Heuristics structure and pervade formal risk assessment

Brian H MacGillivray

macgillivraybh@cardiff.ac.uk

DOI: 10.1111/risa.12136

ABSTRACT

Lay perceptions of risk appear rooted more in heuristics than in reason. A major concern of the risk regulation literature is that such "error-strewn" perceptions may be replicated in policy, as governments respond to the (mis)fears of the citizenry. This has led many to advocate a relatively technocratic approach to regulating risk, characterised by high reliance on formal risk and cost-benefit analysis. However, through two studies of chemicals regulation, we show that the formal assessment of risk is pervaded by its own set of heuristics. These include rules to categorise potential threats, define what constitutes valid data, guide causal inference, and to select and apply formal models. Some of these heuristics lay claim to theoretical or empirical justifications, others are more back-of-the-envelope calculations, whilst still more purport not to reflect some truth but simply to constrain discretion or perform a desk-clearing function. These heuristics can be understood as a way of authenticating or formalising risk assessment as a scientific practice, representing a series of rules for bounding problems, collecting data, and interpreting evidence (a *methodology*). Heuristics are indispensible elements of induction. And so they are not problematic *per se*, but they can become so when treated as laws rather than as contingent and provisional rules. Pitfalls include the potential for systematic error, masking uncertainties, strategic manipulation, and entrenchment. Our central claim is that by studying the rules of risk assessment *qua* rules, we develop a novel *representation* of the methods, conventions, and biases of the prior art.

Keywords: science-policy, risk-regulation, governance, model evaluation, inference rules.

1. INTRODUCTION: THE FALL OF HOMO ECONOMICUS AND THE RISE OF TECHNOCRATIC RISK REGULATION

"For regulation of the risks associated with particulate matter or acid deposition, it makes no sense to consult heuristic-based judgments. The best approach is instead highly quantitative, based on an analysis of the costs and benefits of various possible approaches." (Cass Sunstein, 2008) [1]

A substantial body of research suggests that lay perceptions of risk are rooted more in cognitive heuristics than in reason; *homo economicus* is no longer with us. [2-4] In short, when forming judgements under uncertainty people often answer complex, time consuming questions (*e.g.* what is the risk posed by nuclear power?) by substituting them with simpler ones (*e.g.* how do I *feel* about nuclear power?). Although such heuristics are argued by some to be efficient and often accurate tools for dealing with bounded rationality, [e.g. 5] the focus of the risk regulation literature has been on the systematic and predictable errors that arise from their application. [4, 6-9] The concern is that such "erroneous" risk perceptions may be replicated in law, policy and regulation, as democratic governments respond to the (mis)fears of the citizenry. This has catalysed an influential school of thought which prescribes a relatively technocratic approach to regulating risk, characterised by a high degree of deference to formal risk and cost-benefit analysis in the policy-making process. [e.g. 4, 10-13] These methodologies are intended to serve as "institutional safeguards" [8] that screen out the malign influence of heuristic-based judgments, and thus help ensure rational risk regulation (*i.e.* utility maximising). In short, the argument is that *"even if many heuristics mostly work well in daily life, a sensible government can do much better than to rely on*

them." (Sunstein, 2002 [4]) This school of thought – and its close cousin the Nudge agenda [14] – has intimately shaped policy debates and practices throughout the West.

Nevertheless, the basic argument has been challenged on a variety of fronts. It has been: critiqued as paternalistic or even undemocratic, in that it fails to respect citizen preferences; [15] portrayed as placing too much emphasis on the role of heuristics and biases in shaping risk perceptions to the neglect of factors such as culture; [16] and accused of downplaying the value judgements embedded within technical analysis. Yet the overarching premise – that reliance on risk and cost-benefit analysis is a viable means for screening out heuristic influences from risk regulation – remains broadly unchallenged. To be sure, several scholars [e.g. 17] have emphasised that risk assessors are unlikely to be free of cognitive quirks, yet this *tu quoque* argument rather misses its target. After all, Sunstein and his fellow travellers are not suggesting that policy-making should rely on expert *perceptions* of risk, but on expert *analyses*. For a *tu quoque* argument to have any force, it must be directed at the methodology and practice of formal risk assessment, not at how experts perform in psychometric tasks.

Through two studies of chemicals regulation, this paper takes up that challenge. It shows that the technical assessment of risk is structured and pervaded by its own set of heuristics, which are used to categorise potential threats, define what constitutes valid data, guide causal inference, and select and apply formal models. Some of these heuristics lay claim to theoretical or empirical justification, others are more back-of-the-envelope style calculations, whilst still more purport not to reflect some truth but simply to constrain discretion or perform a sort of desk-clearing function. The *tu quoque* critique turns out to have some mileage, although in a different direction than initially conceived. Of course, a

collective reliance on heuristics does not imply that lay and expert analyses of risk are on the same footing, at least if we are to take the idea of expertise at all seriously. Indeed, the central theme of the paper is that heuristics are *not* problematic *per se* – induction necessarily depends on procedures of inference that cannot be fully justified in a formal sense [18] – but rather that they have certain structural features that carry significant implications for how we think about and design risk assessment processes.

We emphasize that the heuristics we identify are rather different in character from the lay rules of thumb à la Danny Kahneman, [e.g. 2] which are intuitive, fuzzy (*i.e.* based on linguistic rather than mathematical terms), and domain-general. Our rules have more in common with another school of thought on heuristics, which traces its (modern) origins to the work of Polya. [19] Polya saw heuristics as provisional methods and principles of discovery that fell short of demonstration, yet that were indispensible for domains that did not admit of formal logic or proofs. His notion of heuristics heavily influenced the (then) emerging fields of artificial intelligence and expert systems, including pioneers such as Simon, [20] Pearl [21] and Lenat. [22] In these fields, heuristics are conceived and modelled as one of the foundational elements of expert knowledge. They are generally represented as sets of if-then rules that guide and structure inferences, decisions, and problem-solving in systematic ways. Their functions range from interpreting data, to predicting system behaviour, and to system control. [23] These sorts of rules are typically domain specific rather than general purpose, often represented in formal language rather than linguistic variables, and work by leveraging empirical facts or causal relations that are more or less accepted or understood. Although they do not guarantee correct solutions or optimal outcomes, they offer useful guidance through drawing on incomplete or imperfect knowledge and making problems

tractable (*e.g.* through ignoring information or options). These basic properties – formalism, domain-specificity, a focus on induction rather than proof, and explicit justification – suggest that this tradition is a useful framework to bring to the analysis of technical or expert practices such as formal risk assessment. And so we draw on some of its key insights.

Our paper's main contribution is to show that by studying the heuristics of risk assessment *qua* heuristics – that is, by thinking about the particular justifications, functions, and problems of rule-based reasoning – we develop important insights on the form, strengths, and limits of the prior art. Put another way, it offers a novel *representation* of the methods, conventions, and biases of formal risk assessment, and one that offers practical insights on how to improve this crucial process. The body of the paper is split into two parts. Study I offers a detailed but largely theoretical analysis of the heuristics that structure the United States Environmental Protection Agency's (EPA) risk assessment methodology for industrial chemicals. Study II offers a brief empirical proof of concept that such heuristics are applied and interpreted in significant and interesting ways across various policy domains, drawing on judicial reviews of agency decisions.

2. STUDY I: METHODOLOGY

2.1 Case Selection and Rationale

This section of the paper examines the risk assessment of industrial chemicals, focussing on the practices of the EPA. The main logic of this choice was that the four-stage chemical risk assessment paradigm (Figure 1) – or its close cousins – underpins a range of policy-domains (*e.g.* pharmaceuticals, food additives, hazardous waste site evaluations, microbiological risk assessment, *etc.*), and so the findings would lend themselves to generalisation. The second

logic was that the EPA's practices are relatively transparent and well documented. Thirdly, the EPA has pioneered risk assessment across a range of environmental and public health domains, and so can be seen as something of a "hard case" from which to build the thesis. Broadly speaking, the aim of the research was to uncover and critically evaluate the heuristics embedded within the EPA's approach to chemical risk assessment. The narrow goal was to question the claim that a sensible government can do better than rely on rules of thumb, although the broader ambition was to develop insights into the practice of risk assessment by studying those rules *qua* rules.

2.2 Methods and Definitions

Data was collected from various sources describing, evaluating, and providing context or background on the EPA's practices, including: a) risk assessment statutes, guidelines, procedures and outputs; b) critiques and evaluations of those practices from within the scientific and policy communities; and c) primary research papers and reviews providing detail and background on the relevant theories, methods, and assumptions adopted by the agency. The data was then inspected to identify the heuristics embedded within the risk assessment process. The heuristics were next grouped into categories according to the functions they performed (*e.g.* causal inference, weighting lines of evidence, *etc.*). Heuristics are defined here as rules of thumb for inference and choice. There are three key elements to this definition: heuristics are rule-like, in the sense that there is a presumption in favor of following them (*i.e.* they structure or constrain judgment); they are rules of inference or choice (*i.e. if-then*), not simply assumptions; and finally, they are simple, frugal approaches to problem solving which ignore relevant information rather than seeking to optimize. Although heuristics are treated rather differently across and even within different

disciplines – in terms of their normative status, where they reside, and whether they are implicit or explicit – this definition seems to capture a shared understanding of what they are. As we are concerned with the practice of technical risk assessment it should be clear that we do not – in contrast to the mainstream risk literature – see heuristics solely as intuitive tools residing primarily at the subconscious level. Instead, our focus shall be on those heuristics that are socially and scientifically constructed, reside within institutions, and are applied deliberatively.

3. STUDY I: RESULTS AND DISCUSSION

This section places the heuristics identified within the context of a functional typology, and includes discussion of their basic form, origins, and characteristics.

3.1 Screening and Categorisation Heuristics

3.1.1. Context and Heuristics

A basic problem in chemicals regulation is that not all substances of potential regulatory interest can be subject to comprehensive risk assessment, and so some approach has to be devised to search and prioritise the problem space. And so the EPA adopts a tiered approach to evaluation, where the type and degree of scrutiny is matched to the nature and extent of the risk that each chemical poses. For new chemicals, this process involves categorising substances according to shared properties and characteristics, and matching these groups to particular test and assessment requirements. In doing so, the agency draws on a range of evidence, such as physico-chemical properties, (expected) production volumes, existing toxicity or environmental fate and behaviour studies, *etc.* The logic is that these provide cues as to the degree of risk posed by a particular chemical, and to their

expected environmental and biological properties. Historically, the EPA conducted this threat categorisation through what we might be called a *factors-based approach*. Here, the factors to be considered in making a judgment are pre-defined, perhaps along with the relevant sources of evidence, although their relative weights or summing rules are constructed on an *ad hoc* or implicit basis. However, more recently the agency has moved to formalise elements of the process into a set of heuristics. Perhaps most significant are the set of classification rules (essentially decision trees) found in the agency's "chemical categories program." [24] Here, if a chemical is deemed to be a member of a broader class with shared properties, it is matched to generic hazard concerns, "boundaries" which set out any constraints surrounding those concerns, and finally to particular test strategies. One of these decision trees is detailed below, followed by two similar categorisation heuristics found in other agency programs:

- If chemicals pass threshold measures of persistence (*e.g.* greater than 6 months transformation half-life) or bioaccumulation (*e.g.* $\geq$ 5000 BAF* or BCF) they are matched with specific agency actions (*e.g.* "ban pending testing") and test strategies (*e.g.* biodegradability tests).
- Polymers with high molecular weights (> 1,000 daltons), as well as strong mineral acids and bases, are categorised as unlikely to be hormonally active and thus excluded from test requirements under the endocrine disruptor screening program. [25]

* BAF is the ratio of a chemical's concentration in an organism to that in the ambient environment at a steady state, whilst BCF is the ratio of a chemical's concentration in fish to that in water.

- If a new chemical meets one of several criteria that suggest significant or substantial exposure (*e.g.* expected production volume of 100,000 kg/yr), then the agency has the statutory authority to require certain toxicity tests (*e.g.* the Ames test). [26]

3.1.2. Analysis

These heuristics can be seen as an attempt to codify a body of knowledge and to leverage it in a way that reduces the problem space to a more tractable size, and avoids repeat enquiries into similar facts. The rules draw on two different sorts of justifications. The first is to use the identifying properties of a class to categorise chemicals into groups (*e.g.* using the existence of particular functional groups to classify substances). The second sort involves causal knowledge, for example relating to how particular properties of chemicals (*e.g.* functional groups) shape environmental and biological behaviour. Often rules draw on both types of justification. For example, the second rule uses the threshold of > 1,000 daltons as the identifying property to class a substance as heavy. And it is also based on causal assumptions, namely that heavy polymers are unlikely to reach molecular receptors or other target tissues, and that highly reactive chemicals (strong mineral acids and bases) will destroy tissue at the point of entry leading to toxic effects other than through the endocrine system. [25] On the other hand, these heuristics also emerged under pressure from the regulated industry and the courts, who had called for greater predictability, transparency, and consistency in the prioritisation process. [26] From this perspective, the rules serve to *discipline* the process and make it explicit. Heuristics, then, can have multiple origins or types of justifications: ways of justifying specific rules, and also ways of justifying rules *qua* rules.

However, these rules – by virtue of being generalisations – may lead to misclassification in certain instances. Consider the first decision tree. Here, the criterion for classifying chemicals as "very bioaccumulative" or not ($\geq$ 5000 BAF or BCF) is based on the logic that BCF and BAF are proxies for the potential for a chemical to accumulate within organisms and concentrate up the food-chain. Although this generalisation has theoretical and empirical verification, it stems from work on lipophilic, non-ionic, organic substances that undergo minimal metabolism. [27] For substances that do not share those properties, BCF or BAF may not be reliable indicators for bioaccumulation (*e.g.* chemicals where sorption does not depend on hydrophobic interaction). These, then, are the *scope conditions* under which the heuristic is valid, and by explicitly stating them – in apparent contrast to other regimes [28] – the EPA guards against misclassification. This cannot, however, skirt the arbitrariness inherent in using thresholds to classify chemicals according to interval (rather than categorical) variables, a problem compounded by the often substantial uncertainty surrounding measurements of these properties. [28]

3.2. Gatekeeping Heuristics

3.2.1. Context and Heuristics

A key challenge of risk assessment is to winnow down the mass of data that might serve as inputs to the process, whether relating to a chemical's toxicological mode of action, the form of its dose-response curve, or its environmental fate and behaviour. Two evaluation tests are generally applied: is the data relevant, and is it of sufficient quality/validity; our focus is on the latter. One approach to this evaluation is to establish rules for excluding studies deemed to be wholly inadequate in protocol, conduct, or reporting (*i.e.*, gatekeeping rules). For example:

- EPA risk assessors are encouraged to exclude toxicity studies whose test protocols use either excessively or insufficiently high dose selections, according to specified criteria (*e.g.* if neither toxicity nor weight gain is observed in the test population, then the maximum dose is generally considered to have been insufficiently high). [29]

However, such a rule-like approach to determining technical validity is very much the exception at the EPA. This is likely due to the rise of weight-of-evidence approaches in the agency's approach to risk assessment, and the associated principle that even imperfect studies might serve as inputs, albeit with appropriately low weights. This brings with it a focus on weighting the quality or strength of different lines of evidence, rather than a reliance on categorical tests for accepting or rejecting data. However, one long-established and highly significant gatekeeping rule resists this *Zeitgeist*:

- Toxicity studies funded or conducted by industry are deemed invalid for regulatory purposes if they do not adhere to Good Laboratory Practices (GLP) and EPA approved test methods (the "GLP heuristic"). [30]

This heuristic plays a rather more complex function than simply discriminating between valid and invalid research, as we see below.

3.2.2. Analysis

The GLP heuristic acts to exclude, for the purposes of risk assessment, industry toxicity studies that do not conform to the mix of general quality requirements and technical prescriptions contained in GLP and the relevant test methods (*e.g.* covering the care of laboratory animals, instrument calibration, test protocols, and the collection and storage of raw data). Crucially, the principle is not that studies which do not conform to these

requirements are *necessarily* invalid, as can be seen from the fact that it applies only to industry rather than academic research. Indeed the concept of GLP emerged in response to sloppy research practices, research misconduct, and even outright fraud on the part of private research companies operating free from the checks and balances of academic peer review. [31, 32] And so the GLP heuristic was intended as a way of controlling any incentives that industry might have to downplay the risks associated with their products, and to serve as a surrogate for peer review to ensure a basic level of methodological quality. Yet the rule carries with it some unintended consequences. For example, given that the approved test methods are subject to detailed validation and interlaboratory reliability studies, they may come to represent relatively insensitive and outdated methodologies when compared to cutting edge or novel research techniques or protocols. [33] In such instances, the heuristic can serve to ensure that industry studies don't follow state-of-the-art methods, and so might impede the resolution of questions that lie at the centre of risk assessment controversies. [32]

Citing concerns about the reproducibility and validity of research conducted via non-standardized methodologies, some toxicologists (including industry scientists) have implied that academic research should be subject to the same GLP heuristic. [33] The explicit claim is that only GLP compliant work is sufficiently robust and rigorous to serve as a basis for quantitative risk assessment. [e.g. 34] This has been fiercely resisted by some scientists – on the grounds of the high costs of compliance, the fact that academic peer review acts as an analogous (or superior) form of quality control, and the concern that it would exclude novel experimental designs from consideration in risk assessment. [32] The defence, in short, is that such a scope-extension would have little basis in the original justification of the heuristic,

and that in some instances, it may represent a strategic effort on the part of industry to exclude research that may threaten their commercial interests (*i.e.* an attempt to game or bend the rule). Whatever the merits of this dispute, these efforts at scope-extension appear to have found isolated success in other policy domains, with the US Food and Drug Administration's recent evaluation of BPA excluding academic toxicology studies that were not GLP-compliant. [32]

3.3. Causal Inference

Causal inference is broadly concerned with determining whether an association should be treated as causal (*i.e.* "true"), rather than as non-existent or spurious. It can be classed into two categories depending on whether it focuses on the interpretation of a discrete signal, or on the weighting and aggregation of multiple lines of evidence.

3.3.1. Signal vs. Noise

3.3.1.1. Context and Heuristics

The need to discriminate between a signal from a target and noise from a distracter underlies various elements of chemical risk assessment. For example, is an apparent increase in tumour incidence in laboratory animals likely due to exposure to the test chemical, or does it represent random variation; and does the response of an analytical instrument (*i.e.* an apparent "detection") indicate the presence of a substance or rather stem from interference? A standard approach to resolving such questions is to focus on the statistical properties of the signal. One classic version of this is significance testing, which often takes the form of the following heuristic:

- A finding or association is treated as “true” (or chance is rejected as a plausible explanation) if it meets a predefined level of statistical significance (typically, a p-value of 0.05).

At the EPA, variants of this rule are used to:

- Determine whether chance, rather than a treatment-related effect, is a *plausible* explanation for an apparent increase in tumour formation (during hazard identification); [29]
- Inform the derivation of the no-observed-adverse-effect-level (NOAEL) in dose-response analysis (a statistically significant response is sufficient, but not necessary, to reject an exposure level as representing the NOAEL); [35] and
- To discriminate between “true” detections of a chemical substance (analyte) and detections that cannot be reliably distinguished from instrument error or noise (*i.e.* the limit of detection). [36]

3.3.1.2. Analysis

The statistical significance heuristic is a long-standing and controversial approach to causal inference. General critiques of its application within risk assessment have centred on: [37, 38]

1) The fact that it focuses entirely on statistical considerations to the neglect of other relevant factors;
2) The arbitrariness of (typically) selecting a p-value of 0.05 as the threshold that demarcates “true” from “false” associations (*i.e.* that it is a mere convention); and, relatedly

3) A perception that a p-value of 0.05 – whilst perhaps justified in "basic" science (given the preference for false negatives over false positives) – places too high an evidentiary burden on "proving" an effect in a regulatory context (the "conservatism" charge).

In other words, it is not just the content or form of the heuristic that has been seen as problematic; instead, it has been criticised *as a rule*. That is, it has been critiqued as neglecting information that often turns out to be crucial, portrayed as being ritualistically and uncritically adhered to, and attacked as having been adopted in a context where its original justification no longer holds true. These are all general critiques of rules *qua* rules. However, the third element of this critique – the charge of "conservatism" – has limited traction where the EPA's practices are concerned. This is because, for the three variants of the rule listed above, a failure to meet the threshold of statistical significance does *not* mean that the finding or association is *treated* as chance or as noise (which would err on the side of under-estimating risk). For example, in exposure assessment, where observed values (concentration levels) are deemed to be too low to be reliably distinguished from instrument error or noise, they are not treated as zero. Instead, procedures are applied to substitute for this "censored" data (*e.g*. proxy measures are used), to guard against under-estimation. [36] This is broadly analogous to what Pearl [21] called "recovery schemes" – safeguards built into a heuristic's application designed to ensure that it doesn't function in perverse or undesirable ways. This suggests a fairly subtle appreciation of the biases that are intrinsic to rule-based reasoning. Although, as we shall see, this appreciation is sometimes lacking.

3.3.2. Weight of Evidence Heuristics

3.3.2.1. Context and Heuristics

Weight of evidence (WoE) approaches are increasingly prominent at the EPA, following the logic that there are often multiple lines of evidence that bear on a particular causal inference (*e.g.* is a chemical a human carcinogen), which need to be weighted and aggregated prior to making a final determination. This process may in principle be guided by some formal algorithm or set of rules (as proposed in [39]), though in practice typically takes the form of factors-based judgments. A classic example is Bradford-Hill's list of criteria for determining whether a chemical causes disease in a given population, variants of which are set out across EPA guidelines. Here, the factors to be considered in interpreting evidence are pre-defined (*e.g.* the strength of any epidemiological association, its biological plausibility, *etc.*), yet generally without rules for weighting or summing them. [e.g. 29] As an exception to this rulelessness, the agency has set out heuristics that classify the relative strength of different sources of animal and epidemiological evidence:

- In evaluating whether a chemical is potentially mutagenic, the agency places greater weight on tests conducted in germ cells than somatic cells, on in viva rather than in vitro tests, on tests in eukaryotes rather than prokaryotes, and on mammalian rather than sub-mammalian species. [40]
- A similar set of rules exists to guide the evaluation of potential endocrine disruptors (*e.g.* in vitro results from assay systems with metabolic capacity outweigh results from those without, *etc.*). [41, 42]
- In cancer risk assessment, human data is given greater weight than animal data in hazard characterization and dose-response assessment (provided it is of high quality and

adequate power), and amongst epidemiological studies, greater weight is given to those having more specific and precise exposure estimates. [29]

3.3.2.2. Analysis

Most of these hierarchies[†] are based on principle that particular sources of evidence (*e.g.* specific toxicity test systems) generate data of broadly similar accuracy, reliability or relevance. For example, mammalian test systems are given greater weight than bacterial ones because of major differences in their cells (*e.g.* membrane structures, DNA repair capabilities, *etc.*), with the former of course being closer analogues to humans. However, one blindspot of this approach is that it neglects the question of how well designed, conducted, and reported the *individual* studies were. A second blindspot is that such hierarchies don't accommodate the idea that the specific hypothesis under question might alter the strength to be afforded to a particular line of evidence. For example, a *particular* study on a *particular* test system may rank poorly on existing hierarchies (because the test system, in general, has modest sensitivity and specificity), yet nevertheless be best suited to resolve a question regarding a chemical's mode of toxicological action. The point – by now perhaps laboured – is that *rules are generalisations* and thus abstract away from or ignore details that may in some contexts prove critical. It is an open question whether the *ceteris paribus* qualifications that the agency attaches to these heuristics – *e.g.* "assuming appropriate dose and route of exposure" [42] and "provided it is of high quality and adequate power" [29] – sufficiently clarify this point. This is not an issue restricted to evidence hierarchies. It is the broader problem of clarifying the nature and extent of evidence required to depart from heuristics in particular instances. We return to this later.

[†] The latter hierarchy is based on the statistical properties of the study *outputs*, the idea being that studies with high precision (or low variance) are more informative (*e.g.* typically having larger sample sizes, *etc.*).

3.4. Modelling

The heuristics discussed so far have dealt with tasks of classification; we now turn to heuristics involved in the selection and application of dose-response models, looking at both carcinogens and non-carcinogens (Figure 2), and focussing on the use of animal rather than epidemiological data. The EPA adopts different modelling approaches for interpolation and extrapolation, so we address each separately before providing a joint analysis.

3.4.1. Interpolation (analysis in the observed range)

3.4.1.1. Context and Heuristics

The determination of a (potential) causal link between a chemical and harm is the precursor to building formal models of the relationship between dose and response. Interpolation involves structuring and regimenting the raw test data into a dose-response curve (typically covering moderate-high dose levels), with the purpose of deriving a "point of departure" (POD) from which extrapolation to the low-dose range can then be made. [35, 43, 44] This is a complex task with numerous possible approaches. For example, which species and endpoint should be used as the basis for the dose-response relationship? Should a biological or statistical model be used for interpolation? And which parameter estimation technique should be used to apply it (*e.g.* maximum likelihood vs. non-linear least squares)? Or should the dose-response data simply be plotted by hand, and formal modelling eschewed (*e.g.* where the NOAEL or LOAEL are taken as the POD)? These are the classic dilemmas of modelling-for-policy, where model selection is typically underdetermined by the raw data, underlying theories may be contested, and any choice involves a mixture of science and policy judgment. Rather than leave these decisions entirely up to the discretion of risk

assessment teams, the EPA is famous for having adopted presumptive rules to structure the process. [43-45] One such rule, generally perceived to be precautionary, is that:

- The "critical effect"[‡] is the preferred basis for constructing the dose-response model, in situations where there are several datasets to choose from (*e.g.* datasets relating to different endpoints, such as carcinogenic vs. neurotoxic, and perhaps of different species). [35]

Moreover, there are simple rules applied to adjust dose-response data prior to model construction:

- The cross-species scaling heuristic for oral exposure,[§] wherein animal doses are scaled by ¾ power of body-weight to derive the toxicologically equivalent human dose (to account for differences in size and lifespan across species); [46] and
- A special-form of Haber's Rule – where identical products of dose concentration and exposure time are presumed to lead to equivalent biological responses ($C \times t = k$) – is used, where necessary, to adjust from intermittent laboratory exposures to "real world" continuous exposure conditions. [47, 48]

The above two are rules setting out how to adjust the raw dose-response data, and are empirically rooted in the sense that they purport to represent relationships that exist in the real-world, rather than being justified on policy grounds. They embody truth claims, in other words, rather than being *mere* conventions or value judgments, and these claims can be

[‡] "The first adverse effect, or its known precursor, that occurs to the most sensitive species as the dose rate of an agent increases."

[§] This applies to carcinogenic risk assessment (Figure 2). For non-carcinogens, cross-species extrapolation is achieved through applying an adjustment factor (discussed later), although the EPA has recently moved to harmonise the two. Pharmacokinetic modelling is an alternative approach to cross-species scaling, although rarely possible in practice.

held to account. Accordingly, their empirical bases have been subject to critical scrutiny, [46, 48] and indeed one long-standing critique of Haber's Rule is that it has at times been interpreted by toxicologists as a causal *law,* rather than as a statistical approximation subject to certain scope conditions. [49]

We now turn to model selection and implementation, which is largely a curve-fitting exercise, in the sense that there is typically little mechanistic basis to the models being considered. [50] For model implementation – or parameter estimation – the EPA relies on factors-based judgments rather than explicit rules. For example, the factors that may influence the selection of parameter estimation techniques are discussed in the more technically-oriented guidelines, [e.g. 50] although no default method is established, nor is a rule or algorithm established for choosing between them. For model selection, the EPA tends not to rely on default procedures *per se*.** That is to say that there is typically no default model recommended, nor an explicit hierarchy of preferred models, in part because different study designs and datasets shape dose-response patterns. [50] But the agency does use heuristics to guide the *process* of model selection. [50]

- In benchmark dose-response modelling, $\alpha = 0.1$ is recommended as measure of adequate fit, along with a less formal "eyeballing" test.
- Should multiple models pass the above tests, the minimum score on a measure that balances complexity with fit – Akaike's Information Criterion (AIC) – is recommended to select the best model.

3.4.2. *Extrapolation (analysis outside the observed range)*

** An exception is that the multi stage model is the preferred approach to analysing cancer bioassay data, which is a relatively flexible modelling framework.

3.4.2.1. Context and Heuristics

Extrapolation involves the construction of new data points, and, in our context, refers to analysis in the low-dose region. It is the most significant stage of the modelling process, as the low-dose region is relevant to real-world human exposure levels and is thus the basis for regulatory decisions. Perhaps the most controversial heuristics are those guiding the selection of model *structure* at low doses: [44]

- If a chemical is carcinogenic, then a linear non-threshold dose-response model is adopted. The exception to this rule is cases where a) the chemical's mode-of-action is established and b) that mode of action is thought to be non-linear at low doses (*e.g.* cytotoxicity);
- If a chemical is non-carcinogenic, then a threshold dose-response model is applied (*i.e.* below a certain exposure, no response is expected).

The non-threshold linearity assumption for carcinogens originated in radiation biology, and – although there are of course significant differences in how the two types of exposure are absorbed by and interact with organisms – was later mapped across to chemicals regulation. [51] This was on the grounds that it was both theoretically plausible – the idea being that even a single "molecular event" can initiate malignant or neoplastic cell transformation, and so might lead to cancer – and likely to be health protective. [51] By contrast, the presumption of a threshold for non-carcinogens follows the idea that below a certain dose, clearance pathways, cellular defences, and repair processes can minimize damage and thus render exposure practically negligible. [44] However, these rules have been critiqued for retaining a certain ambiguity about when they can be overturned, which we return to later.

The above, then, are rules for selecting the fundamental assumptions or basic structures of the models to be adopted. Let us now turn to rules for applying those model structures to the data. In linear extrapolation, the goal of the modelling exercise is to derive the slope of the dose-response curve, as this (combined with exposure) is the basis for the subsequent risk estimate. The relevant heuristic is as follows:

- The slope is derived by simply drawing a straight line from the point of departure (POD) to the intersect of the two axes, correcting for background.[††]

By contrast, for threshold models, the goal is to derive a "safe" level of human exposure (*i.e.* the threshold for humans, known as the reference dose or reference concentration). This is done through applying a series of uncertainty and adjustment factors (conventionally a value of 10 is used for each, where applicable) to the POD to account for: [(52)]

- Inter-species variability;
- Variations in susceptibility within the human population;
- The difference in sensitivity between chronic and sub-chronic toxicity studies (in the case where sub-chronic studies are the basis for deriving a threshold for a chronic effect);
- The difference between LOAEL and NOAEL (if LOAEL is used as the POD); and
- Gaps or uncertainties in the toxicity database.

These "safety factors" are essentially back-of-the-envelope style heuristics designed to accommodate the various forms of uncertainty – both in terms of variability and ignorance – and adjustments required to extrapolate from the findings of a specific laboratory protocol

[††] Model-based approaches to low dose extrapolation may also be performed (*e.g.* in biologically based dose response modelling), but theoretical and empirical limitations generally preclude this.

to derive results meaningful for human populations. A slightly more refined approach is the use of chemical-specific adjustment factors, where, for example, empirical data on interspecies differences in the uptake or metabolism of a chemical are used to depart from the conventional value of 10. [53]

3.4.3. *Analysis*

There exist a rich range of modelling heuristics, from those selecting what is to be modelled in the first instance (*e.g.* which endpoint, which species), to those adjusting or laundering the raw data to make it fit-for-purpose (*e.g.* adjusting measures of exposure), and even those selecting the fundamental assumptions (*e.g.* threshold vs. non-threshold) and specific functional form of the model to be applied. Of course rules being rules, they inevitably will have their exceptions, yet one major issue is the lack of clarity about what sort of evidence, and how much, is required to depart from them. This has led to several instances when the EPA has been reluctant to depart from them their heuristics even in the face of substantial contradictory empirical evidence. For example, the agency retained the default linear approach to low dose extrapolation in its risk assessment of dioxin, despite clear evidence suggesting that the substance promoted cancer via receptor-mediated pathways that are by nature non-linear. [53] This not only led to a skewed estimate of risk, but also understated uncertainty, as the assessment failed to explore how departing from this particular default would have influenced the analysis outcomes. The very same problem – a lack of clarity on what constitutes "sufficient showing" for departures – is found with the default safety factors. [54] And again, this ambiguity has led EPA risk assessors to rigidly adhere to the default values in cases, such as with dioxin and formaldehyde, where a wealth of data seems to justify departing from them and using chemical-specific adjustment factors. [54]

A second common thread is the tension between adopting heuristics that have a reasonable theoretical and empirical basis on the one hand, and the desire to have rules that are health protective on the other. This is a conflict lying at the heart of many debates over method and policy in risk regulation – when does it make sense to attempt to regulate on the basis of "best" estimates of risk (*e.g.* mean, or median) vs. more precautionary approaches? Thus, critical analyses of these modelling rules have focused not simply on their accuracy or generality, [*e.g. 46, 48*] but more broadly on their perceived tendency to exaggerate risk – particularly when compounded. [*e.g. 55*] This has led the EPA to articulate (in different levels of detail) the circumstances or evidence that might justify departures from some rules generally perceived to be precautionary (*e.g.* specifying the factors that may justify using the species that most *resembles* humans as the basis for dose-response modelling, rather than the "critical effect"), and to establish one other reform that holds particular conceptual interest: [*52*]

- A safeguard against the excessive compounding of uncertainty and adjustment factors has been established – specifically, they must not combine beyond a value of 10,000 in deriving the safe level of human exposure.[¶]

The above has some parallels with Pearl's notion of recovery schemes, which traditionally serves as an "escape valve" from the rigid application of rules when they are perceived to lead in unhelpful or perverse directions. However, in the above case, the logic is that a compounding of factors to 10,000 and beyond – whilst sometimes required by a strict interpretation of the rules – would lead to significant exaggerations of the "true" risk posed by the chemical. And so the distinction is that departures from the rules are justified out of

[¶] The safeguard of 10,000 applies to oral doses; a safeguard of 3,000 is applied to inhaled doses.

concern for the *interaction* of rules, rather than a concern with rules operating in isolation. This is distinct from the usual scope restrictions and "recovery schemes" that come with heuristics, and suggests the importance of analysing heuristics as systems or *chains of rules*, rather than merely as independent objects. (c.f. *56*)

Also worth reflecting on is the distinction between the EPA's mix of factors-based guidance and heuristics for selecting and applying models for interpolation of the dose-response data, compared to the rather more prescriptive and rule-bound approach to extrapolation. In extrapolation, heuristics set out the default model structures to be adopted, and set out in detail the particular rules to be followed in implementing those models. In contrast, in interpolation, model implementation is factors-based rather than rule-bound. And model selection is a mixture of clear rules and informal tests (*e.g.* "eyeballing"). And the rules focus on the processes by which models should be selected (*e.g.* measures of goodness of fit and AIC), rather than prescribing particular outcomes (*i.e.* default models). At first glance this contrast appears surprising or perhaps even incoherent, as the low-dose region is the very area in which theories are most contentious and data largely lacking. Yet an explanation comes to light when we consider that alternative plausible approaches to low dose extrapolation lead to risk estimates that may differ by several orders of magnitude. (57) In other words, the analysis outcomes are particularly sensitive to the choices made during extrapolation, rather than interpolation. The assumption and model-laden nature of extrapolation, and the extent to which it determines outcomes, has long been recognised. And given the general lack of scientific agreement over the most appropriate assumptions, it historically fed concerns about both inconsistency in risk assessment, and that outcomes may even be tailored to meet policy objectives. (54) And so the introduction of ruleness in

this case appears to represent not so much an attempt to codify or formalise existing (albeit provisional) bodies of knowledge, nor simply a desire to err on the side of precaution, but rather an attempt to narrow or constrain the inferences open to risk assessors, and thus enforce a degree of consistency, transparency, and predictability in both process and outcomes. Critics argue that this pursuit of consistency has brought with it blindspots, such as the neglect or marginalisation of low-dose phenomena that do not cohere with the extrapolation rules (*e.g.* hormetic responses characterised by low-dose stimulation and high-dose inhibition, [58] and postulated mechanisms of endocrine disruption where conventional dose-response assumptions are held to be invalid [59]). The claim here is not just that rules have exceptions, but that they can become *entrenched* in ways that make it difficult to displace or circumvent them in practice. For example, *as a consequence of the very existence* of the low-dose extrapolation rules – and the way that they shape the underlying paradigms of toxicity testing – there is rarely case-specific evidence available to question or indeed support their validity (*i.e.* empirical data within the low-dose region is simply not part of conventional test requirements). On this reading, rules can shape practices in the real world in ways that make it less likely for them to be challenged.

4. STUDY II: PROBLEMS WITH RULES IN PRACTICE

The foregoing analysis has is in a sense been largely theoretical, focussed more on the content, structural features, strengths and potential problems of the heuristics of risk assessment, rather than on how they are interpreted and applied in concrete settings. We now turn to make the case that some of those structural features and potential limitations play out in real-world settings in interesting and important ways, and that this is not an issue restricted to industrial chemicals regulation. We do so by analysing six cases of judicial

review of regulatory risk assessments in the United States, in policy domains ranging from air pollution, to water quality standards, to waste management. We identified these cases through entering various search strings intended to reflect or relate to our previously discussed expert heuristics (*e.g.* "risk assessment" AND "default assumptions;" "risk assessment" AND "safety factors"), and scrutinizing the results for relevance. The idea was not to recover a comprehensive listing of all such disputes, but rather to identify those cases which had turned in significant measure on the regulatory agency's application or interpretation of heuristics.

The particular disputes recovered focussed on such things as the application of default model structures, on the use and neglect of safety factors, and on the reasoning behind departing from hierarchies of evidence (Table 1). They reveal instances of the unthinking reliance on rules, such as when an emission standard for an air pollutant was based on a default model that assumed the substance to be a gas, when in fact it was known to be a solid at the relevant temperatures. They highlight examples of the dubious interpretation of rules, such as where the EPA relied on *ad hoc* justifications to retain the default linearity assumption in modelling the risk of a carcinogen, despite having earlier concluded that it followed a nonlinear mode of action (cytotoxic rather than genotoxic). They show instances of unreasoned departures from rules, such as where the EPA and OSHA failed to justify deviating from safety factors designed to account for sensitive populations, which again highlights the dangers of not clarifying what constitutes "sufficient showing" for departures. And they point towards the interpretive flexibility associated with linguistically defined rather than mathematical rules, such as in the dispute over whether a particular toxicology experiment was unethically conducted and so should not be relied on. Taken as a whole,

the cases show that heuristics play important roles in risk assessment across various policy domains, that they are sometimes applied or interpreted in problematic and contested ways, and that authoritative actors (*e.g.* courts) take these issues seriously.

5. CONCLUSIONS

Psychological research has cast doubt on the idea that laypeople use formal logic or statistical approaches to reason about risk issues, suggesting instead that they rely on simple and error-prone rules of thumb. Many scholars have drawn a sharp-distinction between this heuristic-reasoning and the rigour and reliability of formal risk assessment, and by extension called for relatively technocratic approaches to risk regulation as a way of ensuring that policies conform to rational ideals. [1, 4, 7-8, 10-13] Yet as we have seen here, formal risk assessment is structured and pervaded by its own set of heuristics. As a preliminary typology, we distinguished between classification rules that inform priority setting and test strategies; gatekeeping heuristics that filter problematic studies away from the risk assessment process; rules that structure and guide causal inference; and heuristics for model selection and application. Some of these rules lay claim to empirical or theoretical justifications, whilst others are more back-of-the-envelope style estimations, and still more serve policy goals beyond that of simply codifying knowledge. This is not to imply an equivalence between lay and expert risk assessment. Instead, we have tried to suggest that significant aspects of risk assessment can be *represented* as heuristics, and to use this insight to work towards a useful analytical framework for characterising the process.

Put another way, our basic argument is that by analysing the rules of risk assessment *qua* rules, we generate important insights about the nature, strengths, and limits of the prior art, and insights that offer practical ways of improving it. We suggest that the heuristics of risk

assessment are not just an unruly collection of tools. When viewed as an approach to governing risk assessment, heuristics serve to close down or constrain interpretive possibilities, thus conferring a modicum of stability, consistency, and transparency upon a process that might otherwise be (seen as) highly sensitive to the choices of individual analysts. They can be understood as a way of authenticating or formalising risk assessment as a scientific practice, [60, 61] representing as they do a series of rules for bounding problems, collecting data, and making sense out of it (*i.e.* a *methodology*). There are, of course, substantive critiques of heuristics *qua* rules, including that they generate systematic error, may create a veneer of consistency by masking uncertainties, can be manipulated for strategic purposes, and have a tendency to become entrenched. These critiques can be *partially* offset by ensuring that the scope conditions of rules are specified, that the nature and extent of evidence required to justify a departure from them are made clear, that programs exist to periodically re-evaluate the empirical and theoretical support of particular heuristics, that recovery schemes are in place to correct for when they go awry, and that "escape valves" are in place to guard against perverse outcomes stemming from the interaction of rules. The critical idea here is that heuristics are not problematic *per se*, but they can become so when we treat them as *laws* by neglecting their rough and contingent nature. This principle may sound almost self-evident, but as our empirical evidence shows, at times it is more honoured in the breach than the observance. However, regardless of whether these ideals are met in practice, rule-based reasoning has certain intrinsic characteristics, strengths, and weaknesses, and tensions remain between the virtues served by relying on rules compared with those furthered by alternative species of reasoning (*e.g.* factors-based judgments).

A concern with the rules of risk assessment raises questions for future research. One involves developing a set of criteria and methods for evaluating individual heuristics, and this seems in principle possible to do within a formal decision theoretic framework. [*e.g. 62*] Another relates to how heuristics function in the wild [c.f. *63*] – how are they selected, interpreted, applied or neglected in practice? We have taken tentative steps in this direction, but more systematic work could follow. And what rules might exist in the real-world of risk assessment, but not on the books? A complicating issue is that expert heuristics are often not directly accessible or known to the expert themselves, and require dedicated approaches to knowledge elicitation. [*64*] Another avenue is the question of whether and *why* there are significant differences in the degree of ruleness, and in the particular rules relied upon, across regimes and jurisdictions? And other stages of the regulatory process – such as economic analysis and the decision-making process itself – are ripe for exploration. Finally, there is the more general normative issue of when or in what contexts it makes sense to rely on heuristics, versus alternative species of reasoning (*e.g.* factors based judgments, complex algorithms, probabilistic approaches, unconstrained discretion, *etc.*)? [*63*] In the meantime, our present analysis has revealed the pervasive influence of heuristics in structuring formal risk assessment, and cast light on the significance of this for understanding and designing risk regulation regimes.

ACKNOWLEDGEMENTS

This research was partly funded by a Mellon Sawyer Fellowship (University of Cambridge). Thanks are due to the editor (Tony Cox) and reviewers for helpful comments. David Spiegelhalter, Jerry Busby, and Keith Richards were generous in providing guidance and critiques of previous drafts. The usual caveats remain.

REFERENCES

1. Sunstein CR. Adolescent risk-taking and social meaning: a commentary. Dev Rev, 2008; 28(1): 145-152.

2. Tversky A, Kahneman D. Judgement under uncertainty: Heuristics and biases. Science, 1974; 185(4157): 1124–1131.

3. Peters E, Slovic, P. The role of affect and worldviews as orienting dispositions in the perception and acceptance of nuclear power. J Appl Soc Psychol, 1996; 26(16): 1427-1453.

4. Sunstein CR. Laws of fear: the perception of risk. Harvard Law Rev, 2002; 115(4): 1119-1168.

5. Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: Models of bounded rationality. Psychol Rev, 1996; 103(4): 650-669.

6. Noll RG, Krier JE. Some implications of cognitive psychology for risk regulation. J Legal Stud, 1990; 19(2): 747-779.

7. Jolls C, Sunstein CR, Thaler R. A behavioral approach to law and economics. Stanford Law Rev, 1998; 50(5): 1471-1550.

8. Kuran T, Sunstein, CR. Availability cascades and risk regulation. Stanford Law Rev, 1999; 51(4): 683-768.

9. Eskridge WN, Ferejohn J. Structuring lawmaking to reduce cognitive bias: A critical view. Cornell Law Rev, 2002; 87(2): 616-647.

10. Breyer SG. Breaking the vicious circle: Toward effective risk regulation. Harvard University Press, 1993.

11. Graham JD, Wiener JB. Risk versus risk: Tradeoffs in protecting health and the environment. Harvard University Press, 1998.

12. Sunstein CR. Laws of fear: Beyond the precautionary principle. Cambridge University Press, 2005.

13. Baron J. Judgment misguided: Intuition and error in public decision making. New York: Oxford University Press, 1998.

14. Thaler RH, Sunstein CR. Nudge: Improving Decisions About Health, Wealth, and Happiness. New Haven, CT: Yale University Press, 2008.

15. Shrader-Frechette KS. Evaluating the expertise of experts. Risk: Health, Safety and Environment, 1995; 6: 1115-1126.

16. Kahan DM, Slovic P, Braman D, Gastil J. Fear of democracy: A cultural evaluation of Sunstein on risk. Harvard Law Rev, 2006; 119(4): 1071-1109.

17. Posner RA. Rational choice, behavioral economics, and the law. Stanford Law Rev, 1998; 50(5): 1551-1575.

18. Greenland S. Intuitions, simulations, theorems: The role and limits of methodology. Epidemiology, 2012; 23(3): 440-442.

19. Polya G. How to solve it. Princeton University Press, 2004.

20. Simon HA, Newell A. Heuristic problem solving: The next advance in operations research. Operations Research, 1958; 6(1): 1-10.

21. Pearl J. Heuristics: intelligent search strategies for computer problem solving. Addison-Wesley, MA, 1984.

22. Lenat DB. The nature of heuristics. Artificial Intelligence, 1982; 19: 189–249.

23. WJ Clancey. Heuristic classification. Artificial Intelligence, 1985; 27: 289-350.

24. EPA. TSCA new chemicals program (NCP): chemical categories; 2010: http://epa.gov/oppt/newchems/pubs/npcchemicalcategories.pdf

25. Federal Register. Vol. 67, No. 250, December 30, 2002: http://www.epa.gov/endo/pubs/12-02-frnotice.pdf

26. Elkins CL. Chemicals Manufacturers Association Letter. 1988: http://www.epa.gov/oppt/newchems/pubs/cmaxpltr.htm

27. McGeer JC et al. Inverse relationship between bioconcentration factor and exposure concentration for metals: implications for hazard assessment of metals in the aquatic environment. Environ Toxicol Chem, 2003; 22(5): 1017-1037.

28. Arnot JA, Mackay, D. Policies for chemical hazard and risk priority setting: can persistence, bioaccumulation, toxicity, and quantity information be combined? Environ Sci Technol, 2008; 42(13): 4648-4654.

29. EPA. Guidelines for carcinogen risk assessment, 2005: http://www.epa.gov/raf/publications/pdfs/CANCER_GUIDELINES_FINAL_3-25-05.PDF

30. EPA. EPA authorities under TSCA, 2005:

http://www.epa.gov/oppt/npptac/pubs/tscaauthorities71105.pdf

31. EPA. Memorandum: interpretation of the good laboratory practice (GLP) guideline, GLP regulations advisory no. 36, 1991:

http://www.epa.gov/compliance/resources/policies/monitoring/fifra/glp/advisory36.pdf

32. Myers JP et al. Why public health agencies cannot depend on good laboratory practices as a criterion for selecting data: the case of bisphenol A. Environ Health Persp, 2009; 117(3): 309-315.

33. Alcock RE, MacGillivray BH, Busby JS. Understanding the mismatch between the demands of risk assessment and practice of scientists – the case of Deca-BDE. Environ Int, 2011; 37(1): 226-235.

34. Tyl RW. Basic exploratory research versus guideline-compliant studies used for hazard evaluation and risk assessment: Bisphenol A as a case study. Environ Health Perspect, 2009; 117:1644–1651.

35. EPA. An examination of EPA risk assessment principles and practices, 2004:

http://www.epa.gov/osa/pdfs/ratf-final.pdf

36. EPA. Guidelines for exposure assessment, 1992:

http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=15263

37. Kriebel D et al. The precautionary principle in environmental science. Environ Health Persp, 2001; 109(9): 871-876.

38. Suter, GW. Abuse of hypothesis testing statistics in ecological risk assessment. Hum Ecol Risk Assess, 2001; 2(2): 331-347.

39. McCarty LS, Borgert CJ, Mihaich EM. Information quality in regulatory decision-making. Environ Health Persp, in press. http://dx.doi.org/10.1289/ehp.1104277

40. EPA. Guidelines for mutagenicity risk assessment, 1986: http://www.epa.gov/osa/mmoaframework/pdfs/MUTAGEN2.PDF

41. Doull J et al. Framework for use of toxicity screening tools in context-based decision-making. Food Chem Toxicol, 2007; 45(5): 759-796.

42. Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) Final Report, Chapter 5: screening and testing, 1998: http://www.epa.gov/endo/pubs/edstac/chap5v14.pdf

43. National Research Council. Science and Judgment in Risk Assessment. National Academies Press, Washington, DC, 1994.

44. National Research Council. Science and Decisions: Advancing Risk Assessment. National Academies Press, Washington, DC, 2009.

45. National Research Council. Risk Assessment in the Federal Government: Managing the Process. National Academies Press, Washington, DC, 1983.

46. Rhomberg LR, Lewandowski TA. Methods for identifying a default cross-species scaling factor, report for EPA, 2004: http://www.epa.gov/raf/publications/pdfs/RHOMBERGSPAPER.PDF

47. Chou C-HSJ, Holler J, De Rosa CT. Minimal risk levels (MRLs) for hazardous substances. J Clean Technol Environ Toxicol, 1998; 7(1): 1-24.

48. Belkebir E et al. Haber's rule duration adjustments should not be used systematically for risk assessment in public health decision-making. Toxicol Lett, 2011; 204(2-3): 148-155.

49. G Atherley. A Critical Review of Time-Weighted Average as an Index of Exposure and Dose, and of Its Key Elements. Am Ind Hyg Assoc J, 1985; 46(9): 481-487.

50. EPA. Benchmark dose technical guidance document, 2012: http://www.epa.gov/osa/raf/publications/benchmark_dose_guidance.pdf

51. Calabrese EJ. The road to linearity: why linearity at low doses became the basis for carcinogen risk assessment. Arch Toxicol, 2009; 83: 203-225.

52. EPA. A review of the reference dose and reference concentration processes, 2002: http://www.epa.gov/osa/pdfs/ratf-final.pdf

53. National Research Council. Health risks from dioxin and related compounds: evaluation of the EPA reassessment. National Academies Press, Washington, DC, 2006.

54. Institute of Medicine, Environmental decisions in the face of uncertainty. National Academies Press, Washington: DC, 2013.

55. Nichols AL, Zeckhuaser RJ. The perils of prudence: how conservative risk assessments distort regulation. Regul Toxicol Pharm, 1998; 8: 61-75.

56. WJ Clancey. The epistemology of a rule-based expert system – a framework for explanation, Artificial Intelligence, 1983; 20: 215-251.

57. Crump KS, Howe RB. A review of methods for calculating statistical confidence limits in low dose extrapolation, in Toxicological risk assessment: Vol 1. Biological and statistical criteria, eds. DB Clayson, D Krewski, I Munro, CRC Press, Boca Raton, FL, 1985.

58. Calabrese EJ, Baldwin LA. Toxicology rethinks its central belief. Nature, 2003; 421(6294): 691-692.

59. Welshons WV et al. Large effects from small exposures. I. Mechanisms for endocrine-disrupting chemicals with estrogenic activity. Environ Health Persp, 2003; 111(8), 994-1006.

60. Jasanoff S. The Fifth Branch: Science Advisors as Policy Makers. Cambridge, MA: Harvard University Press, 1994.

61. Wynne B. Carving out science (and politics) in the regulatory jungle. Soc Stud Sci, 1992; 22(4): 745-758.

62. Langlotz CP, Shortliffe EH, Fagan LM. Using decision theory to justify heuristics. Proceedings of AAAI, Philadelphia, 215–219, 1986.

63. Gigerenzer G, Todd PM, ABC Research Group. Simple Heuristics That Make Us Smart. Oxford, UK: Oxford University Press, 1999.

64. Feigenbaum E. Knowledge engineering: The applied side of artificial intelligence. Annals of the New York Academy of Sciences, 1984; 426: 91-107.

Case	Description	Comments
Chemical Manufacturers Association vs. EPA (1994), United States Court of Appeals, District of Columbia Circuit. 28 F.3d 1259	EPA's rule designating Methylene diphenyl diisocyanate a high risk air pollutant was challenged on various grounds. One was that the emission standard was based on the agency's default model, which assumed the substance would be emitted as a gas and disperse as such, when in fact it is solid at the (most) relevant temperatures, has a low vapour point, and typically disperses as an aerosol. The decision was overturned as the default, generic model "bore no rational relationship" to the air pollutant in this case.	The court acknowledged trade-off between accuracy and efficiency in risk assessment, but critiqued the EPA for rigid, unthinking adherence to default procedures for model selection (heuristics) when faced with overwhelming evidence that they were unsound in a particular instance.
Leather Industries of America, Inc. *et al.* vs. EPA (1994), United States Court of Appeals, District of Columbia Circuit. 40 F.3d 392	Various petitioners challenged EPA's decisions on sludge sewage disposal. The court critiqued the agency's reliance on default, conservative assumptions about the use and application of heat-dried sludge, when it had specific, empirical data on these variables that were of quite a different order. Also critiqued the agency's use of highly conservative exposure assumptions (high child exposure model), when in fact a high proportion of sewage sludge application is done at sites with limited potential for public (and child) contact. Key elements of the ruling were overturned by the court.	Critique once more focussed on blanket reliance on conservative modelling assumptions and practices when, crucially, there was empirical data available that suggested them to be problematic. A central concern of the court was the lack of justification or explanation given by the agency for relying on heuristics that seemed, on face value, to be unreasonable.
International Union *et al.* vs. OSHA (1994), United States Court of Appeals, District of Columbia Circuit. 878 F.2d 389	Several unions challenged OSHA's emission standards on formaldehyde in the workplace as insufficiently stringent. Central issue was whether the agency adequately explained its reasons for departing from the default assumption of linearity at low doses for the carcinogen (the linearity heuristic). The rule was remanded.	Court held that the agency had not given a properly reasoned basis for departing from the heuristic, and so overturned the rule. As a subsidiary issue, court found no issue with the agency departing from another default rule – when they favoured animal studies rather than epidemiological data in standard setting – on the grounds that this was a reasoned departure from the rule. For

		example, they cited difficulties in getting precise exposure data, and controlling for possible confounders, *etc.*
Chlorine Chemistry Council, *et al.* vs. EPA (2000), United States Court of Appeals for the District of Columbia Circuit 206 F.3d 1286	Industry groups challenged EPA for retaining a maximum contaminant level goal (MCLG) of zero for chloroform. The level had originally been set based on the assumption of linearity in the low dose range (the default heuristic for carcinogens), combined with statutory mandate to set MCLG at level where no known or anticipated adverse effects occur. The groups challenged EPA for failing to revise the MCLG despite having more recently concluded that the substance exhibited a nonlinear mode of carcinogenic action (cytotoxic rather than genotoxic mode of action). EPA deployed various ad hoc arguments for why they had not revised the standard (*e.g.* that it was waiting for deliberations with its advisory board before departing from a long held policy). Court sided with the petitioners.	EPA's decision was overturned on the grounds that it had failed to adhere to one of its own rules of risk assessment. That is, the agency's own guidelines state that when "adequate data on mode of action show that linearity is not the most reasonable working judgment and provide sufficient evidence to support a nonlinear mode of action," the default (heuristic) assumption of linearity should be departed from.
Northwest Coalition *et al.* vs. EPA (2008), United States Court of Appeals for the Ninth Circuit F.3d 1043, 1052	Environmental groups challenged EPA's tolerances for seven pesticides. In particular, they challenged EPA's deviation from default safety factors (the back of the envelope heuristics) designed to be protective of infants and children for acetamiprid, mepiquat, and pymetrozine. The governing statute states that the EPA is allowed to "use a different margin of safety for the pesticide chemical residue only if, on the basis of reliable data, such margin will be safe for infants and children." However, the court found that the EPA had not clearly argued or articulated the link between the underlying toxicological evidence, and the (non-default) safety factors that were ultimately adopted. The decision was remanded on this basis.	The dispute, as the court saw it, turned on whether the EPA's evidence constituted "reliable data" for inferring that the lower margins of safety would be sufficiently protective. The agency had explicitly stated its reasoning behind deviating from (and lowering) its default safety factors when it set the tolerances (*e.g.* claiming data showed no evidence of greater sensitivity to the young, nor of abnormalities in development of foetal nervous system). However, the court found this inadequate, holding that the *particular* deviations – e.g. from a safety factor of 10 to one of 3 for pymetrozine – seemed chosen arbitrarily. This perhaps seems a rather high standard of evidence to

		demand – safety factors are by their nature rather arbitrarily chosen. However, what is interesting is that the court emphasised that what constitutes "reliable data," in other words, what constitutes reasonable grounds for departing from the default rules, is not explicitly defined in the relevant statute. This is the common problem of neglecting to define the nature and extent of evidence that constitutes "sufficient showing" to depart from heuristics.
National Resources Defense Council (NRDC) vs. EPA (2011), United States Court of Appeals for the Second Circuit 658 F.3d 200	NRDC challenged EPA's dichlorvos risk assessment on several grounds. The risk assessment itself had various components, with some relying on a study of human subjects (the "Gledhill study"), and some relying on animal toxicology studies. NRDC claimed that the agency's departure from the presumptive tenfold children's safety factor (the heuristic) was unsound. The court agreed in part, finding that in some components of the risk assessment the agency had failed to explicitly justify this departure. These portions of the risk assessment were vacated. By contrast, in the portions of the risk assessment that relied on animal studies, the EPA provided explicit reasons for departing from the default (*e.g.* "no evidence for increased susceptibility of the rat and rabbit offspring to prenatal or postnatal exposure to dichlorvos"). These portions were upheld. Another grounds for NRDC's challenge was that the EPA's reliance on the Gledhill study violated its own gatekeeping heuristic. The rule, which is restricted to pesticides regulation, forbids the EPA (and FDA) from relying on data from research on human subjects initiated prior to April 7, 2006 in cases where there is "clear and convincing evidence that the research was either:	Once more, (portions of) a risk assessment were overturned in court due to unreasoned departures from heuristics. However, unreasoned here should be interpreted in the literal sense of a failure to provide explicit justifications, rather than necessarily being unwise. The (unresolved) dispute over the gatekeeping rule is interesting in that it highlights the interpretive flexibility contained within linguistically formulated (rather than mathematical) rules.

	[1] fundamentally unethical (e.g., the research was intended to seriously harm participants or failed to obtain informed consent), or [2] significantly deficient relative to the ethical standards prevailing at the time the research was conducted." This dispute turned on whether the human subjects could be considered to have given their informed consent. For example, in parts of the consent forms the pesticide was referred to as a drug. However, the EPA countered that its independent experts found the consent forms ‘‘clearly advised subjects that this was a study involving consuming an insecticide." Court declined to resolve this dispute.	

Table 1: Data showing contested or problematic applications of heuristics in agency decision making, drawn from judicial reviews.