Expert Judgment for Cyber-Security Risk

EuroSPI 2019 Workshop Report by Michael Krisper, Graz University of Technology

In the workshop “Expert Judgment for Cyber-Security Risk” at the EuroSPI conference in September 2019 in Edinburgh, Scotland [1], an expert elicitation was done to assess the risk of a cyber-security scenario involving car theft by hacking the passive keyless entry system. This elicitation was done with 21 volunteers from the conference and conducted by Michael Krisper and Georg Macher from Graz University of Technology. It followed the **IDEA protocol** by Victoria Hemming [2] and used a variation of **structured expert judgment** by Roger Cooke [3], [4].

Key Points:

1. Always use **multiple experts** for assessments. However, randomly chosen experts could lead to highly unprecise results, even when using 20 of them. We recommend a **diverse group** of domain experts (customers), technical experts (engineers) and quality supervisors (consultants).
2. **Combine** the expert response **based on their performance** during calibration – don’t simply trust them. Good experts should get much higher weight than bad experts. Don’t let the bad experts worsen the results.
3. Let the experts first build their **own unbiased opinion**, then let them **discuss and revise it**.

Procedure: The IDEA Protocol and Structured Expert Judgment

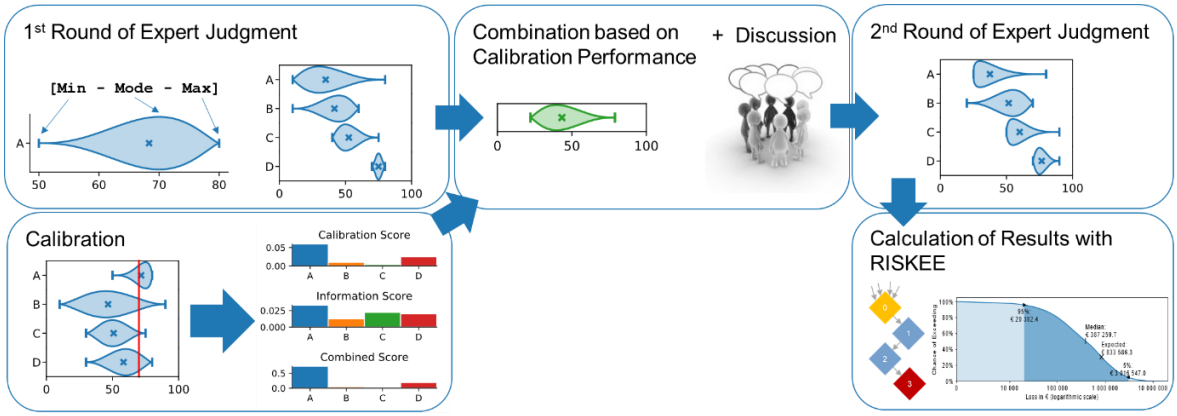


Figure 1: Our version of the IDEA protocol for EuroSPI 2019.

The IDEA protocol by Hemming et al. [2], [5] consists of two rounds of expert judgment with a discussion in between. During the discussion, the aggregated results of the first round are presented to the experts. The aggregation is done via a weighted combination of the responses based on the performance during the calibration questions. The performance is measured via the information and calibration score based on structured expert judgment by Roger Cooke [3]. In the EuroSPI workshop, we brought together 21 experts from different domains and with different background knowledge and did a study to assess the risk of a cyber-security scenario involving hacking and stealing cars. The whole workshop took approximately one hour and consisted of the following steps:

1. **Introduction** (10 Minutes) We described the general process to the workshop participants and explained some background information.
2. **Calibration** (10 Minutes) The calibration was done with a questionnaire that was filled out by the participants. The responses for the actual survey were weighted based on the calibration performance.
3. **Scenario explanation** (10 Minutes) We explained the scenario, which should be evaluated. We gave some background information and showed a possible attack which could be done.
4. **First Round of expert elicitation** (5 Minutes) Participants give their assessments for the scenario in the form of a three-point-estimate (min-mode-max). This was done individually and without prior discussion.
5. **Discussion and presentation of results** (15 Minutes) After the first round, a moderated discussion was done to clarify the scenario, let the experts exchange opinions and arguments and analyze the results of the first round.
6. **Second round of expert elicitation** (5 Minutes) Participants gave a revised second judgment.

Scenario: Fast Furious and Insecure – Hacking the passive-keyless entry system of Tesla Model S cars

The COSIC Group from KU Leuven published an attack, which allows stealing Tesla Model S cars, by exploiting flaws in the passive-keyless entry system [6]. The attack is using consumer hardware to communicate with the car and the key-fob, and a huge database of precomputed security keys. It needs vicinity to the car and the key-fob (not necessarily at the same time). The researchers demonstrated that the attack is feasible by acquiring the wireless communication hardware for a few hundred dollars and computing a rainbow table (which requires 5.4TB of disk space in total). These are acceptable efforts for being able to steal Tesla Model S cars. The attack consists of 4 steps, each taking only a few seconds to accomplish:

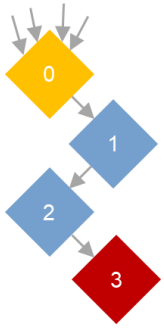


Figure 2: The Attack path.

- *Phase 0*: Adversary acquires the car identifier wirelessly.
- *Phase 1*: Adversary does two constructed challenge-response requests with the key fob.
- *Phase 2*: The adversary recovers the key with the help of a precomputed rainbow table.
- *Phase 3*: With the recovered key, the adversary unlocks the car, starts it, and drives away.

The experts got a predefined sheet with each phase as a step on the attack path (see Figure 2). They had to fill in their estimated values in the form of [minimum-most-likely-maximum] (see Figure 3) for the attack frequency, the vulnerability, and the impact.

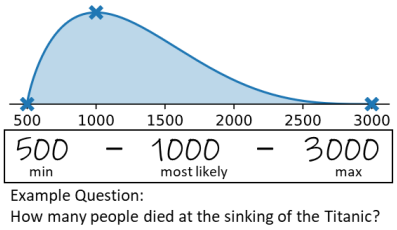


Figure 3: An example three-point estimate [min-mode-max] visualized with the PERT-distribution.

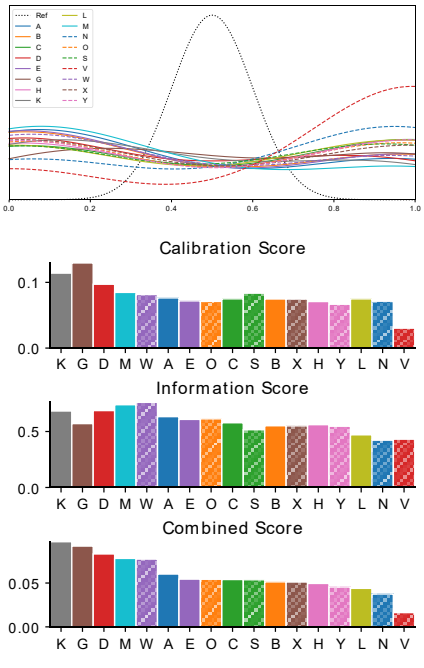
With this knowledge the experts should judge the risk of cyber-security for the following scenario: Imagine you are the CISO (chief information security officer) of a car rental company, having a fleet of 100 Tesla Model S cars in the beautiful city of Leuven (the town of the COSIC research group). **What are the risks concerning the previously mentioned attack?**

Results

First, the calibration results are discussed, and then the actual survey results of the two assessment rounds.

Calibration Results

The calibration results showed a disillusioning image of the experts' performance. The figure on the right shows the calibration distributions for the experts compared to a reference model of the expected outcome. Not a single expert came near the expected results. While this was a staggering finding of the survey, it was no surprise since Colson and Cooke [7] found out that this happens quite often. They looked at 33 independent studies in many domains that applied structured expert judgment, and they found out that in approximately one-third of all studies, less than two good experts were present. In 20% of the studies, not even a single good one took part, which is highly alarming and highlights the need for calibration.



The final weights of the experts had approximately the same magnitude, which means that they performed about equally during the calibration. A good expert would have got more weight than the others by multiple magnitudes.

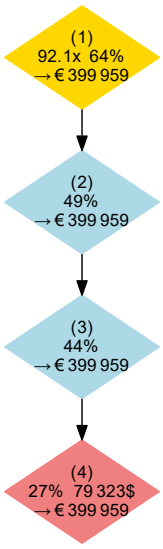
K	G	D	M	W	A	E	O	C	S	B	X	H	Y	L	N	V
10%	9%	8%	8%	8%	6%	5%	5%	5%	5%	5%	5%	5%	5%	4%	4%	2%

Table 1: The final weights for the experts, based on their calibration and information score.

Survey Results

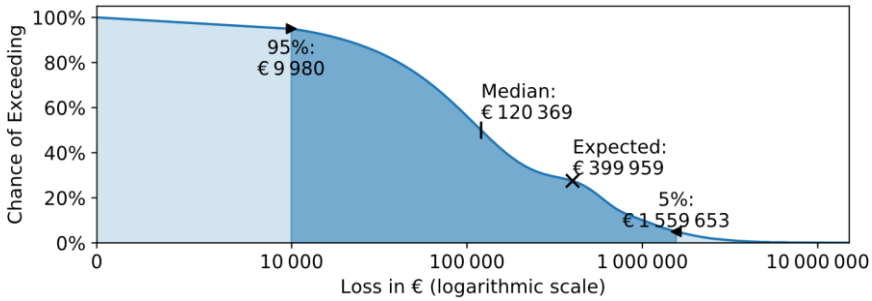
The results were evaluated using RISKEE [8], a tool for propagating and calculating uncertain risk values throughout a graph, which was also presented the first time at the EuroSPI 2019 conference [1].

Round 1



In the first round, the experts did not have the chance to discuss the scenario with each other. They judged the values independently and without external input, and the results are as follows. The average attack frequency was estimated to be about 92 times per year, and the average impact was about € 400 000. With a 95% probability, the loss will exceed about € 10 000, and with 5%, it exceeds € 1 560 000. The value with a 50% probability of exceeding is approximately € 120 000, which would correspond to one stolen car in our fleet of 100 cars.

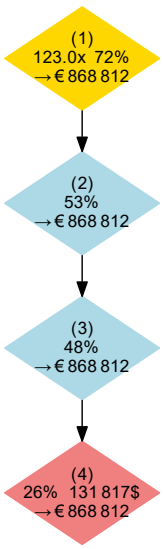
In summary, we can expect the risk to be between € 10 000 and € 1 500 000, and we should prepare for one stolen car per year.



Discussion

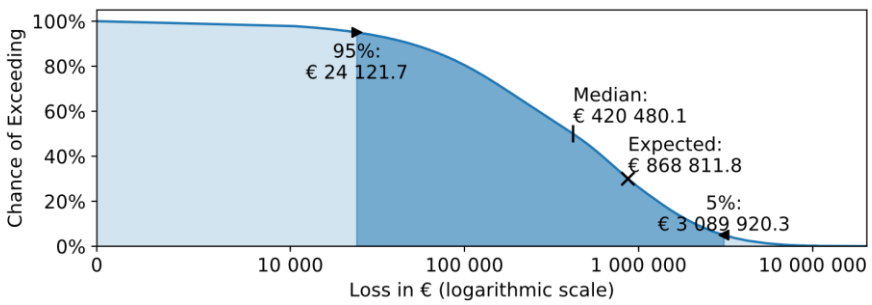
During the discussion, many interesting questions came up. How difficult is the attack? How many people know the details of the attack? How likely is it that attackers can derive the details of the attack from the published materials by the COSIC researchers? What knowledge do the attackers need to calculate the rainbow table? Where are the cars parked? How often are the cars rented?
In hindsight, these were very informative because we could pin down many uncertainties and influence factors which we should investigate if we were doing a real risk assessment.

Round 2



In round two, the experts in general increased all their values, beginning with the attack frequencies as well as the vulnerabilities and the impact. It seems that discussion made them more suspicious and careful, which manifested in the nearly doubled risk values. The average attack frequency was now 123 times, while the average impact was around € 869 000. Also, the confidence interval doubled by being between € 24 000 and a whopping € 3 000 000.

As management, we would have to calculate with a median loss of € 420 000, which corresponds to about four stolen cars per year. This is quite an increase compared to the first round. Also, the 5% exceedance value of € 3 Million is quite worrying, because it would correspond to about 1/3 of the whole car fleet stolen per year. Fortunately, it has a very low probability.



Discussion and Conclusion

The study shows that it is beneficial to assess the quality of experts via calibration and combine their judgments based on that. Quality can be evaluated based on information and calibration score. Information tells us how precise or uncertain a given judgment is, compared to all other judgments. Calibration tells us how well the prediction captured the true value.

In this study, the overall calibration scores were rather low, which indicates that we, unfortunately, cannot put much trust in the results. Without calibration, we would have never known this. One reason for the low calibration scores could be that we just used random volunteers to participate in the study, and not preselected experts. We think that using preselected experts with distinct backgrounds from the following three fields would have led to better results: Firstly, experts with domain knowledge (the customers who know the problems in practice), experts with a technical background (technicians and engineers who have detailed technical knowledge), and generally knowledgeable supervisor (the consultants, who have experience, can ensure the quality, and translate between the others to avoid misunderstandings). Such diverse groups would have added different viewpoints and relevant information to the responses. To support this even further, future research about the ideal composition of expert councils must be done to get more data.

Another interesting finding of the study was that between the first and the second round of assessment, the overall risk more than doubled while the individual responses by the experts just increased slightly. It seems that the discussion ignited concerns in the participants, which in turn led to increasing the adjustment of their values. These adjustments multiplied up and resulted in doubling the total risk values.

Bibliography

[1] A. Walker, R. V. O'Connor, and R. Messnarz, Systems, Software and Services Process Improvement: 26th European Conference, EuroSPI 2019, Edinburgh, UK, September 18-20, 2019, Proceedings. Springer International Publishing, 2019.

[2] V. Hemming, T. V. Walshe, A. M. Hanea, F. Fidler, and M. A. Burgman, 'Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management', PLoS ONE, vol. 13, no. 6, p. e0198468, Jun. 2018, doi: 10.1371/journal.pone.0198468.

[3] R. M. Cooke, Experts in uncertainty: opinion and subjective probability in science. New York: Oxford University Press, 1991.

[4] A. R. Colson and R. M. Cooke, 'Expert Elicitation: Using the Classical Model to Validate Experts' Judgments', Review of Environmental Economics and Policy, vol. 12, no. 1, pp. 113–132, Feb. 2018, doi: 10.1093/reep/rex022.

[5] V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle, 'A practical guide to structured expert elicitation using the IDEA protocol', Methods Ecol Evol, vol. 9, no. 1, pp. 169–180, Jan. 2018, doi: 10.1111/2041-210X.12857.

[6] L. Wouters, E. Marin, T. Ashur, B. Gierlichs, and B. Preneel, 'Fast, Furious and Insecure: Passive Keyless Entry and Start Systems in Modern Supercars', IACR Transactions on Cryptographic Hardware and Embedded Systems, vol. 2019, no. 3, pp. 66–85, May 2019, doi: 10.13154/tches.v2019.i3.66-85.

[7] A. R. Colson and R. M. Cooke, 'Cross validation for the classical model of structured expert judgment', Reliability Engineering & System Safety, vol. 163, pp. 109–120, Jul. 2017, doi: 10.1016/j.ress.2017.02.003.

[8] M. Krisper, J. Dobaj, G. Macher, and C. Schmittner, 'RISKEE: A Risk-Tree Based Method for Assessing Risk in Cyber Security', in Systems, Software and Services Process Improvement, Cham, 2019, vol. 1060, pp. 45–56, doi: 10.1007/978-3-030-28005-5_4.