Multi-Agent Systems

Deontic Logic

Albert-Ludwigs-Universität Freiburg

Bernhard Nebel, Rolf Bergdoll, and Thorsten Engesser

Winter Term 2019/20

Ethics

What is the morally acceptable way to act?

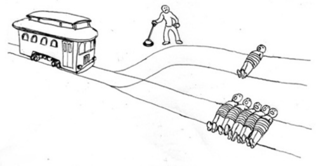
Economical Answer

Maximize expected utility (for society)!

Success Story of AI

New Challenges

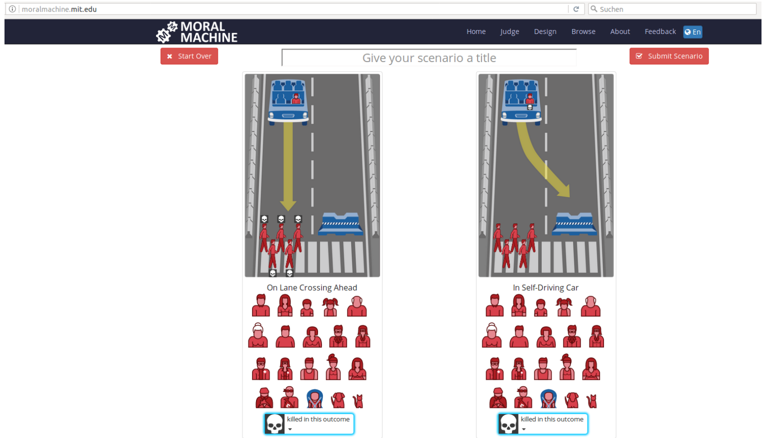
Trolley Problem



Utility Maximization

- The utility-based robot:
  - Goal: Do whatever maximizes utility.
  - Utility function: Negative utility per harmed human being.

MIT: Moral Machine



Alignment Problem

- How the robot acts:
  1. The robot throws the switch.
  2. The robot pushes the man.
  3. The robot sacrifices the life of the passenger.
- Most people agree with (1) but disagree with (2). (Mixed opinion regarding 3.)
- Alignment Problem: Aligning machines' and humans' ethical judgments **Which options are there?**

Asimov's Laws of Robotics

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the first law.
3. A robot must protect its own existence as long as such protection does not conflict with the first or second law.

⇒In case of a dilemma, the first law renders all possible solution inacceptable.

Moral Principles

- Moral principles determine the subset of morally acceptable options from the set of all available options.
- Examples:
  - Utilitarianism (maximize social welfare)
  - Deontology
  - Principle of Double Effect
  - Virtue Ethics
  - ...

Hybrid Ethical Reasoning Robots

Video 2:30

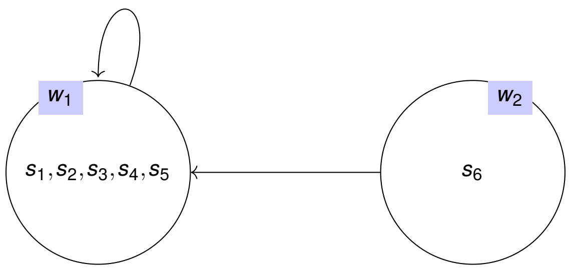
Role of Deontic Logic

- Given that an agent can compute what it should or should not do ...
  - We will not deal with moral decision making in this lecture, but come back to that later ...
- ...deontic logic is a tool to logically represent and reason about what an agent should and should not do.

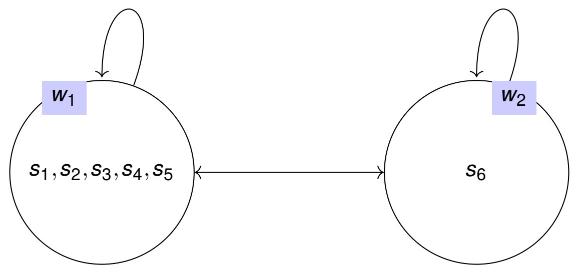
Standard Deontic Logic (SDL): Semantics

- Kripke models for Standard Deontic Logic (SDL)
  - $M = (W, R, V)$
  - Set of possible worlds W
  - Accessibility relation: $R : W \to 2^W$
    - An edge between worlds w and w' means that w' is normatively ideal relative to w.
    - R is assumed to be serial.
  - Valuation: $V : P \to 2^W$

Example: Trolley Case (Utilitarian)



Example: Trolley Case (Kantian)



Language of Deontic Logic

$$\varphi ::= p_i \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid \neg\varphi \mid O\varphi \mid F\varphi \mid P\varphi$$

- E.g., $(a \wedge b), Oa, O(a \vee b), OO(a \rightarrow b)$

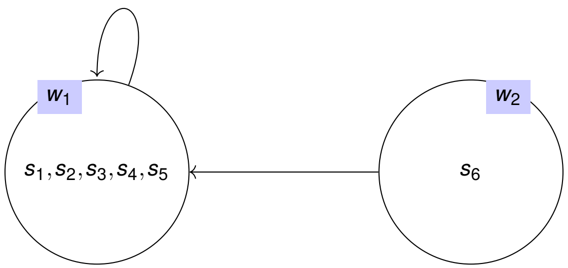
Two readings: Ought-to-be and Ought-to-do

- p := "You help your neighbor."
- Op := "You ought to help your neighbor."
- Ought-to-be: "It ought to be the case that you help your neighbor."
- Ought-to-do: "You ought to execute an action of type helping your neighbor." (How to make sense of OOp?)

Truth Conditions

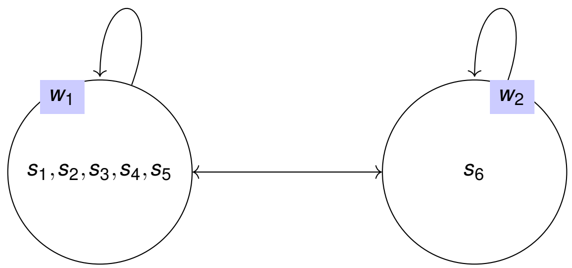
- $M, w \models O\varphi$ iff for all $(w, w') \in R : M, w' \models \varphi$
- Permissible
$$P\varphi \stackrel{\text{def}}{=} \neg O \neg \varphi$$
- Forbidden
$$F\varphi \stackrel{\text{def}}{=} O \neg \varphi$$
- Omissible
$$OM\varphi \stackrel{\text{def}}{=} \neg O\varphi$$
- Optional
$$OP\varphi \stackrel{\text{def}}{=} (\neg O\varphi \wedge \neg O \neg \varphi)$$

Example: Trolley Case (Utilitarian)



$$M, w_1 \models Os_1 \wedge \ldots \wedge Os_5 \wedge O\neg s_6 \wedge P\neg s_6 \wedge Fs_6 \wedge \ldots \text{ (Utilitarian)}$$

Example: Trolley Case (Kantian)



$$M, w_1 \models O(s_1 \vee s_6) \wedge \neg Os_1 \wedge \neg Os_6 \wedge P\neg s_1 \wedge \neg Fs_6 \wedge \ldots \text{ (Kantian)}$$

Axioms

O behaves according to the axioms of the system **KD**:

- $\models \varphi$ for all propositional tautologies φ
- If $\models \varphi \rightarrow \psi$ and $\models \varphi$ then $\models \psi$ (Modus Ponens)
- If $\models \varphi$ then $\models O\varphi$ (Necessity)
- $\models O\varphi \rightarrow \neg O \neg \varphi$ (Seriality)
- $\models O(\varphi \rightarrow \psi) \rightarrow (O\varphi \rightarrow O\psi)$ (K-Axiom)

All-things-considered Obligations

$\models \neg(O\varphi \wedge O\neg\varphi)$ directly follows from seriality: It is impossible to have contradicting obligations.

- Standard deontic logic is about all-things-considered obligations, i.e., it does not allow one to express prima-facie obligations, e.g., that one is at the same time both obliged to go to the lecture and to visit the friend in the hospital.
- But: In such a situation deontic logic permits to express that the agent may do either without prescribing one of the options.

Ought implies Allowed

$\models O\varphi \rightarrow P\varphi$

- The theorem follows from seriality and the definition of permissibility. Accepted as a rationality requirement: If a legal code prescribes something, then it must also permit that something.

Ross Paradox

Ross Paradox (Weakening Rule)

$\models O\varphi \rightarrow O(\varphi \vee \psi)$.

Proof

$\models \varphi \rightarrow (\varphi \vee \psi)$ (Propositional calculus)
$\models O(\varphi \rightarrow (\varphi \vee \psi))$ (Necessitation rule)
$\models O(\varphi \rightarrow (\varphi \vee \psi)) \rightarrow (O(\varphi) \rightarrow O(\varphi \vee \psi))$ (K-Axiom)
$\models O(\varphi) \rightarrow O(\varphi \vee \psi)$ (Modus Ponens)

- If is obligatory that the letter is mailed, then it is obligatory that the letter is mailed or the letter is burned.

Free-Choice Permission

$\not\models P(a \vee b) \rightarrow Pa \wedge Pb$

- What happens if one adds this as an axiom to SDL?
  - $\models O\varphi \rightarrow O(\varphi \vee \psi)$ (Weakening Rule)
  - $\models O(\varphi \vee \psi) \rightarrow P(\varphi \vee \psi)$ (Seriality)
  - $\models O\varphi \rightarrow P(\varphi) \wedge P(\psi)$ (viz., if something is obligatory, then everything is permissible)
- $\Rightarrow$Mind the gap between natural language and propositional logics.

Mixing modalities

The Paradox of Epistemic Obligation (Åqvist 1967)

$\models OK\varphi \rightarrow O\varphi$.

Proof

$\models K\varphi \rightarrow \varphi$ (T-axiom)

$\models O(K\varphi \rightarrow \varphi)$ (Necessitation rule)

$\models O(K\varphi \rightarrow \varphi) \rightarrow (OK\varphi \rightarrow O\varphi)$ (K-axiom)

$\models OK\varphi \rightarrow O\varphi$ (Modus Ponens)

- If it ought to be the case that one knows that Berlin is the capital of Germany, then it ought to be the case that Berlin is the capital of Germany.
- If it does not ought to be the case that Berlin is the capital of Germany, then it it does not ought to be the case that one knows that Berlin is the capital of Germany.

Leibniz-Kangerian-Reduction (LKA)

$$\varphi ::= p_i \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid \neg\varphi \mid O\varphi \mid F\varphi \mid P\varphi \mid \Box\varphi \mid \Diamond\varphi$$

- Defines SDL within alethic modal logic (logic of necessity).
- Its deontic fragment equals SDL plus a new axiom: $O(O\varphi \rightarrow \varphi)$.
- Higher syntactic expressivity due to alethic modality.

Obligation and Necessity

Gottfried Wilhelm Leibniz, 1646–1716

- The permitted is what is possible for a good person to do.
- The obligatory is what is necessary for a good person to do.

Petrus Abaelardus, 1097–1144

- Necessity is what nature demands.
- Possibility is what nature allows.
- Impossibility is what nature forbids.

Leibnizian-Kangerian-Andersonian reduction

- Leibnizian definition of obligation: φ is obligatory iff bringing about φ is necessary for being a good person.
- Can be written as: $O\varphi \overset{\text{def}}{=} \Box(g \rightarrow \varphi)$. The propositional symbol g represents "being a good person".
- Permission can be defined as: $P\varphi \overset{\text{def}}{=} \Diamond(g \wedge \varphi)$.

LKA: Semantics

- Kripke models $M = (W, G, R, V)$
  - Possible worls W
  - Accessibility relation $R : W \rightarrow 2^W$
    - R is reflexive ($\Rightarrow$stronger than the serial relation of SDL models)
  - $G \subseteq W$
  - For every world w there is a w' s. th. $w' \in G$ and $R(w, w')$

$\Rightarrow$New tableaux rule G: Introduce a new world with formula g.

Truth Conditions

- $M, w \models \Box\varphi$ iff $M, w' \models \varphi$ for each w' s.th. $(w, w') \in R$
- $M, w \models \Diamond\varphi$ iff $M, w' \models \varphi$ for some w' s.th. $(w, w') \in R$.
- $M, w \models g$ iff $w \in G$

Definitions

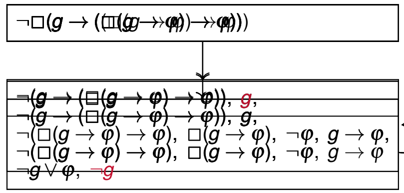
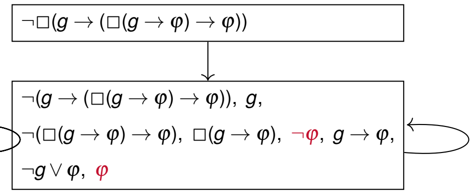
- Obligatory
$$O\varphi \overset{\text{def}}{=} \Box(g \rightarrow \varphi)$$
- Permissible
$$P\varphi \overset{\text{def}}{=} \Diamond(g \wedge \varphi)$$
- Forbidden
$$F\varphi \overset{\text{def}}{=} \Box(g \rightarrow \neg\varphi)$$
- Omissible
$$OM\varphi \overset{\text{def}}{=} \Diamond(g \wedge \neg\varphi)$$
- Optional
$$OP\varphi \overset{\text{def}}{=} \Diamond(g \wedge \varphi) \wedge \Diamond(g \wedge \neg\varphi)$$

Validities

- LKA is a **KT** logic, thus all **KT**-axioms hold.
- $\models \Diamond g$ (for the special “good” proposition)
- All axioms of SDL are valid in the deontic fragment of LKA.
- Addional validity: $\models O(O\varphi \rightarrow \varphi)$
- Mixed validity: $\models O\varphi \rightarrow \Diamond\varphi$ (Kant: Ought implies Can)

$\models O(O\varphi \rightarrow \varphi)$

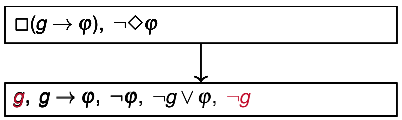
- $\Box(g \rightarrow (\Box(g \rightarrow \varphi) \rightarrow \varphi))$ (Def.)
- Prove $\neg\Box(g \rightarrow (\Box(g \rightarrow \varphi) \rightarrow \varphi))$ unsatisfiable

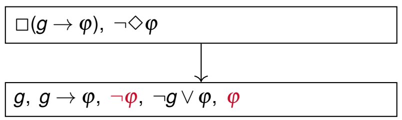



After ¬[I]-Rule.
After NotImpl-Rule.
After NotImpl-Rule (again).
After **T**-Rule.
After [I]-Rule. After Impl-Rule. After Or-Rule. Done.

$\models O\varphi \rightarrow \Diamond\varphi$

- Prove $O\varphi \wedge \neg\Diamond\varphi$ unsatisfiable.





After G-rule application.
After [I]- and ¬<I>-Rules.
After Impl-Rule.
After Or-Rule. Done.

Be good!

Theorem

$\models Og$. It is obligatory to be a good person.

Proof

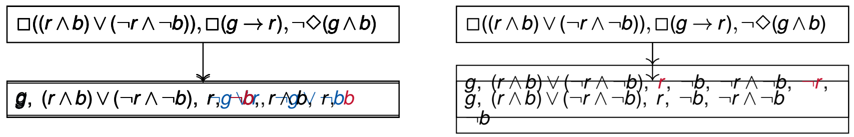
$\models g \rightarrow g$ (Propositonal calculus)
$\models \Box(g \rightarrow g)$ (Necessitation rule)
$\models O(g)$ (Def. of O)

Final Example

- Description of the Situation: A search-and-rescue robot has the choice between rescuing a patient (r) which would involve breaking an expensive vase (b), or refraining from doing so. The robot's decision procedure decides that the patient should be rescued.
  - $\varphi_1 = \Box((r \land b) \lor (\neg r \land \neg b))$
  - $\varphi_2 = Or$
- May the robot break the vase?
  - The answer is "yes" iff $\models (\varphi_1 \land \varphi_2) \to Pb$ can be shown.

Show $\models (\varphi_1 \land \varphi_2) \to Pb$

Show $\Box((r \land b) \lor (\neg r \land \neg b)) \land \Box(g \to r) \land \neg\Diamond(g \land b)$ unsatisfiable:



After G-rule application. After Or-rule application. After simplification.

- Applied: [I]-, ¬<I>-, NotAnd-, and Impl-Rules.
- Slight simplification possible (to save time and space): $((\varphi \lor \psi) \land \neg\varphi) \equiv \psi)$

After And-Rule. Done.

Applications

- Soft Constraints
- Fault-Tolerant Systems
- Analysis of Law (Law & AI)
- Modeling of moral agents

Further Advanced Topics

- Free-Choice Permissions
  - $P(a \lor b) \to Pa \land Pb$
- Conditional Obligations
  - $O(\varphi \mid \psi)$
- Deontic Conflicts
  - Prima facie oughts, allowing $O\varphi \land O\neg\varphi$ or at least $O_i\varphi \land O_j\neg\varphi$
- Multi-Agent Deontic Logics
  - $O_1P_2\varphi$

Literature

- R. Hilpinen, P. McNamara, Deontic logic: A historical survey and introduction, In D. Gabbay, J. Horty, X. Parent, R. van der Meyden, L. cn der Torre (Eds.) Handbook of Deontic Logic and Normative Systems, 2013, College Publications.
- P. McNamara, Deontic Logic, Stanford Encylopedia of Philosophy, `http://plato.stanford.edu/entries/logic-deontic/`
- Trolley Problem Memes on Facebook, `https://www.facebook.com/TrolleyProblemMemes`