD in Britain as a result of eating contaminated beef may be anywhere between a few hundred and several million [99]. For this reason, this inference is probably the most controversial aspect of carcinogenicity claims, after that of animal-to-human inference (Inference-Mode No. 4 above).

8. **FROM An animal dose-level TO A human equivalent.** The discussion of Inference-Mode no. 7 used the phrase "allowing for the relative sizes of the animal and human species". But how is this to be done? Is the

dose extrapolated according to relative body weights of the two species (animal and human); or skin surface area (which may be appropriate for chemicals absorbed through the skin); or relative size of the organ affected? What is appropriate if different organs are affected in different species?

9. **FROM Administered doses TO Environmental exposure.** In order to expedite response times, bioassays may administer the chemical in a manner different to that likely to be experienced by humans exposed to it in their environment. For example, the chemical may be fed via a tube directly into the stomach of the animal subject, which is unlikely to be the case naturally.

10. **FROM A limited number of doses TO Cumulative exposure.** Some chemicals may only produce adverse health effects after a lifetime of accumulated exposure. Body chemistry can be very subtle, and a small number of large doses of a chemical may have a very different impact from a much larger number of smaller doses, even when the total dose received is the same in each case.

11. **FROM A pure chemical substance TO A chemical compound.** Most chemicals to which people are exposed are compounds of several chemicals, not pure substances. Bioassay experiments, however, need to be undertaken with pure substances, so as to eliminate any spurious causal effects. Consequently, a bioassay will not be able to assess any effects due to interactions between substances which occur in a real environment, including any transformations which take place inside the human body.

12. **FROM The human population TO Individual humans.** Individuals vary in their reactions to chemical stimuli, due to factors such as their genetic profiles, lifestyles, and personalities. Risks of carcinogenicity may be much higher or much lower than claimed for specific groups or individuals.

These forms of inference could correspond to the inference rules R_i discussed in the previous section.

To claim human carcinogenicity on the basis of evidence from a bioassay thus depends on a number of different modes of inference, each of which must be valid for the claim to stand. We could write:
"The chemical $\mathcal{X}$ is carcinogenic to humans at dose d based on a bioassay of animal species a if:

- *There is a relationship between administered dose and delivered dose in the bioassay, AND*

- *The sample of animals used for the experiment was selected in a representative manner from the population of animals, AND*

- *The animal population from which the sample was drawn is as genetically diverse as the animal population as a whole, AND*
- *The specific animal physiology and chemistry relevant to the activity of $\mathcal{X}$ is sufficiently similar to human physiology and chemistry,"*

⋮

and so on, through the remaining eight inference steps.

It is important to note that even if all modes of inference were valid in a particular case, our assertion could, strictly speaking, only validly be that the chemical $\mathcal{X}$ is associated with an increase in incidence of the particular cancer. The assertion ϕ does not articulate, nor could a bioassay or epidemiological study prove, a causal pathway from one to the other. There may, for example, be other causal factors leading both to the presence of the chemical in the particular environment and to the observed carcinogenicity.

For the archetypal analysis above, we began with the assumption of just one bioassay being used as evidence to assert a claim for carcinogenicity. In reality, however, there is often evidence from more than one experiment and, if so, statistical meta-analysis may be appropriate [81]. This may involve pooling of results across different animal species, or across both animal and human species.[15] None of these tasks are straightforward, and will generally involve further modes of inference, which we have yet to explore. The situation is further complicated by the fact that most chemical substances which adversely impact the body cause a number of effects—cell mutation, malignant tumours, benign tumours, toxicity to cells, cell death, cell replication, suppression of the immune system, endocrine disturbances and so on. Some of these clearly interact—dead cells cannot then become cancerous, for instance—and the extent of interaction may be a non-linear function of the dose levels delivered. Simple claims about carcinogenicity often ignore these other effects and their interactions with the growth of malignant tumours ("carcinogenicity"). We do not deal with this issue here.

It is possible that working biomedical scientists and scientific risk assessors would consider the list above to be an example of extreme pedantry, and that many of these modes of inferences are no more than assumptions made in order to derive usable results. We have treated them as inference-modes so as to be quite clear about the reasoning processes involved. Our purpose in doing so is to make possible the automation of these processes, which we believe we can do using an argumentation formalism as described above.

For now this still remains to be done. However, we have begun to take some steps in this direction. In [65] we describe a formal system of argumentation that includes in the grounds of an argument the inference rules used (in effect replicating the full chain of reasoning in the grounds). This makes it possible to attack an argument not only in terms of the formulae used in its construction,

[15] The U.S.A. Environmental Protection Agency Guidelines [22] deal, at a high level, with the second issue.

but also the mode of inference. This, in turn, paves the way for argumentation based upon different logics, logics that can capture the different inference-modes described above.

We have also [67] investigated how argumentation may be used to support the process of scientific enquiry. This work shows how a system of argumentation may be used to keep track of different claims, about the toxicity of a chemical for instance, and to summarise the overall belief in the claim at a particular time. In addition, we have shown that this form of argumentation eventually converges on the right prediction about toxicity given the evidence—showing that the system exhibits the necessary soundness of reasoning. Future work is to combine these two pieces of work, allowing the different claims in the latter to be based on the different kinds of reasoning of the former.

5 Summary

The aims of this chapter were to examine the extent to which expert systems can provide a suitable solution to the problem of predicting toxicity, and to provide some pointers to those who want to try such a solution for themselves. These aims were achieved in the following way.

First, we gave a description of the kind of knowledge representation and reasoning possible with production rules and Bayesian networks, two of the main approaches to building expert systems. Then we looked in some detail at two particular expert systems, DEREK and StAR, that predict toxicology risk. The heart of the StAR system is provided by a mechanism for argumentation—a system of reasoning that builds up reasons for and against predictions in order to decide which is best. This kind of reasoning is described in detail, along with the directions in which we are developing the theory. This, we feel, gives a good survey of the way in which expert systems techniques can be applied to toxicology risk prediction. The second aim was achieved by the provision of copious references throughout the chapter.

In summary, our view is that expert systems techniques are a good foundation from which to attack the problem of predicting toxicology risk. We feel that the most promising of these techniques is that of argumentation, a position supported by [39], although argumentation needs to be extended, along the lines described above, before its full value will be realised.

References

[1] K. Appel and W. Haken. Every planar map is four colorable. *Bulletin of the American Mathematics Society*, 82:711–712, 1976.

[2] P. Ayton and E. Pascoe. Bias in human judgement under uncertainty? *Knowledge Engineering Review*, 10:21–41, 1995.

[3] M. D. Barratt and J. J. Langowski. Validation and subsequent development of the derek skin sensitization rulebase by analysis of the BgVV list of contact allergens. *Journal of Chemical Information and Computer Sciences*, 39:294–298, 1999.

[4] D. J. Benford and D. R. Tennant. Food chemical risk assessment. In D. R. Tennant, editor, *Food Chemical Risk Analysis*, pages 21–56. Blackie Academic and Professional, London, UK, 1997.

[5] E. Bishop. *Foundations of Constructive Analysis*. McGraw-Hill, New York City, NY, USA, 1967.

[6] C. P. Boyce. Comparison of approaches for developing distributions for carcinogenic slope factors. *Human and Ecological Risk Assessment*, 4 (2):527–577, 1998.

[7] I. Bratko and S.H. Muggleton. Applications of Inductive Logic Programming. *Communications of the ACM*, 38(11):65–70, 1995.

[8] D. S. Bridges. Can Constructive Mathematics be applied in physics? *Journal of Philosophical Logic*, 28:439–453, 1999.

[9] C. L. Broadhead, R. D. Combes, and M. Balls. Risk assessment: alternatives to animal testing. In D. R. Tennant, editor, *Food Chemical Risk Analysis*, pages 133–162. Blackie Academic and Professional, London, UK, 1997.

[10] B. G. Buchanan and E. H. Shortliffe, editors. *Rule-Based Expert Systems: The* MYCIN *Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984.

[11] B. G. Buchanan, G. L. Sutherland, and E. A. Feigenbaum. Heuristic DENDRAL: A program for generating explanatory hypotheses in organic chemistry. In B. Meltzer, D. Michie, and M Swann, editors, *Machine Intelligence 4*, pages 209–254. Edinburgh University Press, Edinburgh, Scotland, 1969.

[12] W. G. Button, P. N. Judson, A. Long, and J. D. Vessey. Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *Journal of Chemical Information and Computer Sciences*, 43:1371–1377, 2003.

[13] E. Castillo, J. M. Gutiérrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer Verlag, Berlin, Germany, 1997.

[14] `http://www.ghg.net/clips/CLIPS.html`.

[15] G. F. Cooper. The computational complexity of probabilistic inference using belief networks. *Artificial Intelligence*, 42:393–405, 1990.

[16] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer Verlag, Berlin, Germany, 1999.

[17] P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

[18] R. Davis, B. G. Buchanan, and E. H. Shortliffe. Production rules as a representation for a knowledge-based consultation system. *Artificial Intelligence*, 8:15–45, 1977.

[19] R. Davis and J. J. King. An overview of production systems. In E. W. Elcock and D. Michie, editors, *Machine Intelligence 8*, pages 300–334. Wiley, 1969.

[20] L. De Raedt and K. Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5(1):31–48, July 2003.

[21] K. Enslein, V. K. Gombar, and B. W. Blake. Use of SAR in prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutation Research*, 305:47–61, 1994.

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

[23] Federal Drug Administration. *General principles for evaluating the safety of compounds used in food-producing animals: Appendix 1. Carcinogen structure Guide*. FDA, 1986.

[24] E. A. Feigenbaum, B. G. Buchanan, and J. Lederberg. On generality and problem solving: A case study using the DENDRAL program. In B. Meltzer and D. Michie, editors, *Machine Intelligence 6*, pages 165–190. Edinburgh University Press, Edinburgh, Scotland, 1969.

[25] C. L. Forgy. OPS5 users's guide. Technical Report CMU-CS-81-135, Carnegie-Mellon University, Pittsburgh, 1981.

[26] C.L. Forgy. Rete: a fast algorithm for the many pattern/many object pattern match problem. *Artificial Intelligence*, 19:17–37, 1982.

[27] J. Fox. Will it happen? can it happen? *Science and Public Affairs*, Winter '97:45–48, 1997.

[28] J. Fox, D. Hardman, P. Krause, P. Ayton, and P. Judson. Risk assessment and communication: a cognitive engineering approach. In A. Macintosh and C. Cooper, editors, *Applications and Innovations in Expert Systems III*. Cambridge University Press, 1995.

[29] J. Fox, P. Krause, and S. Ambler. Arguments, contradictions and practical reasoning. *Proceedings of the European Conference on Artificial Intelligence 1992, Vienna, Austria*, 1992.

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

[31] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 1300–1309, Stockholm, Sweden, 1999.

[32] J. Goguen. An introduction to algebraic semiotics, with application to user interface design. In C. L. Nehaniv, editor, *Computation for Metaphors, Analogy, and Agents*, pages 242–291. Springer Verlag (Lecture Notes in Artificial Intelligence 1562), Berlin, Germany, 1999.

[33] J. D. Graham. Historical perspective on risk assessment in the Federal government. *Toxicology*, 102:29–52, 1995.

[34] J. D. Graham, L. C. Green, and M. J. Roberts. *In Search of Safety: Chemicals and Cancer Risk*. Harvard University Press, Cambridge, MA, USA, 1988.

[35] 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.

[36] N. Greene. Computer software for risk assessment. *Journal of Chemical Information and Computer Sciences*, 37:148–150, 1997.

[37] N. Greene, P. N. Judson, J. J. Langowski, and C. A. Marchant. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, acrostAR and meteor. *SAR and QSAR in Environmental Research*, 10:299–314, 1999.

[38] C. L. Hamblin. Translation to and from Polish notation. *Computer Journal*, 5:210–213, 1962.

[39] Health and Safety Commission. *Advisory Committee on Dangerous Pathogens (UK), Microbiological risk assessment, Interim report*. HMSO, 1996.

[40] D. E. Heckerman. Probabilistic interpretation of MYCIN's certainty factors. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196. Elsevier, Amsterdam, 1986.

[41] D. E. Heckerman. *Probabilistic Similarity Networks*. MIT Press, 1991.

[42] D. E. Heckerman. A tutorial on learning in Bayesian networks. In M. I. Jordan, editor, *Learning in graphical models*. Kluwer, Dordrecht, The Netherlands, 1998.

[43] D. E. Heckerman, J. S. Breese, and K. Rommelse. Decision-theoretic troubleshooting. *Communications of the ACM*, 38:49–57, 1995.

[44] C. Helma, S. Kramer, and B. Pfahringer. Carcinogenicity prediction for noncongeneric compounds: experiments with the machine learning program SRT and various sets of chemical descriptors. In *Proceedings of the Twelfth European Symposium on Quantitative Structure-Activity Relationships: Molecular modelling and prediction of bioactivity*, Copenhagen, Denmark, 1998.

[45] E. Horvitz, C. Ruokangas, S. Srinivas, and M. Barry. A decision-theoretic approach to the display of information for time-critical decisions: Project Vista. Technical Memorandum 96, Rockwell International Science Center, Palo Alto Laboratory, 1992.

[46] E. J. Horvitz and D. E. Heckerman. The inconsistent use of measures of certainty in artificial intelligence research. In L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 137–151. Elsevier, Amsterdam, 1986.

[47] D. Jamieson. Scientific uncertainty and the political process. *Annals of the American Academy of Political and Social Science*, 545:35–43, 1996.

[48] A. L. Jensen and F. V. Jensen. Midas: An influence diagram for management of mildew in winter wheat. In E. Horvitz and F. V. Jensen, editors, *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 349–356, San Francisco, CA, 1996. Morgan Kaufmann.

[49] `http://herzberg.ca.sandia.gov/jess/`.

[50] P. N. Judson. QSAR and expert systems in the prediction of biological activity. *Pestcide Science*, 36:155–160, 1992.

[51] P. N. Judson. Rule induction for systems predicting biological activity. *Journal of Chemical Information and Computer Sciences*, 34:148–153, 1994.

[52] P. N. Judson and R. D. Coombes. Artificial intelligence systems for predicting toxicity. *Pesticide Outlook*, 7 (4):11–15, 1996.

[53] P. N. Judson, J. Fox, and P. J. Krause. Using new reasoning technology in chemical information systems. *Journal of Chemical Information and Computer Sciences*, 36:621–624, 1996.

[54] P. N. Judson, C. A. Marchant, and J. D. Vessey. Using argumentation for abssolute reasoning about the potential toxicity of chemicals. *Journal of Chemical Information and Computer Sciences*, 43:1364–1370, 2003.

[55] P. N. Judson and J. D. Vessey. A comprehensive approach to argumentation. *Journal of Chemical Information and Computer Sciences*, 43:1356–1363, 2003.

[56] J. J. Kaufman. Strategy for computer-generated theoretical and quantum chemical prediction of toxicity and toxicology (and pharmacology in general). *International Journal of Quantum Chemistry*, 8:419–439, 1981.

[57] G. Klopman. Predicting toxicity through a computer automated structure evaluation program. *Environmental Health Perspectives*, 61:269–274, 1985.

[58] G. Klopman and H. S. Rosenkrantz. Approaches to SAR in carcinogenesis and mutagenesis: prediction of carcinogenicity/mutagenicity using MULTICASE. *Mutation Research*, 305:33–46, 1994.

[59] P. Krause, S. Ambler, M. Elvang-Gøransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11:113–131, 1995.

[60] 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.

[61] P. Krause, P. Judson, and M. Patel. Qualitative risk assessment fulfills a need. In A. Hunter and S. Parsons, editors, *Applications of Uncertainty Formalisms*. Springer Verlag, Berlin, 1998.

[62] C. W. H. Lam, S. Swiercz, and L. Thiel. The non-existence of finite projective planes of order 10. *Canadian Journal of Mathematics*, 41:1117–1123, 1989.

[63] D. P. Lovell and G. Thomas. Quantitative risk assessment. In D. R. Tennant, editor, *Food Chemical Risk Analysis*, pages 57–86. Blackie Academic and Professional, London, UK, 1997.

[64] C. Marshall. *Computer Assisted Design of Organic Synthesis*. PhD thesis, University of Leeds, 1984.

[65] P. McBurney and S. Parsons. Tenacious tortoises: a formalism for argument over rules of inference. In G. Vreeswijk, editor, *Workshop on Computational Dialectics, Fourteenth European Conference on Artificial Intelligence (ECAI2000)*, Berlin, Germany, 2000. ECAI.

[66] P. McBurney and S. Parsons. Dialectical argumentation for reasoning about chemical carcinogenicity. *Logic Journal of the IGPL*, 9(2):191–203, 2001.

[67] 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.

[68] J. McDermott. R1: A rule-based configurer of computer systems. *Artificial Intelligence*, 19(1):41–72, 1982.

[69] G. Miller. The data reduction expert assistant. In A. Heck and F. Murtagh, editors, *Astronomy from Large Databases II*, Hagenau, France, 1992.

[70] M. Minsky. A framework for representing knowledge. In P. H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, New York, 1975.

[71] D. P. Miranker. Treat: A better match algorithm for ai production system matching. In *Proceedings of the National Conference on Artificial Intelligece*, pages 42–47, 1987.

[72] S. H. Moolgavkar. Stochastic models of carcinogenesis. In C. R. Rao and R. Chakraborty, editors, *Handbook of Statistics, Volume 8: Statistical Methods in Biological and Medical Sciences*, pages 373–393. North-Holland, Amsterdam, The Netherlands, 1991.

[73] S.H. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19,20:629–679, 1994.

[74] G. J. Myatt. *Computer Aided Estimation of Synthetic Accessibility*. PhD thesis, University of Leeds, 1994.

[75] N. J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

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

[77] S. Parsons. Reasoning with imperfect information. In C. T. Leondes, editor, *Expert Systems, Volume I*, pages 79–117. Academic Press, New York, NY, USA, 2002.

[78] S. Parsons, J. Fox, and A. Coulson. Argumentation and risk assessment. In *Proceedings of the AAAI Spring Symposium on Predictive Toxicology*, Stanford, March 1999.

[79] J. Pearl. *Probabilistic Reasoning in Intelligent Systems; Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA., 1988.

[80] M. Pera. *The Discourses of Science*. University of Chicago Press, Chicago, IL, USA, 1994.

[81] D. B. Petitti. *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Oxford University Press, Oxford, UK, 1994.

[82] R. A. Pollak. Government risk regulation. *Annals of the American Academy of Political and Social Science*, 545:25–34, 1996.

[83] S. Popkorn. *First Steps in Modal Logic*. Cambridge University Press, Cambridge, UK, 1994.

[84] F. J. Provost and V. Kolluri. A survey of methods for scaling up inductive algorithms. *Data Mining and Knowledge Discovery*, 3(2):131–169, 1999.

[85] M. R. Quillian. A design for an understanding machine. In *Colloqioum on Semantic Problems in Natural Language*, Kings College, Cambridge, 1961.

[86] J. R. Quinlan. Generating production rules from decision trees. In *Proceedings of the Tenth International Joint conference on Artificial intelligence*, pages 304–307, Los Altos, CA, 1997. Morgan Kaufmann.

[87] L. R. Rhomberg. A survey of methods for chemical health risk assessment among Federal regulatory agencies. *Human and Ecological Risk Assessment*, 3 (6):1029–1196, 1997.

[88] J. E. Ridings, M. D. Barratt, R. Cary, C. G. Earnshaw, C. E. Eggington, M. K. Ellis, P. N. Judson, J. J. Langowski, C. A. Marchant, M. P. Payne, W. P. Watson, and T. D. Yith. Computer prediction of possible toxic action from chemical structure; an update on the DEREK system. *Toxicology*, 106:267–279, 1996.

[89] D. M. Sanderson and C. G. Earnshaw. Computer prediction of possible toxic action from chemical structure; the DEREK system. *Human & Experimental Toxicology*, 10:261–273, 1991.

[90] M. E. Shere. The myth of meaningful environmental risk assessment. *Harvard Environmental Law Review*, 19 (2):409–492, 1995.

[91] E. H. Shortliffe. *Computer-Based Medical Consultations:*MYCIN. Elsevier, Amsterdam, The Netherlands, 1976.

[92] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379, 1975.

[93] M. P. Smithing and F. Darvas. HazardExpert, an expert system for predicting chemical toxicity. In J. W. Finley, S. F. Robinson, and D. J. Armstrong, editors, *Food Safety Assessment*, pages 191–200. American Chemical Society, Washington, D.C., 1992.

[94] A. Srinivasan, S. H. Muggleton, M. J. E. Sternberg, and R. D. King. Theories of mutagenicity: a study in first-order and feature-based induction. *Artificial Intelligence*, 85:277–299, 1995.

[95] C. A. G. Tonnelier, J. Fox, P. N. Judson, P. J. Krause, N. Pappas, and M. Patel. Representation of chemical structures in knowledge-based systems: The StAR system. *Journal of Chemical Information and Computer Sciences*, 37:117–123, 1997.

[96] S. E. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, UK, 1958.

[97] A. S. Troelstra and D. van Dalen. *Constructivism in Mathematics: An Introduction (Two Volumes)*. North-Holland, Amsterdam, The Netherlands, 1988.

[98] B. Verheij. Automated argument assistance for lawyers. In *Proceedings of the Seventh International Conference on Artificial Intelligence and Law, Oslo, Norway*, pages 43–52, New York City, NY, USA, 1999. ACM.

[99] N. Watt. Millions still at risk from CJD. *The Guardian (London, UK)*, page 1, 22 September 1999.