in Expert Judgement in Risk and Decision Analysis eds Nane, Hanea, French and Bedford, Springer Nature Switzerland AG, Cham, Switzerland, 2021.

Building on Foundations: the SEJ interview with Roger Cooke

1. **To start off, which applications of Structured Expert Judgment (SEJ) were most important for disseminating the Classical Model (CM), in your opinion?**

   The applications of which I am aware are summarized on my webpage http://rogermcooke.net/ , so I won't give separate references. Christian Preyssl got us started with applications at ESTEC. Certainly the nuclear work in the 1990's was most influential in working out dependence elicitation and probabilistic inversion. The EU Procedures Guide (Cooke and Goossens, 2000) emerged from that work and standardized the methods. There was a complex mating ritual between the European and American teams, but it all worked out in the end. I was not involved in Willy Aspinall's volcano work, but that certainly was very fecund. The fine particulate study led by John Evans was very important foray into public health applications. Willy led several subsequent applications in this area. The recent ice sheet applications are very important in bringing SEJ uncertainty quantification to the climate discussion.

2. **You originally studied Philosophy of Science at Yale – can you tell us a little bit about that, and about how that study influenced the later development of your thinking? For a Philosopher you have done a lot of work on real life problems.**

   I started in Philosophy of Physics. Not only was that program weak at Yale, but I struggled with the basics. Put a glass of water on a table. There are two invisible forces acting, gravity is pulling the glass down and the table is pushing back up with, miraculously, the exact same magnitude in the opposite direction so that nothing happens. Take away the table and the glass falls, take away gravity and nothing happens. To do the simple physics exercises you have to know the code, and know which questions not to ask. I eventually learned the code, but retained the sense that there were flaws in the story. The real puzzle for me was why "force" worked and "phlogiston" didn't. Why "space" and "time" worked but "ether" didn't.

   I switched to Philosophy and spent a lot of time on the Greeks, the Scholastics and Enlightenment philosophers, especially Kant and Hegel. The great philosophers construct a coherent system for understanding everything. In so doing they start with the natural language and progressively re-wire it so that concepts successively acquire new meanings, defined in evolving contexts. You can't understand it piecemeal; you just have to keep going until it all starts making sense. Once you "get it" you can see everything in a new way, like a conversion experience. Most people have at most one conversion experience which they then regard as apodictic. If you study philosophy you go through several….it helps. In retrospect, that's one of the great things I learned in philosophy, that and how to read seemingly unintelligible texts.

   Here's an anecdote: a math colleague and I were trying to learn atmospheric dispersion modelling. Atmospheric chemists have their own code, which mathematicians find inscrutable. We started with an elementary text. The colleague would come to something he didn't understand, stop and look for another text to explain. That sequence doesn't converge. I would just keep reading until their code starts to become intelligible.

   The great Systematic Philosophies have at their core a theory of knowledge. What knowledge is determines what we can know; 'what we know' and 'how we know' are very tightly coupled. For Plato, knowledge was acquired by direct intuition of a soul sufficiently purged of false beliefs. For the Scholastics, knowledge is Reason applied to Divine Revelation. For Kant, Newton's mechanics and Euclidean geometry enjoyed a level of certainty not attainable by induction from observations: they are necessarily imposed on our perceptions of nature by our knowledge

apparatus – or so he thought. He was wrong about that, but he was right that, to turn a philosophical phrase, knowledge and the knowledge of knowledge are inseparable. For Hegel, knowledge is self-consciousness of trans-personal spirit. Another story altogether.

3. Foundations of probability played a big role in your thinking, could you elaborate?

The flaws in classical mechanics began to extrude themselves in the latter 19th century and people felt that language was a big part of the problem. In an effort to separate pure definitions and mathematics from deliverances from experience Heinrich Hertz gave the first axiomatization of classical mechanics in a proto-formal language. He found the notions of force and absolute space-time superfluous and unhelpful. Such formal approaches were cross fertilized with activity at the foundations of mathematics – another story. Ernst Mach invented "semantic analysis" whereby notions must obtain a semantic pedigree tracing them to elementary sensations before they are serviceable to science. It emerged that concepts like phlogiston, force, absolute space-time lacked semantic pedigrees. Propositions assigning them properties are not unknowable, they are meaningless. The power of that insight emerges when you contemplate all the unknowable things people believe. The revolutions of relativity and quantum mechanics drew heavily on semantic de-constructions. Mach himself believed that atoms also lacked a semantic pedigree and that propositions about atoms were therefore meaningless. Atoms, however proved very useful. Indeed where would modern physics be if confined to Mach's semantic strictures?

Philosophy of science emerged as an effort to articulate the scientific method and thereby determine what science is and is not. Is risk analysis science? psychoanalysis? creationism? economics? Terms like leptons and quarks do not have operational meaning in the narrow sense as they are not directly linked to measurements, yet they seem to be ok. What about Freud's id? creationism's intelligent designer? economists' representative consumer? What about randomness? fuzzy membership? degrees of possibility? You see where this is going.

The demarcation of science and non-science is closely bound up with the problem of "theoretical terms": articulate a semantics in which terms without direct operational meaning, nonetheless acquire meaning in a given theory. There is a load of active literature on this, which I have tried to boil down to a simple formula (see Cooke, 2004) "

> The operational meaning of "*degree of possibility*" in the proposition: *"The degree of possibility that the Loch Monster exists is 0.0031416"* is the set of non-tautological propositions not containing "*degree of possibility*" which that proposition implies.

What about "uncertainty", what does it mean? In the natural language it means different things in different context, including ambiguity, ambivalence, confusion, distrust, unpredictability and indecisiveness. Anyone wishing to "represent uncertainty" in a scientific context must do some serious re-wiring. As often happens, a scientific reconstruction of a term in the natural language captures only part of its native meaning. Compare "force" in physics and in the natural language.

L.J. Savage's foundation of subjective probability is a superlative example of rational reconstruction in science. He provides axioms describing rational preference with clear operational meaning for the primitive terms. Strong arguments support his axioms – maybe not as strong as arguments for the axioms of Zermelo Frankel set theory, but very strong nonetheless. He then proves that the preferences of a rational individual can be represented as expected utility, where a personal probability (aka subjective degree of belief) is uniquely determined and the utility function is unique up to a positive affine transformation. All my

students had to learn these proofs, not only to understand uncertainty but also to understand how to extend the purview of science.

Others may protest that uncertainty means much more than subjective probability. Duh. However, if you want to quantify, say ambiguity, you must provide operational meaning telling us whether it is, e.g., positive, invariant under monotone, affine or ratio transformations, etc. Those invariances must be derived from the operational meaning of the primitive terms. At a conference in Paris, a leading light presented his new definition of uncertainty which unbeknownst to him, allowed uncertainty to take negative values. The theologians would love that. There have been many variations on Savage's axioms, just as there have been many variations on Zermelo Frankel set theory, but they all remain variations around a core theory that is suitable for applications. There are also countless "alternative representations" of uncertainty which lack any foundation whatsoever.

4. How did the idea of a rational consensus emerge – can you describe what it is and why you think it its useful to policy and decision makers?

We come to the theme of extending the purview of science. Traditional philosophy of science pretends that, within the context of justification, science deals only with certainties and reasons deterministically. It isn't so. Society is increasingly confronting decisions with large uncertainties with consequences impacting our survival. We all know the myriad ways in which private interests can and will exploit uncertainty to further their own aims. We must bring 'decision making under uncertainty' within the purview of science. Savage provides necessary but not sufficient conditions for rational decision making under uncertainty. Indeed, ANY subjective probability combined with ANY utility is rational in the sense of Savage. Rationality in science, whatever that means, is much more restrictive. The challenge is to bring science based restrictions into Savage's model, at least with respect to probability, such that all subjective probabilities are not equal. Utility is another problem. Validation is not hopeless but much less active than the probability component of rational decision (see Neslo and Cooke, 2011).

I first encountered the term rational consensus in a book by Keith Lehrer and Carl Wagner (Rational Consensus in Science and Society, 1981). It is similar to that of De Groot. M. (1974), discussed in Experts in Uncertainty. Participants assign probabilities to events and weights to each other's probabilities, leading to an equilibrium distribution. There's nothing scientific about it IMO, and it is not remotely practical. Experts are over worked and under paid. They're not going to travel long distances to sit together and reach 'dialectical equilibrium' as prerequisite for weighing each other.

However, the term rational consensus stuck in my mind and I sought a more science-based meaning. The idea is that experts construct their rational consensus. They quantify their degrees of belief as subjective probabilities for both the variables of interest and for calibration variables taken from their field. They are scored as statistical hypotheses with respect to statistical accuracy and informativeness. The theory of strictly proper scoring rules, appropriately generalized, converts their scores into weights. The combination scheme satisfies necessary (not sufficient) conditions for the scientific method. Rational consensus means that experts pre-commit to the results of the combination. They needn't adopt the result as their personal probability. However, withdrawing from the rational consensus imposes a proof burden of showing how the necessary conditions were violated or should be improved. The necessary conditions are traceability, neutrality, fairness and empirical control. The last is of course the most consequential, it implements Popper's idea of falsifiability. Fairness excludes pre-judging

experts, neutrality corresponds to proper scoring rules, traceability means that all steps in the calculation must be open and reproducible.

Tony O'Hagan's question 'is rational consensus a subjective probability, if so whose?' gets the simple answer: it is the personal probability of any rational agent whose preference representation as expected utility has a personal probability agreeing with the rational consensus.

5. Can you tell us something about the types of risk problems that you were thinking about when you started developing your ideas about expert judgement?

The topology of the problems was defined in the Rassmussen Report (USNRC, 1975) and evolved through three generations as described in (Cooke, 2013) . We have panels of order 10 experts assessing up to 100 uncertain quantities. Discrete events are sometimes assessed, but most variables are effectively continuous. The Rasmussen report did a good job on traceability. Publishing all the expert raw data made visible the very large differences between experts, thereby raising the issues of combination and validation. The Rasmussen report selected the distributions used in the report in a rather inscrutable fashion. In the second generation studies, experts' rationales were catalogued and their distributions were combined with equal weighting. The third generation in which I participated added performance measurement, empirical validation, dependence modelling and probabilistic inversion.

6. Do you think over the years research on SEJ methods has focussed on the right areas of EJ? How important do you think the social sciences side of EJ is?

Classical Model (CM) drew heavily on decision science research from 1950 – 1990. Publication of the Delft SEJ database (Cooke and Goossens, 2008) spawned good research, starting with the special of RESS (Cooke, 2008b). Wisse et al. (2008) looked at moment based elicitations, rather than quantile elicitation. Lin and Bier (2008) regressed expert calibration on study parameters and found an 'expert effect', suggesting that differences in expert statistical accuracy are not explained by random fluctuations. Perhaps the most productive was Clemen's critique. In addition to raising all the familiar questions regarding calibration variables, he introduced the issue of cross validation. His method is Remove One At a Time (ROAT): calibration variables are removed one at a time and the recomputed Decision Maker (DM) assesses the excluded calibration variable. Thus, with 10 calibration variables, each is assessed by a different DM using weights based on the non-excluded items and scored for performance on the excluded items. Clemen (2008) analysed 14 cases in this way and found only 9 (62%) in which CM out performed equal weighting (EW), which was not statistically significant. Clemen's numbers checked out and I spent quite a bit of time analysing this. On typical data sets, removing one calibration variable can change an expert's calibration score by a factor 2 or 3, hence ROAT can upweight experts who assessed the excluded item badly. Doing this for ALL variables introduces a significant bias against performance weighting. I finally found a simple example that made this very clear. Colson and Cooke (2017) give a complete discussion of the ROAT cross validation exercises with CM. These exercises used own code which was not benchmarked against our publically available code EXCALIBUR, (http://www.lighttwist.net/wp/excalibur) and contained wildly divergent scores. We spent a lot of time trying to figure out what they were doing, even going so far as to obtain and analyse their codes where possible. Those studies can be bracketed (e.g. Lin and Cheng, 2008, 2012, Flandoli et al., 2011).

Eggstaff, Mazzuchi and Sarkani (2014) performed a very serious cross validation on the 62 studies available at the time. They took every non-trivial subset of calibration variables as a training set

to initialize the CM and scored performance on the complementary set. With 15 calibration variables there are 32,766 splits of training / test sets. Abby Colson and I worked with Lt. Col. Eggstaff for some time until we got exact agreement with EXCALIBUR. This is the only cross validation code for which this has been done, to my knowledge. We used this code for the cross validation of the post 2006 studies, and still use it. There are many subtle issues involved in such studies, but the upshot is that performance weighting (PW) outperforms equal weighting (EW) out-of-sample on 72% of studies, similar to Colson and Cooke (2017). Using the recommend training set size of 80% excludes training sets with very low power and pushes the fraction to 78%. Including the most recent studies brings the number to 84%. The bias in ROAT is very roughly the difference between 62% and 84%. The hypothesis that PW and EW are not statistically distinct is rejected at the 1.6E−6 level. All this activity emerged from Clemen's critique.

Researchers at George Washington University are exploring a new idea. EW is based on the idea that one expert is as good as another. If that were true, then a randomized panel should do just as well as the original panel. In other words, we could construct new experts by randomly scrambling the original expert assessments and it would perform just as well statistically. Initial results roundly reject the random expert hypothesis. This approach is potentially more powerful than cross validation because it doesn't require splitting the calibration variables. We're now comparing median predictions. It turns out that averaging medians (equally or performance weighted) yields markedly higher prediction errors than using medians of equally cq performance weighted combinations. Moreover. performance weighting out performs equal weighting in point predictions. Hence, even if one is only interested in point predictions, it is better to quantify uncertainty, measure performance and performance weight the experts' distributions

These are active mathematical research themes. Other active themes include dependence modelling (Werner et al., 2017) , dependence elicitation (Morales et al., 2008), stakeholder preference and probabilistic inversion (Neslo and Cooke, 2011) . The social sciences have also made enormous contributions to this field, for example the many publications of the Eugene Oregon school, of Kahnemann and Tversky, and of Fischoff. I got updated on the social science themes as lead author for the chapter on Risk and Uncertainty for IPCC AR (5).

Once we know how to measure expert performance, research into best training methods would be very helpful. This would probably link with risk communication research. It's a topic I hope the social scientists will pick up.

I would also like to see a good psychometric experiment that tests the Elsberg paradox where there is no information asymmetry between the experimental subject and the experimenter. The Elsberg paradox shows that people prefer a lottery with objective probability (½, ½), to a lottery with probability unknown to the subject (but known to the experimenter). I suspect that a large part of the "ambiguity aversion" effect is due to "manipulation aversion" when the subject knows that the experimenter knows more than (s)he does. For example, let the subject choose an odds-ratio $(1 - r)/r$, and let a fair coin determine which side of the lottery the subject will play. Now the probability of winning is equally uncertain to subject and experimenter alike. Do subjects still prefer a fair coin toss? By how much? and if you decrease the win on the fair coin from $10 to $9.90?

7. Do you think that the definition of SEJ from your book would need a revision? And are there any methods except the Classical Model (CM) that you would think are part of the SEJ group of methods/protocols?

CM has stayed pretty much the same, the only change from the book is that information is measured as relative information with respect to a background measure instead of inverse entropy – this was just for cosmetics to make the role of the background measure more visible. Relative information is a familiar concept, inverse relative entropy less so. Keeping the model unchanged helped build up a large data base of SEJ applications.

Re other methods: The IDEA protocol for discrete events is a very promising initiative (Hanea et al., 2016). It combines the CM with Delphi-like feedback rounds. Philip Tetlock's good judgement project (Unger et al., 2012) has had success forecasting current events measuring performance with the Brier score and successively eliminating experts until a small subset of "super forecasters" is found. The time and resources (in number of experts) required preclude application to science and engineering problems. Eliminating experts is a form of performance weighting. There have been very many proposals that do not attempt to validate their performance. To all these, I say *Why Not?* It's getting harder to pretend that validation is impossible.

8. Your book and other references contain numerous practical suggestions about performing an elicitation. Have any of these advices (technology, remote, etc.) changed with time?

The book says that elicitations should not exceed on hour. I would now say one-on-one elicitations must not exceed four hours. Four hours is gruelling. Other formats are now employed, including remote elicitation with e-tools.

9. Your Classical Model has been criticised by some because of the way it combines different paradigms to uncertainty – for example in using classical statistical tests as a tool to construct a judgemental probability distribution. Since you have impeccable credentials in philosophy of science you will be entirely aware of this, and of the other "rough and ready" choices you made in designing this approach. Is the classical model grounded in science or is it a part of what some may call "decision engineering"?

Familiarity with foundations teaches that the combination of experts is not a mathematics problem - the axioms of probability will never tell us how to combine experts. Its also not a problem of personal expected utility maximization and Bayesian approaches never achieved lift-off (see Cooke, 1991). The expert problem is akin to an engineering problem. We define the objectives and look for a design that optimizes performance. In our case the objective is to promote rational consensus through science-based uncertainty quantification. We use first principles, the axioms of probability, and second principles, Savage's axioms, but they obviously won't give us a working design. Tertiary principles like the marginalization property leading to the linear pool, and quaternary principles like scoring rules, P-values and Shannon relative information are also needed. Finally, we need to apply common sense. Any arbitrary choice of the analyst should be manipulable in the code and available for robustness analysis. Examples are the choice of background measure, the choice of calibration power and choice of P-value cut-off.

Some mathematicians don't appreciate the difference between mathematics and engineering: a bicycle obeys Newton's laws but doesn't follow from them. The design of a bicycle involves many decisions motivated by different considerations; the wheels could be a millimetre larger, the saddle a millimetre smaller, etc. A design always mixes physics, psychology, economics, etc. Complaints from academics about ad hoc-ness and methods mixing are like someone refusing to ride a bicycle because the optimal bicycle cannot be derived from Newton's laws. Such righteousness is most laudable.

in Expert Judgement in Risk and Decision Analysis eds Nane, Hanea, French and Bedford, Springer Nature Switzerland AG, Cham, Switzerland, 2021.

10. Many decision makers and social scientists are familiar with the measures 'accuracy' and 'precision' . How do those relate to the CM?

CM's performance measures of statistical accuracy and information do not map neatly onto the terms "accuracy" and "precision", which are familiar to social scientists. Accuracy denotes the distance between a true value and a mean or median estimate, and precision denotes a standard deviation. While appropriate for repeated measurements of similar variables, these notions are scale dependent and therefore not useful in aggregating performance across variables on vastly different physical scales. For example, how should one add an error of 10^9 colony forming units of campylobacter infection to an error of 25 micrograms per liter of nitrogen concentration in water? Expert judgments frequently involve different scales, both within one study and between studies. For this reason, the performance measures in the Classical Model are scale invariant. That said, the exhaustive out-of-sample analysis of Eggstaff et al. (2014) found that the realizations were closer to the PW combination's median than the EW combination's median in 74% of the 75 million out-of-sample predictions based on the TU Delft data. Such non-parametric ordinal proximity measures, proposed by Clemen (2008) are not used to score expert performance, as the scores strongly depend on the size of the expert panels.

11. Maybe the biggest criticism of the CM is the lack of representativeness of the seed questions for the questions of interest and the way performance is measured on those seeds.

The claim that performance on calibration variables cannot represent performance on variables of interest is just a bald assumption. Scientists don't traffic in bald assumptions; they look at evidence for or against the statement that performance on the calibration variables predicts performance on the variables of interest. Clemen (2008) is the only critic who used valid code, to my knowledge. Since the representativeness question gets asked on virtually every application, I have a standard answer. Suppose you have two experts, one is very accurate statistically and very informative on the calibration variables, the other is massively overconfident with abysmal statistical accuracy. Would you give them equal weight on the variables of interest? If your answer is "yes", then calibration variables have failed in their function.

The CM is subjected to empirical control in-sample on every application, including leave-one-out robustness on experts and calibration variables. It is validated out-of-sample with cross validation. In some studies the actual variables of interest have been observed post hoc (Goossens and Cooke, 2008). There are new ideas in the pipeline. My hope is that other approaches will also be motivated to address validation. For example, why shouldn't the Delphi method validate itself? The early studies did compare predictions with reality with very uneven results (see Cooke, 1991). Has the record improved? Why wouldn't practitioners of the Sheffield method attempt to validate their results against observations? Why shouldn't proponents of imprecise probabilities say what a good imprecise probability assessment is, and measure how well their methods perform? If one degree of possibility is as good as another, one imprecise probability interval as good as another, one fuzzy membership as good as another, then why go to all the trouble? Just use Happy Numbers, i.e. numbers that make you happy.

In view of all the research into validation, the claim that validation is impossible becomes a bit fatuous. Expert judgment is a raucous field with practitioners from very diverse backgrounds. Applying the CM requires a level of numeracy that many analysts may find challenging. Indeed, the analyst has to understand what a likelihood ratio is, what Shannon information is, what a proper scoring rule is. (S)he has to explain the CM to the experts and write it up intelligibly. Perhaps most importantly, (s)he must be able to explain why (s)he is NOT following any of the

other approaches in circulation…Bayesian averaging, quantile averaging, consensual probabilities, imprecision, fuzziness, degrees of possibility, Delphi, etc. etc. etc. Of course mathematicians know that the CM is not a heavy lift and many scientists and engineers have become adepts. However, to a non-numerate person, it may look like a very heavy lift.

People often ask if experts can game the system. It is theoretically possible and would be easy to spot. In all our applications there was one expert in one panel for whom we suspected that his business interests were informing his assessments.

12. What do you see as the greatest challenges facing the EJ practitioners in the coming years?

Finding enough qualified analysts.

13. One of the comments to the "Structured expert judgment" post from 2015 by Judith Curry said: "Uncertainty, like love, cannot be quantified. There is nothing to measure."

Ha Ha, sounds funny but isn't. Such cavalier attitudes towards uncertainty unwittingly license all the defective modes of dealing with uncertainty with which we now struggle. I'll mention three books to underscore this. The first is Oreskes and Conway (2010) "Merchants of Doubt". They detail the massive resources spent by private interests to create doubt and derail government regulation. Smoking causes cancer? NOT PROVEN!! look at this research sponsored by Tobacco Lobbies. Such tactics work as long as the public is unwilling and unable to reason under uncertainty. The question isn't whether we KNOW that smoking causes cancer, of course we don't. We don't KNOW that F = ma. The question is how much smoking raises the risk of cancer. We see the same small set of "experts" pandering to private interests, whether it is supersonic transport, smoking, air pollution or climate change.

The second book is Malcom Nance (2018) "The Plot to Destroy Democracy". Nance is a veteran cyber security specialist in the US Government and gives a detailed account of Russia's propaganda machine. The story goes back to Lenin, but has taken a new form in the age of internet. Hundreds and hundreds of Russians work 24-7 concocting lies targeting specific groups and pushing them onto the internet. A plausible lie gets a preamble in known facts embellished with things the target wants to believe, then topped off with a complete fabrication. This formula was applied to Pizza-gate (Hilary Clinton ran a pedophile trafficking ring from the basement of a Washington Pizza parlor) of which Michael Flynn Jr wrote "until #PizzaGate is proven false it'll remain a story". Less well known was, e.g. the fake news headline in St. Mary Parish, Louisiana "Toxic fume hazard warning in this area until 1:30PM". Despite being a proven Russian hoax, a Wikipedia page and YouTube video showed ISIS claiming responsibility with Burqa's waiving guns. And then D.J. Trump's "The Art of the Deal" written by Tony Schwartz. Trump reveals his tactic of "truthful hyperbole" (https://www.fastcompany.com/3068552/i-call-it-truthful-hyperbole-the-most-popular-quotes-from-trumps-the-art-of-the-deal). I interpret it thus: any proposition not provably false which suits your interests should be repeated as often as possible, challenging the adversaries (there are always adversaries) to disprove it. Since they can't, your story will win. I'm not saying that Savage can deliver us from all this. I am saying that peoples' unwillingness to reason probabilistically makes it possible to influence their behavior by pushing the proof burden to the side you want to lose. 'You haven't PROVED that smoking causes cancer', 'You haven't PROVED that climate change is real', 'You haven't PROVED that Russia hacked the US election', etc., etc., etc.

in Expert Judgement in Risk and Decision Analysis eds Nane, Hanea, French and Bedford, Springer Nature Switzerland AG, Cham, Switzerland, 2021.

An interesting aside on the Russian story: it was the Dutch AIVD which pinpointed the source of Russian troll farm COZY BEAR to a building in Moscow. They even counter-hacked the security cameras on one particular floor of the building and observed the Russian spies using the system.

14. Finally, if you could organise a dinner party with 3 or 4 'great thinkers' who influenced your development of the classical model, who would you invite and why?

Learning to reason probabilistically will be an event in the cognitive history of Man comparable to the formulation of deterministic reasoning in Aristotle's Logic. The great hero here is Frank P. Ramsey. His "Truth and Probability" (1926) is a bolt of sheer genius. Let's also include John von Neuman (Theory of Games and Economic Behavior, 1944) and Lenard Jimmy Ogashevitz (aka Savage)(The Foundations of Statistics, 1954). But not for dinner - nobody could get along with von Neumann.

The most important people at the inception of CM were Louis Goossens, Max Mendel and Simon French. Early adapters from the first hour were Willy Aspinall, Tim Bedford, Jan van Noortwijk, Matthijs Kok, Dmitri Solomatin, Gordon Woo, Tom Mazzuchi and Christian Preyssl. Follow on forces include Dorota Kurowicka, Anca Hanea, Tina Nane, Oswaldo Morales, Jim Hammitt, John Evans, Abby Colson, John Quigley, Justin Eggstaff, Rene van Dorp, Arie Havelaar, and Ben Ale. These would also need to be invited; we will need a Banquet Hall. Then we can also invite all the colleagues who performed the applications, Kim Thompson, Radboud Duintjer Tebbens, Juoni Tuomisto, Nicole van Elst, Daniel Puig, Frank van Overbeek, Xi Quing, Maurits Bakker, Rabin Neslo, Daniel Lewandowski, Sandy Hoffmann, Matt Gerstenberger, Maart Janssen, Augusto Neri, Eric Jager, Ben Goodheaart, Juliana Lopez de la Cruz, Julie Ryan, Maartin Nauta, Marion Whitmann, David Lodge, John Rothlisberger, Arno Willems, Jim Smith, Fred Harper, Steve Hora, Mark Burgman, Elizabeth Beshearse, Raveem Ismail, Vicki Bier, Bernd Kraan, Ben Koch, Daniela Hanea, Christoph Werner, Bis Bhola, Michael Oppenheimer, Jonathan Bamber, Bob Kok, Monika Forys, Michael Tyshenko, Maartin Nauta, Karin Slijkhuis … with apologies to everyone I forgot.

Recalling all these people and their contributions is quite humbling. BTW, didn't we have just such a banquet in July 2017?

References

Clemen, R. T. (2008). Comment on Cooke's Classical Method. Reliability Engineering & System Safety, Expert Judgement, 93 (5): 760–65. doi:10.1016/j.ress.2008.02.003.

Colson, A. R. and Cooke, R. M., (2017). Cross Validation for the Classical Model of Structured Expert Judgment, Reliability Engineering and System Safety, Volume 163, 109–120.

Cooke, R. M. (1991) Experts in Uncertainty: Opinion and Subjective Probability in Science. New York: Oxford University Press.

Cooke, R. M. (2004). "The anatomy of the Squizzle - the role of operational definitions in science". Reliability Engineering and System Safety 85, 2004, 313-319.

Cooke, R.M. (2008a) Discussion: Response to Discussants." Reliability Engineering & System Safety, Expert Judgement, 93 (5): 775–77.

Cooke, R. M. (2008b) Special issue on expert judgment, Editor's Introduction Reliability Engineering & System Safety, 93(5), Available online 12 March 2007.

Cooke, R.M. (2009). The Reliability of Aggregated Probability Judgments Obtained through Cooke's Classical Model. Journal of Modelling in Management 4 (2): 149–61.

Cooke. R.M. (2012). Pitfalls of ROAT Cross-Validation: Comment on Effects of Overconfidence and Dependence on Aggregated Probability Judgments. Journal of Modelling in Management 7 (1): 20–22.

Cooke, R. M. (2013). Uncertainty analysis comes to integrated assessment models for climate change… and conversely. *Climatic change*, *117*(3), 467-479.

Cooke. R.M. (2015). Messaging Climate Change Uncertainty. Nature Climate Change 5 (1).

Cooke, R.M. and Goossens, L.J.H. (2000). Procedures guide for structured expert judgment Project report EUR 18820EN, Nuclear science and technology, specific programme Nuclear fission safety 1994-98, Report to: European Commission. Luxembourg, Euratom. Also in Radiation Protection Dosimetry Vol. 90 No. 3.2000, 64 7, pp 303-311.

Cooke, R.M., and Goossens. L. H. J. (2008). TU Delft Expert Judgment Data Base. Reliability Engineering & System Safety, Expert Judgement, 93 (5): 657–74.

DeGroot, M. (1974). Reaching consensus, J. Amer. Statis. Assoc. vol.69, pp118 - 121.

Eggstaff, J.W., Mazzuchi, T.A. and Sarkani, S. (2014). The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model. Reliability Engineering & System Safety 121 (January): 72–82.

Flandoli, F., Giorgi, E. , Aspinall, W.P. and Neri, A. (2011). Comparison of a New Expert Elicitation Model with the Classical Model, Equal Weights and Single Experts, Using a Cross-Validation Technique. Reliability Engineering & System Safety 96 (10): 1292–1310.

Hanea, A.M.,. M.F. McBride, M.F., Burgman, M.A. and M.A Wintle, M.A. (2016). Classical meets modern in the IDEA protocol for structured expert judgement, Journal of Risk Research Volume 21, 2018 - Issue 4 , Published online: 09 Aug 2016

Lehrer, K., and Wagner, C. (1981). Rational Consensus in Science and Society, D. Reidel, Dordrecht.

Lin, S. W., and Cheng, C. H. (2008) Can Cooke's Model Sift out Better Experts and Produce Well-Calibrated Aggregated Probabilities?. In IEEE International Conference on Industrial Engineering and Engineering Management, 2008. IEEM 2008, 425–29..

Lin, S. W., and Cheng, C. H. (2012). Effects of Overconfidence and Dependence on Aggregated Probability Judgments. Journal of Modelling in Management 7 (1): 6–22. Lin, S. W. and Vicki M. Bier. (2008) A Study of Expert Overconfidence. Reliability Engineering & System Safety, Expert Judgement, 93 (5): 711–21.

Nance, M. (2018) The plot to destroy democracy, Hachette Book Group, New York.

Neslo, R.E.J. and Cooke, R.M. (2011). Modeling and validating stakeholder preferences with probabilistic inversion, Appl. Stochastic Models Bus. Ind. Games and Decisions in Risk and Reliability Analysis, Volume 27, Issue 2, Pages: 71-171, First published: 04 April 2011.

Morales, O. Kurowicka, D. and Roelen, A. (2008). Eliciting Conditional and Unconditional Rank Correlations from Conditional Probabilities, Reliability Engineering & System Safety, 93, 600-710. Available online 12 March 2007, Volume 93, Issue 5, May 2008.

Oreskes, N., and Conway, E. M. (2010). Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming (1st U.S. ed.). New York: Bloomsbury Press.

Ramsey, F. (1931). Truth and Probability" in Foundations of Mathematics and Other Logical Essays, London, originally written in 1926.

Savage, L.J. (1954). The Foundations of Statistics. John Wiley & Sons., 1954.

U.S. Nuclear Regulatory Commission. 1975. "Reactor Safety Study." WASH-1400, NUREG-75/014. Washington, D.C.

Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012) The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions. 2012 AAAI Fall Symposium Series

von Neumann, J. and Morgenstern, O (1944). Theory of Games and Economic Behavior, Princeton University Press.

in Expert Judgement in Risk and Decision Analysis eds Nane, Hanea, French and Bedford, Springer Nature Switzerland AG, Cham, Switzerland, 2021.

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., & Morales-Nápoles, O. (2017). Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions, European Journal of Operational Research, *258*(3), 801-819.

Wisse, B., Bedford,T., and Quigley, J. (2008) Expert judgement combination using moment methods. Reliability Engineering & System Safety, *93*(5), 675-686