Solving the Allais Paradox by Counterfactual Harm

Marco Zaffalon, Alessandro Antonucci, Oleg Szehr

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) - Lugano (Switzerland)

{marco.zaffalon,alessandro.antonucci,oleg.szher}@idsia.ch

13th International Symposium on Imprecise Probabilities: Theories and Applications

11-14 July, 2023, Oviedo (Spain)

ALLAIS PARADOX (1953)

The paradox is a classical choice problem designed to challenge the supposed rationality of expected utility theory. Two experiments, each involving a choice between two gambles, are considered.

- In the first experiment, it is noticed that **a sure 1M$ reward** is generally **preferred** to a gamble having a 1% chance of zero reward, even if there is a 10% chance of 5M$ and 89% chance remains for 1M$. In terms of expected utility, this tells us that, for most people, $u(1) > 0.89\,u(1) + 0.10\,u(5)$
- In the second experiment, a **1M$ reward with an 11% chance** is generally **NOT preferred to a 5M$ reward with 10% chance**. Thus, $0.11\,u(1) < 0.10\,u(5)$, which is incompatible with the first choice!

First Experiment				Second Experiment			
First gamble ($A=0$)		Second gamble ($A=1$)		First gamble ($A=0$)		Second gamble ($A=1$)	
reward	chance	reward	chance	reward	chance	reward	chance
1M$	100%	1M$	89%	0	89%	0M$	90%
		0	1%	1M$	11%		
		5M$	10%			5M$	10%

$(A=0) \succ (A=1)$ $\quad$ $(A=1) \succ (A=0)$

COUNTERFACTUAL HARM (2022)

Action $A=a$ gives consequence $Y=y$ with a utility function U depending on a (possibly uncertain) context $X=x$

Expected Utility (EU) supports $a^* \coloneqq \arg\max_a E[U|a,x]$

with $E[U|a,x] \coloneqq \int_y P(y|a,x)U(a,x,y)$

EU does not directly take into account the other actions' consequences.

The **(counterfactual) harm** (wrt an alternative action a') is instead:

$$h(a,x,y) \coloneqq \int_{y'} P(Y_{a'}=y'|a,x,y)\max\{0, U(a,x,y)-U(a',x,y')\}$$

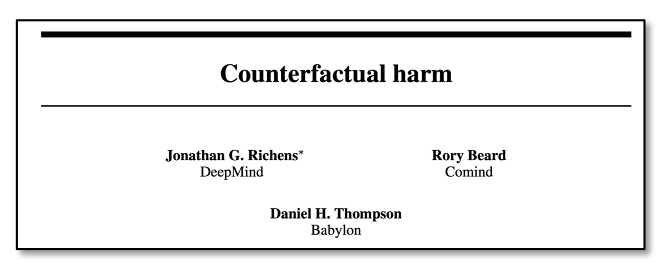
(Non-negative) utility losses are weighted by a probability P mixing the factual (a,y) and counterfactual (a',y') worlds.

A **structural causal model** is needed to compute $P(Y_{a'}=y'|a,x,y)$!

Harm-averse decision-making by harm-penalised utility:

$$V(a,x,y) \coloneqq U(a,x,y) - \lambda\, h(a,x,y)$$

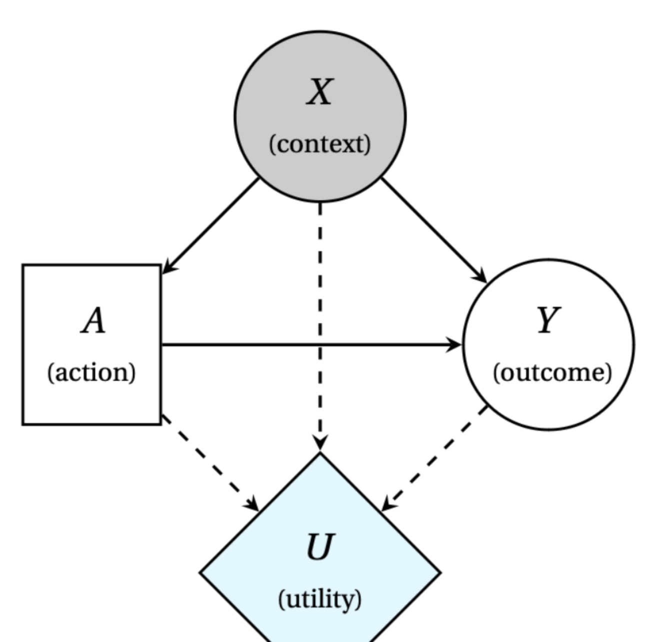
with harm-aversion coefficient $\lambda > 0$


Counterfactual harm

Jonathan G. Richens* DeepMind

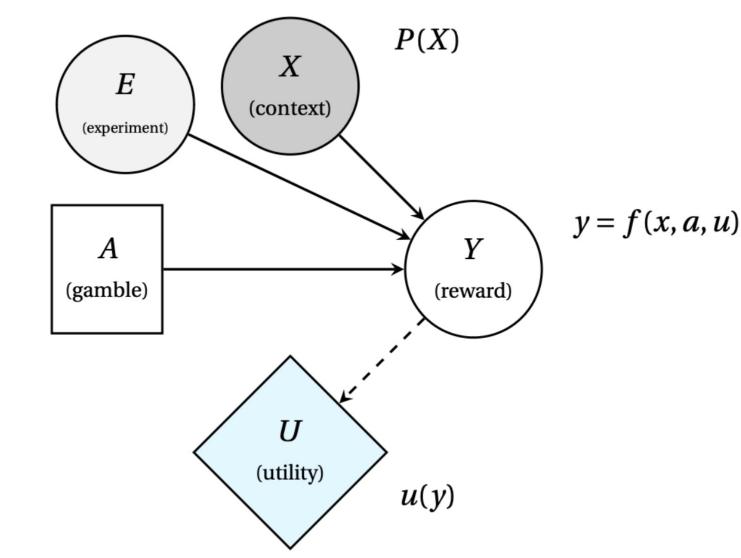
Rory Beard Cumind

Daniel H. Thompson Babylon



ALLAIS CHOICE AS A CAUSAL MODEL (OUR WORK)

- Boolean variables E and A to distinguish the two experiments and gambles
- Context X as a ternary state with chances $P(X=[0,1,2]) = [0.89, 0.01, 0.10]$
- Reward by a structural equation $y = f(a,x,e)$
- Utility U is only determined by the reward $(u(0), u(1), u(5))$



Separately for each experiment, choice between the two gambles ($A=0$ versus $A=1$) described in terms of harm-penalised utility. Let us compute the counterfactual harm by already summing out the context

Reward	$f(a,x,e)$			
Experiment	$E=0$		$E=1$	
Gamble	$A=0$	$A=1$	$A=0$	$A=1$
$X=0$	1	1	0	0
$X=1$	1	0	1	0
$X=2$	1	5	1	5

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

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

The counterfactual probability should be performed in the **twin network** of the structural model with the two worlds duplicated.

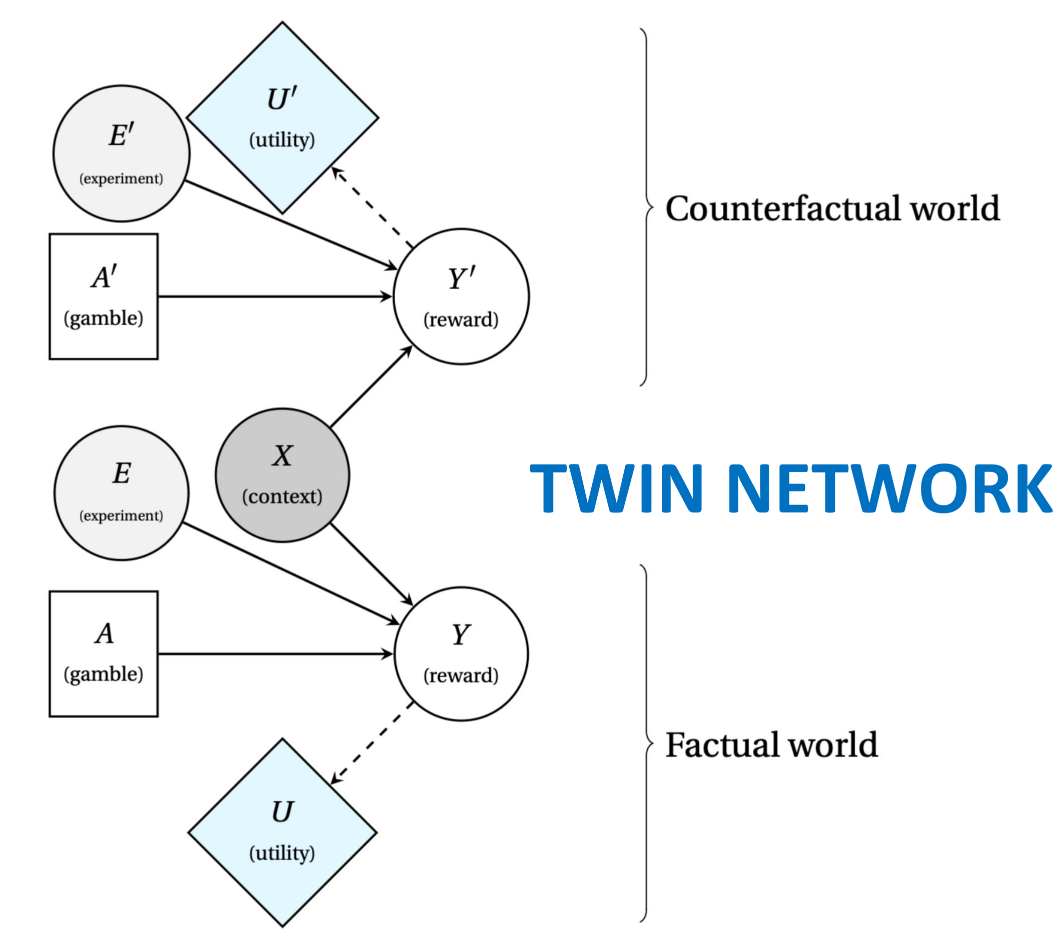
Taking a linear utility (e.g., $u(y)=y$) we get:

$E[h(A=0|E=0)] = 1 > E[h(A=1|E=0)] = 0.4$,

$E[h(A=0|E=1)] = 0.0\bar{1} < E[h(A=1|E=1)] = 3.6\bar{3}$.

If people were to reason counterfactually, there would be no paradox at all.

TWIN NETWORK



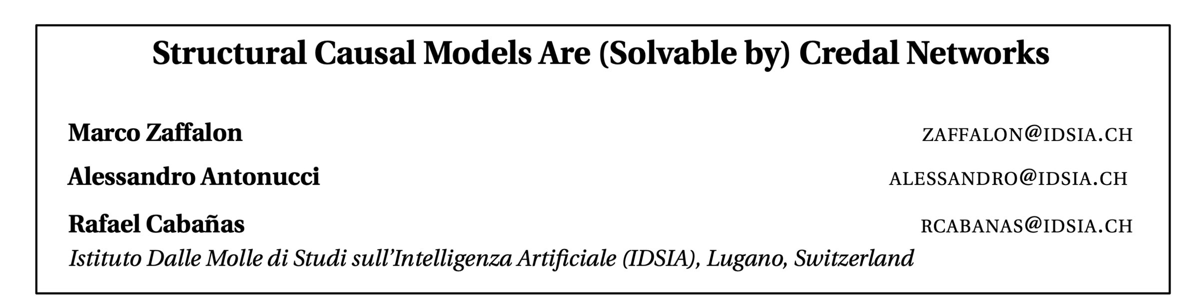
COUNTERFACTUALS ARE IMPRECISE PROBABILISTIC QUERIES (2020)

Causal queries such as those considered by 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.

Solution? A mapping between causal models and credal networks!

E.g., unconditional harm (with a vacuous model over E) gives overlapping intervals, i.e.,

$0.01 \leq \mathbb{E}[h(A=0)] \leq 1.00$ and $0.40 \leq \mathbb{E}[h(A=1)] \leq 3.63$.


Structural Causal Models Are (Solvable by) Credal Networks

Marco Zaffalon ZAFFALON@IDSIA.CH

Alessandro Antonucci ALESSANDRO@IDSIA.CH

Rafael Cabañas RCABANAS@IDSIA.CH

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland

Credici — Credal Inference for Causal Inference

Library for counterfactuals by credal nets and EM

github.com/Idsia/credici