COUNTERFACTUALS IN CRITICAL THINKING,

WITH APPLICATION TO MORALITY

Luís Moniz Pereira and **Ari Saptawijaya**

NOVA LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS (NOVA LINCS)

UNIVERSIDADE NOVA DE LISBOA, PORTUGAL

ABSTRACT

Counterfactuals are conjectures about what would have happened, had an alternative event occurred. "Counterfactual" means contrary to the facts. Counterfactual reasoning involves thoughts on what might have been, what could have happened, had some matter –action, outcome, etc.– been different in the past. It provides lessons for the future by virtue of contemplating alternatives; it permits thought debugging; it supports a justification why different alternatives would have been worse or not better. Typical expressions are: "If only I were taller…", "I could have been a winner…", "I would have passed, were it not for…", "Even if… the same would follow".

Counterfactual thinking covers everyday experiences, like regret: "If only I had told her I love her!", "I should have studied harder"; or guilt responsibility, blame, causation: "If only I had said something sooner, then I could have prevented the accident". The general form is: "If would antecedent, then would consequent".

Are counterfactuals mere fantasies? A waste of time? Not really. They have been well studied in Linguistics, Philosophy, Physics, Ethics, Psychology, Anthropology, and Computation (Collins et al. 2004, Hoerl et al. 2011, Lewis 1973, Pearl 2000, Roese et al. 2009), but not much within Critical Thinking. However, people often think how things that matter to them might have turned out differently (Mandel et al. 2005). Researchers from psychology have asked: Why do people have such a strong tendency to generate counterfactuals? What functions does counterfactual thinking serve? What are the determinants of counterfactual thinking? What are its adaptive and psychological consequences? Human's ability for mental time travel relies on episodic memory. Without it humans would be unable to form a stable concept of self along time, and human cultures would not have been able to evolve.

We believe counterfactual thinking is worth more attention as a means to critical thinking. We illustrate this with an original application to morality, a common concern for critical thinking.

Keywords: Critical Thinking, Counterfactual Reasoning, Abduction, and Morality.

COUNTERFACTUAL REASONING

To enact it, we use our original 3-step logic evaluation procedure for evaluating counterfactuals: (1) ***Abduction***: explain past circumstances in the presence of evidence, i.e. update and fix the circumscribing background information given the evidence. (2) ***Action***: adjust the logical causal model to comply with the antecedent of the counterfactual, i.e. impose the antecedent's hypothesis with an intervention on the model. (3) ***Prediction***: predict if the consequent deductively follows after steps 1 and 2, i.e. compute the truth-value of consequent in the modified intervened model.

We do so using an abductive logic framework (ALP), a triple **<**P, A, I**>** of logic program P, abducibles A, and integrity constraints I. Abduction chooses from available hypotheses –the exogenous variables (Pearl 2000) that constitute the situation's background– those abducibles, or their negation, that best explain the observed evidence **O**. An abduced explanation **E** ⊆ A

mirrors the specific exogenous variables supporting an explanation to the currently observed evidence.

Updating allows the rules to be updated by asserting or retracting abducibles. It represents changes and deals with incomplete information. Updating fixes the initially abduced background context of the counterfactual statement, i.e. updates the rules with a preferred explanation to current observations. Updating also permits causal intervention on the causal model, i.e. with hypothetical updates to the rules, so as to comply with the antecedent of the counterfactual.

An example (Byrne 2007): *Lightning hits a forest, and a devastating forest fire breaks out. The forest was dry, after a long hot summer.* Let's add more causes for forest fire. Two possible alternative causes: storm –implying the lightning– or barbecue. Starred atoms stand for their negations. ALP framework <P, A, I>: A = {storm, barbecue, storm*, barbecue*}, I= {}, logic program P:

fire ← barbecue, dry_leaves. dry_leaves.
fire ← barbecue*, lightning, dry_leaves.
lightning ← storm.

Take counterfactual statement: *If only there had not been lightning, then the forest fire would not have occurred.*

Step 1: Abduce explanations E ⊆A to the two above factual observations: O={lightning, fire}. The observations assure us that both the antecedent and the consequent literals of the counterfactual were *factually* false. Two possible explanations for O (s and b stand for storm and barbecue, resp.): E1={s, b*} and E2={s, b}. Say E1 is preferred for consideration. Then fix that abduced background context for the counterfactual: i.e. update program P with E1.

Step 2: Update program P, to get a new program T, by adding:

make(lightning*). % Intervention: *If only there had not been lightning...*
lightning* ← make(lightning*). % Note that lightning or otherwise now
lightning ← make(lightning). % are available only by intervention.

And, for irrelevancy, deleting: lightning ← storm

Step 3: Prediction tests validity of counterfactual. Verify if the conclusion "*the forest fire would not have occurred*" holds in the semantics, and that I is satisfied. Indeed, 'not fire' holds in T for explanation E1={s, b*} and intervention make(lightning*). The counterfactual is valid.

COUNTERFACTUALS IN MORALITY

Typically, people think critically about what they should or should not have done when they examine decisions in moral situations. It is therefore natural for them to engage in counterfactual thoughts of alternatives in such settings. Counterfactual thinking has been investigated in the context of moral reasoning, notably by psychology experimental studies (Byrne 2007), e.g., to understand the kind of critical counterfactual alternatives people tend to think of in contemplating moral behaviours, and the influence of counterfactual thoughts in moral judgment (Mandel et al. 2005, Roese et al. 2009).

We detail here, to illustrate our general counterfactual method, the issue of moral permissibility from the viewpoint of the Doctrine of Double Effect (DDE). DDE permits actions that cause harm, but bring about a good result, by distinguishing harm as *a mere side effect* versus harm as *a means to*. DDE explains the consistency of judgments, as shared by subjects from diverse populations, on a series of moral dilemmas, via empirical psychological studies (Hauser et al. 2007). DDE formulation: "If some morally wrong effect E is a cause of goal G (which we want to achieve by performing action A), and E is not a mere side-effect of A, then performing A is impermissible." For a critical thinking DDE evaluation in this setting, create a counterfactual to test if E is essential for G: *If not E would have been true, then not G would have been true.*

APPLYING DDE TO MILITARY MORALITY CASES. Consider "Terror Bombing" (Scanlon 2008), or **teb**, for short, which means: Bombing a civilian target during a war, thus killing many civilians, in order to terrorize the enemy, and thereby getting them to end the war. DDE affirms **teb** impermissible. Tactical bombing (**tab**) means: Bombing a military target, which will effectively end the war, but with foreseen consequence of killing the same large number of civilians nearby. DDE affirms **tab** permissible.

MODELLING TERROR BOMBING. Framework ⟨P, A, I⟩, with A={teb, teb*}, I={}, and P:

end_war ← terror_civilians. terror_civilians ← kill_civilians.
kill_civilians ← target_civilians. target_civilians ← teb.

Counterfactual: *If civilians had not been killed, war would not have ended*. Evaluation follows.
Step 1: Observations O = {kill_civilians, end_war} with explanation E = {teb}.
Step 2, produce T, as above**:**

make(kill_civilians*). % Intervention: *If civilians had not been killed...*
kill_civilians * ← make(kill_civilians*). % Killing civilians or otherwise now
kill_civilians ← make(kill_civilians). % available only by intervention.

And, for irrelevancy, deleting: kill_civilians ← target_civilians
Step 3: Counterfactual is valid since conclusion *war would not have ended* is true. Indeed, 'not end_war' holds in the semantics of updated T plus E. Hence, the morally wrong action 'kill_civilians' is an instrument to achieve goal 'end_war.' It is a cause of 'end_war' by performing **teb**, and not a mere side effect of **teb**. Thus **teb** is DDE morally *impermissible*.

MODELLING TACTICAL BOMBING. Framework ⟨P, A, I⟩: A= {tab, tab*}, I={ }, P:

end_war ← target_military. kill_civilians ← tab. target_military ← tab.

Same counterfactual as above. Evaluation follows.
Step 1: Observations O = {kill_civilians, end_war} with explanation E = {tab}.
Step 2, produce T, as above**:**

make(kill_civilians*). % Intervention: *If civilians had not been killed...*
kill_civilians * ← make(kill_civilians*). % Killing civilians or otherwise now
kill_civilians ← make(kill_civilians). % available only by intervention.

And, for irrelevancy, deleting: kill_civilians ← tab.
Step 3: Counterfactual is not valid, since conclusion *war would not have ended* is false. Indeed, 'end_war' holds in the semantics of updated T plus E. Hence, the morally wrong 'kill_civilians' is a just side effect of achieving goal 'end_war'. Thus **tab** is DDE morally *permissible*.

We have shown counterfactual reasoning quite useful for critical thinking, viz. about morality.

REFERENCES

Byrne, R. M. J. (2007). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Collins, J., Hall, N. & Paul, L. A., editors (2004). *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Hauser, M., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). *A dissociation between moral judgments and justifications*. Mind and Language, 22(1):1–21.

Hoerl, C., McCormack, T. & Beck, S. R., editors (2011). *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Oxford: Oxford University Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Mandel, D. R., Hilton, D.J., & Catellani, P (2005). *The Psychology of Counterfactual Thinking*. New York, NY: Routledge.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge, MA: Cambridge University Press.

Roese, N.J. & Olson, J. M., editors (2009). *What Might Have Been: The Social Psychology of Counterfactual Thinking*. New York, NY: Psychology Press.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.