UNIVERSITY of York

This is a repository copy of *Forecasts or fortune-telling:when are expert judgements of safety risk valid?*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/124368/

Version: Accepted Version

Article:
Rae, Andrew and Alexander, Robert David orcid.org/0000-0003-3818-0310 (2017) Forecasts or fortune-telling:when are expert judgements of safety risk valid? Safety science. pp. 156-165. ISSN: 0925-7535

https://doi.org/10.1016/j.ssci.2017.02.018

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Forecasts or fortune-telling: when are expert judgements of safety risk valid?

Andrew Rae[1] and Rob Alexander[2]

[1] Griffith University, Australia
[2] University of York, United Kingdom

Abstract
Safety analysis frequently relies on human estimates of the likelihood of specific events. For this purpose, the opinions of experts are given greater weight than the opinions of non-experts. Combinations of individual judgements are given greater weight than judgements made by a lone expert. Various authors advocate specific techniques for eliciting and combining these judgements. All of these factors – the use of experts, the use of multiple opinions, and the use of elicitation and combination techniques – serve to increase subjective confidence in the safety analysis. But is this confidence justified? Do the factors increase the actual validity of the analysis in proportion to the increase in subjective confidence?

In this paper, by means of a critical synthesis of evidence from multiple disciplines, we argue that it is plausible that expert judgement deserves special standing, but only for well understood local causal mechanisms. We also conclude that expert judgements can be improved by using appropriate elicitation techniques, including by combining judgement from multiple experts. There is, however, no evidence to suggest that fuzzy algorithms, neural networks, or any other form of complicated processing of expert judgement have any advantage over simple combination mechanisms.

1 Why does expert judgement validity matter?

Would you trust a panel of government risk experts who told you that it was safe to build a nuclear waste processing plant in your neighbourhood? How about an international community of scientists predicting climate change? How about a single engineer telling you not to cross a bridge, because their calculations suggested it was unsafe?

Safety analysis has always, to a greater or lesser extent, relied on the opinions of experts. Individuals with specialist domain knowledge, or with superior understanding of risk analysis, are called upon to determine the nature, size, and acceptability of risk. Risk estimates produced by experts are more believable, but this does not necessarily make them more correct.

Our discussion in this paper is concerned with the use of experts for estimating the risk of major accident events. Unlike some risk problems such as population health, where there is a substantial body of recent data on which to base projections, major accidents occur too infrequently for past statistics to be a good indicator of risk.

It is in these situations that expert judgement is most necessary, but also most questionable. A clear understanding of the validity of expert judgements, and of how their validity is influenced by methods of elicitation and combination is essential for good safety practice. It is also important to be able to draw a clear distinction between expert estimates and value judgements. Experts should demand a role in decision-making only to the extent that they have something offer, not because their status confers special privileges.

There is an increasing trend to make use of multiple expert opinions in safety analysis, and to formalise the way these estimates are used. This involves documented methods for how opinions are elicited, how they are combined, and how they are integrated with other facets of the analysis. The trend is manifest in the academic literature - for recent illustrative examples see Zhou (2017), Forteza (2016), and Kokangul (2017) – and in regulatory guidance (Boring et al., 2005).

Practices for forecasting using expert judgement have been heavily studied outside safety science. In particular, there has been extensive work within social psychology and management focussing on group decision-making, and within economics focussing on the mathematics of combining individual probability estimates. There is also a body of large-scale experimental work using prediction markets and competitive forecasting. A lot is known about expert forecasting, but little of this knowledge is employed in safety practice.

In writing this paper we have been motivated by the proliferation of complicated techniques in the academic safety literature for eliciting and combining expert judgments. Of particular concern are papers that make definitive claims about the size and nature of risk based on these methods. Such research takes an unequivocally realist position on the nature of risk, whilst making unwarranted assumptions about the validity of the methods used. For example:

- That the performance of a safety management system has improved
- That human factors make a greater contribution to coal mining accidents than other safety issues
- That there is a particular ordered ranking of risks for cargo ships
- That particular geographic locations are more dangerous than other locations

- That specific companies are safer than other companies[1]

Frankly, we would like this researcher behaviour to stop. Armstrong suggests that the Golden Rule of forecasting is *"be conservative by adhering to cumulative knowledge about the situation and about forecasting methods"* (Armstrong, Green, & Graefe, 2015). In other words, forecasts should take into account what is known about forecasting itself, not just what is known about the problem at hand.

There are two questions that must be answered before expert opinion can be used to make definitive claims about safety risk:

1. What can be currently claimed about the validity of expert estimates as data for the purpose of safety risk estimation?

2. Under what circumstances, and to what extent, do methods for elicitation and combination of expert estimates of safety risk improve their validity?

2 Is there such a thing as a "risk estimation expert"?

2.1 "Expert" is a very ambiguous term for risk assessment and analysis

Predicting the future is a fundamental element of carnival fortune telling, sports betting, religious prophecy, and financial planning. Some types of prediction can be trusted, and others are cannot. Some people are better at making predictions. What does it mean to be an "expert" at predicting the future?

There are two main definitions of experts for the purpose of forecasting.

1. An expert is someone whose judgement is accorded extra weight, due to their qualifications, experience, and other signals of authority (Farrington-Darby & Wilson, 2006).

2. An expert is someone who makes especially accurate forecasts (Mellers et al., 2015).

Each of these descriptions is, in its own way, a fair summary. Which definition applies for risk assessment and analysis depends on how exactly risk is understood.

---

[1] It is not our intention to name and shame individual authors, so we have listed here unreferenced examples of recent definitive claims about risk based on processing of expert judgement.

The realist view (Smith, 2004) maintains that risk is a real, objective and quantifiable truth. The likelihood of an event in the future becomes the frequency of that event with the benefit of hindsight. Whilst we cannot know for certain how accurate risk estimates are at the time they are made, their accuracy may (at least in principle) be knowable at some point in the future.

In contrast, the phenomenological tradition, as explained by Rosa (1998), holds that even if objective risk exists as an abstract idea, there is no way to separate objective risk from our subjective and constructed experience of risk.

Very few researchers or practitioners argue that risk is entirely objective or entirely constructive – strict realism and strict phenomenology are extremes on a theoretical continuum. However, an inclination towards one paradigm or the other determines what is knowable about risk, and therefore what can be "valid". Most practical risk assessment is conducted from a generally realist perspective, whilst acknowledging that some degree of uncertainty is inevitable. Estimating risk, under this perspective, is analogous to guessing the number of marbles in a jar. The estimate is subjective, and the true number may never be known, but it is still possible to make statements about the objective goodness of the estimate. Goodness encompasses accuracy, certainty, and calibration.

An estimate is more "accurate" if it is closer to the true number. For example, if there are 250 marbles, an estimate of 240 is more accurate than an estimate of 230.

An estimate is more "certain" if it provides a narrower range of values. For example, an estimator might say "90% of the time, the true number of marbles will be between 240 and 260". An estimator would be overconfident if statements of this type were correct less than 90% of the time, and under-confident if they were correct more than 90% of the time.

There is no accepted term for correctness of certainty. We will use "calibration"; an estimate is calibrated if it is neither under confident nor over confident. It is better for an estimate to be more certain rather than less, but only if it is also well calibrated.

The applicability of "accuracy", "certainty" and "calibration" obviously depend on how risk is described. Not all descriptions of risk involve quantification (Kristensen, Aven, & Ford, 2006), and not all quantified risk includes separate assessment of certainty.

Not everyone agrees that risk estimate validity can be discussed in terms of accuracy, certainty and calibration at all. For those who believe that risk is primarily constructive, risk assessments and analyses are cultural artefacts. They document rather than determine decisions about risk. Validation comes from

"justifying the choices made in producing statements about risk" (Goerlandt, khakzad, & Reniers, 2016).

In this paper, following the "pragmatic validity" approach of Rae (2012) and Goerlandt (2016), we will evaluate claims about expertise in terms of the ontology used by the people who are making those claims. If risk assessments and risk analyses are being used for the sole purpose of explaining how decisions have been reached – that is, their makers are not intending to make objective statements about risk – then "accuracy" is not a meaningful dimension. These analyses should be validated based on a constructivist understanding. However, when risk estimates are attempts to describe risk as a real objective phenomenon - as they are whenever risk estimates are an input into decision making about further risk treatment – the risk estimates must provide a good description of the thing they purport to measure (Rae et al., 2012).

The combination of realist ontology and the use of experts requires a link between the two definitions of expertise. Experts should be a group of people whose opinions are deserving of extra weight because those opinions can be expected to be some combination of more accurate, more confident, and better calibrated.

Does such a group exist?

There are several plausible ways in which a potential expert could have a systematic advantage in making forecasts.

The first mechanism – private information – is that they could have access to privileged information held only by experts (Morgan, 2014). In economics, it is commonly assumed that given the same information, two people will produce similar forecasts, with small variations due to error and uncertainty. What distinguishes the "expert" from the "lay person" is a store of private information. Private information does not need to be explicit. It is possible that an expert cannot fully articulate exactly what it is they "know" that other people do not. However, they can use this information intuitively to make better predictions.

The second mechanism – domain knowledge – works through deep understanding of the specific causal mechanisms that lead to future outcomes in each particular case. What appears random and unpredictable to a layperson may be obvious to a scientist or engineer who understands the natural laws or technological principles governing the outcome (Farrington-Darby & Wilson, 2006). This type of expertise is more about the ability to process information than the information itself – a civil engineer doesn't just know the relative strengths of various materials, but also how to calculate the integrity of a structure incorporating those materials.

The third mechanism – super forecasting – is that experts may have superior general ability to extrapolate from the past to the future. This may be through pattern-matching skill – either instinctive projection of trends in the in the same way that an elite sports player can judge the future position of a ball, or by mastery of statistical tools for the identification of trends. Unlike the first and second mechanisms, the experts have no private information or domain knowledge, but they are more successful than others at reaching statistical conclusions. They are experts in the generic act of forecasting (Armstrong et al., 2015).

The three mechanisms are not entirely disjoint, but they indicate that there is a spectrum of forecasting expertise specificity. Super forecasting is quite generic and may be applied to a wide range of problems. Forecasting expertise based on knowledge of causal mechanisms is domain specific, but may be applied to most problems within that domain. Forecasting expertise based on private information is limited to those problems where the data is relevant.

Some problems are more or less tractable for each mechanism. Super forecasting and private information are useful where historical data provides a trustworthy (but obscure) indication of the future. Domain expertise is useful for novel situations where historical data does not apply, but where causal models have some predictive power.

All forms of expertise are at their weakest where there is little relevant historical data, and where the causal mechanisms are not well understood. In such cases, are some people still better forecasters than others? If so, under what conditions are they able to make more accurate predictions?

2.2 Expert accuracy is difficult to study

The debate surrounding the "Classical Method" of presenting expert assessments of uncertainty (Aspinall, 2010) provides a useful illustration of the theoretical and empirical difficulties in researching expert estimates. The Classical Method involves asking several experts for their "best estimate", as well as a confidence interval. For example, "How much will the project cost?" and "What is the lowest and highest value such that the true cost will fall between those values 95% of the time?"

Prior to the substantive estimation task, the experts are asked to participate in tasks using "seed variables" known to the elicitors but not the experts. Performance on these tasks is used to differentially weight the experts in the substantive estimation, such that greater weight is given to experts who perform better against the seed variable tasks.

Each step in the Classical Method is well defined, and there is considerable real world experience with the method. Yet, it is a matter of considerable controversy how well the method works (Bolger & Rowe, 2015; Roger M. Cooke, 2015).

The fundamental disagreement is about the relationship between the seed variable tasks and the substantive estimation tasks. Bolger and Rowe argue that since the seed variables represent known values, performance on these tasks is largely determined by skill at describing probabilities. Someone with low domain knowledge but good calibration may outperform a domain expert who is more accurate, but also overconfident or under confident. There is no reason (according to Bolger and Rowe) to believe that this superior performance translates to the substantive estimation task.

Cooke responds by suggesting that the purpose of weighting is not to differentiate experts based on accuracy, but on their calibration. There is (no reason (according to Cooke) to believe that superior calibration demonstrated on the seed variable tasks does not translate to superior calibration on the substantive estimation tasks.

The continuation of this debate 25 years after the Classical Method was first presented shows how difficult it is to provide fully persuasive arguments or empirical evidence about expert risk estimation performance.

To start with, there are many experimenter degrees of freedom in designing the experiment:

Who counts as an "expert"? Performing an experiment requires a sizeable body of experts. Actuaries, finance analysts, safety advisors and nuclear physicists represent very different types of expertise. Studies that show an advantage for one group of experts do not necessarily generalise to other experts.

Who serves as the control group? It has been well established that cultural factors (Wildavsky & Dake, 1990) and demographics (Kahan, Braman, Gastil, Slovic, & Mertz, 2007) influence risk perception. Unless the expert group and control group are demographically matched, it is hard to say whether any observed effect is the result of expertise, rather than of culture and demographics.

What task are the subjects given to perform? "Ecological validity" refers to the extent to which an experimental task matches a real world task. For some experiments, this has been interpreted as needing experts to perform a risk estimation task with which they are familiar. However, such designs actually reduce the generalizability of the experiment to other settings. If experts perform better merely because they have more practice at one specific task, this says little about their generalised ability to estimate risk.

How is performance evaluated?
There are different ways groups of estimates can be compared. On average (i.e. the mean) are their predictions higher or lower than another group? Is the average member of the group (i.e. the median) higher or lower? Is their average error higher or lower? Are their results more spread out, or tightly clustered? Where do most of the answers in the group (i.e. the confidence interval) fall?

How are differences interpreted?
Even when experts make different predictions to lay people, this is not necessarily representative of improved performance. In some cases, there is no correct answer to compare estimates with. Even when there is such an answer, remember that the "correct" answer has been chosen by researchers, who are presumably themselves "risk experts" and therefore have much in common with the expert subjects. What appears to be objectively better performance may be stronger alignment of values and assumptions between the experts and the researchers.

All of these issues make it difficult to draw conclusions about risk forecasting. They also provide numerous avenues for challenging empirical results.

For example, one of the earliest sets of studies on risk perception was conducted by Slovic (1985). These studies held that expert judgements of risk are based on likelihood and consequence, whereas lay judgements are distorted by qualitative factors. This view has been highly influential in subsequent risk research (Sjöberg, 2002).

Critics of the Slovic studies have since pointed out problems with each of the experimenter degrees of freedom (Rowe & Wright, 2001; Sjöberg, 2002): the experts "included a geographer, an environmental policy analyst, an economist, a lawyer, a biologist, a biochemist, and a government regulator of hazardous materials" (Slovic, Fischhoff, & Lichtenstein, 1979); the differences between the experts and the other groups are explainable in terms of their demographics; the task was ambiguous; most of the variance occurred in a small number of items; and the "correct" values chosen by the researchers were not predetermined.

Whilst it is easy, with hindsight, to point out the problems with any particular study, it is also hard to design a study that escapes such criticism.

Rowe and Wright (Rowe & Wright, 2001) review eight other empirical studies comparing experts with lay risk assessors, and point out consistent problems with the selection of the expert group, the design of the task, and the demographic matching of the expert and non-expert groups. Without such matching, they emphasise, any observed difference between lay and expert groups is explainable by factors such as age, gender and education.

Rowe and Wright also note that none of the studies provides an indication of the accuracy of the risk estimates (as opposed to just indicating that the lay and expert groups made different estimates). In absence of direct evidence about accuracy, the apparent lack of reliability (i.e. agreement between the experts) provides a strong suggestion that the experts were not accurate. Reliability is a prerequisite for accuracy (Rae et al., 2012).

2.3 Experts do not have access to privileged information about safety risk

Fischhoff (1982) suggests that any advantage experts have in risk estimation comes, "not in the way they think, but in the substantive knowledge they have at their disposal". When going beyond the available data or outside their realm of expertise, they may in fact operate at a disadvantage by relying on particular types or sources of data that are no longer core to the problem at hand.

All real-world risk estimates that rely on expert judgement operate outside the available data. Experts have a systematic advantage in experimental settings, which almost always involve risk problems for which there is a known answer. For example, Wright (2002) found that insurance underwriters were marginally better at some risk judgements than lay subjects. The experiment was specifically designed to match the types of judgements that the underwriters made in the course of their employment, and asked about population mortality risk, a topic where the underwriters were regularly exposed to the data they were tested against. Even under these ideal conditions the "experts" only performed slightly better than the lay subjects.

This improved performance evaporates when experts are asked to provide estimates in situations where the answer is unknown until after the estimate is made (Goodwin & Wright, 2010). There are two mechanisms undermining the experts. Firstly, they simply do not have enough examples of previous events – a "reference class" – to make accurate judgments. To the extent that they try to use their privileged knowledge about past events, they will be misled because the reference class does not adequately represent the circumstances they are hypothesising about. In fact, whilst their estimates are no better than lay predictions, experts may be overconfident in these estimates due to thinking that they do have useful privileged information (Lin & Bier, 2008). When asked to provide a range of estimates in this fashion, experts are more confident than lay estimators – they give narrower ranges – even though they are not more accurate.

Experts are also undermined by a lack of feedback. The development of expertise requires practice at a learnable task. A learnable task is one with a strong performance-feedback loop. Where there is no such loop – where risk estimators do not receive clear feedback on the accuracy of their predictions – experience does not build expertise (Rowe & Wright, 2001).

We do not mean to imply that historical data is an alternative to expertise, or that access to data makes someone an expert. Thomas (2004) points out that the target failure rates specified in standards for safety critical systems are in fact too low to ever be demonstrated. This is true even for the lowest levels of criticality (SIL 1 in IEC 61508). For these systems, the actual risk will never be known, and so the risk estimators will never have feedback on their performance. Expert judgement about risk under these circumstances is in practice endorsement of the design and the processes used to ensure that the design is safe, rather than an estimation of the residual risk per se, even if the judgements are expressed in the form of probabilistic estimates.

2.4 Domain experts may have superior understanding of causal mechanisms

Richard Feynman, in discussing NASA management culture, famously said "As far as I can tell, 'Engineering Judgement' just means they're going to make up numbers!" (Feynman, 2001, p. 183). In context, Feynman was drawing a distinction between engineering "analysis" – which he considered to be trustworthy and objective– and engineering "judgement" – which he considered to be arbitrary and subjective.

Is this distinction real? Quantitative Risk Assessment has a controversial status as a form of analysis (Aven & Heide, 2009; Rae, Alexander, & McDermid, 2014), but there are many other forms of technical analysis that produce accurate quantitative outputs. Examples of such analysis include:

- Short-term weather forecasting
- Fire and explosion modelling
- Static and dynamic analysis of loaded structures
- Pedestrian and traffic modelling
- Climate change modelling

The common feature of these types of analysis is that they involve the application of scientific and engineering principles to extrapolate the future state of a system from its current state. There is still some uncertainty in this process, since it is still necessary to create or select an appropriate model, and to choose suitable parameters for the influences on the system; however experts may be presumed to have an advantage in performing both of these tasks (Notarianni & Fischbeck, 1999).

This presumed advantage is disputed by Armstrong (1985), who suggests that a small amount of domain knowledge provides a benefit for understanding a forecasting problem, but that beyond this point further domain expertise does not translate into increased forecasting accuracy.

The advantage that experts hold is strongest when the model incorporates a small number of physical laws from a single domain, and weakest when it is

unclear what is or is not within scope of the model (Rae et al., 2014). Predicting the effect that a specific rate of CO2 emissions will have on the global temperature is a very different task from predicting how the international community will respond to climate change.

2.5 Expertise does not confer immunity from bias – but the ability to reduce bias may be a form of expertise

Accuracy in risk estimation can be achieved by reducing either random or systematic error. The previous section suggests that experts do not have an advantage for risk forecasting by reducing random error using information. However, it has also been suggested that experts might have superior skill at reducing systematic error (Slovic et al., 1979).

Unfortunately, it appears that domain experts exhibit the same types of bias as lay people, especially when forced to go beyond the limits of their expertise (Fischhoff et al., 1982; Lin & Bier, 2008; Skjong, Wentworth, & others, 2001). Moreover, there is some evidence that experts may be more prone than non-experts to particular types of bias. For example, experts may tend to structure problems to include existing numerical data and exclude difficult-to-quantify factors (Fischhoff et al., 1982; Slovic et al., 1985). They may also experience overconfidence (Lin & Bier, 2008) and anchoring based on early information or previously expressed opinions (Kinney & Uecker, 1982).

There is, however, some evidence of forecasters who are more expert precisely because they are less prone to reasoning errors. Mellers et al. report on a large scale experiment in the identification and development of "superforecasters" (Mellers et al., 2015). The experiment made use of a forecasting competition to identify individuals who were skilled at real-world predictions. The longitudinal nature of the competition allowed for forecasting tasks where the correct answer was unknown before the study, but was known afterwards (unlike laboratory experiments where it is possible to know in advance the value to be "predicted"). The study presents four hypotheses for the success of the successful individuals:

1. General cognitive abilities and styles such as fluid intelligence, enjoyment of problem solving, and willingness to change their minds
2. Task specific skills, such as the ability to make consistent probabilistic judgements (particularly with respect to conditional probabilities)
3. Motivation and commitment
4. More frequent and nuanced interaction with other forecasters

All four possibilities are plausible based on correlations between the hypothesised success factor and performance in the experiment. However, the project focussed on prediction of newsworthy events, with a particular emphasis on politics. It is not possible at this stage to exclude the possibility that the superforecasters achieved their superior results simply through superior general

knowledge of current affairs and politics, rather than from truly generic non-domain forecasting expertise.

3 Can expert judgements of risk be made more accurate?

3.1 The way questions are asked influences the validity of expert forecasts

Mosleh (1988) makes a distinction between "substantive goodness" and "normative goodness" for expert judgements. Substantive goodness refers to an expert's subject matter knowledge, whilst normative goodness refers to the expert's ability to express that knowledge in probabilistic form. The way in which a problem is put to the expert can significantly change the normative goodness. Different ways to ask the same question can encourage or discourage cognitive bias. They can also create a mismatch between how the expert understands the question, and how the information is going to be used.

Framing bias (Skjong et al., 2001) is where experts are unconsciously steered towards a particular answer by the way they are presented with information. For example, a problem may be presented in a way that has lots of detail about operator error and limited detail on mechanical failure, or vice versa. An unbiased analyst should, in principle, treat the shortage of information in one area as increased uncertainty. In fact, experts are more likely to present an answer dominated by the more detailed topic. They are more sensitive to information that is presented in great detail. The effects of framing bias can be reduced by allowing experts to seek out relevant information for themselves, rather than by providing them with lots of detail in the problem presentation.

Anchoring (Tversky & Kahneman, 1974) is where individuals form a starting estimate (the "anchor") and then insufficiently adjust this estimate based on further information. The starting estimate may come from a recent similar task, the first stage of a mental computation, or information provided in the question. For example, consider an expert asked to estimate the hourly frequency of death for driving a car, and then asked to estimate the hourly frequency of death for riding a motorbike. They are likely to select two values that are closer to each other than the actual historical figures are. Anchoring effects can be reduced by asking for upper and lower bounds rather than for a "best estimate" followed by a judgement of uncertainty (Morgan, 2014)

Equivocation involves multiple meanings for the same term. Walsten (1986) and Wardekker (2008) point out the considerable difficulties with semi-quantitative elicitation of risk judgements. Individual interpretation of probabilistic terms such as "about as likely as not", "medium likelihood", and "to be expected" varies widely, making it unlikely that two experts selecting the same qualitative term are in fact referring to the same underlying quantified range of likelihoods.

Framing bias, anchoring, and equivocation are not the only forms of bias in risk estimation. They are representative examples that demonstrate why the design of appropriate questions, based on up-to-date understanding of cognitive science literature, is important for expert forecast of probabilities.

3.2 Training experts influences the validity of their forecasts

Research on expert elicitation that suggests that particular modes of thinking result in better estimates (Mellers et al., 2015). However, elicitation research is itself mostly only validated for non-forecast probabilistic judgements. Applying this research to the question of forecast validity must assume that there is some degree of substantive goodness in the expert judgements, such that improvements in normative goodness improve the overall goodness. It is also necessary to assume that improving expert performance on non-forecast estimation tasks carries over on to forecast estimation tasks.

For example, systematic under or overestimation can be improved by asking experts to perform sample tasks, and then providing feedback on their performance (Morgan, 2014). This type of training also decreases expert confidence – necessary, since experts are typically overconfident.

There is also some evidence to suggest that asking estimators to make multiple judgements, either at different times or using different assumptions, can improve accuracy (Herzog & Hertwig, 2009; Vul & Pashler, 2008).

3.3 Task decomposition influences the validity of expert forecasts

There is mixed evidence on whether decomposing a risk forecast problem into smaller problems helps or hinders forecast accuracy (Chhibber, Apostolakis, & Okrent, 1992). On the one hand, decomposition of a problem into smaller problems improves transparency, and allows the use of multiple specialists with different areas of expertise. It also allows for the possibility that asking for many smaller estimates allows errors in those estimates to cancel each other out.

On the other hand, decomposing increases the complexity of the forecasting problem. It introduces further source of errors: the decomposition model itself can be wrong, or the experts may misunderstand which part of the problem they are being asked about. Rosqvist (2010) explains that transfer of parameter information from a domain expert to a risk analyst requires shared understanding of the mental model. However, mental models can only be represented by artefacts rather than directly compared, and are not stable over time.

Estimation decomposition may be considered analogous to decision-support tools that ask users a series of simple questions that are algorithmically combined to provide a final answer (Burns & Pearl, 1981). Such systems assume that the users are “expert” in the sense that they hold relevant information, but that the tool is “expert” in its understanding of cause and effect. For risk forecast

problems, decomposition makes sense if the experts have privileged information, but not if they are being relied on for their domain expertise. Domain expertise relies on knowledge of the structure of the problem – asking a domain expert for information in a way that presumes that the questioner knows more about the problem structure than the expert defeats the purpose of expert elicitation.

4 Are multiple experts better than one expert?

4.1 Opinions can be combined socially or mathematically

Do groups make better decisions than individuals, as in the "Wisdom of Crowds" (Surowiecki, 2005) or does the suppression of dissent lead groups astray (Solomon, 2006)? Unsurprisingly, the answer is "It depends". As suggested by Niall Ferguson, "serious students of human psychology will expect as much madness as wisdom from large groups of people" (Ferguson, 2009).

A more useful task is to unpick when, how and why combinations of judgements are better than a single judgement for the purpose of risk forecasting. This task can be undertaken in several ways. All three approaches have the same inputs: a set of individuals with starting opinions, and a process by which the starting opinions are combined. The approaches differ in how they characterise the process.

The first approach, used in social psychology, sees decision making as a form of social behaviour. The research objective is to examine by observation how groups interact to achieve consensus (Kerr & Tindale, 2011). Methodological approaches range from controlled experiments to textual analysis of real-world meetings.

The second approach, common in economics, is to treat group decision making as an information-sharing problem. The researchers model the individual starting opinions, and run simulations to evaluate different algorithms for combination (Clemen, 1989). Each expert is modelled as an assemblage of private information, public information, and noise. On each simulation run their "opinion" is generated using a random value for the noise, and the algorithm is used to establish consensus.

The third approach is to conduct large scale forecasting competitions, and to examine the individual strategies and group dynamics that lead to success (Makridakis & Hibon, 2000). This approach is initially agnostic as to the nature of the processes, but seeks to understand how good forecasters and forecasting groups describe their own processes.

These research methodologies align with the practical methods by which judgements can be combined (R. M. Cooke & Goossens, 2004). Mechanical combination asks experts to make individual estimates, which are used as inputs

to an algorithm to produce a final answer. Mechanical combination involves no social interaction, and so relies on information-sharing via the estimates themselves. As an alternative, behavioural combination allows interaction between experts to produce a judgement. Behavioural combination allows for experts to share information and attempt to persuade each other before reaching a consensus. Whether persuasion is a good or a bad thing depends on whether being right makes someone more persuasive. We will address this question in Section 4.2.

Some approaches use both mechanisms in sequential combination, with individual forecasts followed by group discussion, group discussion followed by individual forecasts, or multiple rounds of forecast and discussion.

4.2 Allowing experts to interact is possibly a good idea

For group-based decision making to improve forecast accuracy, it is necessary that some members will, after interacting with the rest of the group, update their original forecast in the right direction (Wright & Rowe, 2011).

There are certain types of problems, known as "intellective" or "Eureka" problems, where the right answer is obvious once it is known (Laughlin & Ellis, 1986). In a good riddle or cryptic crossword clue there is only one answer that "fits" – a group member who finds this answer can easily persuade the other members.

For other problems (for example knowing the age of a celebrity) a person with the wrong answer is just as likely to be persuasive as someone with the right answer.

Risk estimation is not a Eureka problem. However, there is a possibility that discussion between experts will identify errors in reasoning, and will enhance the information available to each expert (Mellers et al., 2015; Wright & Rowe, 2011)

The counterargument is that groups are predisposed to work towards consensus, rather than making small adjustments to the individual positions (Solomon, 2006). This phenomenon – groupthink – can result in the group selecting the most confident or authoritative opinion. Relying on the most confident opinion is not necessarily a bad thing, and works well for problems where most people know the correct answer (Koriat, 2012). However, for problems where most people are wrong, the least confident person's opinion is more likely to be accurate. It is not possible to know in advance whether the most confident or least confident person is more likely to be right.

Kerr and Tindale (2011) suggest that in purely judgemental tasks (where there is no correct answer) groups tend to coalesce around the majority opinion. However, when the correctness of an answer can be justified, this increases the

chance of a minority opinion prevailing. Unfortunately, it is often possible for an incorrect answer to be justified. Probability calculations are frequently counter-intuitive, and groups may adopt "socially plausible but inappropriate" strategies. This phenomenon is illustrated by the history of the Monty Hall problem, where incorrect solutions are frequently more persuasive than correct solutions (Krauss & T, 2003).

Strategies for group forecasting should take into account the success and failure mechanisms of groups. In particular, group forecasting methods should (Wright & Rowe, 2011):

- Include an opportunity for group members to establish trust in each other as sources of information
- Focus on sharing information, rather than just the individual estimates
- Ensure minority viewpoints are expressed
- Provide individuals an opportunity to reflect on new evidence, and to revise their original position

4.3 Mathematically combining opinions provides a better understanding of both risk and uncertainty

If experts are prohibited from social interaction during the forecasting process, then the efficacy of multiple experts is a mathematical problem with a clear answer. In almost all forecasting situations, combining expert opinions is more accurate than selecting an individual opinion.

Combining individual results to produce a more accurate aggregate result has a long history (Graefe, Armstrong, Jones, & Cuzan, 2013). As early as 1818, Laplace noted that combining results of experiments reduced the effects of random noise. Zajunc (1962) describes the history of statistical aggregation of group estimates in psychology. The original now-famous article by Galton (1907) observed that the median answer in competition to guess the weight of an ox at a county show was also the most accurate answer. Subsequent work replicated this result with other estimation problems, and additionally found that often the mean answer was more accurate than any individual guess.

The main mathematical property at work is now understood as the bracketing effect (Larrick & Soll, 2006). Bracketing works regardless of the sources of individual error, which may arise from:

- Different information held by different individuals
- Different assumptions and models used to create estimates
- Random variation (noise) in individual estimates.

Imagine two or more forecasts, each with some amount of error, such that the true value is bracketed – it lies somewhere within the range of forecasts. Then

create an average of all the forecasts. Mathematically, the error for the average forecast will never be greater than the average expected value of picking a forecast at random. Also, the worst case for picking a forecast at random will always have a larger error than the average forecast.

In other words, where there is a pool of experts with equal skill, it is never better mathematically, and may be considerably worse, to pick an individual than to poll the pool.

What about where bracketing doesn't occur? If there is systematic error in the pool, such that all of the experts are on one side of the correct value, then the average forecast will perform exactly the same as the expected value of choosing a forecast at random (Larrick & Soll, 2006). Under these circumstances there is decided benefit to finding and using the best individual, but there is no way of knowing if, and in which direction, the pool is biased. Trying to anticipate and compensate for all of the experts being biased is a recipe for dramatically increasing rather than decreasing error.

The theoretical advantage of bracketing has been verified in practice through many experiments. Clemen (1989) performed a review of the experimental literature. In those experiments, combining always improved forecasts and often outperformed the best individual forecast.

Once the bracketing effect is understood, the remaining question is whether a better forecast can be created by selecting a smaller, more accurate pool of forecasters – possibly even by choosing the best forecaster.

Goldstein (2014) examined the problem of "smaller, smarter" groupings within a larger pool, by analysing the results of online fantasy soccer competitions. The experiment showed that increasing group size improved performance up to a certain point, after which further increase resulted in less accurate predictions. In other words, adding experts can improve the consensus even when it is predicted that their individual performance is worse than all of the current group members. Group performance comes from the aggregate knowledge, rather the average individual performance. However, additional experts bring noise as well as information. The tipping point occurs when adding an additional expert would contribute more noise than information.

This is why election poll combining systems such as fivethirtyeight.com are able to make credible claims (backed up by a track record of successful predictions) that their model outperforms simply picking the individual polls with the best track record. It also explains why they incorporate (albeit with lower weighting) polls known to be historically less accurate (Graefe et al., 2013).

Combinations of experts are an improvement over individuals not just for the direct forecast task, but also for understanding how reliable the forecast is.

Knowing the degree of consensus between independent experts places a minimum bound on the amount of uncertainty contained in the estimate. Where experts strongly disagree, non-experts should not be confident even in an aggregate forecast.

4.4 There is no extra validity in complicated ways to combine opinions

In 1969, Bates and Granger experimented with various ways of combining forecasts about airline passenger data made by different models. In order to design their experiments, they considered (Bates & Granger, 1969):

1. How should past performance of forecasters be taken into account in combination?
2. How should individual forecasts be transformed before they are combined?
3. Should the combination take into account internal details of the individual forecasting models?
4. After any transformation, how exactly should the forecasts be combined?

In the following decades, hundreds of papers provided different answers to these questions, seeking the ideal mathematical mechanism for combining a pool of forecasts into a single forecast.

The simplest answer to the questions is to calculate a linear average, giving each forecast an equal weight. After a thorough review of twenty years of evidence, Clemen (1989) concluded that this is often the best method of combining.

However, whilst Clemen's conclusions are accorded considerable respect in the forecasting community (Wallis, 2011) there are a number of theoretical arguments why simple averaging with equal linear weights is not expected to produce an optimal answer. The first and most obvious argument is that equal weighting contradicts the starting assumption that some people are more expert than others. If all experts are presumed equal, this is inconsistent with the claim that experts are superior to non-experts.

A second argument is that aggregate forecasts have different mathematical properties to single forecasts. For example, Baron et al. (2014) showed that aggregated probabilities correspond closer to the real world if they are transformed so that they are closer to 0% or 100%. There are both mathematical and psychological reasons for this. Mathematically, it is not possible to have a probability smaller than 0% or greater than 100%. Thus, as an individual estimate approaches 0% or 100%, noise is more likely to take it towards than away from 50%. Psychologically, experts tend to allow for their own ignorance in making estimates, biasing distribution away from extremes. Rather than

cancelling out, as in the case of normally distributed noise, these skews are compounded by forecast aggregation.

The third argument is that individual experts may hold information that is destroyed by simple aggregation. Why not ask experts not just for a final answer, but also for their intermediate calculations? Combining expert opinions for each part of the problem, rather than for the whole, might allow better integration of the full information held by each expert (Kaplan, 1992).

Unfortunately, there is no evidence that any approach based on this third argument provides improved forecast accuracy. Zhang and Tai (2016) provide a review of one such method, Bayesian Belief Networks (BBN). In the field of maritime accidents alone Zhang and Tai identified over thirty published papers presenting methods for risk estimation using BBN with expert opinions. Most of these papers suggest that BBN provides superior objectivity and reduces bias. None provide evidence of increased accuracy.

There is a similar body of work, with similar lack of evidence, covering the Analytic Hierarchy Process (AHP). See for example Wang (2016) claiming increased validity and making managers "feel more confident" but no evidence of increased accuracy. In fact, there are strong arguments that AHP may in fact introduce mathematical anomalies through impermissible mathematical operations, contradictory axioms, and misunderstood scales (Warren, 2004).

BBN and AHP, along with DEMATEL, Neural Networks, and "fuzzy" variants of all four techniques, are mechanisms for aggregating expert opinions to provide risk forecasts. It is beyond the scope of this paper to comment more generally on the usefulness of these techniques for other purposes. For risk forecasting, there is an increasing volume of academic work refining, expanding, and combining the techniques, without any corresponding increase in the body of evidence suggesting that they can produce superior forecasts.

Ball (2002) suggests that the proliferation of complicated risk estimation techniques is a response by the risk estimation community to broader social disputes about risk. Risk estimators are typically mathematicians and engineers - "those who enjoy quiet, meticulous work" - poorly equipped by training and inclination to engage in social and epistemological debate. Instead of responding to broad challenges to risk estimation validity, they concentrate on refining the technical detail of risk estimation methods.

5 The path to improving expert judgement validity is through more description and less quantification

All practitioners of risk assessment adopt ontologies of risk. These ontologies are usually implicit, and adopted without deep consideration, or even any awareness that other ontologies are possible. Risk researchers, on the other hand, are often greatly interested in exactly what risk is, and in subtle distinctions between risk and related concepts such as probability, certainty, and strength of evidence. The result is that critic and practitioner disagreements are often about the nature of validity rather than validity itself.

Experts are current asked to estimate risks in contexts that assume:

1. That actual risk is real and objective
2. That individual estimates of risk have varying performance based on accuracy, certainty and calibration with respect to the actual risk
3. That expert estimates of risk have superior performance
4. That sophisticated elicitation and combination techniques improve expert performance

For the purpose of our review, we granted the first two of these assumptions. This is the principle of pragmatic validity - that risk assessments should be judged based on how they are performed and used, rather than on external ontologies of risk.

After making this concession, we then examined the best available evidence for the remaining two assumptions. We considered evidence about expert forecasting more generally, and risk estimation specifically. Neither assumption can be supported.

It is a mistake to believe that expert status acquired through authority, experience or job description, carries with it an ability to make risk forecasts that are somehow more objectively accurate. Whilst it is plausible that some individuals have superior skill at risk forecasting, there is no method other than past performance to identify such people. And in the case of low frequency high consequence events, there is insufficient past data to make those judgements.

It is also dangerous to suggest that convoluted methods of expert estimate elicitation or complicated mechanisms for estimate combination enhance the validity of expert judgements. Combining multiple forecasts will almost always result in a better forecast, but this improvement is realised by simple averaging of the individual judgements. Providing experts with an opportunity to interact may increase forecast accuracy, but improvement through this method is not guaranteed. Decomposing forecast problems and asking experts to tackle each component in turn has plausible benefits, but also plausible drawbacks.

Against this lack of evidence must be weighted the unquestionable costs of using experts. There is substantial time and money involved in selecting groups of experts and providing them opportunity to interact or participate in multi-stage

elicitation processes. Worse, there is an increase in apparent validity without a corresponding increase in actual validity. Methods that increase cost and apparent validity should be justified by commensurate increases in actual validity.

One thing that is known for sure is that experts tend to be overconfident in the accuracy of their forecasts. Complicated elicitation and combination methods make this worse. Their apparent sophistication increases the appearance of validity without improving actual accuracy. Complexity is bad for transparency. It disguises the effects of modelling assumptions and parameter selection. It risks applying expert opinions in ways that contradict the understanding of those very same experts. It disguises the fundamental weakness of expert risk estimation – that domain experts are being asked questions beyond the limits of their expertise – behind a cloak of algorithmic magic.

Activities that increase the appearance of assurance at the same time as they are unable to provide useful assessment are examples of probative blindness (Rae & Alexander, 2017). Probative blindness contributes to the inability of organisations to appropriately update beliefs about danger, a central theme in theories of accidents such as "drift" (Dekker, 2011) and disaster incubation (Turner, 1976).

What does this mean for conscientious analysts, researchers and decision makers, who want to make use of the best available evidence, but do not wish to overstate the quality and validity of that evidence?

We should all pay attention to experts. When domain experts agree about causal mechanisms, that consensus should form an important element of prediction and planning. Where experts disagree, the contradictions provide important information about structural uncertainty, and tell us to be less confident in the predictions we make.

However, it is clear that expert estimates are currently being elicited and applied – including by researchers – in ways that are not supported by the evidence about what experts are capable of. Expert risk assessments are not fit for purpose when used as truth-engines to measure risk as an objective quantity, so they should not be used for that purpose.

Any solution requires more transparency surrounding the risk assessment process, including clear explanations of the underlying methodological assumptions. What exactly are "experts" being asked to do, and why are they believed to have superior performance at this task?

We, the authors, are agnostic about how this is to be achieved, but we recognise two promising directions.

The first possibility is to focus on risk assessment as a means of describing, rather than quantifying risk. See for example the work of Kristensen (2006), Flage (2009) and Aven (2013). Each of these papers suggests replacing probability as the core dimension of risk with a more nuanced explanation of uncertainty. This is in contrast to current practice, where, when uncertainty is described at all, it is a qualifier or modifier for risk expressed as a probability (Eduljee, 2000).

There are open questions about the best way to communicate information about strength of evidence (Shackley & Wynne, 1996) and whether experts are in fact capable of distinguishing between their estimates and their uncertainties (Bolger & Rowe, 2015).

The second possibility is to encourage practitioner acceptance of the constructivist view - that expert judgements about risk are a tool for communicating rather than quantifying risk. They gain validity through transparency. This view is consistent with standards for risk assessment review based on form and content, but is inconsistent with regulations based on the achievement of specific risk targets.

Ultimately, safety is improved through better physical and organisational conditions, including but not limited to specific hazard controls. This is best served by open discussions about the sources of risk, the current safety measures, and the quality of the evidence informing decisions about further improvements. Expert opinion has an important role to play in these discussions, but this role should not be confused by methods that provide a false assurance of objectivity.

Armstrong, J. S. (1985). *Long-Range Forecasting: From Crystal Ball to Computer*. New York: John Wiley & Sons.

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, *68*(8), 1717–1731. https://doi.org/10.1016/j.jbusres.2015.03.031

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*(7279), 294–295. https://doi.org/10.1038/463294a

Aven, T., & Heide, B. (2009). Reliability and Validity of Risk Analysis. *Reliability Engineering & System Safety*, *94*(11), 1862–1868. https://doi.org/10.1016/j.ress.2009.06.003

Aven, T., & Reniers, G. (2013). How to define and interpret a probability in a risk and safety setting. *Safety Science*, *51*(1), 223–231. https://doi.org/10.1016/j.ssci.2012.06.005

Ball, D. J. (2002). Environmental risk assessment and the intrusion of bias. *Environment International*, *28*(6), 529–544. https://doi.org/10.1016/S0160-4120(02)00061-2

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468. https://doi.org/10.1057/jors.1969.103

Bolger, F., & Rowe, G. (2015). The aggregation of expert judgment: do good things come to those who weight? *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *35*(1), 5–11. https://doi.org/10.1111/risa.12272

Boring, R., Gertman, D., Joe, J., Marble, J., Galyean, W., Blackwood, L., & Blackman, H. (2005). *Simplified Expert Elicitation Guideline For Risk Assessment Of Operating Events* (No. INL/EXT-05-00433). Idaho National Laboratory.

Burns, M., & Pearl, J. (1981). Causal and diagnostic inferences: A comparison of validity. *Organizational Behavior and Human Performance*, *28*(3), 379–394.

Chhibber, S., Apostolakis, G., & Okrent, D. (1992). A taxonomy of issues related to the use of expert judgments in probabilistic safety studies. *Reliability*

Engineering & System Safety, *38*(1), 27–45. https://doi.org/10.1016/0951-8320(92)90103-R

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Cooke, R. M., & Goossens, L. H. J. (2004). Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research*, *7*(6), 643–656. https://doi.org/10.1080/1366987042000192237

Cooke, Roger M. (2015). The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight? *Risk Analysis*, *35*(1), 12–15. https://doi.org/10.1111/risa.12353

Dekker, S. (2011). *Drift into Failure*. Farnham, UK: Ashgate.

Eduljee, G. H. (2000). Trends in risk assessment and risk management. *The Science of the Total Environment*, *249*, 13–23.

Farrington-Darby, T., & Wilson, J. R. (2006). The nature of expertise: A review. *Applied Ergonomics*, *37*(1), 17–32. https://doi.org/10.1016/j.apergo.2005.09.001

Ferguson, N. (2009). *The Ascent of Money: A Financial History of the World* (1 edition). New York: Penguin Books.

Feynman, R. P. (2001). *"What Do You Care What Other People Think?": Further Adventures of a Curious Character*. (R. Leighton, Ed.) (Reprint edition). W. W. Norton & Company.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1982). Lay Foibles and Expert Fables in Judgments about Risk. *The American Statistician*, *36*(3b), 240–255. https://doi.org/10.1080/00031305.1982.10482845

Flage, R., & Aven, T. (2009). Expressing and communicating uncertainty in relation to quantitative risk analysis. *Reliability & Risk Analysis: Theory & Application*, *2*(13), 9–18.

Forteza, F. J., Sesé, A., & Carretero-Gómez, J. M. (2016). CONSRAT. Construction sites risk assessment tool. *Safety Science*, *89*, 338–354. https://doi.org/10.1016/j.ssci.2016.07.012

Galton, F. (1907). Vox Populi. *Nature*, *75*(1949), 450–451.

Goerlandt, F., khakzad, N., & Reniers, G. (2016). Validity and validation of safety-related quantitative risk analysis: A review. *Safety Science*.

Goldstein, D. G., McAfee, R. P., & Suri, S. (2014). The wisdom of smaller, smarter crowds (pp. 471–488). ACM Press. https://doi.org/10.1145/2600057.2602886

Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological Forecasting and Social Change*, *77*(3), 355–368. https://doi.org/10.1016/j.techfore.2009.10.008

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzan, A. G. (2013). *Combining Forecasts: An Application to Elections* (SSRN Scholarly Paper No. ID 1902850). Rochester, NY: Social Science Research Network.

Herzog, S. M., & Hertwig, R. (2009). The Wisdom of Many in One Mind Improving Individual Judgments With Dialectical Bootstrapping. *Psychological Science*, *20*(2), 231–237. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and Identity-Protective Cognition: Explaining the White-Male Effect in Risk

Perception. *Journal of Empirical Legal Studies*, *4*(3), 465–505. https://doi.org/10.1111/j.1740-1461.2007.00097.x

Kaplan, S. (1992). “Expert information” versus “expert opinions”. Another approach to the problem of eliciting/ combining/using expert knowledge in PRA. *Reliability Engineering & System Safety*, *35*(1), 61–72. https://doi.org/10.1016/0951-8320(92)90023-E

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting?: A social psychological analysis. *International Journal of Forecasting*, *27*(1), 14–40. https://doi.org/10.1016/j.ijforecast.2010.02.001

Kinney, W. R., & Uecker, W. C. (1982). Mitigating the Consequences of Anchoring in Auditor Judgments. *The Accounting Review*, *57*(1), 55–69.

Kokangül, A., Polat, U., & Dağsuyu, C. (2017). A new approximation for risk assessment using the AHP and Fine Kinney methodologies. *Safety Science*, *91*, 24–32. https://doi.org/10.1016/j.ssci.2016.07.015

Koriat, A. (2012). When Are Two Heads Better than One and Why? *Science*, *336*(6079), 360–362. https://doi.org/10.1126/science.1216549

Krauss, S., & T, X. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132*(1), 3–22. https://doi.org/10.1037/0096-3445.132.1.3

Kristensen, V., Aven, T., & Ford, D. (2006). A new perspective on Renn and Klinke’s approach to risk evaluation and management. *Reliability Engineering & System Safety*, *91*(4), 421–432. https://doi.org/10.1016/j.ress.2005.02.006

Larrick, R., & Soll, J. (2006). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle. *Management Science*, *52*(1), 111–127.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189. https://doi.org/10.1016/0022-1031(86)90022-3

Lin, S.-W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering & System Safety*, *93*(5), 711–721. https://doi.org/10.1016/j.ress.2007.03.014

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., … Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111

Mosleh, A., Bier, V. M., & Apostolakis, G. (1988). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering & System Safety*, *20*(1), 63–85. https://doi.org/16/0951-8320(88)90006-3

Notarianni, K., & Fischbeck, P. S. (1999). Dealing with Uncertainty to Improve Regulation. Presented at the Second Fire Safety Design in the 21st Century Conference, Worcester, MA.

Rae, A. J., & Alexander, R. D. (2017). Probative blindness and false assurance about safety. *Safety Science*, *92*, 190–204. https://doi.org/10.1016/j.ssci.2016.10.005

Rae, A. J., Alexander, R., & McDermid, J. (2014). Fixing the cracks in the crystal ball: A maturity model for quantitative risk assessment. *Reliability Engineering & System Safety*, *125*, 67–81. https://doi.org/10.1016/j.ress.2013.09.008

Rae, A. J., McDermid, J. A., & Alexander, R. D. (2012). The Science and Superstition of Quantitative Risk Assessment. In *Annual European Safety and Reliability Conference*. Helsinki.

Rosa, E. A. (1998). Metatheoretical foundations for post-normal risk. *Journal of Risk Research*, *1*(1), 15–44. https://doi.org/10.1080/136698798377303

Rosqvist, T. (2010). On the validation of risk analysis—A commentary. *Reliability Engineering & System Safety*, *95*(11), 1261–1265. https://doi.org/10.1016/j.ress.2010.06.002

Rowe, G., & Wright, G. (2001). Differences in Expert and Lay Judgments of Risk: Myth or Reality? *Risk Analysis*, *21*(2), 341–356. https://doi.org/10.1111/0272-4332.212116

Shackley, S., & Wynne, B. (1996). Representing Uncertainty in Global Climate Change Science and Policy: Boundary-Ordering Devices and Authority. *Science, Technology & Human Values*, *21*(3), 275–302. https://doi.org/10.1177/016224399602100302

Sjöberg, L. (2002). The Allegedly Simple Structure of Experts' Risk Perception: An Urban Legend in Risk Research. *Science, Technology, & Human Values*, *27*(4), 443–459.

Skjong, R., Wentworth, B. H., & others. (2001). Expert judgment and risk perception. In *The Eleventh International Offshore and Polar Engineering Conference*. International Society of Offshore and Polar Engineers.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1979). Rating the Risks. *Environment: Science and Policy for Sustainable Development*, *21*(3), 14–39. https://doi.org/10.1080/00139157.1979.9933091

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1985). Rating the Risks: The Structure Of Expert And Lay Perceptions. In V. T. Covello, J. L. Mumpower, P. J. M. Stallen, & V. R. R. Uppuluri (Eds.), *Environmental Impact Assessment, Technology Assessment, and Risk Analysis* (pp. 131–156). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-70634-9_7

Smith, M. J. (2004). Mad Cows and Mad Money: Problems of Risk in the Making and Understanding of Policy1. *The British Journal of Politics & International Relations*, *6*(3), 312–332. https://doi.org/10.1111/j.1467-856X.2004.00142.x

Solomon, M. (2006). Groupthink versus The Wisdom of Crowds: The Social Epistemology of Deliberation and Dissent. *The Southern Journal of Philosophy*, *44*(S1), 28–42. https://doi.org/10.1111/j.2041-6962.2006.tb00028.x

Surowiecki, J. (2005). *The Wisdom of Crowds* (Reprint edition). New York: Anchor.

Thomas, M. (2004). Engineering Judgement. In *Proceedings of the 9th Australian Workshop on Safety Critical Systems and Software - Volume 47* (pp. 43–47). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1082338.1082343

Turner, B. A. (1976). The Organizational and Interorganizational Development of Disasters. *Administrative Science Quarterly*, *21*(3), 378–397. https://doi.org/10.2307/2391850

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Wallis, K. F. (2011). Combining forecasts – forty years later. *Applied Financial Economics*, *21*(1–2), 33–41. https://doi.org/10.1080/09603107.2011.523179

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348–365. https://doi.org/10.1037/0096-3445.115.4.348

Wang, Q., Wang, H., & Qi, Z. (2016). An application of nonlinear fuzzy analytic hierarchy process in safety evaluation of coal mine. *Safety Science*, *86*, 78–87. https://doi.org/10.1016/j.ssci.2016.02.012

Wardekker, J. A., van der Sluijs, J. P., Janssen, P. H. M., Kloprogge, P., & Petersen, A. C. (2008). Uncertainty communication in environmental assessments:

views from the Dutch science-policy interface. *Environmental Science & Policy*, *11*(7), 627–641. https://doi.org/10.1016/j.envsci.2008.05.005

Warren, L. (2004). *Uncertainties in the analytic hierarchy process*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA431022

Wildavsky, A., & Dake, K. (1990). Theories of Risk Perception: Who Fears What and Why? *Daedalus*, *119*(4), 41–60.

Wright, G., Bolger, F., & Rowe, G. (2002). An Empirical Test of the Relative Validity of Expert and Lay Judgments of Risk. *Risk Analysis*, *22*(6), 1107–1122. https://doi.org/10.1111/1539-6924.00276

Wright, G., & Rowe, G. (2011). Group-based judgmental forecasting: An integration of extant knowledge and the development of priorities for a new research agenda. *International Journal of Forecasting*, *27*(1), 1–13. https://doi.org/10.1016/j.ijforecast.2010.05.012

Zajonc, R. B. (1962). A Note on Group Judgements and Group Size. *Human Relations*, *15*(2), 177–180. https://doi.org/10.1177/001872676201500206

Zhang, G., & Thai, V. V. (2016). Expert elicitation and Bayesian Network modeling for shipping accidents: A literature review. *Safety Science*, *87*, 53–62. https://doi.org/10.1016/j.ssci.2016.03.019

Zhou, X., Shi, Y., Deng, X., & Deng, Y. (2017). D-DEMATEL: A new method to identify critical success factors in emergency management. *Safety Science*, *91*, 93–104. https://doi.org/10.1016/j.ssci.2016.06.014