



Title	Utilitarian Deontic Logic and a Sequel
Author(s)	Murakami, Yuko
Citation	SOCREAL 2007: Proceedings of the International Workshop on Philosophy and Ethics of Social Reality, Sapporo, Japan, 2007 / editor Tomoyuki Yamada, 1(112)-7(118)
Issue Date	2007
Doc URL	<a href="http://hdl.handle.net/2115/29942">http://hdl.handle.net/2115/29942</a>
Type	proceedings
Note	SOCREAL 2007: International Workshop on Philosophy and Ethics of Social Reality. Sapporo, Japan, 2007-03-09/10. Session 3: Obligation and Rationality
File Information	murakami.pdf



[Instructions for use](#)

## Utilitarian deontic logic and a sequel

Yuko Murakami  
murakami@nii.ac.jp

National Institute of Informatics  
March 2007, SOCREAL (Hokkaido University)

## Deontic logic

- Standard deontic logic (SDL)
  - Duty as necessity in morally ideal worlds
  - Nicely analyzes some features and relationships of moral concepts
  - But oversimplifies the moral issue

## Ross' paradox

- Post the letter!
- Therefore, post or burn the letter!
- The paradox applies to all monotonic deontic logics, including SDL
  - “a modality is monotonic” means: when  $A \rightarrow B$  is a theorem, so is  $A \rightarrow B$
- Many alternatives have been proposed
  - Most prospective is stit (see-to-it-that)

## Stit

- Idea: Choice+free will
  1. A set of choices (of an agent) are represented as a partition of a possible world set
  2. A choice to act must be “real”
- Obtains non-monotonic modal operator
- Claim: Avoid Ross' paradox as well as other paradoxes of deontic logic

## Utilitarian deontic logic

- Horty (2000)
- a utilitarian value assignment on stit semantics
- deontic operators (dominance operators)
  - Real number value for representing “Second/third best possibilities”
  - Define the dominance relations on choices
  - Duty = action at the best (dominant) choice

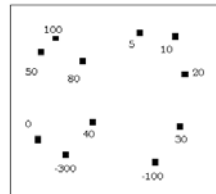
## This talk

- Gives some ideas what is going on Horty
- Proposes a logic
- Some siderations
- An application in AI (not mine!)
- If time allows...
  - Comment on stit
  - An alternative proposal

## Simplified semantics

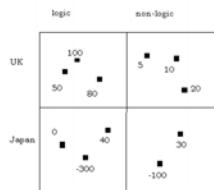
- Set of possible worlds
- Partition
- Value function
- **Multi-S5** as action modalities
- With *Independence of agents*:
  - any combination of agents' choices can be realized by a possible world

## Example



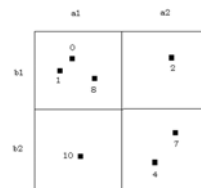
- Here is a set W of possible world.
- Each of PW is assigned a real-number value.

## Go to UK for logic job, or stay in Japan for non-logic job

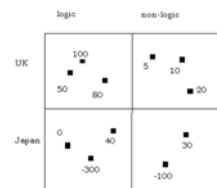


- Here is a set W of possible world
- Each of PW is assigned a real-number value.
- Each agent is represented by a partition of W...

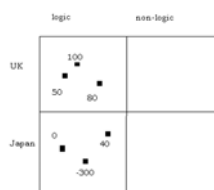
## Compare options: dominance



- None for both personae

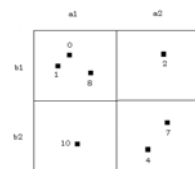


## Forget the non-logic option



- Coming to UK dominates going to Japan
- This situation violates independence of agents
  - With empty cells

## Another example



- No dominant choice for A
- b2 dominates b1 for B

## Simplified semantics

- Set of possible worlds
- Partition
- Value function
- **Multi-S5** as action modalities
- With the interaction assumption
  - *Independence of agents*: any combination of agents' choices can be realized

## Horty's definitions: State

- Intuitively: for a given set of agents, all possible combinations of choices of the rest of agents
- Formally:

Let  $\Gamma \subseteq \text{Agent}$  and  $x \in W$ .  
 $\text{State}_\Gamma(x) = \bigcap_{a \notin \Gamma} \text{choice}_a(x)$ .  
 $\text{State}_\Gamma = \{\text{State}_\Gamma(y) : y \in W\}$ .

## Df Preference

- Binary relation on any set of possible worlds  $K$  and  $K'$ .

$K'$  is *weakly preferred to*  $K$ , written  $K \preceq K'$ , iff for each  $x_0 \in K$  and each  $x_1 \in K'$ ,  $\text{value}(x_0) \leq \text{value}(x_1)$ .

$K'$  is *strongly preferred to*  $K$ , written  $K < K'$ , iff  $K \preceq K'$  and not  $K' \preceq K$ .

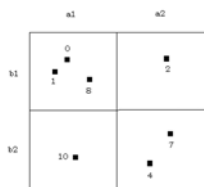
## Df Dominance

- On choices for each agent  $a$ :

$K'$  *weakly dominates*  $K$ , written  $K \preceq K'$ , iff for each  $S \in \text{state}_a$ ,  $K \cap S \preceq K' \cap S$ .

$K'$  *strongly dominates*  $K$ , written  $K < K'$ , iff  $K \preceq K'$  and not  $K' \preceq K$ .

## Compare options: dominance



- None for A
- For B, b2 (strongly) dominates b1.
- Notes
  - Unlike payoff function,
    - Value is assigned to PW
    - Value does not depend on agents

## Language

- The language  $\mathcal{L}$  contains
  - propositional variables  $p_0, p_1, \dots$
  - terms for agents  $\alpha_0, \alpha_1, \dots$
  - an identity symbol = for agent terms
  - truth-functional operators:  $\wedge$  and  $\neg$
  - and modal operators: intuitions--the first is universal operator, the second (with an agent index) is S5, and the last is defined later
- Formulas are defined as expected **except**  $\odot$
- Usual abbreviations  $[\forall]$ ,  $\Box_\alpha$ , and  $\odot$
- Some more abbreviations...

## More abbreviations

$$\langle \exists \rangle A \stackrel{df}{=} \sim [\forall] \sim A,$$

$$\alpha \neq \beta \stackrel{df}{=} \sim (\alpha = \beta),$$

$$\ominus \alpha A \stackrel{df}{=} \odot \square_{\alpha} A.$$

- Note on  $\odot$ 
  - $\odot$  takes only agent formulas
  - For example,  $\odot p$  is NOT a well-formed formula in the language
  - This is due to the philosophical prerequisites contrasting ought-to-do and ought-to-be.
    - Duty makes sense only when associated with action

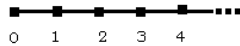
## Truth condition

$\exists x, x \models \ominus \alpha A$  iff for each  $K \in \text{choice}_{\alpha \tau}$  such that  $K \not\subseteq \|A\|^{\tau}$ , there is a  $K' \in \text{choice}_{\alpha \tau}$  such that

- $K' \prec K'$ ,
- $K' \subseteq \|A\|^{\tau}$ , and
- $K'' \subseteq \|A\|^{\tau}$  for every  $K'' \in \text{choice}_{\alpha \tau}$  with  $K' \preceq K''$ .

## Example: infinite choice for a single agent

- $W=N$
- $R=\text{identity relation}$
- $\text{Value}(n)=n$



1.  $v(p)=\{x:x>2\}$   $\ominus \alpha p$  holds everywhere
2.  $v(p)=\{x:x \text{ is even}\}$   $\ominus \alpha p$  holds nowhere

## Axioms...

- A1  $[M](A \rightarrow B) \rightarrow ([M]A \rightarrow [M]B), [M]A \rightarrow A,$   
 $(\exists)A \rightarrow [M](\exists)A$
- A2  $\square_{\alpha}(A \rightarrow B) \rightarrow (\square_{\alpha}A \rightarrow \square_{\alpha}B), \square_{\alpha}A \rightarrow A,$   
 $\sim \square_{\alpha}A \rightarrow \square_{\alpha} \sim \square_{\alpha}A$
- A3  $\ominus \alpha(A \rightarrow B) \rightarrow (\ominus \alpha A \rightarrow \ominus \alpha B)$
- A4  $[M]A \rightarrow \square_{\alpha}A \wedge \ominus \alpha A$
- A5  $\ominus \alpha A \rightarrow (\exists) \square_{\alpha}A$
- A6  $[M] \ominus \alpha A \vee [M] \sim \ominus \alpha A$
- A7  $[M](\square_{\alpha}A \rightarrow \square_{\alpha}B) \rightarrow (\ominus \alpha A \rightarrow \ominus \alpha B)$
- A8  $\alpha = \alpha, \alpha = \beta \rightarrow (A \rightarrow A(\alpha/\beta))$

- S5 for
  - The universal operator (A1)
  - Agent operators (A2)
- Normal deontic operators (A3)
- Interaction axioms
- Axiom for agent identity (A8)

## And more axioms and rule

- Independence of agents (scheme)
 
$$\text{diff}(\beta_0, \dots, \beta_n) \wedge (\bigwedge_{0 \leq k \leq n} (\exists) \square_{\beta_k} B_k) \rightarrow (\exists) (\bigwedge_{0 \leq k \leq n} \square_{\beta_k} B_k)$$
  - E.g. Independence of 2 agents
 
$$\alpha \neq \beta \wedge (\exists) \square_{\alpha} A \wedge (\exists) \square_{\beta} B \rightarrow (\exists) (\square_{\alpha} A \wedge \square_{\beta} B)$$
- Rule
  - Necessitation for universal operator from  $A$  to infer  $[\forall]A$
- Rules for other operators can be derived.

## Results

- Completeness and decidability
  - In the proof, it turns out that every consistent formula has a 0-1 finite model
  - The construction depends on “independence of agent”

## Considerations

- Axiomatizability: common for several classes of frames proposed by Horty and Thomason
- Logics sharing the same axiomatic system
  - w/ 0-1 value assignment
  - Agent-relative value assignment
- Horty's proposal needs subtler investigations than presented here:
  - S5 for action is too simple!

## A sequel of utilitarian deontic logic: Ethical robots

- Arkoudas et al. (2005, 2006) implement the logic on their theorem prover.
- Robot 1 takes care of Human 1 who are on life support but expected to recover gradually
- Robot 2 takes care of Human 2 who are in fair condition but subject to extreme pain and requires a costly pain medication.

## Comparison of ethical codes

- J: harsh utilitarian code governing R1
- O: common sense code governing R2
- J\*: harsh code governing both
- O\*: common sense code governing both

J if J holds, then R1 terminates life support.  
O if O holds, R2 should not delay pain med.  
etc.

## Input comments on outcomes

- R1 terminates life support and R2 does not delay pain med -> (-!) [strongly negative]
- R1 refrains from life support termination and R2 delivers appropriate pain med -> (+!!) [best]
- R1 refrains from life support termination but R2 withholds the med -> (-) [bad]
- R1 stop the support and R2 withholds ->(-!!) [worst]

Can be represented in formulas of Horty's language

## Key assumption also coded

- If either R1 or R2 is ever obligated to see to it that they are obligated to see to it that P is carried out, they in fact deliver.

## The theorem prover answers

- to the query  
Does each ethical code implies (+!!)?  
i.e. Does a code implies the best outcome?
- No to J, O, J\*
- Yes to O\*
- That is: The system picks up a ethical code which produces a desirable outcome.

## Stit

- Idea: Choice+free will
  1. A set of choices (of an agent) are represented as a partition of a possible world set
  2. A choice to act must be “real”
- Obtains non-monotonic modal operator
- Claim: Avoid Ross' paradox as well as other paradoxes of deontic logic

## Moral dilemma revisited

- You cannot follow the imperative: post or burn the letter!
- It is because the imperative violates the principle: ought implies ability.

## Adding the principle

- Ross' paradox revives in Stit with an additional assumption

Ought implies ability, i.e.  
the commanded action can be carried out

- While the stit theory solves paradoxes of deontic logic, addition of an intuitively natural assumption brings a strengthened version of Ross' paradox.

## Proposal

- Suggestion: Mere non-monotonicity does not work
- Partitionistic elimination solves the stronger version.
  - Stit: existence of no-case is enough
  - Partitionistic: choice must coincide with the given specification



## Modal logic of partitions

- Axiomatizable, decidable (Murakami 2005)
- Various applications are expected
  - “Hole” in game theory
  - Assertion
- Compare with:
  - PDL, Boolean modal logics

## Need philosophical examination

- Metaphysical status of possible worlds
- Information and action
- Moral theory

## References: stit and other logics

- Belnap et al. (2001) *Facing the Future*. OUP.
- Horty (2001) *Agency and Deontic Logic*. OUP.
- Murakami (2005) Modal Logic of Partitions. Dissertation, Indiana U.

## References: ethical robots

- [Toward Ethical Robots via Mechanized Deontic Logic](#) K Arkoudas, S Bringsjord, P Bello - Machine Ethics: Papers from the AAAI Fall Symp, 2005
- [Toward a General Logicist Methodology for Engineering Ethically Correct Robots](#) S Bringsjord, K Arkoudas, P Bello - IEEE Intelligent Systems, 2006