A counterfactual is a conditional statement representing what would be the case if its premise were true (but it is not true in fact). “If COMMA 2014 were not held in Scotland, you would not be here now.”

A formal model of counterfactuals has been studied by Stalnaker (1968) and Lewis (1973) based on the possible world semantics.

Counterfactuals are popularly used in dialogue, argument or dispute, while little attention has been paid on it in formal argumentation.

The purpose of this study is to provide an argumentation-theoretic interpretation of counterfactuals.
Let $U$ be the **universe** of all possible arguments. An **argumentation framework (AF)** is a pair $(Ar, att)$ where $Ar$ is a finite subset of $U$ and $att \subseteq Ar \times Ar$. An argument $A$ **attacks** an argument $B$ iff $(A,B) \in att$.

- **A indirectly attacks** (resp. **indirectly defends**) $B$ if there is an odd-length (resp. even-length, non-zero) path from $A$ to $B$ in a directed graph associated with $AF$.

- A **labelling** of $AF$ is a function $L: Ar \rightarrow \{ in, out, undec \}$.

- **Complete, (semi-)stable, grounded,** and **preferred labelling** are defined as usual. We simply say “labelling” to indicate one of these 5 labellings.
COUNTERFACTUAL REASONING IN ARGUMENTATION FRAMEWORK

“If A were rejected, then B would be accepted.”
“If A were rejected, then C would be rejected.”
“If B were accepted, then C would be rejected.”

Modify AF to AF’ in a way that
an argument A that is accepted in AF is rejected in AF’;
or an argument B that is rejected in AF is accepted in AF’.
MODIFICATION OF AF

Let $AF=(Ar, att)$ and $A \in Ar$. Define

$$AF^c_{+_A} = (Ar, att \setminus \{(X, A) | X \in Ar\})$$

(removing every attack relation attacking $A$)

$$AF^c_{-_A} = (Ar \cup \{X\}, att \cup \{(X,A)\}) \text{ where } X \in U \setminus Ar \text{ and } U \setminus Ar \neq \emptyset$$

(introducing an argument that attacks $A$)

- $\mathcal{L}(A)=in$ for every labelling $\mathcal{L}$, if any, in $AF^c_{+_A}$

- $\mathcal{L}(A)=out$ for every labelling $\mathcal{L}$, if any, in $AF^c_{-_A}$

- $AF^c_{+_A}$ and $AF^c_{-_A}$ are simply written $AF^c$ if an argument $A$ is clear in the context.
COUNTERFACTUALS (CF) IN AF

Let $AF = (Ar, att)$ and $A, B \in Ar$, and $\ell \in \{\text{in, out}\}$. Define

- $\text{in}(A) \square \to \ell(B)$ is true in $AF$ if $\mathcal{L}(B) = \ell$ in every labelling $\mathcal{L}$ of $AF^c_{+A}$
- $\text{in}(A) \lozenge \to \ell(B)$ is true in $AF$ if $\mathcal{L}(B) = \ell$ in some labelling $\mathcal{L}$ of $AF^c_{+A}$
- $\text{out}(A) \square \to \ell(B)$ is true in $AF$ if $\mathcal{L}(B) = \ell$ in every labelling $\mathcal{L}$ of $AF^c_{-A}$
- $\text{out}(A) \lozenge \to \ell(B)$ is true in $AF$ if $\mathcal{L}(B) = \ell$ in some labelling $\mathcal{L}$ of $AF^c_{-A}$

where "labelling" means one of the 5 labellings of $AF$.

- $\text{in}(A) \square \to \text{in}(B)$ is read "if $A$ were accepted then $B$ would be accepted."
- $\text{in}(A) \square \to \text{out}(B)$ is read "if $A$ were accepted then $B$ would be rejected."
- $\text{in}(A) \lozenge \to \text{in}(B)$ is read "if $A$ were accepted then $B$ might be accepted."
- $\text{in}(A) \lozenge \to \text{out}(B)$ is read "if $A$ were accepted then $B$ might be rejected."
**EXAMPLE**

- **AF** has the complete labelling: \{ *in*(A), *out*(B), *in*(C), *out*(D), *out*(E) \}.

- The following CFs hold in **AF**:

  - *out*(A) □→ *in*(B)
  - *in*(B) □→ *out*(C)
  - *out*(C) ◇→ *in*(D)
  - *out*(C) ◇→ *out*(E)
  - *out*(C) ◇→ *in*(E)
FORMAL PROPERTIES (1)

Let $AF=(Ar, att)$ and $A, B \in Ar$, and $\ell \in \{ \text{in, out} \}$.

- $\ell_1(A) \Box \rightarrow \ell_2(B)$ implies $\ell_1(A) \lozenge \rightarrow \ell_2(B)$

- CFs are reflexive:
  - $\ell(A) \Box \rightarrow \ell(A)$ is true in $AF$.
  - $\ell(A) \lozenge \rightarrow \ell(A)$ is true in $AF$ whenever $AF^c$ has a labelling.

- CFs with true antecedent:
  - If $\mathcal{L}(A) = \ell_1$ and $\mathcal{L}(B) = \ell_2$ in every complete labelling $\mathcal{L}$ of $AF$, then $\ell_1(A) \Box \rightarrow \ell_2(B)$ is true in $AF$.
  - If $\mathcal{L}(A) = \ell_1$ in every complete labelling $\mathcal{L}$ of $AF$ and $\mathcal{L}(B) = \ell_2$ in some complete labelling $\mathcal{L}$ of $AF$, then $\ell_1(A) \lozenge \rightarrow \ell_2(B)$ is true in $AF$. 
FORMAL PROPERTIES (2)

- **Modus Ponens:**
  - If $\mathcal{L}(A) = \ell_1$ in every complete labelling $\mathcal{L}$ of $AF$ and $\ell_1(A) \square \rightarrow \ell_2(B)$ is true in $AF$, then $\mathcal{L}'(B) = \ell_2$ in every complete labelling $\mathcal{L}'$ of $AF$.
  - If $\mathcal{L}(A) = \ell_1$ in every complete labelling $\mathcal{L}$ of $AF$ and $\ell_1(A) \lozenge \rightarrow \ell_2(B)$ is true in $AF$, then $\mathcal{L}'(B) = \ell_2$ in some complete labelling $\mathcal{L}'$ of $AF$.

- **Modus Tollens:**
  If $\ell_1(A) \square \rightarrow \ell_2(B)$ is true in $AF$ and $\mathcal{L}(B) \neq \ell_2$ for any complete labelling $\mathcal{L}$ of $AF$, then $\mathcal{L}'(A) \neq \ell_1$ for any complete labelling $\mathcal{L}'$ of $AF$.

- **Others**
  - If $\ell_1(A) \square \rightarrow \ell_2(B)$ is true in $AF$, then $\ell_1(A) \lozenge \rightarrow \ell_2(B)$ is true in $AF$.
  - If $\ell_1(A) \lozenge \rightarrow \ell_2(B)$ is true in $AF$, then $\ell_1(A) \triangle \rightarrow \ell_2(B)$ is true in $AF$.

where $\overline{\ell}$ is “out” (resp. “in”) if $\ell$ is “in” (resp. “out”).
COUNTERFACTUAL FALLACIES

- Fallacy of strengthening the antecedent:

\[ \ell_1 (A) \square \rightarrow \ell_2 (B) \] in AF does not imply \[ \ell_1 (A) \land \ell_3 (C) \square \rightarrow \ell_2 (B) \] in AF in general.

- Fallacy of transitivity:

\[ \ell_1 (A) \square \rightarrow \ell_2 (B) \] and \[ \ell_2 (B) \square \rightarrow \ell_3 (C) \] in AF do not imply \[ \ell_1 (A) \square \rightarrow \ell_3 (C) \] in AF in general.

- Fallacy of contraposition:

\[ \ell_1 (A) \square \rightarrow \ell_2 (B) \] in AF does not imply \[ \ell_2 (B) \square \rightarrow \ell_1 (A) \] in AF in general.

The above results also hold by replacing \( \square \rightarrow \) with \( \diamond \rightarrow \).
COUNTERFACTUAL DEPENDENCIES

- "An event $\varphi$ depends causally on another event $\psi$ iff both $\varphi \square \rightarrow \psi$ and $\neg \varphi \square \rightarrow \neg \psi$ hold" (Lewis 1973).

- Counterfactual dependencies

  $\ell_1 (A) \square^c \rightarrow \ell_2 (B)$ is true in $AF$
  if both $\ell_1 (A) \square \rightarrow \ell_2 (B)$ and $\ell_1 (A) \square \rightarrow \ell_2 (B)$ hold in $AF$
  where $\ell_1, \ell_2 \in \{ \text{in, out} \}$.

- Formal properties

  - $\ell_1 (A) \square^c \rightarrow \ell_2 (B)$ implies $\ell_1 (A) \square \rightarrow \ell_2 (B)$, but not vice versa.
  - If $\ell (A) \square \rightarrow \ell (B)$ is true in $AF$, then $A$ indirectly defends $B$.
  - If $\ell (A) \square^c \rightarrow \ell (B)$ is true in $AF$, then $A$ indirectly attacks $B$. 
PREEMPTION

- $\text{in}(A) \square \rightarrow \text{out}(B)$ but $\text{out}(A) \square \downarrow \rightarrow \text{in}(B)$ imply $\text{in}(A) \square \rightarrow \text{out}(B)$

- $\text{in}(C) \square \rightarrow \text{out}(B)$ but $\text{out}(C) \square \downarrow \rightarrow \text{in}(B)$ imply $\text{in}(C) \square \rightarrow \text{out}(B)$

- Thus $\text{out}(B)$ causally depends on neither $\text{in}(A)$ nor $\text{in}(C)$.

- The existence of $A$ is the actual cause of rejecting $B$. On the other hand, if $A$ were rejected then $C$ would be accepted, which results in rejecting $B$. Thus, $C$ is a “potential” alternative cause of rejecting $B$.

- $\text{in}(C) \square \rightarrow \text{out}(B)$ represents that $\text{in}(C)$ is a potential cause of $\text{out}(B)$ but is preempted by the actual cause $\text{in}(A)$.
MODAL INTERPRETATION

- □ℓ(A) \text{ def } ℓ(A) \square \rightarrow \text{in}(⊥) \quad \text{and} \quad ◇ℓ(A) \text{ def } ℓ(A) \square \neg \rightarrow \text{in}(⊥)

- For instance, □\text{in}(A) ("A is necessarily accepted")
  iff out(A) \square \rightarrow \text{in}(⊥)
  ("inconsistency would be accepted if A were rejected")

- Example

A=\("\sqrt{2} \text{ is an irrational number}\"

The validity of the argument A is proven by showing inconsistency under the assumption that \(\sqrt{2}\) were a rational number.

⇒ □\text{in}(A) is proved by showing out(A) \square \rightarrow \text{in}(⊥)\)
COMPLEXITY

- Given an $AF$, the problem of deciding whether $\text{in}(A) \square \rightarrow \ell(B)$ (resp. $\text{in}(A) \diamond \rightarrow \ell(B)$) holds or not in $AF$ is equivalent to deciding whether $L(B) = \ell$ in every (resp. some) labelling $L$ of $AF_{c+A}^c$.

- Likewise, deciding whether $\text{out}(A) \square \rightarrow \ell(B)$ (resp. $\text{out}(A) \diamond \rightarrow \ell(B)$) holds or not in $AF$ is equivalent to deciding whether $L(B) = \ell$ in every (resp. some) labelling $L$ of $AF_{c-A}^c$.

- Therefore, complexities of CF under the operator $\square \rightarrow$ (resp. $\diamond \rightarrow$) are equivalent to those of skeptical (resp. credulous) reasoning of an argument under argumentation semantics.
POTENTIAL APPLICATIONS

- CFs are popularly used in dialogue or dispute, so the framework would be useful for realizing CF reasoning in dialogue systems based on AF.

- CFs are used in diagnosis in which assumptions are introduced for explaining the observed misbehavior of a device. Thus, CFs could be used for analytic tools in AF.

- When one wants to have a desired outcome which would not be achieved by true arguments, one would reason counterfactually. In this case, CFs would be used for building false arguments in debate/discussion games.

- CF studied here is based on abstract AF and it will be applied to instantiated AFs based on particular representation languages.