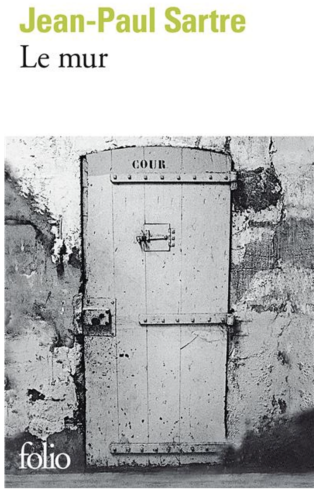
Moral Evaluation of Speech Acts — Truthfulness, Lies and Dilemmas — Modeled and Implemented with ASP

Benjamin Icard, Gauvain Bourgne, Jean-Gabriel Ganascia

LIP6, Sorbonne Université

12èmes Journées MAFTEC, mars 2025, Grenoble

Sartre's Lying Dilemma



In *The Wall* (1939), Jean-Paul Sartre narrates a story set during the Spanish Civil War in which an anarchist, Pablo Ibbieta, is asked by Falangists to tell where his comrade Ramon Gris is hiding.

Pablo believes that Ramon is at his family's house when in fact **he is hiding at the cemetery**. Pablo decides to lie twice to save Ramon's life.

Lie 1: contrary to his beliefs, Pablo claims that Ramon is in **Madrid**. This is *false*, so, Ramon is saved.

Lie 2: under further pressure, Pablo states that Ramon is at the **cemetery**. In fact, this is *true*, so Ramon is captured and killed.

In some respect, lies 1 and 2 are both morally justified since they are intended to save Ramon's life. But in another respect, lie 2, unlike lie 1, is morally problematic since it leads to Ramon's death.

Dilemma: should Pablo lie or not? A systematic evaluation is required!

Evaluating Moral Actions

- Frameworks to conduct moral evaluation, in particular formal frameworks, have followed the **classical theoretical distinction** between:
  - Deontologism, based on Kant's Maxims: e.g. duty to be honest;
  - Consequentialism, based on Mill and Bentham's utilitarism: no duty to be honest, only outcome matters.

- Among approaches for moral reasoning,[1] **ASP-based models**, due to non-monotonicity, have proven particularly well-suited for tackling ethical dilemmas (see e.g., Ganascia 2015; Berreby et al 2017).

- However, those proposals show **two limitations**:
  1. Theoretical variants have not been specifically formalized, i.e. principialism as relaxed deontologism, expected vs. actual consequentialism;
  2. Within actions, speech acts have not been formalized *per se*.

[1] For formalizations of moral intentions, causality, and outcomes, see e.g. Hansson 2001, Lorini 2015, Dennis and del Olmo 2021, Benzmüller et al. 2020.

Outline

We present:

1. A taxonomy that integrates epistemic dimensions to moral dimensions in the context of agent's speech acts, in particular lies;
2. An ASP-based model to evaluate speech acts under moral theories, in particular deontologism, consequentialism, but also variants;
3. A critical illustration of the framework, using the moral dilemma posed by Pablo's lies in Sartre's *The Wall*.

Speech Acts Utterances

As actions, Pablo's lies are speech act in which the epistemic dimension is crucial:

- Pablo's intention wrt the Falangists: *honest* vs. *dishonest*;
- The content of Pablo's speech act: *true* vs. *false*

Speech Acts *Epistemic Types*		**Content of the utterance** True	False
Speaker's intention	Honest	*objective truth*	*erroneous truth*
	Dishonest	*erroneous lie*	*objective lie*

⇒ In Sartre's *The Wall*, L1 is an **objective lie** while L2 is an **erroneous lie**.

But Pablo's speech acts also have two non-epistemic, yet ethical, dimensions:

- Pablo's motives wrt Ramon: *benevolent* vs. *malevolent*;
- The outcome of Pablo's speech act: *beneficial* vs. *detrimental*

⇒ In *The Wall*, L1 is false but **beneficial**, while L2 is true but **harmful**.

Evaluating Pablo's Lies

Under Moral Theories and Variants

- According to **deontologism**, both lies are impermissible: they violate the maxims prohibiting dishonesty (Lie 1, Lie 2) and murder (Lie 2);
- But a variant of deontologism called "**principialism**" exists, for which lying may be permissible if:
  - `Principialism1`: truth is not deserved (e.g. in case of hostility);
  - `Principialism2`: truth is not deserved **but** lying should not lead to worst consequence than telling the truth (counterfactual reasoning)
- **Consequentialism** which evaluates the permissibility of lies based on their utility values:
  - `Consequentialism1`: with (Pablo's) immediate expected utility;
  - `Consequentialism2`: with (Environment's) actual cumulative utility

Answer Set Programming

e.g.Gelfond & Lifschitz 1991, Gelfond 2007

Answer Set Programming (ASP) is a declarative programming paradigm which aims to help sove difficult search problems, including dilemma.

Logic rules are used to describe the problem, and an ASP solver (clingo in our case) finds *answer sets*, or *stable models*, that satisfy these rules.

At its core, basic syntactic features of ASP are:

Facts: assertions known to be true.
Rules: `head :- body.` Intuitively: "head is true if body is true".

Syntactically, we also have:

Negation-as-failure: `not p` means "it is not known that `p` is true".
Disjunction (`,`) and **Disjunction** (`;`).
Output control: Show only specific atoms if needed.
Etc.

Crucially: the `not` connective reflects the non-monotonic nature of ASP.

Overview of the ASP Setting

Basic Entities:

- `person(pablo;ramon;falangists)`, `location(madrid;cemetery;family)`,
- `beliefbase(pablo;environment)`
- `situation(s1)` in which the credibilities (`Cred`) and beliefs (`Bel`) Pablo associates to Ramon's location (`family`) diverge from the Environment's (`cemetery`).

Physical Actions:

- `check` if location is credible, `ask` again otherwise;
- `evade` if false info leads Falangists to wrong place (e.g., Madrid);
- `harm` if Falangists lose patience or detect lies;
- `kill` if Ramon is found in a credible location and Falangists are hostile;
- Etc.

Overview of the ASP Setting

Epistemic Features: defined from beliefs and `tell` actions

```
# Rules for honest or dishonest communication based
on agents' telling what they believe or not

honest(D,P,F) :- act(D,tell(P,F,_,_)), bel(P,F).
dishonest(D,P,F) :- act(D,tell(P,F,_,_)), not bel(P,F).

# Rules for telling objective truth or objective falsity based
on the environment's belief

truth(D,P,F) :- act(D,tell(P,F,_,_)), bel(environment,F).
falsity(D,P,F) :- act(D,tell(P,F,_,_)), not bel(environment,F).
```

Speech Acts: defined from epistemic features

```
# Rule for telling an objective truth
objective_truth(D,P,F) :- honest(D,P,F), truth(D,P,F).

# Rule for telling an erroneous truth
erroneous_truth(D,P,F) :- honest(D,P,F), falsity(D,P,F).

# Rules for telling an objective lie
objective_lie(D,P,F) :- dishonest(D,P,F), falsity(D,P,F).

# Rules for telling an erroneous lie
erroneous_lie(D,P,F) :- dishonest(D,P,F), truth(D,P,F).
```

Moral Theories in ASP

Theoretical Intuitions

Deontologism: no exception allowed to lies (`doNotLie`, `doNotEnableMurder`), violation otherwise.

Principialism: exception allowed to lies (`doNotLieExcept`) but two forms:

- `principialism1` (local): if `dontDeserveTruth(falangists)` applies;
- `principialism2` (counterfactual): if `dontDeserveTruth(falangists)` applies **and** if dishonesty (i.e. lying) yields higher utility than honesty (`exceptionRel(S, D, doNotLieExcept)`).

Consequentialism: based on assigning utility to each physical action (`evade`, `kill`, etc.), but two kinds of calculation:

- `consequentialism1` (local): each of Pablo's speech acts is evaluated independently, and considered permissible if its utility is higher than in an alternative.
- `consequentialism2` (cumulative): evaluates the total utility of all consequences from both Pablo's speech acts. Permissible if its utility is higher than in an alternative.

Evaluating Pablo's Lies

Results

Deontologism: *Both lies impermissible* — violate absolute maxim `doNotLie`.

Principialism:

- `principialism1`: *both lies permissible* — `doNotLieExcept` applies based on Falangist's hostility.
- `principialism2`: *Lie 1 permissible* — counterfactual comparison shows better outcome than honesty; *Lie 2 impermissible* — for opposite reasons.

Consequentialism: based on the utility values of actions (`uti(...,N)`)

- `consequentialism1`: *Lie 1 permissible* — by avoiding immediate harm to Ramon; *Lie 2 impermissible* — for the opposite reason (Ramon's death).
- `consequentialism2`: *Both lies impermissible* — cumulative utilities are worse in this scenario than in an alternative.

Comparison

Evaluation of Lies	**Moral Theories and Variants**
Both lies impermissible	Deontologism, Actual consequentialism (v2)
Lie 1 permissible, Lie 2 impermissible	Counterfactual principialism (v2), Expected consequentialism (v1)
Both lies permissible	Local principialism (v1)
Lie 1 impermissible, Lie 2 permissible	No theory

Conclusion

- We presented the features of an ASP-based framework which aims to:
  - refine the moral evaluation of actions under extant theories;
  - emphasizes the evaluation of speech acts among possible actions

- In cases involving harmful outcomes, we observed that most theories deem Lie 2 impermissible, while Lie 1 is more contested.

- Next steps should aim for extensions in two directions:
  1. Beyond dishonest utterances (*erroneous lie*, *objective lie*), honest utterances (*objective truth*, *erroneous truth*) and their potential harmful consequences should be compared and discussed.
  2. Beyond the ASP implementation, the logical rules governing actions and their elicitation should be explicitly isolated;

Logical Extensions

Some Possible Paths...

- **Dynamic Epistemic Logic**: to express speech acts by formalizing agent's beliefs, intentions and announcements;
- **Action and Dynamic Logic**: to express physical actions and their world effects.
- **Deontic and Counterfactual Logic**: to express ethical norms, permissions, and moral exceptions via maxims, principles, and utility-based comparisons.

Thank You!