What is the morally acceptable way to act?

Economical Answer

Maximize expected utility (for society)!
New Challenges

Utility Maximization

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

Trolley Problem

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

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey any orders given to it by human beings, except where such orders would conflict with the first law.
- 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 unacceptable.

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)
  - $\mathcal{M} = (W, R, V)$
  - 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

\[ \varphi ::= p \mid \varphi \land \varphi \mid \varphi \lor \varphi \mid \varphi \rightarrow \varphi \mid \neg \varphi \mid O \varphi \mid F \varphi \mid P \varphi \]

- E.g., \((a \land b), Oa, O(a \lor b), OO(a \rightarrow b)\)

Two readings: Ought-to-be and Ought-to-do
- \(p \) := “You help your neighbor.”
- \(O p \) := “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 \(OOp\)?)

Example: Trolley Case (Utilitarian)

\[ M, w_1 \models Os_1 \land \ldots \land Os_6 \land O\neg s_6 \land P\neg s_6 \land F\neg s_6 \land \ldots \text{ (Utilitarian)} \]

Truth Conditions

- \( M, w \models O\varphi \) iff for all \((w, w') \in R : M, w' \models \varphi \)
- **Permissible**
  \[ P\varphi \overset{\text{def}}{=} \neg O\neg \varphi \]
- **Forbidden**
  \[ F\varphi \overset{\text{def}}{=} O\neg \varphi \]
- **Omissible**
  \[ O\neg \varphi \overset{\text{def}}{=} \neg O\varphi \]
- **Optional**
  \[ O\neg \varphi \overset{\text{def}}{=} (\neg O\varphi \land \neg \neg \varphi) \]

Example: Trolley Case (Kantian)

\[ M, w_1 \models O(s_1 \lor s_6) \land \neg Os_1 \land \neg Os_6 \land P\neg s_1 \land F\neg s_6 \land \ldots \text{ (Kantian)} \]
**Axioms**

O behaves according to the axioms of the system KD:

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

**All-things-considered Obligations**

$\models \neg(O\phi \land O\neg \phi)$ 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**

$\models O\phi \rightarrow P\phi$

- 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)**

$\models O\phi \rightarrow O(\phi \lor \psi)$.

**Proof**

$\models \phi \rightarrow (\phi \lor \psi)$ (Propositional calculus)

$\models O(\phi \rightarrow (\phi \lor \psi))$ (Necessitation rule)

$\models O(\phi \rightarrow (\phi \lor \psi)) \rightarrow (O(\phi) \rightarrow O(\phi \lor \psi))$ (K-Axiom)

$\models O(\phi) \rightarrow O(\phi \lor \psi)$ (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

\[ \neg P(a \lor b) \rightarrow Pa \land Pb \]

What happens if one adds this as an axiom to SDL?

- \[ \vdash O\varphi \rightarrow O(\varphi \lor \psi) \] (Weakening Rule)
- \[ \vdash O(\varphi \lor \psi) \rightarrow P(\varphi \lor \psi) \] (Seriality)
- \[ \vdash O\varphi \rightarrow P(\varphi) \land P(\psi) \] (viz., if something is obligatory, then everything is permissible)

⇒ Mind the gap between natural language and propositional logics.

Leibniz-Kangerian-Reduction (LKA)

\[ \varphi ::= p_i \mid \varphi \land \varphi \mid \varphi \lor \varphi \mid \varphi \rightarrow \varphi \mid \neg \varphi \mid O\varphi \mid F\varphi \mid P\varphi \mid \Box\varphi \mid \Diamond \varphi \]

- 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**: \( \varphi \) is obligatory iff bringing about \( \varphi \) 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: \( P \varphi \overset{\text{def}}{=} \Diamond(g \land \varphi) \).

LKA: Semantics

- **Kripke models** \( M = (W, G, R, V) \)
  - Possible worlds \( W \)
  - Accessibility relation \( R : W \rightarrow 2^W \)
    - \( R \) is reflexive (⇒ 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 \( R(w, w') \)

\[ \Rightarrow \text{New tableaux rule } G: \text{Introduce a new world with formula } g. \]

Truth Conditions

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

Definitions

- **Obligatory**
  \[ O \varphi \overset{\text{def}}{=} \Box(g \rightarrow \varphi) \]
- **Permissible**
  \[ P \varphi \overset{\text{def}}{=} \Diamond(g \land \varphi) \]
- **Forbidden**
  \[ F \varphi \overset{\text{def}}{=} \Box(g \rightarrow \neg \varphi) \]
- **Omissible**
  \[ OM \varphi \overset{\text{def}}{=} \Diamond(g \land \neg \varphi) \]
- **Optional**
  \[ OP \varphi \overset{\text{def}}{=} \Diamond(g \land \varphi) \land \Diamond(g \land \neg \varphi) \]
Validities

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

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

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

Be good!

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

Proof
- $\models g \rightarrow g$ (Propositional 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.

- \( \varphi_1 = \square((r \land b) \lor (\neg r \land \neg b)) \)
- \( \varphi_2 = Or \)

May the robot break the vase?

- The answer is “yes” iff \( (\varphi_1 \land \varphi_2) \rightarrow Pb \) can be shown.

Further Advanced Topics

- Free-Choice Permissions
  - \( P(a \lor b) \rightarrow Pa \land Pb \)
- Conditional Obligations
  - \( O(\varphi \mid \psi) \)
- Deontic Conflicts
  - Prima facie oughts, allowing \( O\varphi \land O\neg \varphi \) or at least \( O\varphi \land O\neg \varphi \)
- Multi-Agent Deontic Logics
  - \( O_1 P_2 \varphi \)


Trolley Problem Memes on Facebook, https://www.facebook.com/TrolleyProblemMemes