Causality, Harm, and Elections

Jérôme Lang
LAMSADE, CNRS, Université Paris-Dauphine PSL

May 6, 2026

Joe Halpern and I started about causality and harm in voting contexts first in July 2025 in Düsseldorf, and longer in December 2025 in Paris, where Joe had been invited to give a seminar, which was *excellent*. We talked about votiing and causality for several hours and he promised to send me his latest draft [7], which he did on December 28.

For a few years, Joe had been actively working on modelling *harm*. He and his coauthors came up with three successive definitions, which I recall informally.

In [2, 4], harm is defined from two main ingredients: contrastive causation (an outcome being caused rather than another outcome) and an agent's default utility. *An event $\vec{X} = \vec{x}$ harms an agent* (in a causal model and a context) if there exists another event $\vec{X} = \vec{x'}$ and two outcomes o, o' such that

1. $\vec{X} = \vec{x}$ rather than $\vec{X} = \vec{x'}$ causes o rather than o';
2. o' gives the agent a higher utility than o;
3. o gives the agent a utility lower than their default utility.

More informally: an event causes harms to an agent in a model if there is some outcome o such that (1) the event caused o; (2) o gives the agent a utility lower than their default utility; (3) a different event would have resulted in an outcome with higher utility than o.

The definition is quite flexible given that the way the default utility is defined depends very much on the context; in particular, it can be defined on statistical grounds or on prescriptive grounds.

This notion is made quantitative in [3]: in a deterministic setting, the amount of harm caused by an event is the maximum difference between (a) the default utility or the utility of the contrastive outcome o', whichever is lowed, and (b) the utility of the actual outcome (except when this difference is negative or if the event did not cause the actual outcome, in which case the amount of harm is 0).

A crucial question is where the default utility comes from. In [7], the default utility derives from a normality ordering on worlds: it can be taken to be *the agent's utility in their least preferred worlds among the most normal ones.*

Therefore, the agent is harmed if the outcome o is worse than the utility she would have obtain in the worst of all most normal worlds.

It was known for long that normality plays an important role in determining causality [8, 6], bearing on events rather than on worlds: a more abnormal event is typically viewed as a "better" (a more plausible) cause. What [7] argue is that it can be used *also* for determining harm. Normality can be interpreted either statistically or prescriptively, or sometimes both at the same time [6].

Joe Halpern was of course well aware that causation and harm applies to elections and more generally to social choice. Causation (and responsibility and blame) in voting contexts had been discussed more than twenty years ago, cf. examples in [5]. The ideas I list below (informally) are based on the discussions we had with Joe in July and December.

Let C be a finite set of potential candidates (who may decide to run or not). Given $C' \subseteq C$, a *(voting) profile* over C' is a collection of votes over a subset of potential candidates (whose format can vary: rankings, subsets etc.). A voting profile is just a voting profile over C' for an arbitrary C'. A (single-winner, resolute) voting rule f maps each profile P to a winning candidate $f(P)$. The definition of voting rules is general enough to accommodate almost all known single-winner electoral systems, including those based on several 'rounds' such as plurality with runoff. An election consists of a voting rule and a profile.

In voting, causation can appear at several levels, at least these two:

- candidacy: an intervention consists of a candidate deciding to run or not to run (strategic candidacy). We may also consider candidates choosing their campaigning strategy, e.g., their political platform.
- votes: a voter may decide to vote or to abstain; and if they vote they can consider deviating fro their sincere preferences (strategic voting).

For the sake of time I'll focus on candidacy. The structural equations model the voting rule, that is, the functional link between the input profile and the winning candidate. The exogenous variables can correspond to the running candidates and/or the votes, depending on what we want to focus on. Rather than delving into a formalisation, I'll simply give examples.

Example 1 (USA 2000: Bush, Gore, and Nader)**.** In the famous 2000 US President election, Bush won by 271 electoral votes against 266 for Gore (and won although Gore won the national popular vote). Nader got a 2.88 million votes (2.7 %); if he had not run, most of his votes (though not all) would probably have gone to Gore, and many political analysts believe it would have been more than enough for Gore to win Florida (Nader received about 97,000 votes, and the margin was only 537), whose 25 electoral votes decided the election, and therefore to win the overall election. This is however not so simple, because while we know for sure the votes that were cast[1], we have no way to know for

[1]Even if they had to be recounted in Florida...

sure what would have been the outcome if Nader had not run. Political scientists typically estimated that Gore would probably have won, *but there is no certainty*, for several reasons: if Nader had not run, some Nader voters would have abstained, campaign dynamics would have changed, turnout and media attention might have evolved differently. Therefore, analysing causation on a real-life election involves comparing a fully known scenario with a counterfactual scenario for which we have only beliefs (in other words, the structural equations would have to accommodate beliefs and uncertainty, and not be functional).

Anyway: if we take for granted that Gore would have won if Nader hadn't run, then Nader caused Bush to win (B) for various notions of causality. For the normality-based version of causality, it is still the case if we consider Nader running as less normal than him not running, but not if we consider it as most normal that he runs (the argument being that he had run already in 1996). Now, did Nader's decision to run harm anyone, and whom? If we take the default utility to be that of the "Gore winning" (G) outcome, then yes, it harmed Gore, and Gore's supporters, but more interestingly, it harmed most of his own supporters.[2] This is what we may call *backfire harm*.

If we don't take if for granted that Gore would have won if Nader hadn't run, then this is more complex. There are two most normal outcomes: G and B. The default utility (let's say, for Gore's voters), if we follow [7], is the minimum of the utilities of these two outcomes, that is, $u(B)$, so there is no harm. This is not what we expect: if Nader runs, the nondeterministic outcome set is $\{B, G\}$, which under any reasonable set extension is preferred to $\{B\}$. A recent paper directly addressing nondeterministic causality is [1] and the way to handle this is probably to be found there (I haven't had the time to read it yet).

Example 2 (France 2002: Chirac, Jospin, and Le Pen)**.** The French presidential election uses a two-round system (plurality with runoff). In the 2002 edition, it was expected that the finalists would be the incubent president Jacques Chirac and the socialist candidate Lionel Jospin. This first-round outcome was considered normal because this is what polls predicted, but also because in all previous elections since 1974, the finalists were a socialist candidate and a right-wing candidate. The outcome of the second round between Chirac and Jospin was very uncertain: both Chirac winning (C) and Jospin winning (J) were considered as normal. Now, the finalists ended up to be Chirac and the extreme-right candidate Jean-Marie Le Pen (with 16.86 %), and Jospin came third with 16.18 %. (Chirac won the second round with a huge majority.) Among the remaining candidates, quite a lot came from the left or the center, and stole some votes from Jospin, including Bayrou (center, 6.84 %), Chevènement (Eurosceptic left, 5.33 %), Mamère (ecologist, 5.25 %), Hue (communist, 3.37%), and Taubira (2.32 %). It is thought that if *any* of these had decided not to run, Jospin would have been to the second round. If Le Pen had decided not to run, the other

[2]But probably not all (and probably not Nader himself), even among those who wold ave voted for Gore if Nader had not run, because they may derive a positive utility in seeing Nader's arguments deserving more attention in the campaign than usual, and this utility gain could, for some of them, be larger than the utility difference between G and B.

extreme-right candidate (Mégret) would not have made it to the second round. What caused the left wing to be absent from the second round $\overline{L}$)? When asked, most peopled at the time (and even now) named Christiane Taubira's candidacy (T) a *the main cause*. An old-style counterfactual definition would have indeed said that T was a cause, but also Mamère, Bayrou, Chevènement, Hue... and Le Pen. Normality helps. Mamère was expected to run: a world where no ecologist candidate would have been quite abnormal (so that if he had not run, another ecologist wold have); this also applies to Bayrou and Hue, and of course to Le Pen too, whose absence would have been highly abnormal. T was viewed as a quite abnormal event: Taubira belonged to a small party, very close to the Socialist party, which usually does not support a candidate because they make an alliance with the Socialist Party; her program was quite close from Jospin's and pleased the very same voters. Note that I said that T caused $\overline{L}$. I did not, and should not, say that T caused C, because if the second round had been between Chirac and Jospin, nobody knows who would have won, so here we wold have the nondeterministic outcome $\{J, C\}$, which anyway is preferred to C by left-wing voters.

Whom did T harm? This again depends on agents' preferences and their default utility. We can consider a more complex outcome as composed of the two finalists, and the winner. In the actual world, this is $((C, LP), C)$. If we take the most normal complex outcomes to be $(C, J), C)$ and $(C, J), J)$, and take the default utility of supporters of Jospin (and Taubira) to be their minimal utility between these two complex outcomes (which as I said above is not entirely satisfactory) then it is $u((C, J), C)$, which is certainly much lower than $u((C, LP), C)$, even though thee final winner is the same in both. So, yes, even with this (perhaps too strong) notion, T *caused backfire harm*. If we are talking about quantitative harm, (most) voters of Jospin and Taubira weer certainly more harmed than, for instance, those of Bayrou's or Chevènement, because for (most of) those the utility difference was certainly smaller.

Example 3 (France 2017: Fillon, Macron, Le Pen[3])**.** Fillon was the right-wing candidate. A few months before the election the normal world was the one where he wins (F). However, a few weeks before the first round, a scandal broke out after revelations that his wife had been paid for a no-show job. He was asked to withdraw in favor of another right-wing candidate (Juppé, who had lost the primary against Fillon) but he refused. He did not make it to the second round (because of the scandal), which took place between Emmanuel Macron and Marine Le Pen, and was won by Macron. What harmed whom and how much? Was there any backfire harm? You can easily find yourself. (There are many more interesting causations and harm to discuss about this election but I lack time and space.)

[3]Note the same Le Pen as in 2002: his daughter.

References

[1] Sander Beckers. Nondeterministic causal models. In Biwei Huang and Mathias Drton, editors, *Causal Learning and Reasoning, Lausanne, Switzerland, 7-9 May 2025*, Proceedings of Machine Learning Research, pages 1532–1554. PMLR, 2025.

[2] Sander Beckers, Hana Chockler, and Joseph Y. Halpern. A causal analysis of harm. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[3] Sander Beckers, Hana Chockler, and Joseph Y. Halpern. Quantifying harm. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 363–371. ijcai.org, 2023.

[4] Sander Beckers, Hana Chockler, and Joseph Y. Halpern. A causal analysis of harm. *Minds Mach.*, 34(3):34, 2024.

[5] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Intell. Res.*, 22:93–115, 2004.

[6] Joseph Y. Halpern and Christopher Hitchcock. Graded causation and defaults. *CoRR*, abs/1309.1226, 2013.

[7] Joseph Y. Halpern and Spencer van Koevering. Causality, normality, and harm, 2025. Manuscript.

[8] Daniel Kahneman and Dale T. Miller. Norm theory: Comparing reality to its alternatives. *Psychological Review*, 2(93):136–153, 1986.