SOCREAL 2010

Proceedings of the
2nd International Workshop
on
Philosophy and Ethics of
Social Reality

27–28 March 2010
Hokkaido University
Sapporo, JAPAN

Co-Chairs
Johan van Benthem
Tomoyuki Yamada

Table of Contents

Preface

In the past two decades, a number of logics and game theoretical analyses have been proposed and combined to model various aspects of social interaction among agents including individual agents, organizations, and individuals representing organizations. The aim of SOCREAL Workshop is to bring together researchers working on diverse aspects of such interaction in logic, philosophy, ethics, computer science, cognitive science and related fields in order to share issues, ideas, techniques, and results.

The first SOCREAL Workshop was held in 9 - 10 March 2007 under the auspices of GPAE (Graduate Program in Applied Ethics, Graduate School of Letters, Hokkaido University) sponsored by the Ministry of Education, Culture, Sports, Science and Technology (MEXT). Building upon the success of SOCREAL 2007, its second edition, SOCREAL 2010, will be held under the auspices of CAEP (Center for Applied Ethics and Philosophy, Graduate School of Letters, Hokkaido University).

SOCREAL 2010 will consist of lectures by invited speakers and presentations of submitted papers. The present volume contains extended abstracts of the contributed papers to be presented at SOCREAL 2010. These papers are selected out of about 30 submitted papers. We thank all the researchers who had submitted their papers for their interest in SOCREAL 2010, and the members of program committee for their reviews. We also thank CAEP and Graduate School of Letters of Hokkaido University for their support.

Johan van Benthem
Tomoyuki Yamada

The Program

DAY 1: 27 March 2010

09:30 - 09:50 Registration

09:50 - 10:00 Opening

Keynote Lecture 1

10:00 - 11:00 Johan van Benthem
(University of Amsterdam, Netherlands, Stanford University, USA)
Joining Information and Evaluation: a dynamic logical perspective

Session 1: Obligations and Preferences

11:15 - 12:00 Allard Tamminga (University of Groningen, Netherlands)
Nash Equilibria in Multi-agent Deontic Logic
12:15 - 13:00 Satoru Suzuki (Komazawa University, Japan)
Measurement-Theoretic Foundation of Threshold Utility
Maximiser's Preference Logic

13:00 - 15:00 Lunch Break

Session 2: Normative Systems

15:00 - 15:45 Berislav Žarnić (University of Split, Croatia)
A Logical Typology of Normative Systems
16:00 - 16:45 Yasuo Nakayama (Osaka University, Japan)
Logical Framework for Normative Systems

Keynote Lecture 2

17:00 - 18:00 Fenrong Liu (Tsinghua University, China)
Understanding Deontics from a Preference Perspective

DAY 2: 28 March 2010

Session 3: Responsibility and Collective Agency

10:00 - 10:45 Matthew Braham & Martin van Hees
(University of Groningen, Netherlands)
Responsibility in Games
11:00 - 11:45 Biswanath Swain
(Indian Institute of Technology Kanpur, India)
Is Intention sufficient to explicate Collective Agency?
12:00 - 12:45 Ryoji Fujimoto & Choi Chang-bong
(Hokkaido University, Japan)
Interpretation of Action and Sociality of Action

12:45 - 14:45 Lunch Break

Session 4: Argumentations and Speech Acts

14:45 - 15:30 Hajime Sawamura (Niigata University, Japan)
Syncretic Argumentation by means of Lattice Homomorphism and Fusion
15:45 - 16:30 Miranda del Corral
(Universidad Nacional de Educación a Distancia, Spain)
Social Commitments: Expectations, Obligations and Entitlements

Keynote Lecture 3

16:45 - 17:45 Tomoyuki Yamada (Hokkaido University, Japan)
Scorekeeping and Dynamic Logics of Speech Acts

18:00 - 18:10 Closing

WORKSHOP CO-CHAIRS

Johan van Benthem
(University of Amsterdam, The Netherlands, and Stanford University, USA)
Tomoyuki Yamada (Hokkaido University, Japan)

PROGRAM COMMITTEE

Johan van Benthem
(University of Amsterdam, The Netherlands, and Stanford University, USA)
Jose Carmo (Universidade da Madeira, Portugal)
Fenrong Liu (Tsinghua University, China)
Jun Miyoshi (Kanto Gakuin University, Japan)
Yuko Murakami (Tohoku University, Japan)
Yasuo Nakayama (Osaka University, Japan)
Manuel Rebuschi (Nancy University, France)
Nobuyuki Takahashi (Hokkaido University, Japan)
Allard Tamminga (Rijksuniversiteit Groningen, The Netherlands)
Tomoyuki Yamada (Hokkaido University, Japan)
Berislav Žarnić (University of Split, Croatia)

LOCAL ORGANIZING COMMITTEE

Nobuo Kurata (Hokkaido University, Japan)
Shunzo Majima (Hokkaido University, Japan)
Koji Nakatogawa (Hokkaido University, Japan)
Yoshihiko Ono (Hokkaido University, Japan)
Tomoyuki Yamada (Hokkaido University)

Nash Equilibria in Multi-agent Deontic Logic

Allard Tamminga[*]

Abstract

We develop a multi-agent deontic action logic to study the logical behaviour of two types of permissions: (1) absolute permissions, having the form "In group $\mathcal{F}$'s interest, group $\mathcal{G}$ may to perform action $\alpha_{\mathcal{G}}$" and (2) conditional permissions, having the form "If group $\mathcal{H}$ were to perform action $\alpha_{\mathcal{H}}$, then, in group $\mathcal{F}$'s interest, group $\mathcal{G}$ may perform action $\alpha_{\mathcal{G}}$". First, we define a formal language for multi-agent deontic action logic and a class of consequentialist models to interpret the formulas of the language. Second, we define a transformation that converts any strategic game into a consequentialist model. Third, we show that an outcome a^* is a Nash equilibrium of a strategic game if and only if a conjunction of certain conditional permissions is true in the consequentialist model that results from the transformation of that strategic game.

1 Introduction

Deontic logic concerns the formal study of obligations, permissions, and prohibitions. Since its inception, it has been a lively and fruitful branch of philosophical logic. It has long been largely confined, however, to the formal study of norms within single-agent or even agentless contexts. The recent development of a multi-agent logic of agency has finally made it possible to transpose deontic logic from single-agent to multi-agent settings. As a result, *multi-agent deontic logic* studies obligations, permissions, and prohibitions within the context of formal models of strategic interaction between (groups of) agents with different preferences.

Strategic interaction between (coalitions of) players with different preferences is, of course, also studied in *game theory*. Hence, multi-agent deontic logic and game theory both study multi-agent phenomena that are largely comparable, athough they approach them from widely diverging perspectives: whereas deontic logicians concentrate on the formal structure of moral obligations, game theorists focus on the mathematics of canons of instrumental rationality. The question of how to establish connections between this new multi-agent deontic logic and game theory is therefore both natural and pressing.

To bring deontic logic and game theory together conceptually, we have to be more explicit on (1) the type of moral theory that gives rise to the obligations and permissions we set out to formalize, and (2) the type of preferences that figure in our deontic logic as the evaluative basis for the moral rightness of actions. Let us address the latter point first. At first sight, obligatory actions and preferred actions are worlds apart: it is perfectly possible that I have the obligation to do X, but at the same time prefer not to do X. Things begin to look different, however, as soon as we make a distinction

[*] Faculty of Philosophy, University of Groningen, The Netherlands, A.M.Tamminga@rug.nl.

between *extrinsic* preferences (which are the result of a previous judgment of betterness on the basis of reasons) and *intrinsic* preferences (which reflect the unreasoned subjective likings of the agents concerned) – see (von Wright 1963, p. 14). Now, given the distinction between extrinsic and intrinsic preferences, it still makes perfect sense that I have the obligation to do X and at the same time *intrinsically* prefer not to do X (I just don't feel like it). Some intellectual effort is needed, however, to imagine a situation where I have the obligation to do X and at the same time *extrinsically* prefer not to do X. Hence, a first step in bringing deontic logic and game theory together, is to assume that the preferences that figure in our deontic logic are extrinsic.

To make the conceptual match between deontic logic and game theory even closer, we also have to be specific on the type of moral theory that gives rise to the obligations and permissions we aim to formalize. From a deontological perspective, it still might be that I have the obligation to do X (I promised to do so) and at the same time extrinsically prefer not to do X. This possibility is minimized once we adopt a version of *act consequentialism* as our moral theory. In evaluative act consequentialism, the moral rightness of an action only depends on the value of its consequences. It will be seen below that the modelling of obligations by way of a formal framework inspired by evaluative act consequentialism using extrinsic preferences makes it plausible that an action is obligatory if and only if that action is extrinsically preferred. The consequentialist models we shall use to interpret the formulas of multi-agent deontic action logic thus establish a strong conceptual bond between deontic logic and game theory. The main purpose of this paper, however, is to establish formal connections between multi-agent deontic action logic and game theory.

The set-up of the paper is as follows. First, we define a formal language for multi-agent deontic action logic and a class of consequentialist models to formally interpret the formulas of that language. Second, we give standard definitions of strategic games and Nash equilibria. Third, we define a transformation $\mathfrak{T}$ and a valuation function v that convert any strategic game G into a consequentialist model $\langle \mathfrak{T}(G), v\rangle$. Fourth, we show, as a benchmark case for establishing formal connections between deontic logic and game theory, that an outcome a^* is a Nash equilibrium of strategic game G if and only if a finite conjunction of certain conditional permissions is true in the consequentialist model $\langle \mathfrak{T}(G), v\rangle$.

2 Multi-agent Deontic Logic

Our present deontic logic studies the logical behavior of two types of permissions: (1) *absolute permissions* of the form "In group $\mathcal{F}$'s interest, group $\mathcal{G}$ may to perform action $\alpha_\mathcal{G}$" (abbreviated as $\mathsf{P}^{\mathcal{F}}_{\mathcal{G}}\alpha_\mathcal{G}$) and (2) *conditional permissions* of the form "If group $\mathcal{H}$ were to perform action $\alpha_\mathcal{H}$, then, in group $\mathcal{F}$'s interest, group $\mathcal{G}$ may perform action $\alpha_\mathcal{G}$" (abbreviated as $\mathsf{P}^{\mathcal{F}}_{\mathcal{G}}(\alpha_\mathcal{G}/\alpha_\mathcal{H})$).

2.1 Language

Our modal language $\mathfrak{L}$ is built from a countable set $\mathfrak{A} = \{\alpha^n_\mathcal{G} : \mathcal{G} \subseteq \mathcal{N} \text{ and } n \in \mathbb{N}\}$ of atomic propositions, where $\mathcal{N}$ is a finite set of agents and $\mathbb{N}$ is the set of natural numbers. Thus, for each group $\mathcal{G}$ of agents there is a countable set $\mathfrak{A}_\mathcal{G} = \{\alpha^1_\mathcal{G}, \alpha^2_\mathcal{G}, \ldots\}$ of atomic propositions. We use $\alpha_\mathcal{G}$ and $\alpha_\mathcal{H}$ as variables for atomic propositions in $\mathfrak{A}$. The formal language $\mathfrak{L}$ is the smallest set satisfying the conditions (i) through (vi):

(i) $\mathfrak{A} \subseteq \mathfrak{L}$
(ii) If $\varphi \in \mathfrak{L}$, then $\neg\varphi \in \mathfrak{L}$
(iii) If $\varphi, \psi \in \mathfrak{L}$, then $\varphi \wedge \psi \in \mathfrak{L}$
(iv) If $\varphi \in \mathfrak{L}$, then $\Diamond\varphi \in \mathfrak{L}$
(v) If $\alpha_{\mathcal{G}} \in \mathfrak{A}$ and $\mathcal{F} \subseteq \mathcal{N}$, then $\mathsf{P}^{\mathcal{F}}_{\mathcal{G}}\alpha_{\mathcal{G}} \in \mathfrak{L}$
(vi) If $\alpha_{\mathcal{G}}, \alpha_{\mathcal{H}} \in \mathfrak{A}$ and $\mathcal{F} \subseteq \mathcal{N}$ and $\mathcal{H} \subseteq \mathcal{N} - \mathcal{G}$, then $\mathsf{P}^{\mathcal{F}}_{\mathcal{G}}(\alpha_{\mathcal{G}}/\alpha_{\mathcal{H}}) \in \mathfrak{L}$.

We interpret the formulas in $\mathfrak{L}$ in terms of consequentialist models.

2.2 Consequentialist Models

Consequentialist models are Kripke-style possible worlds models, built from a non-empty set of possible worlds and a finite set of agents. For the sake of formal simplicity, we adopt a reductionist stand on group actions, that is, we define group actions straightforwardly in terms of actions of the individual agents who are the group's members. Each group of agents is assigned its own choice set of options for acting. A group of agents performs an action by choosing an option from its choice set. Each choice set is modelled as a partition of the total set of possible worlds, and hence a group of agents performs an action by restricting the total set of possible worlds to those worlds that are elements of the option that corresponds to the action being performed. Each group of agents has its own preference relation over the total set of possible worlds. These preference relations guide a group of agents in choosing the most advantageous option(s) from its choice set.[1] The following definitions make these ideas precise:

Definition 1 A *consequentialist frame* $\mathfrak{F}$ is a quadruple $\langle \mathcal{W}, \mathcal{N}, Choice, (\succeq_{\mathcal{F}})\rangle$, where $\mathcal{W}$ is a non-empty set of possible worlds, $\mathcal{N}$ is a finite set of agents, $Choice$ is a choice function, and $\succeq_{\mathcal{F}}$ is a reflexive, transitive, and complete relation on $\mathcal{W}$ for each $\mathcal{F} \subseteq \mathcal{N}$.

Choice sets of *individual agents* are given by a function $Choice : \mathcal{N} \to \wp(\wp(\mathcal{W}))$ that meets two conditions: (1) for each individual agent $i \in \mathcal{N}$ it holds that $Choice(i)$ is a partition of $\mathcal{W}$, and (2) for each selection function s assigning to each individual agent $i \in \mathcal{N}$ a set of possible worlds $s(i)$ such that $s(i) \in Choice(i)$ it holds that $\bigcap_{i \in \mathcal{N}} s(i)$ is non-empty.

Next, we extend the choice function for individual agents to a function $Choice : \wp(\mathcal{N}) \to \wp(\wp(\mathcal{W}))$ for *groups of agents*. Let $Select$ be the set of all selection functions s assigning to each individual agent $i \in \mathcal{N}$ an option $s(i) \in Choice(i)$. Then

$$Choice(\mathcal{G}) = \{\bigcap_{i \in \mathcal{G}} s(i) : s \in Select\},$$

if $\mathcal{G}$ is non-empty. Otherwise, $Choice(\mathcal{G}) = \{\mathcal{W}\}$.

Definition 2 A *consequentialist model* $\mathfrak{M}$ is an ordered pair $\langle \mathfrak{F}, v\rangle$, where $\mathfrak{F}$ is a consequentialist frame and v a valuation function that for each $\mathcal{G} \subseteq \mathcal{N}$ assigns to each atomic proposition $\alpha_{\mathcal{G}} \in \mathfrak{A}$ an action $K \in Choice(\mathcal{G})$.

[1] Our consequentialist models are closely related to the models set forth in (Horty 2001; Kooi & Tamminga 2008).

2.3 Absolute and Conditional $\mathcal{F}$-Dominance

In search of a formal interpretation of absolute and conditional permissions, we start from Apostel's dictum that "an act is permissible if it can be considered as the application of a strategy such that there is no better one (there may be many equally good)" (Apostel 1960, p. 75). In the present context, this means that for each group $\mathcal{G}$ of agents we need to order the actions available to it in terms of the preference relations $\succeq_{\mathcal{F}}$ over the total set of possible worlds. Hence, we need to transform each preference relation $\succeq_{\mathcal{F}}$ into an $\mathcal{F}$-ordering of the set $Choice(\mathcal{G})$. We adopt the notion of weak dominance and adapt it to the present situation.

Definition 3 Let $\mathfrak{F}$ be a consequentialist frame. Let $\mathcal{F}, \mathcal{G} \subseteq \mathcal{N}$ and $\mathcal{H} \subseteq \mathcal{N} - \mathcal{G}$. Let $K, K' \in Choice(\mathcal{G})$ and $L \in Choice(\mathcal{H})$. Then

$K \geq^{\mathcal{F}}_{\mathcal{G}} K'$ iff for all $S \in Choice(\mathcal{N} - \mathcal{G})$ and for all $w, w' \in \mathcal{W}$ it holds that if $w \in K \cap S$ and $w' \in K' \cap S$, then $w \succeq_{\mathcal{F}} w'$.

$K \geq^{\mathcal{F}}_{(\mathcal{G}/\mathcal{H},L)} K'$ iff for all $S \in Choice((\mathcal{N} - \mathcal{G}) - \mathcal{H})$ and for all $w, w' \in \mathcal{W}$ it holds that if $w \in K \cap L \cap S$ and $w' \in K' \cap L \cap S$, then $w \succeq_{\mathcal{F}} w'$.

2.4 Semantics

Now that we have defined the notions of a consequentialist model, of absolute and conditional $\mathcal{F}$-dominance, we are in a position to provide the semantical rules to interpret the formulas in $\mathfrak{L}$.

Definition 4 (Semantical Rules) Let $\mathfrak{M} = \langle \mathfrak{F}, v \rangle$ be a consequentialist model. Let $\mathcal{F}, \mathcal{G} \subseteq \mathcal{N}$ and let $\mathcal{H} \subseteq \mathcal{N} - \mathcal{G}$. Let $w \in \mathcal{W}$ and let $\alpha_{\mathcal{G}}, \alpha_{\mathcal{H}} \in \mathfrak{A}$ and $\varphi, \psi \in \mathfrak{L}$. Then

$\mathfrak{M}, w \models \alpha_{\mathcal{G}}$ iff $w \in v(\alpha_{\mathcal{G}})$
$\mathfrak{M}, w \models \neg\varphi$ iff $\mathfrak{M}, w \not\models \varphi$
$\mathfrak{M}, w \models \varphi \wedge \psi$ iff $\mathfrak{M}, w \models \varphi$ and $\mathfrak{M}, w \models \psi$
$\mathfrak{M}, w \models \Diamond\varphi$ iff there is a w' such that $\mathfrak{M}, w' \models \varphi$
$\mathfrak{M}, w \models \mathsf{P}^{\mathcal{F}}_{\mathcal{G}}\alpha_{\mathcal{G}}$ iff for all K in $Choice(\mathcal{G})$ with $K \neq v(\alpha_{\mathcal{G}})$ it holds that $v(\alpha_{\mathcal{G}}) \geq^{\mathcal{F}}_{\mathcal{G}} K$
$\mathfrak{M}, w \models \mathsf{P}^{\mathcal{F}}_{\mathcal{G}}(\alpha_{\mathcal{G}}/\alpha_{\mathcal{H}})$ iff for all K in $Choice(\mathcal{G})$ with $K \neq v(\alpha_{\mathcal{G}})$ it holds that $v(\alpha_{\mathcal{G}}) \geq^{\mathcal{F}}_{(\mathcal{G}/\mathcal{H},v(\alpha_{\mathcal{H}}))} K$.

We write $\mathfrak{M} \models \varphi$, if for all possible worlds w in $\mathcal{W}$ it holds that $\mathfrak{M}, w \models \varphi$.

It should be noted that our formal semantics provides truth conditions for a wide variety of conditional permissions. We can distinguish at least four different types of conditional permissions: (1) conditional permissions where $\mathcal{G}$ is a non-singleton group of agents, (2) conditional permissions where $\mathcal{G}$ and $\mathcal{H}$ do not partition the grand coalition, (3) conditional permissions where the acting group $\mathcal{G}$ does not coincide with the interest group $\mathcal{F}$, and (4) conditional permissions of the form $\mathsf{P}^{i}_{i}(\alpha_i/\alpha_{\mathcal{N}-i})$. We only need the last type to characterize Nash equilibria of strategic games.

3 Nash Equilibria of Strategic Games

The following definitions of strategic games and Nash equilibria are provided by (Osborne & Rubinstein 1994). We also adopt their notational conventions.[2]

[2] See (Osborne & Rubinstein 1994, Section 1.7).

Definition 5 A *strategic game* G is a triple $\langle N, (A_i), (\succsim_i)\rangle$, where N is a finite set of players, for each player $i \in N$ it holds that A_i is a non-empty set of actions available to player i, and for each player $i \in N$ it holds that $\succsim_i$ is a preference relation on the set of outcomes $A = \times_{i \in N} A_i$.

We assume each A_i to be finite or countably infinite. Preference relations $\succsim_i$ are assumed to be reflexive, transitive, and complete. We use a_i and a_i^* as variables for actions in A_i. Likewise, a and a^* are variables for outcomes in A.

Given a strategic game $\langle N, (A_i), (\succsim_i)\rangle$, for each non-empty coalition $\mathcal{G} \subseteq N$ we define the set $A_\mathcal{G}$ of actions available to coalition $\mathcal{G}$ as $A_\mathcal{G} = \times_{i \in \mathcal{G}} A_i$. We use $a_\mathcal{G}$ and $a_\mathcal{G}^*$ as variables for actions in $A_\mathcal{G}$.

Definition 6 An outcome $a^* \in A$ is a *Nash equilibrium* of a strategic game $G = \langle N, (A_i), (\succsim_i)\rangle$ if and only if for each player $i \in N$ it holds that

$$(a_{-i}^*, a_i^*) \succsim_i (a_{-i}^*, a_i) \text{ for all } a_i \in A_i.$$

3.1 From Strategic Games to Consequentialist Models

Any strategic game can be converted into a consequentialist model. We first define a transformation $\mathfrak{T}$ that converts any strategic game G into a consequentialist frame $\mathfrak{T}(G)$. To obtain an appropriate consequentialist model $\langle \mathfrak{T}(G), v\rangle$ from this frame, we then define a suitable valuation function v.

Definition 7 Let $G = \langle N, (A_i), (\succsim_i)\rangle$ be a strategic game. The quadruple $\mathfrak{T}(G) = \langle \mathcal{W}, \mathcal{N}, \mathit{Choice}, (\succeq_\mathcal{F})\rangle$ is defined as follows:

(i) $\mathcal{W} = A$

(ii) $\mathcal{N} = N$

(iii) $\mathit{Choice}(\mathcal{G}) = \begin{cases} \{\{(a_\mathcal{G}, a_{-\mathcal{G}}) \in A : a_{-\mathcal{G}} \in A_{-\mathcal{G}}\} : a_\mathcal{G} \in A_\mathcal{G}\}, & \text{if } \mathcal{G} \neq \emptyset \\ \{\mathcal{W}\}, & \text{otherwise} \end{cases}$

(iv) $\succeq_\mathcal{F} = \begin{cases} \succsim_i, & \text{if } \mathcal{F} = \{i\} \\ \mathcal{W} \times \mathcal{W}, & \text{otherwise.} \end{cases}$

The operator $\mathfrak{T}$ transforms any strategic game into a consequentialist frame:

Theorem 1 Let G be a strategic game. Then $\mathfrak{T}(G)$ is a consequentialist frame.

We now must define a valuation function v to obtain a consequentialist model $\langle \mathfrak{T}(G), v\rangle$. To establish a formal connection between Nash equilibria and conditional permissions, we need to keep track of which atomic proposition $\alpha_\mathcal{G}$ in $\mathfrak{A}_\mathcal{G}$ is validated by the performance of which action $a_\mathcal{G}$ in $A_\mathcal{G}$.

To ensure this, we use an injective map f that for each $\mathcal{G} \subseteq \mathcal{N}$ assigns to each action $a_\mathcal{G}$ in each $A_\mathcal{G}$ an atomic proposition $\alpha_\mathcal{G}$ in $\mathfrak{A}_\mathcal{G}$. If there is an action $a_\mathcal{G}$ in $A_\mathcal{G}$ such that $f(a_\mathcal{G}) = \alpha_\mathcal{G}$, then we define $v_f(\alpha_\mathcal{G}) = \{(a_\mathcal{G}, a_{-\mathcal{G}}) \in A : a_{-\mathcal{G}} \in A_{-\mathcal{G}}\}$ (note that $a_\mathcal{G}$ is unique, since f is injective). If there is no action $a_\mathcal{G}$ in $A_\mathcal{G}$ such that $f(a_\mathcal{G}) = \alpha_\mathcal{G}$, then we simply put $v_f(\alpha_\mathcal{G}) = K$ for some unique designated $K \in \mathit{Choice}(\mathcal{G})$. Any valuation function v_f for $\mathfrak{T}(G)$ that is based on such an injection f will henceforth be called a *suitable valuation function*.

4 Nash Equilibria and Conditional Permissions

Conditional permissions enable us to give a formal characterization of Nash equilibria of strategic games in terms of conditional permissions:

Theorem 2 Let G be a strategic game and let v_f be a suitable valuation function for $\mathfrak{T}(G)$. Then

$$a^* \text{ is a Nash equilibrium of } G \quad \text{iff} \quad \langle \mathfrak{T}(G), v_f \rangle \models \textstyle\bigwedge_{i \in \mathcal{N}} \mathsf{P}^i_i(f(a^*_i)/f(a^*_{-i})).$$

References

Apostel, L. (1960). Game theory and the interpretation of deontic logic. *Logique & Analyse*, 3, 70–90.

Belnap, N., M. Perloff, & M. Xu (2001). *Facing the Future*. New York: Oxford University Press.

Broersen, J., R. Mastop, J.-J. C. Meyer, & P. Turrini (2008). A deontic logic for socially optimal norms. In R. van der Meyden & L. van der Torre (eds.). *Deontic Logic in Computer Science* (pp. 218–232). Berlin: Springer.

Horty, J. F. (2001). *Agency and Deontic Logic*. New York: Oxford University Press.

Kooi, B. P. & A. M. Tamminga (2008). Moral conflicts between groups of agents. *Journal of Philosophical Logic*, 37, 1–21.

Osborne, M. J. & A. Rubinstein (1994). *A Course in Game Theory*. Cambridge, MA: The MIT Press.

von Wright, G. H. (1963). *The Logic of Preference*. Edinburgh: Edinburgh University Press.

Measurement-Theoretic Foundation of Threshold Utility Maximiser's Preference Logic (Extended Abstract)

Satoru Suzuki
Faculty of Arts and Sciences, Komazawa University
1-23-1, Komazawa, Setagaya-ku, Tokyo, 154-8525 Japan
bxs05253@nifty.com

Abstract

The first problem of this paper is as follows: what kind of logic can formalise inferences in which the Sorites Paradox in preference can be avoided? (Formalisation Problem) The aim of this paper is to propose a new version of complete and decidable extrinsic preference logic–threshold utility maximiser's preference logic (TUMPL) that can solve the Formalisation Problem. Generally, preference logics are in danger of inviting the following problem: almost every principle which has been proposed as fundamental to one preference logic has been rejected by another one. (Fundamental Problem of Intrinsic Preference) The Scott-Suppes theorem in measurement theory enables TUMPL to avoid the Fundamental Problem of Intrinsic Preference.

Key Words: preference logic, semiorder, Sorites Paradox, threshold utility maximisation, bounded rationality, measurement theory, representation theorem.

The economist Armstrong ([1]) was one of the first to argue that *indifference* is not always *transitive.* Luce gave the following counterexample to the transitivity of indifference:

Example 1 (Avoidance of Sorites Paradox) *If indifference were transitive, then he would be unable to detect any weight differences, however great, which is patently false.... Find a subject who prefers a cup of coffee with one cube of sugar to one with five cubes.... Now prepare 401 cups of coffee with* $(1 + \frac{i}{100})x$ *grams of sugar,* $i = 0, 1, \dots, 400$*, where* x *is the weight of one cube of sugar. It is evident that he will be indifferent between cup* i *and cup* $i+1$*, for any* i*, but by choice he is not indifferent between* $i = 0$ *and* $i = 400$*. [[3]: 179]* ■

This example shows a situation where we would face the *Sorites Paradox* in preference if indifference were transitive. The first problem now arises:

Problem 1 (Formalisation Problem) *What kind of logic can formalise inferences in which the Sorites Paradox in preference can be avoided?* ■

We call it the *Formalisation Problem.* The aim of this paper is to propose a new version of complete and decidable *extrinsic* preference logic–*threshold utility maximiser's preference logic* (TUMPL) that can solve the Explanation Problem. In order to solve it, we resort to *measurement theory.*[1] There are two fundamental problems with measurement theory:

1. the representation problem–justifying the assignment of numbers to objects or propositions,

2. the uniqueness problem–specifying the transformation up to which this assignment is unique.

A solution to the former can be furnished by a *representation theorem*, which establishes that the chosen numerical system preserves the relations of the relational system. The standard model of economics is based on *global rationality* that requires an *optimising behavior. Utility maximisation* is a typical example of an optimising behavior. Cantor ([2]) proved the representation theorem for utility maximisation.

Theorem 1 (Representation for Utility Maximisation, Cantor ([2])) *Suppose* $\mathbf{A}$ *is a countable set and* $\succeq$ *is a binary relation on* $\mathbf{A}$. *Then* $\succeq$ *is a weak order (transitive and connected) iff there is a function* $u : \mathbf{A} \to \mathbb{R}$ *such that for any* $x, y \in \mathbf{A}$,

$$x \succeq y \text{ iff } u(x) \geq u(y).$$

■

But according to Simon ([9]), cognitive and information-processing constrains on the capabilities of agents, together with the complexity of their environment, render an optimising behavior an unattainable ideal. He dismissed the idea that agents should exhibit global rationality and suggested that they in fact exhibit *bounded rationality* that allows a *satisficing behavior.*[2] One explanation for Example 1 is that the nontransitivity of indifference results from the fact that we cannot generally discriminate very close quantities. The concept of a *semiorder* was introduced by Luce ([3]) to construct a model to interpret situations like Example 1 of nontransitive indifference with a *threshold* of discrimination. Scott and Suppes defined ([[6]: 117]) a semiorder as follows:

Definition 1 (Semiorder) $\succ$ *on* $\mathbf{A}$ *is called a semiorder if, for any* $w, x, y, z \in \mathbf{A}$, *the following conditions are satisfied:*

1. $x \not\succ x$. *(Irreflexivity),*

2. *If* $w \succ x$ *and* $y \succ z$, *then* $w \succ z$ *or* $y \succ x$. *(Intervality),*

3. *If* $w \succ x$ *and* $x \succ y$, *then* $w \succ z$ *or* $z \succ y$. *(Semitransitivity).*

■

[1] [5] gives a comprehensive survey of measurement theory.

[2] We learned from van Rooij ([10]) the relation between the Sorites Paradox and bounded rationality.

Threshold utility maximisation is a typical example of a satisficing behavior. Scott and Suppes ([6]) proved a representation theorem for threshold utility maximisation when $\mathbf{A}$ is *finite.*

Theorem 2 (Representation for Threshold Utility Maximisation, Scott and Suppes ([6])) *Suppose that $\succ$ is a binary relation on a finite set $\mathbf{A}$ and δ is a positive number. Then $\succ$ is a semiorder iff there is a function $u : \mathbf{A} \to \mathbb{R}$ such that for any $x, y \in \mathbf{A}$,*

$$x \succ y \textit{ iff } u(x) > u(y) + \delta.$$

■

Remark 1 *Scott ([7]) simplified the Scott-Suppes theorem in terms of the solvability of finite system of linear inequalities.* ■

In order to construct a model of TUMPL, we wish to use preference relations on Boolean algebras. Since A is an arbitrary finite set, the next corollary follows directly from Theorem 2.

Corollary 1 (Representation on Finite Boolean Algebra) *Suppose that $\mathbf{W}$ is a finite set of possible worlds and $\mathcal{F}$ is a finite Boolean algebra of subsets of $\mathbf{W}$ and $\succ$ is a binary relation on $\mathcal{F}$, and δ is a positive number. Then $\succ$ is a semiorder iff there is a function $u : \mathcal{F} \to \mathbb{R}$ such that for any $\alpha, \beta \in \mathcal{F}$,*

$$\alpha \succ \beta \textit{ iff } u(\alpha) > u(\beta) + \delta.$$

■

Remark 2 *Corollary 1 can guarantee that $\succ$ on $\mathcal{F}$ is a threshold utility maximiser's preference relation.* ■

Generally, preference logics are in danger of inviting the following problem. Von Wright ([11]) divided preferences into two categories: *extrinsic* and *intrinsic* preference. An agent is said to prefer φ_1 extrinsically to φ_2 if φ_1 is better than φ_2 in some explicit respect. So we can explain extrinsic preference from some explicit point of view. If we cannot explain preference from any explicit point of view, we call it intrinsic. Most preference logics that have been proposed are intrinsic but little attention has been paid to extrinsic preference. Von Wright ([12]) posed the following fundamental problem intrinsic preference logics faced.

Problem 2 (Fundamental Problem of Intrinsic Preference) *The development of a satisfactory logic of preference has turned out to be unexpectedly problematic. The evidence for this lies in the fact that almost every principle which has been proposed as fundamental to one preference logic has been rejected by another one.* ■

We call it the *Fundamental Problem of Intrinsic Preference.* According to Mullen ([4]), we can analyse the cause of the Fundamental Problem as follows. The adequacy criteria for intrinsic preference principles considered by preference logicians have been whether the principles are consistent with our *intuitions* of reasonableness. But these intuitions invite the Fundamental Problem. Different theories, such as ethics, welfare economics, consumer demand theory, game theory and decision theory make different demands upon the fundamental properties of preference. Preference logic should be constructed not from intuition but from a fixed *theory.* So preference logic should be *extrinsic.* When we provide TUMPL with a model based on semiorders, by virtue of Corollary 1, we can adopt *threshold utility maximisation* as a theory that makes demands upon the fundamental properties of preference, which can avoid the Fundamental Problem of Intrinsic Preference.

We define the language $\mathcal{L}_{\mathsf{TUMPL}}$ of TUMPL.

Definition 2 (Language) *Let* $\mathbf{S}$ *denote a set of sentential variables,* $\Box$ *a necessity operator,* $\mathbf{SPR}$ *a strict preference relation symbol. The language* $\mathcal{L}_{\mathsf{TUMPL}}$ *of* TUMPL *is given by the following rule:*

$$\varphi ::= s \mid \top \mid \neg\varphi \mid \varphi_1 \& \varphi_2 \mid \Box\varphi \mid \mathbf{SPR}(\varphi_1, \varphi_2),$$

where $s \in \mathbf{S}$*, and nestings of* $\mathbf{SPR}$ *do not occur.* $\bot, \vee, \rightarrow, \leftrightarrow$ *and* $\Diamond$ *are introduced by the standard definitions. Both an indifference relation symbol* $\mathbf{IND}$ *and a weak preference relation symbol* $\mathbf{WPR}$ *are also introduced by the standard definitions. The set of all well-formed formulae of* $\mathcal{L}_{\mathsf{TUMPL}}$ *will be denoted by* $\Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$*.* ■

We define a structured Kripke model $\mathcal{M}$ for TUMPL.

Definition 3 (Model) $\mathcal{M}$ *is a quadruple* $(\mathbf{W}, R, V, \rho)$*, where:*

- $\mathbf{W}$ *is a nonempty set of possible worlds,*
- R *is a binary relation on* $\mathbf{W}$*,*
- V *is a truth assignment to each* $s \in \mathbf{S}$ *for each* $w \in \mathbf{W}$*,*
- ρ *is a preference space assignment that assigns to each* $w \in \mathbf{W}$ *a preference space* $(\mathcal{F}_w, \succ_w)$ *such that* $\mathcal{F}_w$ *is a Boolean* σ*-algebra of subsets of* $\{w' \in \mathbf{W} : R(w, w')\}$ *and* $\succ_w$ *on* $\mathcal{F}_w$ *is a semiorder.*

■

We provide TUMPL with the following truth definition relative to $\mathcal{M}$:

Definition 4 (Truth) *The notion of* $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$ *being true at* $w \in W$ *in* $\mathcal{M}$*, in symbols* $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \varphi$ *is inductively defined as follows:*

- $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} s$ *iff* $V(w)(s) = \mathbf{true}$,
- $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \top$,
- $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \varphi_1 \& \varphi_2$ *iff* $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \varphi_1$ *and* $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \varphi_2$,
- $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \neg\varphi$ *iff* $(\mathcal{M}, w) \not\models_{\mathsf{TUMPL}} \varphi$,
- $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \Box\varphi$ *iff, for any* w' *such that* $R(w, w')$, $(\mathcal{M}, w') \models_{\mathsf{TUMPL}} \varphi$,
- $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \mathbf{SPR}(\varphi_1, \varphi_2)$ *iff* $[\![\varphi_1]\!]^{\mathcal{M}}_w \succ_w [\![\varphi_2]\!]^{\mathcal{M}}_w$,

where $[\![\varphi]\!]^{\mathcal{M}}_w := \{w' \in \mathbf{W} : R(w, w') \text{ and } (\mathcal{M}, w') \models_{\mathsf{TUMPL}} \varphi\}$*. If* $(\mathcal{M}, w) \models_{\mathsf{TUMPL}} \varphi$ *for all* $w \in \mathbf{W}$*, we write* $\mathcal{M} \models_{\mathsf{TUMPL}} \varphi$ *and say that* φ *is valid in* $\mathcal{M}$*. If* φ *is valid in all structured Kripke models for* TUMPL*, we write* $\models_{\mathsf{TUMPL}} \varphi$ *and say that* φ *is valid.* ■

We provide TUMPL with a proof system.

Definition 5 (Proof System) *The proof system of* TUMPL *consists of the following:*

1. *all tautologies of classical sentential logic,*
2. $\Box(\varphi_1 \to \varphi_2) \to (\Box\varphi_1 \to \Box\varphi_2) \qquad (K)$,
3. $\Box(\varphi_1 \leftrightarrow \varphi_2) \& \Box(\psi_1 \leftrightarrow \psi_2) \to (\mathbf{SPR}(\varphi_1, \psi_1) \leftrightarrow \mathbf{SPR}(\varphi_2, \psi_2))$ *(Replacement of Necessary Equivalents),*
4. $\neg\mathbf{SPR}(\varphi, \varphi)$ *(Syntactic Counterpart of Irreflexivity),*
5. $(\mathbf{SPR}(\varphi_1, \varphi_2) \wedge \mathbf{SPR}(\varphi_3, \varphi_4)) \to (\mathbf{SPR}(\varphi_1, \varphi_4) \vee \mathbf{SPR}(\varphi_3, \varphi_2))$ *(Syntactic Counterpart of Intervality),*
6. $(\mathbf{SPR}(\varphi_1, \varphi_2) \wedge \mathbf{SPR}(\varphi_2, \varphi_3)) \to (\mathbf{SPR}(\varphi_1, \varphi_4) \vee \mathbf{SPR}(\varphi_4, \varphi_3))$ *(Syntactic Counterpart of Semitransitivity),*
7. *Modus Ponens,*
8. *Necessitation.*

A proof of $\varphi \in \Phi_{\mathsf{TUMPL}}$ *is a finite sequence of* $\mathcal{L}_{\mathsf{TUMPL}}$*-formulae having* φ *as the last formula such that either each formula is an instance of an axiom, or it can be obtained from formulae that appear earlier in the sequence by applying an inference rule. If there is a proof of* φ*, we write* $\vdash_{\mathsf{TUMPL}} \varphi$*.* ■

We prove the metatheorems of TUMPL.

Theorem 3 (Soundness) *For any* $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$*, if* $\vdash_{\mathsf{TUMPL}} \varphi$*, then* $\models_{\mathsf{TUMPL}} \varphi$*.* ■

We prove the completeness of TUMPL by using the idea of Segerberg ([8]) that we modify *filtration theory* in such a way that completeness can be established by a *representation theorem* in measurement theory.

Theorem 4 (Completeness) *For any* $\varphi \in \Phi_{\mathcal{L}_{\mathsf{TUMPL}}}$*, if* $\models_{\mathsf{TUMPL}} \varphi$*, then* $\vdash_{\mathsf{TUMPL}} \varphi$*.* ■

Proof *We wish to outline the proof.*

1. *We define the canonical model* $\mathcal{U}^{C^-}$ *for the modal logical part of* TUMPL *and define a restriction* $\mathcal{U}$ *of* $\mathcal{U}^{C^-}$*.*
2. *We define a filtration* $\mathcal{U}^{\equiv}$ *of* $\mathcal{U}$*, where the universe* $\mathbf{W}^{\equiv}$ *of* $\mathcal{U}^{\equiv}$ *is finite.*
3. *We prove, in* $\mathcal{U}^{\equiv}$*, the Truth Lemma for all formulae of* $\mathcal{L}_{\mathsf{TUMPL}}$ *that do not contain* $\mathbf{SPR}$*.*
4. *We prove that, for any* $\xi \in \mathbf{W}^{\equiv}$*,* $\mathcal{F}_\xi$ *is a finite Boolean algebra of subsets of* $\{\eta \in \mathbf{W}^{\equiv} : R^{\equiv}(\xi, \eta)\}$
5. *We prove that, for any* φ, ψ *containing no* $\mathbf{SPR}$ *and any* $\xi \in \mathbf{W}^{\equiv}$*,* $[\![\varphi]\!]_\xi^{\mathcal{U}^{\equiv}} \succ_\xi [\![\psi]\!]_\xi^{\mathcal{U}^{\equiv}}$ *iff, for any* $\Gamma \in \xi$*,* $\mathbf{SPR}(\varphi, \psi) \in \Gamma$*.*

6. *We prove that, for any $\xi \in \mathbf{W}^{\equiv}$, $\succ_{\xi}$ on $\mathcal{F}_{\xi}$ satisfies irreflexivity, intervality and semitransitivity.*

7. *We define $\rho^{\equiv}$ as a preference space assignment that assigns to each $\xi \in \mathbf{W}^{\equiv}$ a preference space $(\mathcal{F}_{\xi}, \succ_{\xi})$ and define a model $\mathcal{U}^{\equiv}_{\sharp}$ for* TUMPL *having all entries of $\mathcal{U}^{\equiv}$ and $\rho^{\equiv}$.*

8. *We prove, in $\mathcal{U}^{\equiv}_{\sharp}$, the Truth Lemma for all formulae of $\mathcal{L}_{\mathsf{TUMPL}}$.*

■

We prove the decidability of TUMPL in terms of the finite model property that every non-theorem of TUMPL fails in a structured Kripke model for TUMPL with only finitely many elements.

Theorem 5 (Decidability) TUMPL *is decidable.* ■

References

[1] Armstrong, E. W.: The Determinateness of the Utility Function. Economic Journal **49** (1939) 453–467.

[2] Cantor, G.: Beiträge zur Begründung der Transfiniten Mengenlehre I. Mathematische Annalen **46** (1895) 481–512.

[3] Luce, D.: Semiorders and a Theory of Utility Discrimination. Econometrica **24** (1956) 178–191.

[4] Mullen, J. D.: Does the Logic of Preference Rest on a Mistake?. Metaphilosophy **10** (1979) 247–255.

[5] Roberts, F. S.: Measurement Theory. Addison-Wesley, Reading (1979).

[6] Scott, D. and Suppes, P.: Foundational Aspects of Theories of Measurement. Journal of Symbolic Logic **3** (1958) 113–128.

[7] Scott, D.: Measurement Structures and Linear Inequalities. Journal of Mathematical Psychology **1** (1964) 233–247.

[8] Segerberg, K.: Qualitative Probability in a Modal Setting. In: Fenstad, J. E. (ed.): Proceedings of the Second Scandinavian Logic Symposium. North-Holland, Amsterdam (1971) 341–352.

[9] Simon, H. A.: Models of Bounded Rationality. MIT Press, Cambridge, Mass. (1982).

[10] Van Rooij, R.: Revealed Preference and Satisficing Behavior. preprint.

[11] Von Wright, G. H.: The Logic of Preference. Edinburgh UP, Edinburgh (1963).

[12] Von Wright, G. H.: The Logic of Preference Reconsidered. Theory and Decision **3** (1972) 140–169.

A Logical Typology of Normative Systems

Berislav Žarnić
University of Split, Croatia

This paper gives a first order formalization of the proposal put forward by John Broome[1] [2] and develops a typology on that basis. The three-place code function $\mathrm{k} : \mathrm{S} \times \mathrm{A} \times W \longrightarrow \wp\mathcal{L}_{\mathrm{n}}$ delivers the set $\mathrm{k}_{\mathsf{s}}(i, w) \subseteq \mathcal{L}_{\mathrm{n}}$ of propositions in the normative language $\mathcal{L}_{\mathrm{n}}$ that a normative source $\mathsf{s} \in \mathrm{S}$ requires of an agent $i \in \mathrm{A}$ in a world $w \in W$. The value of the code function $\mathrm{k}_{\mathsf{s}}(i, w)$ will be termed the 'set of requirements'. The vocabulary of the normative language $\mathcal{L}_{\mathrm{n}}$ will contain modal operators for belief, B, desire, D, and intention, I. The worlds are construed as subsets of normative language $\mathcal{L}_{\mathrm{n}}$ which are maximal consistent in propositional logic. Possible worlds may violate the laws of modal logics of intentionality according to the philosophical thesis that the essence of the mental is to be subject to norms, not to conform to them (Zangwill [6]).

Definition 1 *The normative language $\mathcal{L}_{\mathrm{n}}$ is built over the base language of propositional logic $\mathcal{L}_{PL}$. Let $i \in A$, $\mathrm{X} = \mathrm{B}, \mathrm{D}, \mathrm{I}$, and $p \in \mathcal{L}_{PL}$*

$$\textit{Sentences of } \mathcal{L}_{\mathrm{n}} ::= p \mid [\mathrm{X}_i]\varphi \mid \neg\varphi \mid (\varphi \wedge \psi)$$

The set of quasi-literals is the set of propositional letters and their negations, and modal formulas and their negations.

The T axiom ($\Box p \rightarrow p$) poses a serious threat to this kind of modeling that keeps modality and world apart. If modalities obeying axiom T were allowed (e.g. epistemic or praxeologic), then possible worlds, being defined as maximal consistent sets in propositional logic, would become intuitively impossible[2]. Since the corresponding T axioms seem to constitute an important part of the meaning of verbs of knowledge and of action, epistemic and praxeologic modalities must be excluded from the language of norms $\mathcal{L}_{\mathrm{n}}$. Von Wright [4] defined 'content of a norm' as "that which ought to or may or must not be or be done". The normative language $\mathcal{L}_{\mathrm{n}}$ departs from von Wright's definition by taking norm-content to be *the psychological state or relation of psychological states that ought to or may or must not be present in the mind of the norm addressee on a particular occasion.* The reduction and the switch may seem

[1] "We must allow for the possibility that the requirements you are under depend on your circumstances. ... There is a set of worlds, at each of which propositions have a truth value. The values of all propositions at a particular world conform to the axioms of propositional calculus. For each source of requirements s, each person i and each world w, there is a set of propositions $\mathrm{k}_{\mathsf{s}}(i, w)$, which is to be interpreted as the set of things that s requires of i at w. Each proposition in the set is a required proposition. The function k_{s} from i and w to $\mathrm{k}_{\mathsf{s}}(i, w)$ I shall call s's *code* of requirements". (Broome [2], p. 14) The symbols in the citation have been changed to match the symbols used in this paper.

[2] For example, although $\{\neg p, [\mathrm{K}]_i p\}$ is pl-consistent set, we do not want to have it included in any world since no false sentence can be known to be true.

drastic but there is a rationale for it. The requirement that agent i knows that p could be replaced by $p \to [\mathrm{B}_i]p$; a required action to see to it that p could be replaced by the required intention, i.e. $[\mathrm{I}_i]p$.

In order to achieve technical clarity we define a first-order metanormative many-sorted language $\mathcal{L}_{\text{meta}}$ with the following extralogical **vocabulary** – individual constants for normative sources, agents and worlds: $\mathrm{s}, \mathrm{s}_1, \ldots, \mathrm{a}, \mathrm{a}_1, \ldots, \mathrm{v}, \mathrm{v}_1, \ldots$; function symbols for code of requirement, propositional logic consequence, and logic function: k^3, Cn^1, l^1; function symbols for the sentential forms: neg^1, conj^2, and a set of symbols of the type $\mathrm{mod}^1_{\mathrm{X}i}$; monadic predicate symbols expressing properties of being a normative source, an agent, a sentence in $\mathcal{L}_{\mathrm{n}}$, a possible world: $\mathrm{Sr}^1, Ag^1, \mathrm{Sen}^1, W^1$, and dyadic predicates expressing relations of an agent having i-th normative property (corresponding to i-th code of requirements) in a world, and relation of membership: $\mathrm{K}^2_{s_1}, \mathrm{K}^2_{s_2}, \ldots, \mathrm{K}^2_{s_n}, \in^2$. The **structures** $\mathfrak{M}_{\text{meta}} = \langle \mathrm{D}, \mathcal{I} \rangle$ are built over the domain $\mathrm{D} = \mathrm{S} \cup \mathrm{A} \cup \mathcal{L}_{\mathrm{n}} \cup \wp\mathcal{L}_{\mathrm{n}}$ where S and A are non-empty and disjoint sets, and $\mathcal{L}_{\mathrm{n}}$ is the set already defined (Definition 1). We use variables $w, w_1, \ldots$ to range over worlds; variables $p, p_1, \ldots, q, q_1, \ldots$ to range over sentences in $\mathcal{L}_{\mathrm{n}}$; variables $i, i_1, \ldots$ to range over agents; and variables $x, y, \ldots$ to range over everything. The shorthand notation for sentential form functions uses "Quine quotes", e.g. the shorthand notation for $\mathrm{neg}(x)$ is $\ulcorner \neg x \urcorner$. For the ease of reading, the universal closure of the formula will be notated by formula with free variables. The interpretation of the nonlogical vocabulary is almost straightforward. More complex cases are:

- interpretation of sentential form functions, which we introduce by the way of example – $\mathcal{I}(\mathrm{neg})$ is a function: $\mathcal{L}_{\mathrm{n}} \to \mathcal{L}_{\mathrm{n}}$ such that

$$\mathcal{I}(\mathrm{neg})([\![x]\!]_g^{\mathfrak{M}_{\text{meta}}}) = \begin{cases} \neg \frown [\![x]\!]_g^{\mathfrak{M}_{\text{meta}}} \text{ if } [\![x]\!]_g^{\mathfrak{M}_{\text{meta}}} \in \mathcal{L}_{\mathrm{n}}, \\ \text{undefined}, \quad \text{otherwise.} \end{cases}$$

  where g is an assignment function and $\frown$ is concatenation operation;

- interpretation of logic function l is function $\mathcal{I}(\mathrm{l}) : \wp\mathcal{L}_{\mathrm{n}} \to \wp\mathcal{L}_{\mathrm{n}}$ such that $\mathcal{I}(\mathrm{l})([\![x]\!]_g^{\mathfrak{M}_{\text{meta}}})$ is the set of all substitutional instances of the formula $[\![y]\!]_g^{\mathfrak{M}_{\text{meta}}} \in \mathcal{L}_{\mathrm{n}}$ for each $y \in x$;

- interpretation of consequence function Cn is a set of consequences in classical propositional logic for a given set, i.e.

$$\mathcal{I}(\mathrm{Cn})([\![x]\!]_g^{\mathfrak{M}_{\text{meta}}}) = \begin{cases} \{y \in \mathcal{L}_{\mathrm{n}} \mid [\![x]\!]_g^{\mathfrak{M}_{\text{meta}}} \vdash_{\mathrm{pl}} y\} \text{ if } [\![x]\!]_g^{\mathfrak{M}_{\text{meta}}} \subseteq \mathcal{L}_{\mathrm{n}}, \\ \text{undefined}, \quad \text{otherwise.} \end{cases}$$

Definitions 2 *Quantifications over different argument positions in the code function enable a number of interesting type distinctions, some of which will be introduced below using a $\mathcal{L}_{\text{meta}}$ formula in the definiens.*

- k_{s} *is a* pl-*congruent code iff* $\ulcorner p \leftrightarrow q \urcorner \in \mathrm{Cn}(\varnothing) \to (p \in \mathrm{k}_{\mathsf{s}}(i, w) \leftrightarrow q \in \mathrm{k}_{\mathsf{s}}(i, w))$*;*

- k_{s} *is a* pl-*consistent code iff* $\exists w_2 \, \mathrm{k}_{\mathsf{s}}(i, w_1) \subseteq w_2$*;*

- k_{s} *is an achievable code iff* $\exists w \, \mathrm{k}_{\mathsf{s}}(i, w) \subseteq w$*;*

- k_{s} *is a* pl-*deductively closed iff* $\mathrm{k}_{\mathsf{s}}(i, w) = \mathrm{Cn}(\mathrm{k}_{\mathsf{s}}(i, w))$*;*

- k_s *is a relativistic code iff* $\exists i \exists w_1 \exists w_2\, \mathrm{k}_\mathsf{s}(i, w_1) \neq \mathrm{k}_\mathsf{s}(i, w_2)$*;*
- *a code is absolute iff it is not relativistic;*
- k_s *is a socially consistent code iff* $\exists w_2\, \mathrm{k}_\mathsf{s}(i_1, w_1) \cup \mathrm{k}_\mathsf{s}(i_2, w_1) \subseteq w_2$*;*
- *codes* k_x *and* k_y *are realization-equivalent iff* $\mathrm{k}_\mathsf{x}(i, w) \subseteq w \leftrightarrow \mathrm{k}_\mathsf{y}(i, w) \subseteq w$*;*
- *codes* k_x *and* k_y *are compatible iff* $\exists w_2\, \mathrm{k}_\mathsf{x}(i, w_1) \cup \mathrm{k}_\mathsf{y}(i, w_1) \subseteq w_2$*.*

Consistent and deductively closed codes seem to play an important role in our understanding of the basic normative concepts. For example, deontic KD logic without iterated deontic modalities may be conceived as logic of the specific type of code, namely of consistent pl-deductively closed code.

Definition 3 *Let* $p \in \mathcal{L}_{PL}$ *be a formula of propositional logic:*

$$\textit{Formulas of } \mathcal{L}^{\mathrm{O}}_{\mathrm{KD}} ::= p \mid \mathrm{O}p \mid \mathrm{P}p \mid \neg\varphi \mid (\varphi \wedge \psi)$$

Let us introduce the translation τ^1 from the restricted language $\mathcal{L}^{\mathrm{O}}_{\mathrm{KD}}$ to the metanormative language $\mathcal{L}_{\mathrm{meta}}$, with $\mathrm{O}\,p$ and $\mathrm{P}\,p$ standing for 'i in v has s-obligation (s-permission) to p'.

Definition 4 *Function* τ *maps sentences from the fragment* $\mathcal{L}^{\mathrm{O}}_{\mathrm{KD}} \cap \mathcal{L}_{PL}$ *to the set of sentential variables and sentential function terms of* $\mathcal{L}_{\mathrm{meta}}$*:*

$$\begin{array}{rcll} \tau(l) & \in & \{p, p_1, \ldots, q, q_1, \ldots\} & \textit{for propositional letters } l \in \mathcal{L}_{PL} \\ \tau(\neg\varphi) & = & \ulcorner \neg\tau(\varphi) \urcorner & \\ \tau(\varphi \wedge \psi) & = & \ulcorner \tau(\varphi) \wedge \tau(\psi) \urcorner & \end{array}$$

Definition 5 *Translation* $\tau^1 : \mathcal{L}^{\mathrm{O}}_{\mathrm{KD}} \to \mathcal{L}_{\mathrm{meta}}$

$$\begin{array}{rcll} \tau^1(p) & = & \tau(p) \in \mathrm{v} & \textit{if } p \in \mathcal{L}_{PL} \\ \tau^1(\mathrm{O}\varphi) & = & \tau(\varphi) \in \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}) & \\ \tau^1(\mathrm{P}\varphi) & = & \tau(\neg\varphi) \notin \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}) & \\ \tau^1(\neg\varphi) & = & \neg\tau^1(\varphi) & \\ \tau^1(\varphi \wedge \psi) & = & \tau^1(\varphi) \wedge \tau^1(\psi) & \end{array}$$

The principles of the standard deontic logic[3] hold under the translation τ^1:

- "gaplessness" condition $\mathrm{P}p \vee \mathrm{O}\neg p$ translates to $\ulcorner \neg p \urcorner \notin \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}) \vee \ulcorner \neg p \urcorner \in \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v})$ and that property obviously holds for any set of requirements;
- K axiom becomes $\ulcorner p \to q \urcorner \in \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}) \to (p \in \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}) \to q \in \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}))$ and that property holds for any pl-deductively closed set;
- D axiom becomes $p \in \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v}) \to \ulcorner \neg p \urcorner \notin \mathrm{k}_\mathsf{s}(\mathrm{a}, \mathrm{v})$ and that is just another way of stating pl-consistency;

[3] "...classical deontic logic, on the descriptive interpretation of its formulas, pictures a gapless and contradiction-free system of norms". (Von Wright [5] p. 32)
According to our translation scheme von Wright's claim should be appended: classical deontic logic "pictures a system of norms" that is deductively closed too.

- mutual definability, $\mathrm{P}_1 p \leftrightarrow \neg\mathrm{O}\neg p$ holds if the set of requirements is congruent.

Although iterated deontic operators receive no translation in the scheme proposed above, one may extend the line of thought by giving additional translation rules for language of standard deontic $\mathcal{L}_{\mathrm{KD}}^{\mathrm{OO}}$ restricted to the maximum of two iterations of deontic operators, treating iterated deontic modalities as a sequence of heterogenous operators and introducing the distinction into the syntax:

$$\mathcal{L}_{\mathrm{KD}}^{\mathrm{O_2O}} ::= p \in \mathcal{L}_{\mathrm{KD}}^{\mathrm{O}} \mid \mathrm{O}_2 p \mid \mathrm{P}_2 p \mid \neg\varphi \mid (\varphi \wedge \psi)$$

Definition 6 *Let $Sub(\varphi)_{[\frac{c_1}{x_1} \dots \frac{c_n}{x_n}]}$ denote substitutional instance of $\varphi \in \mathcal{L}_{\mathrm{meta}}$ in which constants $c_1, ..., c_n$ are replaced by variables $x_1, ..., x_n$. Translation $\tau^2 : \mathcal{L}_{\mathrm{KD}}^{\mathrm{O_2O}} \rightarrow \mathcal{L}_{\mathrm{meta}}$*

$$\begin{array}{rcll} \tau^2(\mathrm{O}_2 p) & = & \forall i \forall w\ Sub(\tau^1(p))_{[\frac{\mathsf{a}}{i}\,\frac{\mathsf{v}}{w}]} & \quad \textit{for } p \in \mathcal{L}_{\mathrm{KD}}^{\mathrm{O}} \\ \tau^2(\mathrm{P}_2 p) & = & \exists i \exists w\ Sub(\tau^1(p))_{[\frac{\mathsf{a}}{i}\,\frac{\mathsf{v}}{w}]} & \quad \textit{for } p \in \mathcal{L}_{\mathrm{KD}}^{\mathrm{O}} \\ \tau^2(\neg\varphi) & = & \neg\tau^2(\varphi) & \\ \tau^2(\varphi \wedge \psi) & = & \tau^2(\varphi) \wedge \tau^2(\psi) & \end{array}$$

Such an approach to iterated deontic modalities departs from von Wright's [5] "second order descriptive interpretation" where e.g. O_2 would stand for existence of "normative demands on normative systems" ("norms for the norm givers"). The "first order" translation τ^1 as well as the "second order" translation τ^2 give us statements in metanormative language $\mathcal{L}_{\mathrm{meta}}$ both of which may "picture" some type of "normative system". The difference lies in the fact that τ^1 gives a local picture of a set of requirements (for a particular source, agent and world) while τ^2 gives a more global picture of a code function. In the second case the properties depicted are the properties of a code function for a particular source with respect to any agent and any world.

Let us consider KD45 deontic logic! The τ^2 translations of reinterpreted axioms 4, $\mathrm{O}_1 p \rightarrow \mathrm{O}_2\mathrm{O}_1 p$ and 5, $\mathrm{P}_1 p \rightarrow \mathrm{O}_2\mathrm{P}_1 p$ amount to stating that any s-obligation and any s-permission holds universally. So, the reinterpreted axioms will hold only if s-code is absolute.

Definition 7 *An agent i at world w has an "all-or-nothing" normative property K_{s} that corresponds to the source s iff the set of requirements $\mathrm{k}_{\mathsf{s}}(i, w)$ is satisfied in w, i.e. $\mathrm{K}_{\mathsf{s}}(i, w) \leftrightarrow \mathrm{k}_{\mathsf{s}}(i, w) \subseteq w$.*

If the only way to satisfy some relativistic code and some absolute code is to satisfy them simultaneously, then these codes define the same normative property. The question arises as to whether (non)absoluteness of a code function introduces a difference with respect to normative properties. The next theorem provides a negative answer.

Theorem 8 *For any code there is a realization equivalent absolute code.*

The proof requires extension of the normative language $\mathcal{L}_{\mathrm{n}}$ to the language $\mathcal{L}_{\mathrm{n}(\omega_1)}$ of a variant of infinitary logic which has the same vocabulary as $\mathcal{L}_{\mathrm{n}}$, but in $\mathcal{L}_{\mathrm{n}(\omega_1)}$ the conjunction symbol $\bigwedge$ may be applied to subsets of the set of *quasi-literals.* A function $\mathrm{k}_{\mathsf{s}}^{cond}$ is a conditionalized variant of a code k_{s} iff

$$\forall p \forall w_1 (p \in \mathrm{k}_{\mathsf{s}}^{cond}(i, w_1) \leftrightarrow \exists q \exists w_2 (p = \ulcorner \bigwedge \mathrm{lit}(w_2) \rightarrow q \urcorner \wedge q \in \mathrm{k}_{\mathsf{s}}(i, w_2)))$$

where $\mathrm{lit}(w_2)$ is the set of all *quasi-literals* belonging to w_2. The existence of conditionalized variant for any code proves the theorem. In the light of theorem 8, world and agent generalizing translation of axioms 4 and 5 do not introduce distinctions into logical typology of normative properties.

There are several plausible principles of intentionality and normativity: intentionality is normative, i.e. subjected to norms of different sources (e.g. [6]); rationality is one of the normative sources; some norms of rationality are based on logic of psychological modalities. If we accept these principles, then the codes that deliver some "logical" set of sentences deserve our attention. A number of authors take the closure under equivalence to be either unproblematic (e.g. [2]) or at least plausible minimal logical property of a code. In other words, the codes inherit some of the easily noticeable logical properties of the language in which norm-contents are stated. But then a question arises as to which properties are to be preserved in any code. E.g. if the truth-functional equivalence should be inherited, should not the modal congruence[4] be inherited as well, especially in the light of the widely accepted principle that propositions, and not sentences, are the objects of intentionality?

Definition 9 *Let $x \subseteq \mathcal{L}_{\mathrm{n}}$. The set of sentences $\mathrm{l}(x) \subseteq \mathcal{L}_{\mathrm{n}}$ is an axiomatic basis for a set of modal operators occurring in sentences in x ($\mathrm{l}(x)$ is the set of all the substitutional instances of sentences in the set x).*

Let us suppose that $\mathrm{l}(x)$ is also an axiomatic basis for the set of modal operators occurring in sentences in the sets of requirements delivered by k_{s}. Then we may distinguish several interesting types of codes that do not violate a logic of the modal part of its language:

- code is consistent with respect to $\mathrm{l}(x)$ iff $\exists w_2\ \mathrm{Cn}(\mathrm{l}(x) \cup \mathrm{k}_{\mathsf{s}}(i, w_1)) \subseteq w_2$;
- code is a logic iff $\exists x\ \mathrm{k}_{\mathsf{s}}(i, w) = \mathrm{Cn}(\mathrm{l}(x))$;
- code is "more than a logic" iff $\exists x \exists y(\neg y \subseteq \mathrm{Cn}(\mathrm{l}(x)) \wedge \mathrm{k}_{\mathsf{s}}(i, w) = \mathrm{Cn}(\mathrm{l}(x) \cup y))$;
- code is "less than a logic" iff $\exists x \exists y(\neg y \subseteq \mathrm{Cn}(\mathrm{l}(x)) \wedge \mathrm{k}_{\mathsf{s}}(i, w) = \mathrm{Cn}(\mathrm{l}(x) \cup y) - \mathrm{Cn}(\mathrm{l}(x)))$.

The second type of the logical code could be termed 'formal code', the third and the fourth — 'material codes'. All the four types exhibit some kind of "internal logicality".

One may distinguish two types of logical properties that a code may have. On the one hand, there are external properties of sets of requirements and code functions, like those given in the definitions 2. On the other hand, there is also an internal logicality of a code, pertaining to the modal logic of the code contents.

This approach relaxes the burden of unrealistic logical models of intentionality by their relocation to the normative side; e.g. it is nonsensical to attribute logical omniscience to real agents with finite resources available for reasoning, but one might argue that it is not nonsensical to consider logical omniscience as a normative requirement. The interaction between the normative and the real takes place on the level of agent properties. The widely accepted "ought

[4] If p and q are truth-functionally equivalent, then $[\mathrm{X}_i]p \in \mathrm{k}_{\mathsf{s}}(i, w)$ iff $[\mathrm{X}_i]q \in \mathrm{k}_{\mathsf{s}}(i, w)$.

implies can" principle holds if a code is achievable, i.e. if it is possible for an agent to have the normative property that corresponds to to the source of the code. A straightforward definition of the "all-or-nothing" normative property has been proposed by Broome (see definition 7 above). It is commonly held that rationality as a normative property is not all-or-nothing property but a matter of degree (e.g. Davidson [3]). Therefore, the set of requirements satisfied by an agent having the property of rationality need not include all the requirements delivered by rationality as a normative source. Consequently, the definition of achievability of the code should be modified for "extensive properties".

Further work. The typology of normative systems seems to need a supplementary typology of normative properties, most notably of those that are defined in terms of partial satisfaction. The motivation for the theory of belief revision came from legal context. AGM theory *inter alia* described the logical ways in which consistency of a theory should be maintained. The logical properties that define the state of equilibrium for "homeostatic dynamics" of normative codes should be determined. *Prima facie*, a number of other properties besides "external consistency" like social consistency, achievability, "internal consistency" should be included[5].

References

[1] Carlos E. Alchourrón and Eugenio Bulygin. The expressive conception of norms. In R. Hilpinen (ed.)*New Studies in Deontic Logic*, pp. 95–125, D. Reidel Publishing Company, Dordrecht, 1981.

[2] John Broome. Requirements. In T. Rønnow-Rasmussen, B. Petersson, J. Josefsson, and D. Egonsson, editors,*Homage a Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*, pages 1–41. Lunds universitet, Lund, 2007. `http://www.fil.lu.se/hommageawlodek`.

[3] Donald Davidson. *Problems of Rationality.* Clarendon Press, Oxford, 2004.

[4] Georg Henrik von Wright. *Norm and Action : A Logical Enquiry.* Routledge and Kegan Paul, London, 1963.

[5] Georg Henrik von Wright. Deontic logic: a personal view. *Ratio Juris*, **12**: 26–38, 1999.

[6] Nick Zangwill. The normativity of the mental. *Philosophical Explorations*, **8**: 1–19, 2005.

[5] The author would like to thank anonymous referees for helpful comments. The investigation presented in the paper is supported by MZOŠ project "Logical structures and intentionality".

Logical Framework for Normative Systems

Yasuo NAKAYAMA
Osaka University, Graduate School of Human Sciences
1-2 Yamada-oka, Suita, Osaka, Japan
nakayama@hus.osaka-u.ac.jp

In this paper, I propose a new logical framework that can be used to analyze normative phenomena in general. I call this framework a *Logic for Normative Systems* (LNS). I also demonstrate how to solve some paradoxes of Standard Deontic Logic (SDL). A characteristic of LNS is its dynamic behavior. LNS is flexible, hence it can be applied to describe complex normative problems including ethical problems.

1 Definition of Normative Systems

A normative system can be defined as follows, where $\vdash$ means the inference in the first-order logic:

(1a) Let T and OB be sets of sentences having the property that no sentence in OB follows from T. A pair $\langle T, OB\rangle$ consisting of *propositional system* T and *obligation space* OB is called a *normative system* (NS).

(1b) A sentence p belongs to the *propositional context* of normative system $\langle T, OB\rangle$ if and only if (*iff*) $T \vdash p$.

(1c) A sentence p belongs to the *obligation context* of normative system $\langle T, OB\rangle$ (abbreviated as $\mathbf{O}_{\langle T,OB\rangle}p$) *iff* $T \cup OB \vdash p$ & $T \nvdash p$.

(1d) A sentence p belongs to the *prohibition context* of normative system $\langle T, OB\rangle$ (abbreviated as $\mathbf{F}_{\langle T,OB\rangle}p$) *iff* $\mathbf{O}_{\langle T,OB\rangle}\neg p$.

(1e) A sentence p *belongs to the permission context of normative system* $\langle T, OB\rangle$ (abbreviated as $\mathbf{P}_{\langle T,OB\rangle}p$) *iff* $T \cup OB \cup \{p\} \nvdash \bot$ & $T \nvdash p$.

(1f) A group G has (*normative*) *power* of doing act_1 in $\langle T, OB\rangle$ *iff*
$T \vdash \forall x(member(x, G) \rightarrow agent(x))$ &
$\mathbf{P}_{\langle T,OB\rangle}\forall x(member(x, G) \rightarrow do(x, act_1))$ &
$\mathbf{F}_{\langle T,OB\rangle}\forall x(\neg member(x, G) \rightarrow do(x, act_1))$. [1]

These definitions presuppose that we insert *what we believe to be true* into the propositional system and *what we believe ought to be done* into the obligation space.

[1] The notion of (legal) *power* plays an essential role in Hart (1961). This shows that this notion is inevitable for describing legal systems. According to definition (1f), a group G has power of doing act_1 *iff* all members of G and only members of G are allowed to perform act_1.

From these definitions immediately follow the following three fundamental characterizations for LNS.

(2a) If $\mathbf{O}_{\langle T,OB\rangle}p$, then $\mathbf{P}_{\langle T,OB\rangle}p$. [2]

(2b) If $\mathbf{F}_{\langle T,OB\rangle}p$, then not $\mathbf{P}_{\langle T,OB\rangle}p$.

(2c) The propositional context of a NS is independent of its obligation space, while the obligation context depends on the propositional system.

To see how LNS works, let us consider an example.

(3) You should not kill any human beings. Peter is a human being. So you should not kill Peter.

According to our presupposition mentioned before, the first sentence in (3) expresses the content of one component of OB_1 and the second sentence expresses the content of one component in T_1 of normative system $\langle T_1, OB_1\rangle$:

(4a) $\forall x \forall y(agent(x) \wedge human(y) \rightarrow \neg kill(x,y))$ is an element of OB_1.

(4b) $human(Peter)$ is an element of T_1.

From (1a), (1c), and (4a) immediately follow (4c) and (4d), where (4c) corresponds to the conclusion of (3).

(4c) $\mathbf{O}_{\langle T_1,OB_1\rangle}\forall x(agent(x) \rightarrow \neg kill(x, Peter))$.

(4d) For any agent A, $\mathbf{O}_{\langle T_1,OB_1\rangle}\neg kill(A, Peter)$.

This result can be summarized as follows:

(4e) If $\{human(Peter), agent(A)\} \subseteq T_1$ & $\{\forall x \forall y(agent(x) \wedge human(y) \rightarrow \neg kill(x,y))\} \subseteq OB_1$, then $\mathbf{O}_{\langle T_1,OB_1\rangle}\forall x(agent(x) \rightarrow \neg kill(x, Peter))$ & $\mathbf{O}_{\langle T_1,OB_1\rangle}\neg kill(A, Peter)$ & $\mathbf{F}_{\langle T_1,OB_1\rangle}kill(A, Peter)$.

In this way, the informal reasoning in (3) can be formally justified within LNS.

2 Paradoxes of Deontic Logic and Their Solutions

In this section, I propose how to solve some paradoxes of SDL. First, let us consider Ross's paradox (Ross (1941), McNamara (2006) sec. 4.3, Åqvist (2002) sec. 6):

[2] This characterization corresponds to an axiom of SDL, namely $\mathbf{O}p \rightarrow \neg\mathbf{O}\neg p$. However, within LNS the iteration of normative sentences is not possible. This is no failure of LNS, because many normative systems in our ordinary life are expressible without iterative normative expressions. It is interesting that LNS fulfills the *principle of deontic contingency* required by Von Wright (1951). Because of (1c), we can easily prove that no tautology belongs to an obligation context.

(5a) It is obligatory that the letter is mailed.

(5b) It is obligatory that the letter is mailed or the letter is burned.

Within SDL, (5b) seems to follow from (5a), because $\mathbf{O}m \rightarrow \mathbf{O}(m \vee b)$ is a theorem of SDL, where $\mathbf{O}p$ means "It is obligatory that p". However, it seems rather odd to say that an obligation to mail the letter entails an obligation that can be fulfilled by burning the letter. Within LNS, a similar inference is valid:

(6*) If $\mathbf{O}_{\langle T,OB \rangle}p$ & $T \nvdash p \vee q$, then $\mathbf{O}_{\langle T,OB \rangle}(p \vee q)$.

The source of the paradoxical appearance of this example consists in ignoring the incompatibility among two types of actions. In this case, it will be appropriate to assume that mailing a letter is incompatible with burning it. This fact justifies to accept that $\forall x(\mathit{letter}(x) \rightarrow \neg(\mathit{mailed}(x) \wedge \mathit{burned}(x)))$ is a component of the propositional system of the given normative system $\langle T_2, OB_2 \rangle$.

(6a) $\mathit{mailed}(l_1)$ is an element of OB_2. (From (5a))

(6b) $T_2 \nvdash \mathit{mailed}(l_1) \vee \mathit{burned}(l_1)$. (Observation)

(6c) $\mathit{letter}(l_1)$ and $\forall x(\mathit{letter}(x) \rightarrow \neg(\mathit{mailed}(x) \wedge \mathit{burned}(x)))$ are elements of T_2. (Observation)

From these conditions follows that $\neg \mathit{burned}(l_1)$ belongs to the obligation context of $\langle T_2, OB_2 \rangle$ (see (6d)). Thus, it is forbidden to burn the letter. [3] [4]

(6d) If $\{\mathit{letter}(l_1), \forall x(\mathit{letter}(x) \rightarrow \neg(\mathit{mailed}(x) \wedge \mathit{burned}(x)))\} \subseteq T_2$ & $\{\mathit{mailed}(l_1)\} \subseteq OB_2$ & $T_2 \nvdash (\mathit{mailed}(l_1) \vee \mathit{burned}(l_1))$, then $T_2 \vdash \neg(\mathit{mailed}(l_1) \wedge \mathit{burned}(l_1))$ & $\mathbf{O}_{\langle T_2,OB_2 \rangle}(\mathit{mailed}(l_1) \vee \mathit{burned}(l_1))$ & $\mathbf{O}_{\langle T_2,OB_2 \rangle}\neg \mathit{burned}(l_1)$ & $\mathbf{F}_{\langle T_2,OB_2 \rangle}\mathit{burned}(l_1)$.

The *Good Samaritan Paradox* pointed out by Prior (1958) can be solved in a similar way. Let us consider the following sentences:

(7a) It ought to be the case that John helps Smith who has been robbed.

(7b) John helps Smith who has been robbed *iff* John helps Smith and Smith has been robbed.

$\mathbf{O}(h \wedge r) \rightarrow \mathbf{O}(h)$ is a theorem of SDL. Thus, if we represent sentence (7a) as $\mathbf{O}(h \wedge r)$, (7c) follows from (7a) and (7b). However, (7c) seems hardly right.

(7c) It ought to be the case that Smith has been robbed.

Within LNS, (7a) can be analyzed as the combination of two conditions (8a) and (8c), where (8c) follows from (8a) and (8b).

[3] Belnap et al (2001) also discusses how to solve Ross's paradox within Stit logic (p. 84f). Stit logic can take future developments into consideration and they solve Ross's paradox using this property of Stit logic. They demonstrate that $[\alpha\ \mathit{stit} : p \vee q]$ does not follow from $[\alpha\ \mathit{stit} : p]$, where $[\alpha\ \mathit{stit} : p]$ is an abbreviation of $[\alpha$ sees to it that $p]$.

[4] One of the reviewers pointed out my misinterpretation of Ross's paradox in the first draft of this paper. According to him, Ross's paradox is $\mathbf{O}m \rightarrow \mathbf{O}(m \vee b)$ itself. In this case, I will state that (6*) is not so bad, because realizing p remains still as an obligation after realizing q.

(8a) $agent(John)$ and $robbed(Smith)$ are elements of T_3.

(8b) $\forall x \forall y(agent(x) \land robbed(y) \rightarrow help(x, y))$ is an element of OB_3.

(8c) $\mathbf{O}_{\langle T_3, OB_3 \rangle}(robbed(Smith) \rightarrow help(John, Smith))$.

(8d) If $\{agent(John), robbed(Smith)\} \subseteq T_3$ &
$\{\forall x \forall y(agent(x) \land robbed(y) \rightarrow help(x, y))\} \subseteq OB_3$, then
$\mathbf{O}_{\langle T_3, OB_3 \rangle}(robbed(Smith) \rightarrow help(John, Smith))$ &
$\mathbf{O}_{\langle T_3, OB_3 \rangle} help(John, Smith)$.

From (8a) and (8c), follows that "John helps Smith" belongs to the obligation context of $\langle T_3, OB_3 \rangle$ ((8d)). This result means that *It ought to be the case that John helps Smith*, which is the result we sought.

Note that a normative system can express explicitly who the bearers of an obligation are, where we assume that they always try to fulfill their obligations, if they accept them:

(9a) Given a normative system $\langle T, OB \rangle$, "Teachers should prepare for their lectures" can be expressed as follows: $\mathbf{O}_{\langle T, OB \rangle}\ \forall x \forall y(agent(x) \land teacher(x) \land lecture\text{-}of(y, x) \rightarrow prepare(x, y))$.

(9b) Given a normative system $\langle T, OB \rangle$, "Students should study hard" can be expressed as follows: $\mathbf{O}_{\langle T, OB \rangle}\ \forall x(agent(x) \land student(x) \rightarrow study\text{-}hard(x))$.

3 Conflicts in Normative Systems

In normative system $\langle T, OB \rangle$, two kinds of contradictions are distinguished: the contradiction in propositional system T and that in $\langle T, OB \rangle$. T is *inconsistent as the propositional system of* $\langle T, OB \rangle$ *iff* T is inconsistent. However, $\langle T, OB \rangle$ is *inconsistent iff* $T \cup OB$ is inconsistent.

A characteristic of NSs is the property that any violation of the presupposed obligations produces inconsistency in NSs. Let us consider the following example:

(10a) It ought to be that Jones does go (to the assistance of his neighbors).

(10b) Jones doesn't go.

This situation can be represented as follows:

(10c) $\{agent(Jones), \neg go(Jones)\} \subseteq T_4$ & $\{go(Jones)\} \subseteq OB_4$.

This kind of violation of obligations could threaten the significance of a NS. If people perform their actions without respecting a given NS, that system can lose its significance. However, a small violation of a NS can sometimes be managed through taking the dynamic aspect of the reality into consideration (see section 4).

Next, let us consider a case of conflicts in an obligation space. Suppose that Tom is required to do act_1 as a member of group A. Furthermore, suppose that he is also required not to do act_1 as a member of group B. This situation produces inconsistency in the presupposed normative system $\langle T_5, OB_5 \rangle$:

(11) If $\{agent(Tom), member(Tom, A), member(Tom, B)\} \subseteq T_5$ &
$\{\forall x(agent(x) \wedge member(x, A) \rightarrow do(x, act_1)),$
$\forall x(agent(x) \wedge member(x, B) \rightarrow \neg do(x, act_1))\} \subseteq OB_5$, then
$\mathbf{O}_{\langle T_5, OB_5 \rangle} do(Tom, act_1)$ & $\mathbf{O}_{\langle T_5, OB_5 \rangle} \neg do(Tom, act_1)$.

One possible solution for Tom is to drop out from one of these groups. In that case, this decision reproduces a consistent NS. For example, if Tom drops out from group B, he need no more consider the prohibition of doing act_1.

4 Dynamic Aspects and Future Orientation of LNS

A normative system $\langle T, OB \rangle$ can function like a *conversational score* proposed by David Lewis (1979). A normative system can be updated to describe a development of a situation. Let us reconsider the case of a violating action described in (10a) and (10b). To describe the shift of time, we introduce here the *past-tense operator* P. Then, we can consider the shift of time and update normative system $\langle T_4, OB_4 \rangle$ and create $\langle T_{4up}, OB_4 \rangle$:

(12a) $\{agent(Jones), \neg go(Jones)\} \subseteq T_4$ & $\{go(Jones)\} \subseteq OB_4$.

(12b) $T_{4up} = (T_4 - \{\neg go(Jones)\}) \cup \{P(\neg go(Jones))\}$. (Information updated)

This example shows that a violated NS can sometimes automatically recover its consistency when its propositional system is updated.

Next, let us consider a case where an obligation becomes applicable through a change of the given situation.

(13a) We should help suffering neighbors.

(13b) Mary, who is a neighbor of John, was not suffering, but is now suffering.

(13c) So John should help Mary now.

Within LNS, the informal inference "(13c) follows from (13a) and (13b)" can be explicitly described as the inference of (13f) from (13d) and (13e):

(13d) $\{agent(John, neighbor(Mary, John)\} \subseteq T_6$ &
$\{\forall x \forall y(agent(x) \wedge neighbor(y, x) \wedge suffering(y) \rightarrow help(x, y))\} \subseteq OB_6$.

(13e) $T_{6up} = T_6 \cup \{suffering(Mary)\}$. (Information updated)

(13f) $\mathbf{O}_{\langle T_{6up}, OB_6 \rangle} help(John, Mary)$.

Normative sentences are normally future oriented. Some problems can be solved by considering this property. Consider the following sentences: [5]

(14a) It should be the case that from now on any two countries have peaceful relations ($\{\forall t \forall x \forall y(t_{now} \leq t \wedge country(x) \wedge country(y) \wedge x \neq y \rightarrow peaceful(x, y, t))\} \subseteq OB_7$).

[5] This problem is pointed out by one of the reviewers.

(14b) A and B are countries ($\{country(A) \wedge country(B)\} \subseteq T_7$).

(14c1) A and B have now peaceful relations ($\{peaceful(A, B, t_{now})\} \subseteq T_7$).

(14c2) A and B have always peaceful relations ($\{\forall t\ peaceful(A, B, t)\} \subseteq T_7$)

(14d) It should be the case that from now on A and B have peaceful relations ($\mathbf{O}_{\langle T_7, OB_7 \rangle} \forall t(t_{now} \leq t \rightarrow peaceful(A, B, t))$).

Within LNS, (14d) follows from (14a), (14b), and (14c1), while (14d) does not follow from (14a), (14b), and (14c2). However, the second invalidity can be justified, because (14c2) expresses that the goal of (14d) has been already satisfied. [6]

5 Conclusions

It is well known that SDL has many theoretical difficulties (Åqvist (2002), McNamara (2006)). Recently, Stit logic was proposed and researchers have shown many interesting results (Belnap et al (2001), Horty (2001)). LNS is an alternative framework that can explicitly express both propositional and normative constraints. This property of explicitness makes LNS applicable to numerous classes of normative problems. For example, a legal system could be described as a normative system in the sense of LNS. The method developed in this paper can be modified to explain inferences among speech acts. However, this remains a future task. [7]

References

[1] Åqvist, L. (2002) "Deontic Logic", D. Gabbay and F. Guenthner (eds.) *Handbook of Philosophical Logic*, Vol. 8, Kluwer Academic Pub., pp. 147-264.
[2] Belnap, N., Perloff, M. and Xu, M. (2001) *Facing the Future: Agents and Choices in Our Indeterminist World*, Oxford University Press.
[3] Hart, H. L. A. (1961) *The Concept of Law*, Clarendon Press.
[4] Horty, J. F. (2001) *Agency and Deontic Logic*, Oxford University Press.
[5] Lewis, D. (1979) "Scorekeeping in a Language Game", *Journal of Philosophical Logic* 8, pp. 339-359.
[6] McNamara, P. (2006) "Deontic Logic", *Stanford Encyclopedia of Philosophy.*
[7] Prior, A. N. (1958) "Escapism: The Logical Basis of Ethics", In A.I. Melden (1958), *Essays in Moral Philosophy*, University of Washington Press, pp. 135-146.
[8] Ross, A. (1941) "Imperatives and Logic", *Theoria* 7, pp. 53-71.
[9] Von Wright, G. H. (1951) "Deontic Logic", *Mind* 60, pp. 1-15.

[6] As this section shows, LNS can deal with interactions between deontic states and temporal developments. Thus, LNS can be seen as a framework that satisfies the following requirements mentioned in Åqvist (2002): "for any serious perposes of application, the expressive resources of deontic languages must be enriched so as to include temporal and quantificational ones" (p. 150).

[7] I would like to thank two reviewers for many insightful comments.

Responsibility in Games

Matthew Braham
Faculty of Philosophy
University of Groningen
The Netherlands
m.braham@rug.nl

Martin van Hees
Faculty of Philosophy
University of Groningen
The Netherlands
Martin.van.Hees@rug.nl

January 10, 2010

When outcomes result from the joint actions of two or more people it is generally believed that we face major difficulties in ascribing responsibility. Thompson (1980) has christened it the 'many hands problem'. One part of the problem is determining the causal contributions given that individual efforts may be like strands in a rope: together each strand makes up the rope but each particular strand may be dispensable. In joint actions, the role of each individual in bringing about an outcome appears to be lost in a complex process. Another part of the problem is that even when causal contributions can be determined, there may not actually be anything wrong with each of the actions per se. Determining 'whose hands' will not necessarily pick out 'whose hands were dirty', i.e. the set of individuals who should – as is generally the case – be punished, held liable, or subjected to moral criticism.

The problems are germane and pervasive. Feinberg (1968) discusses the example of the 'Jesse James train robbery' in which an armed man holds up a car full of passengers. If the passengers had risen up as one and rushed the robber, one or two of them would have perhaps been shot, but collectively they would have overwhelmed him and saved their own and other's property. Feinberg asks to what extent are any of the passengers responsible for the loss of their property given that none alone could have prevented Jesse James walking off with it? Copp (2006) and Pettit (2007) have discussed a quirk of collective decision-making known as the 'discursive paradox' in which members of a committee each have their reasons to reject a particular proposal put before them but the proposal passes nevertheless given the way in which the decision procedure works. In what way can the committee members be held responsible for the outcome?

A more down to earth example concerns the group of managers and engineers at MacDonnell-Douglas who knew of the design faults in early DC-10s that led to these planes dropping out of the sky in the 1970s but nevertheless allowed these planes to, fatally, go into service. As it turned out no single individual was declared as having been directly decisive for the harm that occurred. Then there are the My Lai or Sebrenica massacres; the infamous murder of Kitty Genovese in New York in which onlookers watched her slow death; and the run of recent bank collapses that have greatly damaged the international financial system. The list is endless (for a catalogue of further cases see Bovens, 1998).

One solution to the problem has been the introduction of the concept of *collective responsibility.* The idea here is to argue that the outcomes are the result of a form of 'collective agency', which supervenes on individual members of a group. The task has been to show first which types of groups can be treated as 'moral agents' and then assign any personal responsibility on the basis of voluntary membership of the collective agent. Proponents of this idea of collective responsibility are French (1984) and Pettit (2007). The concept of collective responsibility is, however, not undisputed. The hypostasization of groups is anything but straightforward. It raises a host of metaphysical quandaries about the ontological status of agency. Normatively speaking there are equally difficult problems, one of which is that is that the 'membership' or 'shared attitudes' criteria has the highly unpalatable consequence that it can result in holding people responsible for states of affairs for which they played no part in bring about.

The purpose of this paper is to tackle the problem anew and demonstrate that it is in fact possible in principle to ascribe moral responsibility to individual agents in complex joint activities. To do so, we assume an agent can be held responsible for the realization of a state of affairs A if the following criteria are satisfied:

Agency Condition — The person is an autonomous, intentional, and planning agent who is capable of distinguishing right and wrong and good and bad.

Causal Relevancy Condition — There should be a causal relation between the action of the agent and the resultant state of affairs.

Avoidance Opportunity Condition — The agent should have had a reasonable opportunity to have done otherwise.

Building on previous work (Braham and Holler, 2008; Braham, 2008; Braham and van Hees, 2009), we show how these conditions can be recast in a very natural habitat: a game theoretic framework. We then define the components of what we call a 'responsibility game' and examine how different allocations of responsibility can be associated with different classes of responsibility games. Subsequently we show that, from a theoretical viewpoint, the 'many hands problem' is not as severe as it may at first appear.

References

Bovens, M. (1998). *The Quest for Responsibility*. Cambridge: Cambridge University Press.

Braham, M. (2008). Social Power and Social Causation: Towards a Formal Synthesis. In Braham, M. and Steffen, F. (eds), *Power, Freedom, and Voting*. Heidelberg: Springer.

Braham, M. and Hees, M. van (2009). Degrees of Causation. *Erkenntnis* 73: 323–344.

Braham, M. and Holler, M. J. (2008). Distributing Causal Responsibility in Collectivities. In Boylan, T. and Gekker, R. (eds), *Economics, Rational Choice, and Normative Philosophy*. London: Routledge.

Copp, D. (2006). On the Agency of Certain Collective Entities: An Argument from “Normative Autonomy”. *Midwest Studies in Philosophy* 30: 194–221.

Feinberg, J. (1968). Collective Responsibility. *Journal of Philosophy* 65: 674–688.

French, P. A. (1984). *Collective and Corporate Responsibility*. New York: Columbia University Press.

Pettit, P. (2007). Responsibility Incorporated. *Ethics* 117: 171–201.

Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review* 74: 905–916.

Is Intention sufficient to explicate Collective Agency?

Biswanath Swain

Doctoral Scholar
Department of Humanities and Social Sciences
Indian Institute of Technology Kanpur, India
E-Mail: biswanath80@gmail.com

Abstract

In this paper, I would critically examine the intention-based account of collective agency, and argue for its radical insufficiency. The proposal, I intend to proffer here is that efficacious collective agency is a far more robust notion than what comes forth in the intention-based approach. Collective agency must be predicated upon the collective 's ability to form an internal evaluative mechanism for the execution of its intention to accomplish the desired goal that it sets for itself. In the point of fact, it is the presence of this mechanism that is responsible for endowing cohesiveness and robustness to the collective as a proper agent.

Keywords: Agency; Will; Intention; Action; Internality Constraint; Rationality; Social Beings; Autonomy; Collective Agency, Plural Agents; Normativity; Goal; Shared (collective) Intention.

My 'being' in the world around as an active individual person can be defined by two contrary agential features namely, *in-dependence* and *inter-dependence*. There are numerous ways in which I individually, of course independently, bring about desired changes in the world outside. At other times, my ways of bringing about such changes in the world necessarily requires me to depend on other individuals. So I start explicating this sort of actions in terms of certain states of mind that are framed within, and attitudes or behaviors that are shown by other individual agents. The explanatory mental state is purely internal to the concerned individual subjects of the action in question. We might call this the 'internality constraint' of the explanation of action with a reference to intention. While the intentional explanation of independent action evidently conforms to the internality constraint, but it is far from obvious whether an intentional explanation of interdependent action can be carried through to conform to this constraint account of intention. Since doing an interdependent action involves the participation of more than one individual person, so there is no way of attributing the explanatory intentional state to any single member engaged in accomplishing the interdependent act.

If we attach importance to the internality constraint in explaining an action, independent or interdependent, then it would appear that the idea of an interdependent action is unintelligible from the explanatory standpoint. There is, however, no question of de-recognizing the reality of interdependent actions, which are also considered as genuine action-types that define our very-existence as *social* beings – where we witness the *gregarious* instinct outrightly. It therefore means that there has to be some complex account of the possibility of human beings acting interdependently, which makes the idea of co-operative, social action intelligible without conflicting seriously with the intuitive plausibility of the internality constraint.

The idea of co-operative, collective action has already been an issue of fervent philosophical discussion in recent trend of philosophy i.e., philosophy of mind and action. The crux of the discussion has been there to analyze the idea of collective action in terms of the possibility of collective intention. Prominent among the philosophers who have dealt seriously with this issue are Searle, Tuomela, Bratman, Gilbert, and Velleman. They have addressed themselves to the question of *what makes our social life possible*, or, more specifically, *what it is for us to intend to do something together.* Irrespective of their individual differences on the notion of collective intention, they are unreservedly unanimous on the point that the existence of collective intention is a *necessary condition* for the inception of collective agency. But what remains still a moot question is whether collective intention is a *sufficient condition* for collective agency or not. And this latter question deserves to be critically examined. Indeed, the question of sufficiency raises the issue of whether collective intentional goal-directedness has to be supplemented by an evaluative conception of the goal as being worthy of pursuit on the part of the agent.

Acknowledgement
I must owe my deep sense of gratitude to my supervisor Prof. C. A. Tomy, my teacher Prof. B.H. Boruah, and my mentor Prof. M. Jha, and the members of Philosophy Discussion Group at Indian Institute of Technology Kanpur. This paper has grown out of the discussions that I have done many a time with those people. A deep bow of thanks and appreciation are due to the organizers, and the anonymous reviewers of my manuscript.

References

[1] Bratman, M. E. (1999). *Faces of Intention.* Cambridge, MA: Cambridge University Press.

[2] Dennett, D. (1987). *The Intentional Stance*, Cambridge, MA: MIT Press.

[3] Davidson, D. (1980). " Mental Events. " in his *Essays on Actions and Events.* Oxford: Clarendon Press.

[4] Gilbert, M. (1989). *On Social Facts.* New Jersey: Princeton University Press.

[5] —, (1996). *Living Together: Rationality, sociality and Obligation.* Lanhan, MD: Rowman & Littlefield.

[6] —, (2003). " The Structure of the Social Atom: Joint Commitment as the Foundation of Human Social Behavior. " in Schmitt, F. *Socializing Metaphysics*, Lanham, MD: Rowman and Littlefield, P. 39-64.

[7] —, (2005). " Rationality in Collective Action. "*Philosophy of the Social Sciences*, Volume 36, No. 1, March, pp. 3-17.

[8] —, (2006). " Who 's to Blame? Collective Moral responsibility and Its Implication for Group Members. "*Midwest Studies in Philosophy*, XXX, 94-114.

[9] Helm, B. W. (2008). " Plural Agents. " *Nous*, 42, (1), pp. 17-49.

[10] Pettit, P. (1993). *The Common Mind: An Essay On Psychology, Society and Politics*, Paperback ed. New York: Oxford University Press.

[11] Pettit, P. and Schweikard, D. (2006), " Joint Action and Group Agents. " *Philosophy of the Social Sciences*, Vol. 36. Number 1, March 2006, pp. 18-39.

[12] Petit, P. (2003)." Groups with Minds of Their Own. "in Schmitt, F.F.(ed.), *Socializing Metaphysics: The nature of Social Reality*, Lanham, ML: Rowman and Littelfield, pp. 167-194.

[13] Pettit, P. (2007)." Responsibility Incorporated. "*Ethics*, 117, January, pp. 171-201.

[14] Schmid, H.B. (2008). " Plural Agents. "*Philosophy of the Social Sciences*, Vol. 38, Number 1, March 2008, pp.25-54.

[15] Searle, J. (1990). "Collective Intentions and Actions." In *Intentions in Communication*, P.Cohen, J. Morgan, and M.E. Pollack, eds. Cambridge, MA: Bradford Books, MIT press.

[16] Searle, J. (1995). *The Construction of Social Reality.* New York, N.Y.: Free Press.

[17] Tuomela, R. (1977). *Human Action and its Explanation: A Study on the Philosophical Foundations of Psychology*, Dordrecht, Boston: D. Reidel Publishing Company.

[18] —, (1984). *A Theory of Social Action*, Dordrecht, Boston and Lancaster: D. Reidel Publishing Company.

[19] —, (1995). *The Importance of Us: A Philosophical Study of Basic Social Notions*, Stanford: Stanford University Press.

[20] Velleman, D. 1997. "How to Share an Intention." *Philosophy and Phenomenological Research* LVII: 29-50.

[21] Wittgenstein, L. (1958). *Philosophical Investigations*, translated by Anscombe, G.E.M. Englewood Cliffs/N.J: Prentice Hall.

Interpretation of Action and Sociality of Action

Ryoji Fujimoto and Choi Chang-Bong

Hokkaido University

In this presentation, we attempt an analysis of social actions in a way that it provides a clear sketch of the relations between the ingredients of actions and the responsibilities of them. So, we try to get an analysis of social actions in a way that it meets following simple criterion on a relation between actions and their responsibilities.

(A) A person P is responsible for an action (including social one) X or results of X only if P is constituents of X.

In social actions, there may be actions which depend on social conventions or shared-intentions of agents or the like, which are often treated as salient features of social actions, but we do not exclude other types of social actions from our analysis. For, if we accept too ‘ strong ’ criterions such as,

(B) An action X is social only if X is made by obeying some social conventions,

or,

(C) An action X is social only if agents of X have shared-intentions,

or the like, we cannot maintain criterion (A) since there are various cases which do not suffice (A) or (B) and nevertheless we can naturally say that a person who is involved in the case is responsible for what he did. These cases makes difficult to analyze responsibility of action unless we abandon criterion (A) or (B). So we have to find suitable criterions for what makes an action social (see below (T1) and (T2)). However, this does not mean that shared-intensions or social conventions do not play any roles in explanation of any (social) actions, and even the task to analyze these factors provides philosophically attractive issue. But, when one takes not only (social) actions but their responsibilities into consideration, we have to investigate carefully where we should place these factors in the course of explanation of actions. So we will begin with the question ‘what makes an action social’, and show that even when an action depends on these factors, what makes it social are not these factors themselves but various attitudes of agents towards them. Then, our first thesis on social actions will be as follows.

(T1) What makes an action social are agents’ attitudes towards other persons such as an expectation or a prediction of being interpreted correctly.

And we also claim following thesis as supplement of (T1), (T2) Social conventions or shared-intentions [1] themselves make an action social, but when the agents of an action have suitable attitudes towards them, they can have the functions such as increasing probability of achievement of correct interpretation of an action.

For example, it is natural to think that stopping a taxi by raising arm with the intention to stop the taxi is a social action if the driver of the taxi actually stops his car with correct interpretation of the signal which is made by the bodily movement (raising arm). And it is natural to think that the case in which the driver is actually aware of the signal but does not stop his car and passes away is also, at least in some sense, social action. In the latter case, for example, criterion (B) would be confronted with some difficulty. For, in this case, the social convention which consists of raising arm and stopping taxi does not work. So, we cannot simply say like "Social actions are actions which are done by following some social convention". We should say that the overall event which consists of sending signal and ignoring it and passing away is related to agents' attitudes to the convention rather than to the convention itself. And here, it is plausible to say that the function of the convention is increasing probability of achievement of correct interpretation of the signal rather than letting him stop his car. But then, we can ask as a matter of course whether social conventions play essential roles to make an action social. Our answer to this question is no. If one can act in some situations with expectation of being interpreted correctly, we can act socially without social conventions. And we will argue that there surely are such situations. So, we will conclude that what make an action social are not social conventions but some types of attitudes such as expectations, predictions.

However, we do not preclude the existence of some *types* of social actions which essentially depend on social conventions or rules. We can create a new board-game with explicit rules and can play it. In such a case, an act in violation of the rules can make the game itself invalid. So, when we play a game, we can say that the rules of it play constructive roles *in the game*. But this does not mean that the rules play constructive roles *to make an action social*. Even if we play a game incorrectly, in some sense, we can say that we still act socially. Violation of rules may destruct validity of games but do not demolish sociality of playing games as far as players expect their moves as to be interpreted by each other. In other words, violation of constructive rules changes the *type* of an action, but it can be still social as far as agents have suitable attitudes. So, yet we can hold that (T1) and (T2) are correct.

Here, we can point out that these elements such as expectations, predications or the like also play important roles in the analysis of responsibility. We often take the question of one's moral responsibilities for his action as the question under the rules of bivalence. But, we believe that this is not correct. Unlike moral responsibility, if you look at criminal court, you will certainly find that estimating the degrees of responsibilities play important roles to hold criminal liabilities. And we will hold that in the analysis of moral responsibilities we also have to take quantitative aspects of it into consideration. Here, one may

[1] Shared-intentions are, at least we think, usually 'higher' intention in the sense that they usually represent not how to make bodily movements or the like but 'overall plan' of entire action or 'the end' of it. So we take shared-intentions as regulative or constructive like conventions.

naturally ask what quantitative aspects of moral responsibilities are. We can point out two; the seriousness of the matter caused by relevant action and the degrees of possibilities of predicting or expecting results of relevant action. And the latter is what we have already seen in the analysis of actions.

On the basis of these considerations, we, then, will examine several examples of typical social actions which have various patterns about the degrees of two quantitative factors in the analysis of responsibilities. Through the analysis of these examples, it will be clear that our 'modest' position (like (T2)) on the roles of rules, conventions, or shared-intentions will be helpful to estimate the responsibilities of social actions.

References

[1] Bratman, M. 1992: "Shared Cooperative Activity", *Philosophical Review* 101: 327-341.

[2] Bratman, M. 1993: "Shared Intention", *Ethics* 104: 97-113.

[3] Bratman, M. 1999: *Faces of Intention: Selected Essays on Intention and Agency.* Cambridge University Press.

[4] Gilbert, M. 1987: "Modelling Collective Belief", *Synthese* 73: 185-204.

[5] Gilbert, M. 1990: "Walking Together: A Paradigmatic Social Phenomenon", *Midwest Studies* 15: 1-14.

[6] Searle, J. 1990: "Collective Intentions and Actions", in Cohen, P., Morgan , J and Pollack, M. (eds.): *Intentions in Communication.* The MIT Press.: 401-415.

[7] Searle, John R. 1995: *The Construction of Social Reality.* Penguin Books.

[8] Tuomela, R. and Miller, K. 1988: "We-Intentions", *Philosophical Studies* 53: 115-137.

[9] Tuomela, R. 1991: "We Will Do It: An Analysis of Group-Intentions, *Philosophy and Phenomenological Research* 51: 249-277.

[10] Velleman, J. 1997: "How to Share an Intention?", *Philosophy and Phenomenological Research* 57: 29-50.

Syncretic Argumentation by means of Lattice Homomorphism and Fusion

Hajime Sawamura
Institute of Science and Technology, Niigata University, Japan
Email: sawamura@ie.niigata-u.ac.jp

In his influential work on the abstract argumentation framework [5], Dung introduced the notion of "acceptability" of arguments that has played the most significant role in specifying the various kinds of semantics for argumentation: admissible, stable, preferred, grounded, complete. The abstract argumentation framework is specified as follows.

Definition 1 (**Argumentation Framework** [5]) *An argumentation framework is a pair* $AF =< AR, attacks >$ *where* AR *is a set of arguments, and* $attacks$ *is a binary relation on* AR*, i. e.,* $attacks \subseteq AR \times AR$*.*

In Dung's theory of argumentation, we are not concerned with the internal structure of arguments and why and how arguments attack others. Everything is abstracted away in this way. This abstraction, however, was a good starting point for developing the formal argumentation semantics that is to capture what acceptable or admissible arguments are and the whole of justified arguments.

Definition 2 (**Acceptability and Admissibility** [5])

1. *An argument* $A \in AR$ *is acceptable w.r.t. a set* S *of arguments iff for each argument* $B \in AR$*: if* B *attacks* A *then* B *is attacked by* S*.*

2. *A conflict-free set of arguments* S *is admissible iff each argument in* S *is acceptable w.r.t.* S*.*

The notion of acceptability is a counterpart of the phenomenon observed in our daily argumentation and originates from an old saying, "The one who has the last word laughs best", as stated by Dung. It is an empirical social truth or wisdom that has been evolved in various cultural sphere over generations and considered useful by people. It is remarkable and suggestive that Dung's theory of argumentation had started from such a daily but philosophical observation. This might be because argumentation is humans' most normal but intelligent action for thought and communication by language.

There, however, can be a plurality of sets of justified arguments in argumentation as mentioned above, contrasting with the semantics of an ordinary logic that is to be uniquely given by the Tarskian semantics, for example. Naturally, this reflects a figure of argumentation, a decisive difference from a logic. The preferred semantics, for example, is defined as follows

Definition 3 (**Preferred Extension** [5]) *A preferred extension of an argumentation framework AF is a maximal (w.r.t. set inclusion) admissible set of AF.*

We developed the Logic of Multiple-valued Argumentation (LMA) [13] that is a variant of Dung's abstract argumentation framework concretized in such a way that the arguments are represented in terms of the knowledge representation language, Extended Annotated Logic Programming (EALP) and the attack relation consists of various sorts of attack such as rebuttal, undercut, defeat, etc. with three kinds of negation: ontological negation ($\tilde{}$), default negation (**not**), and epistemological negation ($\neg$) that play a role of momentum in argumentation. EALP is an extension of ELP (Extended Logic Programming), and a very expressive knowledge representation language in which agents can express their knowledge and belief with annotations as truth-values that allow to represent various kinds of uncertainty of information. In a word, LMA is an argumentation framework that allows agents to participate in uncertain argumentation under uncertain knowledge bases if once the common annotation is shared among agents. Put it differently, agents are assumed to have a homogeneous recognition for propositions with the same annotation as truth-values.

In this paper, we make a clean break with this assumption, directing to a more natural but complex settings of argumentation named "Syncretic Argumentation". By the term "syncretic argumentation", it is meant to be such an argumentation that each agent can have its own knowledge base, based on its own epistemology, and participate in argumentation with it. More specifically, each agent can attend the argumentation in which arguments are represented in EALP and annotated with its own truth-values which are assumed to represent modes of truth or epistemic states of propositions [13]. The syncretic argumentation is a new framework that allows agents to argue about issues of mutual interest even when they have their own annotations, for example, agent A has two values $\mathcal{TWO} = \{f, t\}$ as annotation (this is typical in the Occident), and agent B has 4-values $\mathcal{FOUR} = \{\bot, \mathbf{t}, \mathbf{f}, \top\}$ as annotation (this is called tetralemma in the early philosophical literature and text of Buddhism [11][12]). This reflects an attitude against unilateralism, so that one agent world may not be forced to assimilate to another unilaterally. We realize the goal by means of the lattice homomorphism since the mathematical structure of annotations is a complete lattice and the homomorphism is a mathematical apparatus convenient to syncretize the difference of epistemic states of propositions.

Definition 4 (Homomorphism [4]) *Let $< L, \vee_L, \wedge_L, \leq_L >$ and $< K, \vee_K, \wedge_K, \leq_K >$ be complete lattices. A map $h : L \to K$ is said to be a homomorphism if h satisfies the following conditions: for all $a, b \in L$,*

- $h(a \vee_L b) = h(a) \vee_K h(b)$
- $h(a \wedge_L b) = h(a) \wedge_K h(b)$
- $h(0_L) = 0_K$ *for the least element*
- $h(1_L) = 1_K$ *for the greatest element*

Example 1 *Let us consider two typical lattices: the two-valued complete lattice $\mathcal{TWO}$ and the four-valued one $\mathcal{FOUR}$. The former is typical in the West,* *and the latter in the early philosophical literature and text of Buddhism [11].* $\mathcal{TWO} =< \{f, t\}, \vee, \wedge, \leq >$, *where* $f \leq t$, *and* $\mathcal{FOUR} =< \{\bot, \boldsymbol{t}, \boldsymbol{f}, \top\}, \vee, \wedge, \leq >$, *where* $\forall x, y \in \{\bot, \boldsymbol{t}, \boldsymbol{f}, \top\} \;\; x \leq y \;\Leftrightarrow\; x = y \;\vee\; x = \bot \;\vee\; y = \top$.

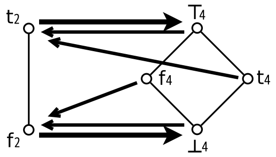

Fig. 1: Homomorphism: $h1 : \mathcal{TWO} \rightarrow \mathcal{FOUR}$ and $h2 : \mathcal{FOUR} \rightarrow \mathcal{TWO}$

With the lattice homomorphism above, we will illustrate how agents who have their own epistemology can reach an agreement and accept arguments through the grounded semantics or the dialectical proof theory of LMA [13].

Example 2 *Suppose two agents A and B have the following knowledge bases respectively.*

$K_A = \{\;\; a : t_2 \leftarrow, \;\; \sim b : t_2 \leftarrow, \;\; c : t_2 \leftarrow, \;\; \sim d : t_2 \leftarrow \;\; \}$

$K_B = \{\;\; \sim a : t_4 \leftarrow, \;\; b : t_4 \leftarrow, \;\; \sim c : \top_4 \leftarrow, \;\; d : \bot_4 \leftarrow, \;\; e : t_4 \leftarrow g : f_4, \;\; g : t_4 \leftarrow \;\; \}$

Then the agents A and B can make the following set of arguments $Args_{K_A}$ *and* $Args_{K_B}$ *from their knowledge bases respectively. (See [13] for the precise definition of arguments in LMA.)*

$Args_{K_A} = \{\; [a : t_2 \leftarrow], \;\; [\sim b : t_2 \leftarrow], \;\; [c : t_2 \leftarrow], \;\; [\sim d : t_2 \leftarrow] \;\}$

$Args_{K_B} = \{\; [\sim a : t_4 \leftarrow], \;\; [b : t_4 \leftarrow], \;\; [\sim c : \top_4 \leftarrow], \;\; [d : \bot_4 \leftarrow], \;\; [g : t_4 \leftarrow] \;\}$

The agents first assimilate their knowledge bases above to each other by the lattice homomorphism in Fig. 1, and compute justified arguments from them using the grounded semantics or the dialectical proof theory [13], in each direction of the homomorphism as follows.

[1] *Lattice homomorphism* $h1$*:* $\mathcal{TWO} \rightarrow \mathcal{FOUR}$ *(simply written as* $\mathcal{T} \rightarrow \mathcal{F}$*)*

$h1(K_A) = \{\;\; a : \top_4 \leftarrow, \;\; \sim b : \top_4 \leftarrow, \;\; c : \top_4 \leftarrow, \;\; \sim d : \top_4 \leftarrow \;\; \}$

$K_B = \{\;\; \sim a : t_4 \leftarrow, \;\; b : t_4 \leftarrow, \;\; \sim c : \top_4 \leftarrow, \;\; d : \bot_4 \leftarrow, \;\; e : t_4 \leftarrow g : f_4, \;\; g : t_4 \leftarrow \;\; \}$

$Args_{h1(K_A)} = \{\; [a : \top_4 \leftarrow], \;\; [\sim b : \top_4 \leftarrow], \;\; [c : \top_4 \leftarrow], \;\; [\sim d : \top_4 \leftarrow] \;\}$

$Args_{K_B} = \{\; [\sim a : t_4 \leftarrow], \;\; [b : t_4 \leftarrow], \;\; [\sim c : \top_4 \leftarrow], \;\; [d : \bot_4 \leftarrow], \;\; [g : t_4 \leftarrow]\}$

Note that $Args_{h1(K_A)} = h1(Args_{K_A})$ *since the homomorphism preserves the lattice ordering. From these argument sets, the agents can have the following set of justified arguments.*

$Justified\_Args_{\mathcal{T} \rightarrow \mathcal{F}} = \{\; [\sim b : \top_4 \leftarrow], \;\; [\sim d : \top_4 \leftarrow], \;\; [b : t_4 \leftarrow], \;\; [d : \bot_4 \leftarrow], \;\; [g : t_4 \leftarrow] \;\}$

[2] *Lattice homomorphism h2:* $\mathcal{FOUR} \rightarrow \mathcal{TWO}$ *(simply written as* $\mathcal{F} \rightarrow \mathcal{T}$*)*

$K_A = \{\;\; a : t_2 \leftarrow, \;\; \sim b : t_2 \leftarrow, \;\; c : t_2 \leftarrow, \;\; \sim d : t_2 \leftarrow \;\; \}$

$h2(K_B) = \{\;\; \sim a : t_2 \leftarrow, \;\; b : t_2 \leftarrow, \;\; \sim c : t_2 \leftarrow, \;\; d : f_2 \leftarrow, \;\; e : t_2 \leftarrow g : f_2, \;\; g : t_2 \leftarrow \;\; \}$

$Args_{K_A} = \{\ [a : t_2 \leftarrow],\ \ [\sim b : t_2 \leftarrow],\ \ [c : t_2 \leftarrow],\ \ [\sim d : t_2 \leftarrow]\ \}$

$Args_{h2(K_B)} = \{\ [\sim a : t_2 \leftarrow],\ \ [b : t_2 \leftarrow],\ \ [\sim c : t_2 \leftarrow],\ \ [d : f_2 \leftarrow],\ \ [g : t_2 \leftarrow],$
$[e : t_2 \leftarrow g : f_2,\ \ g : t_2 \leftarrow]\}$

Note that $Args_{h2(K_B)} \neq h2(Args_{K_B})$ in case of the homomorphism $h2$ since $[e : t_2 \leftarrow g : f_2,\ \ g : t_2 \leftarrow]$ has been qualified as an argument by $h2$ although its original form $[e : t_4 \leftarrow g : f_4,\ \ g : t_4 \leftarrow]$ in K_B is not an argument. From these argument sets, the agents can have the following set of justified arguments.

$Justified\_Args_{\mathcal{F}\rightarrow\mathcal{T}} = \{\ [\sim d : t_2 \leftarrow],\ \ [d : f_2 \leftarrow],\ \ [g : t_2 \leftarrow],\ \ [e : t_2 \leftarrow g : f_2,$
$g : t_2 \leftarrow]\ \}$

Through the two-way homomorphism, we had two different sets of justified arguments: $Justified\_Args_{\mathcal{T}\rightarrow\mathcal{F}}$ and $Justified\_Args_{\mathcal{F}\rightarrow\mathcal{T}}$. Next, we are interested in defining a set of justified arguments as a "common good" that is acceptable for both agents. Actually, we have three kinds of agent attitudes or criteria to chose it from among two different sets of justified arguments[7]. The following is the notion of skeptically justified arguments.

Definition 5 (Skeptically justified arguments)

- *An argument a in $Args_{K_A}$ is skeptically justified iff $a \in Justified\_Args_{\mathcal{F}\rightarrow\mathcal{T}}$ and h1(a) $\in Justified\_Args_{\mathcal{T}\rightarrow\mathcal{F}}$.*
- *An argument a in $Args_{K_B}$ is skeptically justified iff $a \in Justified\_Args_{\mathcal{T}\rightarrow\mathcal{F}}$ and h2(a) $\in Justified\_Args_{\mathcal{F}\rightarrow\mathcal{T}}$.*

This is a fair and unbiased notion of justified arguments in the sense that the both sides can attain a perfect consensus by the two-way homomorphism. Morally, it reflects such a compassionate attitude that agents look from the other agents' viewpoint, or place themselves in the other agents' position.

The syncretic argumentation is obviously a radical departure from the past argumentation frameworks [1][9] [10] in the sense that they are basically frameworks using two-valued knowledge base, or simply a fixed multi-valued one [2]. Here we should emphasize that our approach to the syncretic argumentation is not only technically new but also has a profound philosophy that underlies our syncretic argumentation. They are,

- Golden Rule in the ethics of reciprocity(of positive form): "Treat others (only) as you consent to being treated in the same situation." [6]
- Confucius' Golden Rule(of negative form): "Never impose on others what you would not choose for yourself". " [3]

and may be said to be ethical in contrast with Dung's background idea on the acceptability.

Next we turn to another construction of syncretic argumentation since there are cases where lattice homomorphism does not exist. We devise the new notions: the lattice fusion operator and fusion lattice that are induced through the lattice product, and can be considered as providing a natural way to syncretize the difference of epistemic states of propositions. Figure 3 shows an example of the fusion lattice constructed from two lattices: $\mathcal{TWO}$ and $\mathcal{FOUR}$, via. their

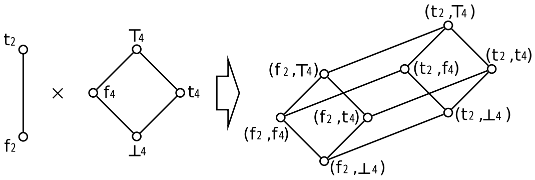

Fig. 2: Product of $\mathcal{TWO}$ and $\mathcal{FOUR}$

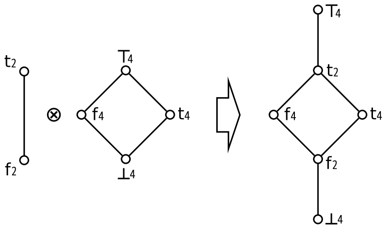

Fig. 3: Fusion of $\mathcal{TWO}$ and $\mathcal{FOUR}$

product depicted in Figure 2. The fusion lattice provides for agents a common argumentation field where agents can start syncretic argumentation using their knowledge bases with annotation specified in the fusion lattice. Our approach to fusing lattices has such advantages as majority principle, order preserving and commutativity (for the details, see [8]).

Agents have to live in the multi-cultural computer-networked virtual society as well as humans living in the multi-cultural society. This implies that agents also get involved in arguing about issues of mutual interest on the basis of their own belief and knowledge. But, if they insisted only on their epistemology, we would lose chances to interact or communicate with each other. The enterprise in this paper is an attempt to avoid such a cul-de-sac appearing even in argument-based problem solving.

There has been no work on argumentation frameworks in which each agent has its own knowledge representation language, its own epistemology, and its own argumentation framework. They have been all common to agents who participate in argumentaion. Our work goes to the polar opposite direction from the perspective of the past works.

The general golden rule has its roots in a wide range of world cultures: ancient Greece, ancient Egypt, ancient China, etc. and almost all religion and philosophy such as Buddhism, Christianity, Islam, Judaism, Confucianism, etc. The human history accepts it as a universal standard with which we resolve conflicts among different civilization and culture. Although the Golden Rule has had its critics on the one hand, the key element of it is that a person attempting to live by this rule should treat all people, not just members of his or her in-group, with consideration and compassion. Therefore it is reasonable for us to employ it and formalize the syncretic argumentation under the general golden rule as the rationale of our attempt. Our bi-directional homomorphism (operation) between different annotations and the fusion lattice approach could realize the key and may be said to the general golden rule itself in the syncretic argumentation. We hope that the syncretic argumentation could lead to overcome and bridge the gulf of incommensurability among different cultural agents, and result in fair and equal argumentation without unilateral imposition.

References

[1] C. I. Chesñevar, G. Maguitman, and R. P. Loui. Logical models of argument. *ACM Computing Surveys*, 32:337–383, 2000.

[2] C. I. Chesñevar, G. Simari, T. Alsinet, and L. Godo. A logic programming framework for possibilistic argumentation with vague knowledge. In *Proc. of the Intl. Conference on Uncertainty in Artificial Intelligence (UAI2004)*, 2004.

[3] Confucius. *The Analects, translated by D. Hinton.* Counterpoint, 1998.

[4] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order.* Cambridge, 2002.

[5] P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logics programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[6] H. J. Gensler. *Formal Ethics.* Routledge, 1996.

[7] T. Hasegawa, S. Abbas, and H. Sawamura. Syncretic argumentation by means of lattice homomorphism. In *PRIMA*, volume 5925 of *Lecture Notes in Computer Science*, pages 159–174. Springer, 2009.

[8] T. Hasegawa and H. Sawamura. Syncretic argumentation by means of lattice fusion. In *JURISIN*, pages 159–174. JSAI, 2009.

[9] H. Prakken and G. Vreeswijk. Logical systems for defeasible argumentation. In *In D. Gabbay and F. Guenther, editors, Handbook of Philosophical Logic*, pages 219–318. Kluwer, 2002.

[10] I. Rahwan and G. R. E. Simari. *Argumentation in Artificial Intelligence, with a forward by John van Benthem.* Springer, 2009.

[11] H. Sawamura and E. Mares. How agents should exploit tetralemma with an eastern mind in argumentation. In *Mike Barley and Nik Kasabov (eds.): Intelligent Agents and Multi-Agent Systems VII, LNAI 3371, Springer*, pages 259–278, 2004.

[12] H. Sawamura and T. Takahashi. Applying logic of multiple-valued argumentation to eastern arguments. *IEICE Transactions*, 88-D(9):2021–2030, 2005.

[13] T. Takahashi and H. Sawamura. A logic of multiple-valued argumentation. In *Proceedings of the third international joint conference on Autonomous Agents and Multi Agent Systems (AAMAS'2004)*, pages 800–807. ACM, 2004.

Social Commitments: Expectations, Obligations and Entitlements

Miranda del Corral
UNED, Spain

The concept of commitment has experienced a growing use in social sciences and in philosophy of action during the past two decades; however, no deep analysis of the concept has been offered since its first formulations (Castelfranchi, 1995, Tuomela, 2007, Graffeo, 2009). Particularly, commitment plays an important role in the explanation of altruist behavior, cooperation and collective action and other social interactions, such as promises, threats and agreements. However, the use of the concept of commitment is far from being homogeneous. Commitment usually plays a secondary role in explaining individual and social behavior, and different definitions arise within each theoretical approach in the literature. The aim of this paper is to propose a framework for the analysis of social commitments, attending to three elements which, from our point of view, are constitutive of this kind of social interaction.

When attempting to analyze social commitments, it is important to discuss their specificity, this is, what differences them from other kinds of social interactions. We believe that three main features may differentiate social commitments from other kinds of social interaction: their impact on empirical and normative expectations, their creation and attribution processes, and the set of possible operations that can be performed on them.

Expectations. Social commitments generate expectations, both in the agent who commits herself and in the agent with whom the former is committed to. The difference between empirical and normative expectations[1] is crucial in understanding social interactions (Bicchieri, 2006), and we will argue that while the former are a precondition for the commitment to take place (since they relate to trust and credibility), the latter are generated with the creation of the commitment, and apply only to the commitment's creator. Normative beliefs about the fulfillment of the commitment usually come to existence after the commitment has been made, except for the cases in which the content of the commitment derives from a social norm, or a moral principle (which we consider to be a subclass of social norms). Once that normative expectations are present, social norms regarding its violation (or

[1] We refer here to normative and empirical expectations in Bicchieri's sense: empirical expectations are what we think others will do, and normative expectations are what we think that others believe we should do (2006). In the literature, it is more frequent to define normative expectations as what we think others should do (Sudgen, 1998), but Bicchieri refers to this belief as “normative belief” (Bicchieri, 2009).

fulfillment) regulate what actions may be performed, such as lack of trust, punishment, reward, or a increase of the agent's credibility. Expectations are also important for understanding the persuasive and dissuasive roles of some social commitments, although, against Schelling's account (2007), we will argue that manipulation does not have to be the necessary function of commitments.

Commitment creation and attribution. Social commitments begin at some point, although they may not have a clear (temporally speaking) end. To set a social commitment, there must be, in principle, a creator, who undertakes the responsibility for performing the content of the commitment, and an attributer, who accepts this responsibility. If a speaker tries to make a promise, but the hearer does not trust her, and does not attribute the responsibility of performing that action to the speaker, the commitment has not been set. In this case, there would be no social obligation of performing the action, and no rights over the speaker's actions have been granted to the hearer. Commitment attribution may also explain implicit commitments: since they are not created by a communication process, but have the same features as explicit commitments, we could say that commitment attribution and common knowledge are sufficient to explain them. Social norms are necessary to explain commitment attribution, insofar as they regulate social obligations (Miller, 2006). Some authors argue that social commitment is a form of goal adoption (Castelfranchi, 1995). We agree in that becoming committed implies becoming responsible for a goal (the content of the commitment), but we do not consider that this goal has to be desired or wanted by the other agent. Other reasons could justify that an agent accepts a social commitment (this is, attributes it to the creator of a commitment).

Operations on commitments. When a commitment is created, a complex set of conditions determine the state of the commitment. Following Singh (1999), commitments can be created, discharged (fulfilled), canceled, released, delegated or assigned. Both external conditions (the context) and internal conditions (such as the agent's beliefs and expectations) are able to operate on commitments and to change their state. Except commitment cancellation, which can be freely performed by the committed agent, the other possible operations are subject to regulation: the conditions under which a commitment may change its state are mutually known by the agents. These conditions are usually regulated upon social norms[2], although the content of the commitment can impose limitations. When a social commitment is made, it creates a set of obligations and entitlements, which include the capability of modifying the commitment.

Taking into account these three features of social commitments, we propose the following definition. A social commitment (SC) is a kind of social interaction involving, at least, two agents: a creator (*x*), and an attributer (*y*), who mutually know that the commitment exists, this is, know that the following conditions exist (or at least do not deny their existence). If *x* is socially committed to *y* to perform the action

[2] However, it would be interesting to analyze the relation between “good reasons” for canceling (or revoking) a social commitment and “good reasons” for dropping an intention, in the sense that it is not seen as weakness of the will (Holton, 1999)

z, then:

1. *y* has empirical expectations concerning *x*'s performance of *z*.
2. *y* has a normative belief concerning *x*'s performance of *z*
3. *x* has normative expectations concerning her performance of *z*
4. Commitments can change their state since their creation. The conditions of change are mutually known by *x* and *y*, and they can be subject to negotiation[3].

Condition 1 means that y has a belief about *x*'s actions, but not necessarily the belief that *z* will be attained. If y believes that *x* will not perform *z*, it would be a case of "self defeating commitment", because y would be accepting a commitment upon its violation.

Why aren't conditions 1 and 2 applied to *x*? If *x*'s empirical expectations and *x*'s normative beliefs about her performance of *z* were necessary, dishonest commitments would not be commitments at all. We consider that it is not necessary for a commitment to exist neither that *x* believes that *z* will be performed by her, nor *x*'s normative beliefs about her action. However, it is necessary that *x* knows that she is entering into a commitment, even if dishonest.

Note that conditions 2 and 3 depend on social norms defining what kind of promises, for instance, should be kept. As we will argue later on this paper, we cannot trace the distinction between promise and threat (including promises and excluding threats from the category of social commitments) on the basis that no normative expectation affects threats: some of them indeed fall into normative beliefs and expectations.

Conditions 2 and 3 refer to the mandatory character of *z*. To say that "*z* should be done" means that there is an recognized and accepted reason for *z* to be done. Social commitments constitute social obligations, this is, socially acknowledged reasons for action (Miller, 2006). The fact of having committed oneself to perform action *z* is a socially accepted reason for performing action *z*, independently of whatever reasons motivated the agent for entering into the commitment, and independently of other reasons the agent may have for performing *z*.

Thus, condition 3, against Searle's account of commitment (defined as desire independent reason for action)[4], does not entail that the *x*'s reason for action is the commitment itself (*x* did *z* because she promised to), but implies that *x* considers the commitment a social obligation in the sense mentioned above.

We will now consider the problem of delimiting social commitments, in order to differentiate them from other social interactions. Particularly, we will analyze the relation between social commitments and social norms, and, on the other hand, the relation between commitments and other similar speech acts, such as a declaration of intentions.

[3] For instance, a commitment may be dropped if *x* comes to believe that *z* is unachievable; this would be a valid condition for dropping the commitment. However, the reasons that have led the agent to that believe may also be subject to negotiation with *y*. Social commitments, in general, cannot be unilaterally changed.

[4] See Searle, 2001, 2008.

Social norms. In our analysis of social commitments, social norms are present almost in every step in the commitment creation process. It is important, however, to distinguish social commitments from other social interactions consisting in the application, violation or attribution of a social norm. Conditions 2 and 3 from our analysis can be present in both cases; this is so because these conditions are related to the concept of social obligation, which is caused by social norms. Miller (2006) has argued that social obligations exist because there is a social norm that apply to that social interaction: for instance, he claims, the social norm “promises ought to be kept” is the source of normativity of social commitments of this kind. However, we do not agree with this claim. The example of transcultural social norm used by Miller is, from our point of view, too general to have explanatory power. The fact that some promises are legitimate and some are not show that social norms have to do not only with the act of committing itself, but with the content of the commitment. In fact, some commitments are not established because of their content (Davis, 2009). Miller is right in pointing out that the norm “promises ought to be kept” may indeed generate an obligation; however, the difference between what constitutes a promise and what does not is also a matter of social norms and other pragmatic circumstances. Recognizing a social interaction as a social commitment, and creating a social obligation, cannot only rely on a definitional norm about promising.

Then, what makes social commitments different from other interactions involving operations with social norms? The attribution of social norms seems very similar to an implicit commitment, this is, a social commitment described as above except that there is no communication between the agents. However, we think that social norms are, as Castelfranchi and Conte (2006) put it, two-sided objects, both internal (or mental) and external (social). Thus, for instance, the four conditions presented above would describe a situation in which both agents know that the norm exist and that it is being applied to one of them, which is not necessarily the case. We would then say that social commitments are a conscious and voluntary actions that pursue a goal or intention of the agent, and that the conditions mentioned above are necessarily present.

Speech acts. The relation between social commitments and social norms tend to be ignored in the analyses that focus on the communicative aspects of commitments, such as trust, manipulation and persuasion (Kurzban et.al., 2001; Müller, 2007; Schelling, 2007). It can be argued that, considering the Gricean maxim of quality a norm, every speech act may be analyzed as a social commitment. However, not every speech act generates a social obligation, this is, a reason for a future action. A declaration of intentions differs from a social commitment in that it does not attribute rights and obligations (related to the speaker's actions) to the speaker and the hearer, and thus the declaration itself does not bind the speaker in a social sense, although the speaker can feel internally committed. This is not, however, the kind of commitments we have tried to analyze. Some authors claim that, while promises constitute genuine social commitments, threats are mere declaration of intentions, because they do not generate normative beliefs or expectations (Castelfranchi, 1995; Miller, 2006). However, some threats can be considered social commitments (Castelfranchi and

Guerini, 2007); for instance, some conditional threats involve a promise of not performing the threat if certain conditions are satisfied. On the other hand, excluding threats as a form of commitment could lead to confusion when dealing with more sophisticated situations, in which the content of the commitment is not necessarily desired by all the agents performing it, such as multilateral contracts.

A social commitment exceeds the scope of the speech act that generates it. As we have argued, the speaker can propose a commitment, and the hearer can reject it, thus canceling the commitment creation process. Pragmatic and social conditions play an important role defining the creation, the status and the end of a social commitment.

Conclusions

Theories on commitment usually define the binding between the agent and the action set by a promise as a kind of internal commitment, a goal that the agent adopts. We have argued that this binding is better defined as a social obligation, this is, a socially accepted reason for action, and not necessarily the agent's reason for action. The mandatory character of social commitments can be thus explained by the interaction between social norms and the communication process (e.g., the act of promising) which generate a set of obligations and entitlements for the creator of the commitment, and for its attributer.

References

Bicchieri, C. (2006): *The Grammar of Society*, Cambridge Univ. Press

Castelfranchi, C. (1995): “Commitments: from Individual Intentions to Groups and Organizations”, in *ICMAS96*, AAAI/MIT Press, Cambridge, MA.

Castelfranchi, C. and Guerini, M. (2007): "Is it a Promise or a Threat?", *Pragmatics and Cognition*, 15 (2), 277-311

Conte, R. and Castelfranchi, C. (2006): “The Mental Path of Norms”, *Ratio Juris*, 19 (4), pp. 501-517

Davis, M. (2009): “Fourteen Kinds of Social Contract”, *Journal of Applied Ethics and Philosophy*, 1 (1), pp. 8-19

Graffeo, M, Savadori, L, Tentori, K., Bonini, N. and Rumiati, R. (2009): "Consumer Decision in the Context of a Food Hazard: the Effect of Commitment", *Mind & Society*, 8 (1), pp. 59-76

Miller, K. (2006): “Social Obligation as a Reason for Action”, *Cognitive Systems Research*, 7 (2-3), pp. 273-285

Müller, T. (2008): “Living up to one's Commitments: Agency, Strategies and Trust”, *Journal of Applied Logic*, 6 (2), pp. 251-266

Holton, R. (1999): “Intention and Weakness of Will”, *Journal of Philosophy*, 96 (5), pp. 241-262

Kurzban, R., McCabe, K., Smith, V. L. and Wilson, B. J.: 2001, “Incremental Commitment and Reciprocity in a Real-Time Public Goods Game”, *Personality and*

Social Psychology Bulletin, 27 (12), pp. 1662-1673.

Schelling, T.C. (2007): *Strategies of Commitment and Other Essays*, Harvard Univ. Press

Searle, J. (2001): *Rationality in Action*, Cambridge, MA, MIT Press

Searle, J. (2008): “Language and Social Ontology”, *Theory and Society*, 37 (5), pp. 443-459

Singh, M.P. (1999): “An Ontology for Commitments in Multiagent Systems”, *Artificial Intelligence* and Law, 7 (1), pp. 97-113

Sudgen, R. (1998): “Normative Expectations: the Simultaneous Evolution of Institutions and Norms”, in Ben-Ner, A. and Putterman, L. (eds.): *Economics, Values, and Organization*, Cambridge Univ. Press

Tuomela, R. (2007): *The Philosophy of Sociality: the Shared Point of View*, Oxford Univ. Press