Solving the Allais Paradox by Counterfactual Harm

Marco Zaffalon MARCO.ZAFFALON@IDSIA.CH
Alessandro Antonucci ALESSANDRO.ANTONUCCI@IDSIA.CH
Oleg Szehr OLEG.SZEHR@IDSIA.CH
IDSIA, Switzerland

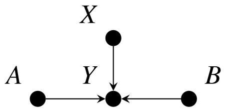

Figure 1: A causal graph.

The so-called *Allais paradox* is a classical choice problem designed to challenge the supposed rationality of expected utility theory [1]. We formulate the paradox with the causal graph in Fig. 1. A Boolean variable B is used to distinguish two experiments, each of which presents a choice between two gambles, indexed by a Boolean variable A. The state of nature is described by a variable X taking values in $\{0, 1, 2\}$ with $P(X) = [0.89, 0.01, 0.10]$. Gambles are represented via variable $Y := f(X, A, B)$, whose possible values are $\{0, 1, 5\}$ million dollars, as determined by the structural equations $f(X, A = 0, B = 0) = [1, 1, 1]$, $f(X, A = 1, B = 0) = [1, 0, 5]$, $f(X, A = 0, B = 1) = [0, 1, 1]$ and $f(X, A = 1, B = 1) = [0, 0, 5]$. In the first experiment ($B = 0$), people typically prefer $A = 0$ over $A = 1$, which means that earning one million dollars for sure is preferred to a lottery that pays nothing with probability 0.01, one million with probability 0.89 and five millions with probability 0.10. The opposite happens in the second experiment ($B = 1$), where the winning of five millions with probability 0.10 for $A = 1$ is preferred to the winning of one million with probability 0.11 for $A = 0$. Given a utility function u, the preference of $A = 0$ over $A = 1$ for $B = 0$ corresponds to $\mathbb{E}[u(A = 0|B = 0)] > \mathbb{E}[u(A = 1|B = 0)]$, which implies $0.11u(Y = 1) > 0.01u(Y = 0) + 0.1u(Y = 5)$. Vice versa, $\mathbb{E}[u(A = 0|B = 1)] < \mathbb{E}[u(A = 1|B = 1)]$ implies $0.11u(Y = 1) < 0.01u(Y = 0) + 0.1u(Y = 5)$. Thus, no utility function is compatible with the preference of $A = 0$ over $A = 1$ in $B = 0$ and, simultaneously, of $A = 1$ over $A = 0$ in $B = 1$.

We argue that the incompatibility above can be resolved by reasoning counterfactually, following the approach recently proposed by Richens et al. in [2]. To see how, we need to compute the expected *harm* (i.e. utility drop) obtained by comparing the factual utility received for a given choice with the utility one would counterfactually receive for the alternative choice. Taking as alternative $A = 1$, the counterfactual harm h caused by $A = 0$ for the winning $Y = y$ in experiment $B = b$ is:

$$h(A = 0, Y = y|B = b) = \textstyle\sum_{y'=0,1,5} P(y'_{A=1}|Y = y, A = 0, B = b) \max\{0, u(Y = y') - u(Y = y)\} ,$$

where the standard counterfactual notation is used to denote interventions as subscripts. By standard computations in the causal model, we obtain $h(A = 0, Y = 5|B = 0) = 0.1\,(u(Y = 5) - u(Y = 1))$ and, since no other outcome produces harm, this is also the expected harm $\mathbb{E}[h(A = 0|B = 0)]$. Similarly, we obtain $\mathbb{E}[h(A = 1|B = 0)] = (u(Y = 1) - u(Y = 0))$, $\mathbb{E}[h(A = 0|B = 1)] = 0.0\overline{1}\,(u(Y = 1) - u(Y = 0))$ and $\mathbb{E}[h(A = 1|B = 1)] = 0.\overline{90}\,(u(Y = 1) - u(Y = 0))$. Taking a linear utility $u(Y) = Y$, for example, we get $\mathbb{E}[h(A = 0|B = 0)] = 1.0 > \mathbb{E}[h(A = 1|B = 0)] = 0.4$ and $\mathbb{E}[h(A = 1|B = 1)] = 3.\overline{63} > \mathbb{E}[h(A = 0|B = 1)] = 0.0\overline{1}$, which solves the paradox. This indicates that if people were to reason counterfactually, there would be no paradox at all. Causal queries such as those considered by the above notion of counterfactual harm might suffer from *partial identifiability* issues: this means that, unlike the case in our example, a precise computation of the query is not possible, and the model specification only allows to compute bounds. The work by Zaffalon et al. in [3] clearly formalises this by providing a mapping between causal models and credal networks. As an example, we consider the unconditional harm. As $P(B)$ is unavailable, we describe it as a vacuous model yielding bounds to the expectations. The computation for this case gives overlapping intervals, i.e., $0.0\overline{1} \leq \mathbb{E}[h(A = 0)] \leq 1.00$ and $0.40 \leq \mathbb{E}[h(A = 1)] \leq 3.\overline{63}$. Although such an overlap might reflect a condition of *indecision* between the options, the decision criteria derived within the imprecise probability community (e.g., *maximality* or E-admissibility) could be considered in order to reduce the indecision.

References

[1] Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the econometric society*, 21(4):503–546, 1953. doi:10.2307/1907921.

[2] Jonathan Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm. In *Proceedings of NeurIPS*, 2022.

[3] Marco Zaffalon, Alessandro Antonucci, and Rafael Cabañas. Structural causal models are (solvable by) credal networks. In *Proceedings of PGM*, 2020.