Preface

In the past two decades, various logics and game theoretical analyses are proposed and combined to model various aspects of social interaction among agents including individual agents, organizations, and individuals having certain positions within organizations. The aim of SOCREAL 2007 (International Workshop on Philosophy and Ethics of Social Reality 2007) was 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 and ideas and get them into perspective. This was an eminent success.

SOCREAL 2007 consisted of one special guest lecture, seven invited lectures and presentations of three submitted papers. It was held in Sapporo, Japan, on March 9 - 10, 2007. It was organized by GPAE (Graduate Program in Applied Ethics, Graduate School of Letters, Hokkaido University). This volume is the proceedings of the workshop. It contains extended abstracts, slides, and full papers submitted by the participants.

Our thanks go to all the authors and participants for their stimulating lectures and lively discussions. We are also grateful to the anonymous reviewers for their valuable work in reviewing submitted abstracts.

September 2007

On behalf of the Organizing Committee
Tomoyuki Yamada
Organization

WORKSHOP CO-CHAIRS

Jose Carmo (Universidade da Madeira)
Tomoyuki Yamada (Hokkaido University)

ORGANIZING COMMITTEE

Jose Carmo (Universidade da Madeira)
Kohji Ishihara (Hokkaido University)
Nobuo Kurata (Hokkaido University)
Koji Nakatogawa (Hokkaido University)
Tomoyuki Yamada (Hokkaido University)

Supported by:

GPAE (Graduate Program in Applied Ethics)
Graduate School of Letters, Hokkaido University
Contents

Session 1: Collectivity

Logics for modeling (organized) collective agency
Jose Carmo (University of Madeira) .................................................................1(1)-21(21)

An Attempt to Secure the Status of Collectivities
Ryoji Fujimoto (Hokkaido University) ..............................................................1(22)-8(29)

A New Analysis of the Structure of Social Reality
Yasuo Nakayama (Osaka University) ..............................................................1(30)-10(39)

Session 2: Logical Dynamics of Social Interaction

Substructuralized Modal Logics Applied to the Two girls Puzzle
Hisashi Kitamura, Koji Nakatogawa, and Yohei Fukayama (Hokkaido University)
........................................................................................................................... 1(40)-14(53)

Logical Dynamics of Speech Acts
Tomoyuki Yamada (Hokkaido University) ......................................................1(54)-12(65)

Institutions in Channel Theory
Stijn De Saeger (Japan Advanced Institute of Science and Technology) ........1(66)-5(70)

Special Guest Lecture:

Adaptive Bases of Human Rationality
Nobuyuki Takahashi (Hokkaido University, The Center for the Study of Cultural and Ecological Foundations of the Mind)
.........................................................................................................................1(71)-34(104)

Session 3: Obligation and Rationality

Moral Conflicts between Groups of Agents
Allard Tamminga (University of Groningen) ..............................................1(105)-7(111)

Utilitarian Deontic Logic and a Sequel
Yuko Murakami (National Institute of Informatics) .................................1(112)-7(118)

A Strategic View of Promising
Jun Miyoshi (Kanto Gakuin University) .....................................................1(119)-13(131)

Limits of Logic in Ethical Decision-Making
Christopher Melley (Fachhochschule Kaiserslautern) .........................1(132)-8(139)

1 A full paper version of this contribution has been published as 'Moral Conflicts between Groups of Agents', Journal of Philosophical Logic, 37, 2008, pp. 1-21 (with Barteld Kooi).
Logics for modeling (organized) collective agency

José Carmo
Engineering and Mathematics Department
University of Madeira (Madeira, Portugal)

jcc@uma.pt

My thanks to the GPAE and to Tomoyuki Yamada

Part of this talk is based on some joint works with Olga Pacheco

Goal: to use and combine deontic and action modal operators, plus the counts-as operator, in order to characterize some concepts that are essential for the understanding of (organized) collective agency and agent’s acting and interacting in general.

We start by making a brief review of these logics, that we want to use, extend and combine.

Notation used:
Capital Latin letters (D, E, F, H, O, P) will be reserved for modal operators; a, b, ... for agents; φ, ψ, ... for sentences (assertions about the states of affairs, etc.); α, β for actions; s for the relevant normative/legal system (or for the “society”); X, Y for groups of agents; o for organizations and collective agents; r for roles; and (in the semantics) w, v, ... for worlds (states). Propositional (Boolean) connectives: ¬, ∧, ∨, →, ↔.
Part I - Preliminaries: modal logics we want to use and combine
Deontic logic

• The traditional approach sees deontic logic as a branch of modal logic
  1. the necessity operator is interpreted as meaning obligation, and
denoted by $O$
  2. the dual of $O$ ($\neg O\neg$) is then denoted by $P$ and interpreted as meaning
  permission
  3. and prohibition is expressed by an operator $F$, defined as $O\neg$

Standard deontic logic (SDL, for short)

• The models are the standard modal models $M=(W,R,V)$, where
  $W \cup \emptyset$ is the set of worlds (or states)
  $R$ is the accessibility relation (assumed serial): $wRv$ means $v$ is an ideal version of $w$
  $V$ assigns to each atomic sentence $p$ the set of worlds where $p$ is true
  $M \models_w O\varphi$ iff $\forall v (if \ wRv then M \models_v \varphi)$
• We get $O$ as a normal KD modality (using Chellas classification), that is:
  (PC) All instances of tautologies
  (K) $O(\varphi \rightarrow \psi) \rightarrow (O\varphi \rightarrow O\psi)$
  (D) $O\varphi \rightarrow P\varphi$
  $O$-necessitation (rule): if $\vdash \varphi$, then $\vdash O\varphi$
  plus Modus Ponens (MP)
SDL gives raise to a set of paradoxes
part of them related with the closure of the O-operator under logical consequence:

(RM)-rule: $\vdash \phi \rightarrow \psi$ then $\vdash O\phi \rightarrow O\psi$

And many other proposals have been made to try to solve the paradoxes:

✓ consider ideal and sub-ideal worlds + ideal and sub-ideal transitions between worlds
✓ consider e.g. variants of the minimal models of Chellas of the form: $M=(W,f_O,V)$, where

$f_O:2^W \rightarrow 2^W$ (denotes the set of worlds where proposition Z is obligatory) and

$M \models w O\phi$ iff $w \in f_O(||\phi||)$ where $||\phi|| = \{v: M \models = v \phi\}$
✓ consider dyadic obligation operators and a temporal dimension, or a preference (ideality) ordering of the worlds, or contexts

In this talk we do not have time to enter in details on this issue
(also the deontic component, in isolation, is not our main focus here)

There have been also proposals for defining deontic logic on the top of dynamic logic - briefly:

• Dynamic logic appeared in Computer Science related with the correctness proofs of programs.

• In dynamic logic, associated to each action $\alpha$ there exists a modal operator $[\alpha]$, and expressions of

the form $[\alpha]\phi$ are read "if $\alpha$ is executed, then after action $\alpha$, $\phi$ is the case"

• Models: $M=<W, \{R^{\alpha} : \text{for each action } \alpha\}, V>$

$w R^{\alpha} v$ if we can obtain v executing $\alpha$ on w

$M \models w [\alpha]\phi$ iff $\forall v (\text{if } w R^{\alpha} v \text{ then } M \models v \phi)$

The relations $R^\alpha$ associated to some forms of composition of actions (like sequencing, choice and

iterations) are then constrained in order to get the intuitive desired meaning. Particular complications

occur when we consider actions (like some programs) that may not terminate, or when we consider

an action that corresponds to the “negation of an action $\alpha$".

• J.-J. Meyer proposed to define deontic logic on the top of dynamic logic, as follows:

$F(\alpha) \leftrightarrow [\alpha]V$, $P(\alpha) \leftrightarrow \neg F\alpha \leftrightarrow \langle \alpha \rightarrow \neg \rangle V$ and $O(\alpha) \leftrightarrow F(\neg \alpha) \leftrightarrow \langle \neg \rangle V$

where $V$ denotes a special violation atom, $\langle \alpha \rangle = \neg (\langle \alpha \rangle \neg \alpha)$ (the “negation” of $\alpha$) expresses the

non-performance the action $\alpha$. (The deontic operators are applied to action terms)
The “brings it about” action/agency logic

- Kanger, Pörn and Lindahl have combined deontic and action logics as basic building blocks to describe social interaction and complex normative concepts. Their logics have sufficient expressive power to be able to articulate several distinctions at an appropriate abstract level, mainly in virtue of the modal logic of action they employ.
- They introduce an action operator (E) that relates an agent (α) with the effects of his action (ϕ), omitting details about the specific action that was performed and setting aside temporal aspects.
- The expression Eaϕ can be read as:
  - “the agent α brings it about that ϕ” (“agent α has brought it about that ϕ”)
  - or “agent α sees to it that ϕ is the case”
  - or “agent α is responsible for its being the case that ϕ”

For the characterization of the Ea operator we find different approaches:

1. Kanger and Pörn define Ea as Boolean combinations of two normal modalities. Pörn’s proposal can be seen as follows: consider standard models with two accessibility relations, associated to each agent α, R1α and R2α:

   (w,v) ∈ R1α iff everything which α brings about in w is the case in v
   (w,v) ∈ R2α iff not everything that α brings about in w is the case in v

   M |=w Eaϕ iff ∀v( if w R1α v then M |=v ϕ) and ∃v( w R2α v and M |=v ϕ)

   Idea: Eaϕ iff it is necessary for something α does that ϕ,
   but for α’s action it might have been the case that ¬ϕ

2. Dag Elgesem and others consider Ea as primitive, and define its semantics using variants of the minimal models (e.g. Santos & Carmo consider, for each agent α, a function fα such that fα(Z) denotes the set of worlds where agent α brings it about Z)
3. A more elaborated semantics can be found within the “stit theory” of Nuel Belnap and Michael Perloff. They use STITₜ ϕ instead of Eₜϕ and for them “agent a sees to it that ϕ” if “the present fact that ϕ is guaranteed by a prior choice of a”. Their models use histories, time, moments and agent’s choices:

M |= h₁, t Eₐ ϕ iff there is a time (a choice point) t₁ < t such that

(1) M |= h₁₁, t ϕ for every history h₁ belonging to the same choice set of h at time t₁

and (2) M |= h₁₂, t ϕ for some history h₂ with the same past and present that h at time t₁

Drawback: According to Belnap’s initial proposal, the schema ¬Eₐ Eₐ p (for a ≠ b) is valid, which excludes (at least some) notions of direct control.

• Axiomatic presentation

With few exceptions, all the approaches see Eₐ as a non-normal modality, satisfying

(Tautologies and Modus Ponens +)

(T) Eₐ ϕ → ϕ (Eₐ is a “success” operator)

(C) (Eₐ ϕ ∧ Eₐ ψ) → Eₐ (ϕ ∧ ψ)

(RE)-rule: If ⊢ ϕ ↔ ψ then ⊢ Eₐϕ ↔ Eₐψ

+ (No) ¬Eₐ T (where T denotes a tautology)

Schema (No) is used to capture the concept of agency itself:
when Eₐϕ is the case, the state of affairs ϕ is, in some sense, caused by or is the result of actions by agent a. No agent can meaningfully bring about what is logically true or, more generally, what was unavoidable.
The “brings it about” approach to the logic of action offers an expressive power rather different from others action logics (that is particularly useful in some contexts):

✓ one can express several different positions in which an agent \( a \) might be with respect to a certain state of affairs \( \phi \), such as \( E_a \phi \) (did), \( E_a \neg \phi \) (averted) and \( \neg E_a \phi \neg E_a \phi \) (remained passive)

✓ as well as notions of control of other agents, like \( E_a \neg E_b \phi \) (a made \( b \) do), \( E_a \neg E_b \phi \) (a made \( b \) avoid)

✓ Moreover we can combine the action operators with the deontic operators and use that to express various kinds of normative relations between agents.

Suppose e.g. that \( rmr(x,y) \) denotes “\( x \) reads medical record of \( y \)”, \( d \) denotes a doctor and \( p \) a patient. Suppose also that according to some Hospital regulation “Patients do not have the right to access to his or her own record”. What does this exactly mean?

\[
PE_{drmr}(d,p) \land \neg PE_{prmr}(p,p) \quad \text{or} \quad PE_{drmr}(d,p) \land \neg PE_{prmr}(p,p) \land OE_d \neg E_{prmr}(p,p) \quad \text{or} \ldots
\]

Stig Kanger and Lars Lindahl have used these kind of combinations to express legal concepts and relations like rights, duties, etc, and Ingmar Pörn used them to the study of “control and influence” relations in social interactions

Comparing with the dynamic logic approach

✓ the dynamic operator \([\alpha]\) is a kind of conditional operator, whereas \( E_a \) is a "success" operator;

✓ \([\alpha]\) is centered on the actions, whereas \( E_a \) is centered on the results (abstracting from the particular actions performed);

✓ \([\alpha]\phi \) is evaluated before the hypothetical execution of the referred action, whereas \( E_a \phi \) is evaluated after the execution of the relevant action(s);

✓ Dynamic logic is particularly suited to describe state changes and the effects of actions, whereas \( E_a \) is static

✓ Central to the “brings it about” concept is the notion of agency and causation / responsibility.
Some extensions and refinements of these action logics:

Santos, Jones & Carmo proposed a non-necessarily successful action operator $H_a$, with the following informal meaning:

$$H_a \phi$$ means that agent $a$ has attempted to bring about that $\phi$.

(Although we may bring about something without intention, the concept of attempting clearly presupposes intention.)

Naturally, $H_a$ does not verify the (T) schema.

We may combine both operators on actions.

For instance, we may say that an agent $a$ has an effective (or total) control with respect to the state of affairs $\phi$, if $(H_a \phi \rightarrow E_a \phi)$ is the case.

Some other extensions and refinements of these action logics:

Sometimes it is also useful to distinguish between the direct and immediate effect’s of agent $a$’s actions, and those indirect effects that follow from such direct effects, either sometime later by some causal connection, or by institutional connection (or by other reasons), effects that can also be attributed to the responsibility of $a$’s actions (and described through the $E_a$ operator),

Herein we will use $D_a \phi$ to state that

“agent $a$ has directly (just) brought it about that $\phi$ is the case”

and we will continue to use the operator $E_a$ to express states of affairs that are generically brought about by agent $a$, in any generic broad sense.
Preliminaries: Counts-as

- Andre Jones & Marek Sergot (1996):
  “within institutions, organizations, or other normative systems, there are rules that state that some acts, or some state of affairs count as, or are to be classified as acts, or state of affairs of a different kind (rules that may differ from system to system)”

- To express such count as (“meaning”) relations they propose the use of a conditional modal operator, denoted by $\Rightarrow_s$ (where the index $s$ refers the relevant system of analysis)

- Although they do not restrict the use of $\Rightarrow_s$ connected to act-descriptions, that is their main goal:
  $$E_a \varphi \Rightarrow_s E_b \psi$$
  “$a$’s act of bringing about $\varphi$ counts, within institution $s$, as a means by which agent $b$ (who may but need not be the institution $s$ itself) establishes state of affairs $\psi$; $a$ may be said to act on behalf of, or as an agent of $b$”

- Jones & Sergot also introduce a normal modal operator, $D_s$, where $D_s \varphi$ may be read as follows: “it is incompatible with the rules operating in institution $s$ that $\varphi$ is not the case”.

  Herein we use $D$ for direct action. Thus we will use $R_s$ instead of $D_s$, and read $R_s \varphi$ as meaning that “according to the rules operating/accepted in institution $s$, $\varphi$ is the case” or “it is recognized/accepted by institution $s$ that $\varphi$ is the case”

- For $\Rightarrow_s$ they proposed a conditional logic of type CE, containing the rules of replacement of logical equivalents both in the antecedent and in the consequent, plus the following schemas (but not their converses):
  $$(\varphi \Rightarrow_s \psi_1) \wedge (\varphi \Rightarrow_s \psi_2) \rightarrow (\varphi \Rightarrow_s (\psi_1 \wedge \psi_2))$$
  $$(\varphi_1 \Rightarrow_s \psi) \wedge (\varphi_2 \Rightarrow_s \psi) \rightarrow ((\varphi_1 \lor \varphi_2) \Rightarrow_s \psi)$$

- For the operator $R_s$ they propose it as a normal modality of type KD, rejecting both the T-schema $R_s \varphi \rightarrow \varphi$ and its converse $\varphi \rightarrow R_s \varphi$

- And as bridging principles, they propose
  1. $(\varphi \Rightarrow_s \psi) \rightarrow R_s (\varphi \rightarrow \psi)$
  2. they also adopt $(\varphi \Rightarrow_s \psi) \rightarrow (\varphi \rightarrow R_s \psi)$

  as a means of securing the stronger detachment principle $(\varphi \Rightarrow_s \psi) \rightarrow (\varphi \rightarrow R_s \psi)$
Deontic logic again:

It is natural to try to use deontic operators to direct individual personal agency (by forbidding some actions, making others obligatory, etc.), and we could conceive explicit personal deontic operators, like:

\[ O_\alpha \phi : \text{\(\alpha\) is under an obligation of doing} \ \phi \]

\[ P_\alpha \phi : \text{\(\alpha\) is permitted to do} \ \phi \]

\[ F_\alpha \phi : \text{\(\alpha\) is forbidden to do} \ \phi \]

- A natural question is if such personal deontic operators need to be primitive, or can be defined in terms of other operators. As we have seen, we can combine and iterate impersonal deontic operators and action operators, and it seems natural to define:

\[ O_\alpha \phi = O \ E_\alpha \phi \]

\[ P_\alpha \phi = P \ E_\alpha \phi \]

\[ F_\alpha \phi = F \ E_\alpha \phi \]

- However, we should be aware that these definitions have some consequences:

First we lose the interdefinability between the personal deontic operators:

We still get \(P_\alpha \phi \leftrightarrow \neg F_\alpha \phi\) as well as \(O_\alpha \phi \rightarrow P_\alpha \phi\) (if \(O\) satisfies the D-schema), but \(F_\alpha \phi \leftrightarrow O_\alpha \neg \phi\) is no longer valid (since \(O_\alpha \neg \phi = O \ E_\alpha \neg \phi\) is stronger that \(F_\alpha \phi = O \neg E_\alpha \phi\))

We do not think that this is a problem!

- But we will have problems, if \(O\) verifies the (RM)-rule: if \(|\vdash \phi \rightarrow \psi\), then \(|\vdash O \phi \rightarrow O \psi\)

- Main problem - the problem of transmission of obligations:

\[ \vdash E_\alpha B \rightarrow E_\alpha B \]

by the (RM)-rule:

\[ \vdash O_\alpha \phi \rightarrow O_\alpha \phi \]

(which is clearly unacceptable)

\[ \vdash E_\alpha \phi \rightarrow E_\beta \neg \phi \]

by the (RM)-rule:

\[ \vdash O_\alpha \phi \rightarrow F_\beta \neg \phi \]

making impossible to express conflicts of obligations between different agents.

What the deontic status of \(\alpha\) has to do with the deontic status of \(\beta\)?

- However, as we have already referred, we can define the \(O\) operator in such a way that it does not verify the (RM)-rule (possibly satisfying some weaker versions of it)

- But how to define and characterize deontic operators when we want to use them to direct the agency of groups of agents, organizations, etc.?
Part II - Organized Collective Agency

Agents acting

- Agents may be subject of obligations and in order to fulfill them they must act. And by acting and interacting, the agents can modify the relevant state of affairs, as well as create new obligations (for instance making contracts).
- And, as we have already seen, we may combine deontic and action operators (for instance of the brings it about type) to describe the obligations that apply to each agent, and the effects of his actions on the relevant state of affairs.
- But if we want to describe social interaction we cannot avoid joint actions and collective agency.

Joint action

- Two or more agents can jointly act in order to do some task (e.g. to move a very heavy table, to make a contract, etc), and we can extend the brings it about operator to a set $X$ of agents in order to capture that (as was done by Lindahl):
  $$E_{X}\psi : \text{the set of agents described in } X \text{ jointly cooperate to bring about that } \psi \text{ is the case}$$
- In this way we can express some notions of collective agency, and define logics where formulas of the form $E_{\{a,b\}}\psi \land \neg E_{a}\psi \land \neg E_{b}\psi$ (with $a \neq b$) can be consistent.
- However, by obvious reasons, we reject a general principle of the form $E_{X}\psi \rightarrow E_{Y}\psi$, for $X \subseteq Y$.
- Using this operator, we can describe the establishment of contracts. For instance, $a$ and $b$ may establish a contract between them, by which $a$ becomes under the obligation of doing $\psi$ and $b$ becomes under the obligation of doing $\varphi$:
  $$E_{\{a,b\}}(O_{a}\psi \land O_{b}\varphi)$$
Collective agency

- Suppose now that a set $X$ of agents wants to act collectively in a more or less permanent basis
- First hypothesis: whenever the group $X$ wants to act, all the members of $X$ meet and jointly act
- In such case we could use the previous action operators to express such situations.
- But, if we assume that the group $X$ wants to act collectively in a more or less permanent basis, probably it will interact with other agents and groups, making contracts, etc., and in this way creating obligations on the proper group $X$ of agents.
- The question now is how we can characterize such kind of collective obligation: $O_X$
  - Since we are assuming that the group always act through a joint act of all its members, it is natural to assume that the obligation $O_X \phi$ will have the form of an obligation on a joint action of the group $X$, which we may represent by an expression like $O_{EX} \phi$
  - However, this does not solve our problem.

  We defend that only agent acting can be deontically qualified, because if an obligation is not fulfilled we must know who is potentially subject to punishment.

  - According to this point of view, such collective obligation must be defined in terms of obligations on the agents, and, at a first sight, in terms of individual obligations of the members of $X$.
    Two options seem then apparently natural
    1) A “general (or universal) obligation”, where an obligation on the group corresponds to an obligation on each of its member
       \[ O_{X} \phi = \forall_{ag \in X} O_{ag} \phi \]
    2) A “weak general obligation”, where an obligation on the group corresponds to an obligation on some of its members
       \[ O_{X} \phi = \exists_{ag \in X} O_{ag} \phi \]
    - Option 2) does not serve: Not only it validates $O_{X} \phi \rightarrow O_{Y} \phi$, for $X \subseteq Y$, as it is of little interest in practice. If the obligation is not fulfilled, who is the responsible?
    - In the case under analysis, where group $X$ always act through the joint agency of all its members, option 1) seems a good option, but we must be careful in stating the obligation that applies to each of its members. We do not want to say that $\forall_{ag \in X} O_{ag} \phi$ (suppose e.g. that $X$ is a football group/team and $\phi$ is “to mark (at least) five goals on today’s game”)
    - The most natural interpretation of $O_{EX} \phi$ would be, in this case
      \[ \forall_{ag \in X} O_{ag} E_{EX} \phi \text{ or, using the attempt operator, } \forall_{ag \in X} O_{ag} E_{EX} \phi \]
Collective agency – Acting in the name of, counts-as and direct acts

• Continue to suppose that a group X wants to act collectively in a more or less permanent basis
• However, the previous case where the group always act through a joint act of all its members is not the most usual case
• Usually, the group will organize its activities in some stable way, allowing that some acts may be performed in the name of the group by some of its members. In that case, the group will create a statute (or an internal code) stating that such is the case.
• Suppose for instance that the group decides that:
  i) its member $a$ may bring about certain type of states of affairs $\phi$ in the name of the group;
  ii) or that any of its members may bring about certain type of states of affairs $\phi$ in the name of the group (suppose that X is a group of Mafia killers and they have a code that “when one kills, all kill”)
• If we want to characterize such situation, what it will be required ?

One hypothesis would be to state:

(*i) $E_X (E_a \phi \rightarrow E_X \phi)$ and (**i) $E_X \forall_{ag} (E_{ag} \phi \rightarrow E_X \phi)$

However, these formalizations have some imprecision’s and drawbacks:

Suppose we have: (**ii) $\forall X, E_a \phi$ and $\phi$ is-dead($b$).
Should we derive that $E_X$ is-dead($b$) ?
First, $\forall_{ag} (E_{ag} \phi \rightarrow E_X \phi)$ may describe only an internal agreement of the members of X, not necessarily accepted by the “external world” (the “society” $s$).
The normative system may not recognized that $E_X$ is-dead($b$) is the case, if $a$ has killed $b$ alone.
(The situation is not the same as when $E_X \forall_{ag} $is-dead($b$) and $a$ kills $b$).

In such case we should replace the material implication by the counts-as operator

(**ii) $E_X (E_a \phi \Rightarrow_X E_X \phi)$ and (**ii) $E_X \forall_{ag} (E_{ag} \phi \Rightarrow_X E_X \phi)$

Then, when $E_a \phi$ is the case, although we cannot derive that $E_X \phi$ is the case, we can derive $R_X E_X \phi$, where an expression like $R_X \psi$ could be read as follows: “(according to the rules accepted by X – by the members of X) , it is recognized/accepted by X (interpreted as it is recognized/accepted by all the members of X) that $\psi$ is the case”.

12
• Of course, there may exist cases where the legal system gives power to X to allow someone to act on its name (for instance, by signing an appropriate document, a family X may give power to someone to sell the family’s house).

In such cases we could write e.g.

\[(***) \quad E_X (E_a \varphi \Rightarrow E_X \varphi)\]

and if we do not want to talk about different legal systems, we may identify truth (of some kind of statements) with its recognition by the relevant “legal system” (and, for practical purposes, we can dispense the “counts-as” operator) and simply write

\[(*) \quad E_X (E_a \varphi \rightarrow E_X \varphi)\]

• However, there are still other reasons by which the formula \(E_a \varphi \rightarrow E_X \varphi\) still does not represent exactly the state of affairs that the group X wants to create.

In fact, not all acts of \(a\) are made in the name of the group X and it is possible that \(a\) may bring about \(\varphi\) for himself, and not in the name of the group X (suppose e.g. that \(\varphi\) represents “buy a Fiat car”), in which case we do not want to derive \(E_X \varphi\) from \(E_a \varphi\).

• An agent can do a similar direct act playing different roles, but to know the effects of such act and its deontic classification, we must know in which role it was played.

For instance, an administrator of a company may be permitted to drive a company’s car when on duty – i.e. when he is acting in the quality of administrator – but may be not permitted to use that car when on holiday;

and even if he is permitted to drive that car on holiday, if he has a car accident the responsibility of repairing the damage caused will depend on the role he was playing when he had the car accident (probably the company will be responsible for repairing the damage, if, and only if, he was on duty).

• Thus, it becomes necessary to express the quality in which \(a\) has acted (the role \(a\) was playing) when he or she brought about \(\varphi\).

• Using \(E_{a,X} \varphi\) (or \(E_{a \ as \ X} \varphi\)) to denote that \(a\) as brought about \(\varphi\) as if it was \(X\) that has acted (\(a\) as brought about \(\varphi\) in the name of \(X\), or in the role of representative of \(X\)), then we can write (where \(E_{X,X} \varphi\) may be read “\(X\) as brought about \(\varphi\) as itself”)

\[E_{a,X} \varphi \rightarrow E_{X,X} \varphi\]
• Naturally, we can question how we know that $\alpha$ as brought about $\varphi$ as himself or in the name of the group $X$. We will discuss such issue later.

• Another question is how we can discriminate in our language the cases where

1) $X$ brought about $\varphi$, because some agent has brought about $\varphi$ in the name of $X$

2) from the cases where $X$ brought about $\varphi$ by itself (by a joint act of the group $X$)

• We can introduce a notation to express 1), like

$E_{X/\alpha}\varphi$

(or more precisely,

$(\text{read “according to the system s, } X \text{ has brought about } \varphi \text{ through } \alpha \text{’s acting”})$

as an abbreviation of

$E_{s,X}\varphi \land (E_{s,X}\varphi \Rightarrow E_{X/\alpha}\varphi)$

• Other option is to consider the direct action operator, $D$, to describe the direct acts made by an agent (or the direct joint acts made by a group $X$ of agents), and discriminate 1) and 2) as follows:

1) $E_{X/\alpha}\varphi \land \neg D_X\varphi$

2) $E_{X/\alpha}\varphi \land D_X\varphi$

• Finally, how should we characterize now a “collective obligation” $O_X\varphi$ for the kind of group $X$ of agents under analysis?

• Since we are not in a presence of a group $X$ that always act through a joint act of all its members, allowing that some members can act on his name, we should not interpret $O_X\varphi$ as meaning $O_{DX}\varphi$.

A reasonable interpretation might be $O_{EX}:\varphi$

• But now this kind of collective obligation $O_X\varphi$ should not be seen as a general obligation of some form like

$O_X\varphi = \forall_{ag \in X} O_{ag}\varphi$

• Although the “weak general obligation” $O_X\varphi = \exists_{ag \in X} O_{ag}\varphi$ still does not is what we want (by the reasons already explained), in some sense it gives us a kind of “meta-rule” stating what the group $X$ must do (when it creates its statute). The group must guarantee that

$O_X\varphi \Rightarrow \exists_{ag \in X} O_{ag}\varphi$

And, for this, the group will state in its statute, something like

$O_X\varphi \Rightarrow OE_{X/\alpha}\varphi$

where the particular agent $\alpha$ may be dependent on the type of statement $\varphi$ the obligation refers to (and also giving representative powers to such acts)
Collective agency – collective agents and roles

- Suppose now that, as before, a group X of agents have common interests and wants to act collectively in a more or less permanent basis, but such intended common activity should proceed even if some member of the group is not (in some moment) anymore interested in it, or when it is possible to aggregate later to the group other members.
- In such cases, the natural way for the group X to proceed is to create an entity with a proper identity (like it is the case of an organization). The members of the group will be related with such entity by special relationships, like member-of (associate-of, or whatsoever), but such entity will persist even when the set of its members will change.
- Of course, such entity that is created by the group X needs to act, and so it is an agent. We have called it an “institutionalized agent; herein we will call it a “collective agent” (even if it may happen that there exists only a unique person related with it).
- In some cases (like when we talk about organizations), the Law will recognize such entity as a “real” agent (an "artificial person"), having juridical personality and legal competence, as any natural person. In other cases, that we do not want to exclude here, such “collective agent” may be more informal, without a legal recognized status.

- Naturally, such “collective agent” needs to act, but it cannot act directly! Thus, someone needs to act on its name. When some group X creates a “collective agent” o, a statute is also created where the main normative rules of o's activity are defined, and where, in particular, it is stated the rules by which someone can act in the name of the “collective agent”.
- The main difference for the previous case is that such rules do not state who is the concrete person that can act in the name of the collective agent. The “collective agent” (the organization) is usually structured in terms of what we may call posts, or roles within the organization (we may call them structural roles, meaning that they correspond to roles defined in the structure of the organization), and the statute of the organization describes who has the power to act in the name of the organization. But this description is abstract, in the sense that it does not say which particular person can act in the name of the organization; it attributes such power to the holders of some roles (the titulars of some posts), that have a representative nature.
- Normally, there is an individual post (in the sense that have only one holder), like president-of, that can act in the name of the organization, and when the organization it is created not only the statute defines the rules for electing/choosing the holder of such position, but it also defines a (usually provisory) initial titular of such position. The statute also defines if the titular of such post has power to delegate the power of acting in the name of the organization, and in which conditions.
• As we said, to some roles are associated representative powers.

Suppose that \( r : \text{REP}(o, \varphi) \) means that "the role \( r \) is a representative role (a role with representative powers) of the collective agent \( o \), with respect to a state of affairs \( \varphi \) (the scope of the representation)".

The notation \( r : \text{REP}(o, \varphi) \) does not mean that the role \( r \) can act in the name of \( o \)

• We think that only agents can act, and roles are not agents (thus, roles do not act)

What \( r : \text{REP}(o, \varphi) \) does mean is that when someone, playing the role \( r \), brings about \( \varphi \), that can be seen as an act in the name of \( o \), which can be expressed by

\[
\forall ag \ (E_{ag;r}\varphi \to E_{o:o}\varphi)
\]

where \( E_{ag;r}\varphi \) means "\( ag \) acting in the role \( r \) brings it about that \( \varphi \)" and \( E_{o:o}\varphi \) means "\( o \) acting in the role of itself brings it about that \( \varphi \)"

• Using the counts-as operator, we can reformulate the previous definition as follows, assuming that such representative status is recognized by the "society" \( s \) (the relevant normative system), and write

\[
\forall ag \ (E_{ag;r}\varphi \Rightarrow s E_{o:o}\varphi)
\]

• Naturally the organization has some general duties from the start, and the agents that can act in the name of an organization may establish new obligations for the organization through their acts, for instance by establishing contracts with other agents (persons, organizations, etc.). And such duties of the organization must be distributed among the different posts, specifying the norms that apply to those that occupy such positions (that hold such roles), and usually attributing to them the power to act in the name of the organization, with respect to the fulfillment of such duties.

• And in this way we have a dynamic of obligations, where the obligations flow from the organization to the holders of some roles, and these, through their acts, create new obligations in the organization.

• The statute of the organization normally distributes the "general duties" of the organization among the different posts, and give power to the holders of some posts to make a similar distribution of the concrete duties that are attached to the organization through its normal activity (by the acts of those that act in the name of the organization).
• The attribution of a duty of an organization \( o \) to a particular role \( r \) of its structure can be described through formulas of the form

\[
O_o \varphi \rightarrow O_r \varphi
\]

or more precisely (as before, such obligation \( O_o \varphi \) on a collective agent \( o \) should be interpreted as \( OE_{o(o)} \varphi \))

\[
OE_{o(o)} \varphi \rightarrow O_r \varphi
\]

where the attribution of deontic notions (obligations, permissions and prohibitions) to roles is defined as follows

\[
O_r \varphi = \text{def} \forall ag \in X (\text{qual}(ag:r) \rightarrow OE_{ag:r} \varphi)
\]

\[
P_r \varphi = \text{def} \forall ag \in X (\text{qual}(ag:r) \rightarrow PE_{ag:r} \varphi)
\]

\[
F_r \varphi = \text{def} \forall ag \in X (\text{qual}(ag:r) \rightarrow FE_{ag:r} \varphi)
\]

where \( \text{qual}(ag,r) \) is true if and only if the agent \( ag \) holds the role \( r \) – \( ag \) is qualified to play the role \( r \)

In general, given a role \( r(...) \), \( \text{qual}(ag,r(...)) \) is a predicate that can be described as \( \text{is}(ag,...) \) :

\[
\text{qual}(a, \text{president_of}(o)) = \text{is-president_of}(a,o)
\]

• Examples:

\[
F_{\text{administrator-of}(o)} \varphi \] informally means that:

all the administrators of \( o \) are forbidden of doing \( \varphi \), when acting in the quality of administrators of \( o \)

\[
OE_{o(o)} \varphi \rightarrow O_{\text{president-of}(o)} \varphi \] informally means that:

the president of \( o \) inherits, from the organization \( o \), the obligation of doing \( \varphi \)

• The previous definitions allow the deontic characterization of roles independently from the agents that hold them in a particular moment.

Even in the (more or less frequent) cases where we have a role that can have only one holder (like \( \text{president-of}(o) \)), it is still useful to attach deontic notions to the role (defined as above), instead of directly to the actual holder of such role, since this one can change.

• However, in such cases (of a role \( r \) with only one holder) it still might be useful to consider abbreviations to refer to its actual holder, like (for instance) the \( r \)
• Using such abbreviation, and defining (similarly to what we have done before)
  \[ E_{b:r1} \land \psi \] 
  (or, simply, assuming \( s \) implicit, \( E_{b:r1/a:r} \psi \))

  (read “according to the normative system \( s \), agent \( b \) has brought about \( \psi \) playing role \( r1 \), through \( a \)’s acting playing role \( r' \)” as an abbreviation of
  \[ E_{a:r} \psi \land (E_{a:r} \psi \Rightarrow E_{b:r1} \psi) \]

  we can express a possible policy of an organization \( o \), like “(according to its statute, recognized by the normative system \( s \)) the organization \( o \) always acts through its president”, as follows:
  \[ E_{o:o} \psi \rightarrow E_{o:o/the-president-of(o):president-of(o)} \psi \]

• In previous works, we have defined a formal language and logic where it is possible to define precisely all these things, including the description of roles and the possibility of defining some relations between roles (incompatibility, sub-roles, etc.), as well as a more informal language for the specification of organizations and interactions between agents (individual and collective) through contracts.

---

**A little more on roles**

• Special relationships are created between an organization and other agents, to which are associated norms that describe the ideal expected behavior of the agents engaged in such relationships and the consequences of the acts made by them. To such relationships correspond *roles* that agents can play.

• But roles are not agents (although it is natural to informally identify a role with its holder, when this is unique), neither sets of agents, although associated to a role we have the (possible singular) set of agents that are qualified to play such role.

  Not only the same set of agents may correspond to the set of holders of two distinct roles (\( a \) may be the president of two distinct organizations, and in order to know the effects of his acts we must know the role \( a \) was playing), as the set of holders of some role may (and usually do) change through time

• Within the context of organized collective agency, roles are used as a high-level mechanism for structuring the desired behaviors, by the association to roles of deontic notions, powers, etc.

  But roles should not be confused with their normative characterization, i.e. roles should not be reduced to mere sets of obligations, permissions or other normative concepts (that apply to the holders of such roles). In particular, the normative characterization of some role may change with time.
• Roles are a very important high-level structural mechanism for abstractly specifying how a collective agent acts and interacts.

• But the concept of role, and of acting in a role, is relevant in many contexts, and not only within the context of the behavior of organizations.

• In our opinion, roles may be seen as corresponding to qualities that the agents might exercise that are relevant when we describe agents acting and interacting with others. Such qualities may either express properties of an agent, independently of others (like being father, owner of a building, etc.), or express relationships with other agents (like being president-of, associated-of, etc.).

• Naturally, in practice, we only associate a role to a property if the fact that someone has that property may be relevant for some of his acts (because, for instance, they are permitted only for persons having that property): e.g., only the owner of the building $xpto$ can sell building $xpto$.

Which acts count as acts in a particular role?

• An agent can be the holder of different roles within the same organization or in different organizations (being the subject of potentially conflicting obligations), and can act by playing different roles. And in order to know the (legal, institutional, ...) effects of his acts (and its deontic qualification) we must know in which role they were played.

• Therefore it is fundamental to know which acts count as acts in a particular role.

• On the other hand, in reality what we have is agents directly acting (which we are expressing through formulas of the form $D \phi$)

• Thus, the precise question seems to be in what conditions some acts will be recognized as acts in some role by the environment, the organization, the society, the normative system, etc.? The fact that a direct act is recognized as an act in some role $r$ must be deduced from some information related with such act (and the agent involved). For instance:

✓ We may have a (possibly implicit) rule that states that a notary, when performing certain kind of acts $\phi$, like signing legal documents, always act in the quality of notary (for the normative system $s$):

$$D_a \phi \land \text{is-notary}(a) \Rightarrow E_a : \text{notary} \theta$$
Analogously, we may have that any (relevant) state of affairs $\phi$ brought about by the president of an organization $o$, inside its building, counts as if it was brought about by the president, on that quality:

$$D_o \phi \land \text{is-president-of}(a,o) \land \text{is-in-the-building-of}(a,o) \Rightarrow E_o \phi$$

In some cases, we may have that an act of an agent $a$ will be considered as an act in any role he can play (or an attempt to act, similarly to what is considered within a calculus for access control in DS)

$$D_a \phi \land \text{qual}(a;r) \Rightarrow E_a:r \phi$$

or that such is the case only when it is also permitted to bring about that $\phi$ when acting in that role:

$$D_a \phi \land \text{qual}(a;r) \land \text{Pr} \phi \Rightarrow E_a:r \phi$$

In the previous examples, as in most cases, the qualifications of an agent to play some role does not change in consequence of his acts. When this is not the case, it seems that the real relevant moment to evaluate the qualification is immediately before the act, and for the recognition of an direct act as an act playing some role what may be relevant is (in some cases) the particular type of action performed and the circumstances when the agent initiated the relevant act.

In general, there exist some conventional signals that an agent must exhibit in order to guarantee that the act that he will perform will be recognized as an act playing some role. However, these action logics do not provide us the means to talk about such two states (before and after acting).

• One hypothesis is to try to combine these action operators with dynamic operators, and a possibility (among others) is to allow to write expressions of the form $[D_a \phi] \psi$ informally interpreted as

$$\forall \alpha \left[\alpha: a \right] (D_a \phi \rightarrow \psi)$$

Then we could state, for instance, that

$$D_a \text{doc-own}(a,\text{xpto}) \rightarrow [D_a \phi] R_a E_a: \text{owner-of}(\text{xpto}) \psi$$

where $\text{doc-own}(a,\text{xpto})$ = “$a$ has exhibited the document of ownership of building xpto”

meaning that when an agent $a$ shows a document that certicate his ownership of building xpto, then any direct (relevant) act that he will made in that moment (like sign a document selling that building) will be recognized (or assumed) as an act playing the role of owner of that building.

And in this example we could even write that (where $<D_a \phi> \equiv \neg[D_a \phi]$)

$$<D_a \phi> R_a E_a: \text{owner-of}(\text{xpto}) \psi \rightarrow D_a \text{doc-own}(a,\text{xpto})$$

sentence that would mean that if it is possible for $a$ to directly bring it about some state of affairs $\phi$ (e.g “sign a document”), that will be recognized as an act in the role of owner of building xpto, for bringing about (the same or other possibly different) state of affairs $\psi$ (e.g. “selling xpto”), this means that $a$ has directly exhibited the respective document of ownership of xpto. This would be a way to express the exact act that is expected as an authentication to act as owner of building xpto.
Conclusions

- The logical framework here considered, that takes an approach to the treatment of actions which is based on expressions of the form “$a$ brings it about that $\phi$ is the case”, focus only on the agent concerned and the states of affairs that result from his or her actions. This framework is intended primarily as a formal tool for the characterization of static phenomena.

- The proposed level of abstraction seems to be appropriate when we want to characterize and model human and organizations interaction, at an abstract level, where we do not know yet or we do not care about the exact type of actions that can be executed, and where we want to concentrate and characterize how the obligations flow from an organization to the holders of some posts and how some of the acts of the latter count as acts of the organization.

- However, the proposed level of abstraction, providing no resources for representing and reasoning about the temporal dimension, the effects of state change and specific actions, also has severe limitations, that limits its applicability.

- The combination of these actions logics with dynamic logic operators may be a way to overcome these limitations.

End

and

*Arigato gozai masu*
I cannot provide a list of necessary and sufficient conditions for holding a collective responsible for the faulty actions of some of its members, but I can perhaps do some of the reflection, which would precede the creation of such a list.

_Terrorism and Collective Responsibility_, Burleigh T. Wilkins

0. Introduction

During the last decade philosophers have paid much attention to the structure of social reality. Social reality comprises social conventions and norms, social roles and relations, social institutions and artifacts and so on. A good deal of attention has focused on the nature of collectivities, i.e. social groups, associations, and corporations. And some philosophers have explored collective responsibility and group rights.

In this regard, the key question is whether a social collectivity amounts to something significantly over and above the simple sum of individual property. Individualists say that social collectivities are determined by individuals and supervene on individual properties. This priority given individual agency is taken by some to require showing that social attitudes can be reduced to individual attitudes. Therefore, they have a negative view of the collective responsibility. Others are opposed to the individualistic view, claiming that a collective action or intentionality isn’t reducible to an individual action or intentionality.

In this paper I will consider the above mentioned controversy. My aim in what follows is to undermine individualist assumptions about the explanatory priority of individual agency. I shall do this by arguing that there is as much reason to ascribe action and responsibility to social collectivities as to individual agents. To bring to light the problems involved in collectivities and show the relation to the real ethical issue, I will take up the issue of collective responsibility. Then we are faced with the problem of whether collectives count as moral agents. I give an affirmative to this problem and

---

1 According to Tuomela(1989), actions by collectives “supervene” on the actions of the operative members of the collective in such way that the properties of particular collectives are “embodied in” and “determined by” the perspective of the properties of individual members.
provide a warrant for my position.

1. Distributive and Non-distributive Collective Responsibility

Let us begin with a clarification of the notion of collective responsibility. According to Feinberg (1968), people often have some different models in mind when they speak of “collective responsibility”. He classifies the notion of collective responsibility into some groups. These are: group liability without fault, group liability with noncontributory fault, contributory group fault: collective and distributive, and contributory group fault: collective but not distributive. It is beyond the scope of this paper to scrutinize his classification. But, it is important to notice the distinction between the notion of distributive or shared collective responsibility and the notion of non-distributive collective responsibility. In the case of distributive collective responsibility, we can attribute moral responsibility or blameworthiness to a whole group because of the contributory fault of each and every member. Since each individual brings about the harm in question purposefully, no one’s responsibility is vicarious. After all, the distributive collective responsibility is simply the sum of all the individual responsibility. In contrast, the notion of non-distributive collective responsibility amounts to something significantly over and above the simple sum of individual responsibility. It is this notion that arouses the philosophical controversy.

Traditionally, it is understood that the notion of moral responsibility grounds moral blameworthiness in the wills of discrete individuals. Most people agree that the collective responsibility would make sense if it were merely a sum of individual responsibility. Because it can be reduced to individual responsibility, they accept the notion of distributive collective responsibility. But they disagree notably about whether collective responsibility makes sense as a non-distributive phenomenon, i.e., as a phenomenon that transcends the contributions of each member of the group.

Defenders of non-distributive collective responsibility argue that the notion of moral responsibility can be associated with groups, as distinct from their individual members, and located in the collective actions taken by groups. They construe groups as moral agents in their own right. And they frequently maintain as follows: While groups understood as collectives generally act through their individual members, their actions don’t coincide with their member’s actions. Nor is their moral agency merely the moral agency of their members or the moral agency of group representatives. Instead, such agency is an agency that is attached to the collective itself and hence not the kind of thing that can be distributed across group members. (Unless otherwise noted, in what follows the term of “collective responsibility” refers to the notion of non-distributive
collective responsibility.)

The notion of non-distributive collective responsibility does not fit into the prevailing philosophical idea on moral responsibility. H.D. Lewis, one of the methodological individualists, challenges the possibility of associating moral agency with groups understood as collectives. According to Lewis (1948), collective responsibility makes no sense because groups, unlike individuals, cannot formulate intentions of the kind normally thought to be necessary to actions and hence groups cannot be understood to act or to cause harm qua groups. He says that “it is the individual who is the sole bearer of moral responsibility”. In addition, Lewis worries about the fairness of ascribing collective responsibility to individuals who don’t themselves directly cause harm. So, he claims that “no one is morally guilty except in relation to some conduct which he himself considered to be wrong”.

These claims each can be related to two philosophical controversies. The first controversy concentrates on the metaphysical foundations of collective responsibility e.g., is it possible for groups to have intentions, for instance. The second controversy focuses on the consequences of ascribing collective responsibility in practice e.g., how can we ascribe moral responsibility to groups for harms that only a few of its members directly caused without violating individual responsibility.

2. Intentionality

Methodological individualist’s claim is based on the assumption that all actions necessarily begin with intentions. This assumption is very useful to critics because it enables them to write group intentions into the definition of collective action itself and hence render group intentions a necessary condition of collective responsibility. They say that a collective action must be caused by the beliefs and desires of the collective itself. And how, critics ask, can groups be understood to have intentions? Lewis here claims that actions are associated exclusively with individuals, not groups, and groups, which don’t have minds of their own, cannot hold beliefs in the sense required by the formulation of intentions.

Defenders of collective responsibility rely on a variety of strategies to debunk the above claim. Interestingly enough, many defenders don’t go as far as to assert that all kinds of groups are capable of acting and intending collectively. Instead, they assert that only particular kinds of groups are capable of acting and intending collectively and being

---

2 Sverdlik (1987) says that “it would be unfair, whether we are considering a result produced by more than one person’s action or by a single person, to blame a person for a result that he or she did not intend to produce.”

3 For example, see Corlett’s definition of a collective action (Corlett 2001, p.575).
collectively responsible for harm. In this regard, French's classification of collectives is useful. He distinguishes between aggregate and conglomerate collectives. According to French, the former are random collections of persons e.g. the crowd, mob, rioter, whereas the latter are organized e.g. nations, corporations. Conglomerate collectives have a characteristic of an "organization of individuals such that its identity is not exhausted by the conjunction of the identities of the persons in the organization"(French 1984, p.13).

French considers conglomerate collectives to be appropriate bearers of moral responsibility because of some features that they share. In other words, he think that some features signal the existence of collective intentional actions that are capable of rendering groups collectively responsible for harm. It is the most important feature of them that is a series of well-ordered decision-making mechanisms through which collective actions can be chosen. A series of well-ordered decision-making mechanisms should permit a conglomerate collective to admit and expel members at will according to its valid rule system. This affords the conglomerate collective the freedom to determine the boundaries of its own membership. By virtue of the well-ordered decision-making mechanisms, conglomerate collectives are able to demonstrate two things that are assumed to be necessary to collective responsibility. The first is a set of group actions that have an identifiable moral agent, e.g. a governing board or a representative body. The second is a set of decisions that are made self-consciously on a rational basis by the group that take the form of group intentions.

In addition, when we find a fault of a member of groups, defenders of collective responsibility can respond more effectively to the fault than methodological individualists. Under the individualistic view, there is nothing to the individual fault but the individual fault itself. But, defenders of collective responsibility can separate the fault resulting from playing his role in group with the structural defect from his individual fault itself. And we can cope with the fault in a coordinated way. We may make every effort to improve a structure of the group. In contrast, individualist may punish the individual, but they will never hold a view on the reformation of the group. After all, individualists leave the group problematic and another member repeats the same mistake made by his predecessor.

4 Other features are (1) a set of enforced standards of conduct for individuals that are more stringent than those usually thought to apply in the larger community of individuals, and (2) a configuration of defined roles by which individuals can exercise certain powers(French 1984).
Some peoples including French argue that collectives are moral agents. David Copp (1984), for example, says that certain collectives, namely conglomerate collectives such as nations and corporations that have rule-governed and highly structured decision-making capacities, can be intentional agents. There are difficulties, however, that confront any claim to the effect that such collectives do act intentionally such that they might qualify as moral agents.

It might be thought that conglomerate collectives do not act intentionally for two reasons. First, they do not act intentionally as individual humans often do. Instead, individual constituents of the conglomerate collective act on its behalf. Another reason why it might be thought that conglomerate collectives do not act intentionally is that it is possible to reduce ascriptions of collective agency to attributions of individual agency in congruence with recognized rule systems. These claims would lead one to adopt some version of moral responsibility individualism.

It seems clear that aggregate collectives do not act intentionally because they do not function according to a recognized formal or informal rule system. Aggregate collectives such as crowds and rioters lack common goals and interests, though their respective members can possess goals and interests. Crowds and rioters are aggregate collectives, and aggregate collectives are simply a loose collection of individual human persons. It seems, then, that such collectives are not plausible candidates for intentional action attributions.

But perhaps conglomerate collectives can act intentionally. In fact, nations and corporation act, though not in a primary way. They are “secondary agents”. A primary agent is one who has the capacity to act on her own intentionally. A secondary agent is one for whom another acts according to a legal or moral rule system intentionally. In the case of secondary agency, both the one on behalf of whom the action is performed and the one performing the action in her name are intentional agents, but in different respects. The secondary agent, for whom the act is carried out, must have the capacity to have the action carried out according to her beliefs and desires. The one acting in her name must be capable of performing her action such that her own doings are caused by her wants and beliefs. For example, an attorney acts on behalf of her clients, and states, corporations. This makes at least some conglomerate collectives secondary agents to the extent that their agents properly represent their putative aims and purposes according to a rule system.

There is no doubt that nations and corporations (and other conglomerate collectives) often behave (without intention) according to official rules of their respective systems. But it is unclear that behavior resulting from such decision making is the result of the
intentionality of the conglomerate collectives themselves or whether it is the consequence of the intentionality of certain powerful decision-makers in those collectives. If a corporation’s attorneys successfully defend that corporation against all suits brought against it for its alleged corporate wrongdoing, then it does make some sense to say that the corporation vindicated itself from charges. But, it does not necessarily follow from this that what the corporation did constitutes an action. At best, it is a doing-related event. Action entail intentionality, doing do not. So, we require an independent argument to show that what a conglomerate collective does amounts to an action, that is, that what it does is caused by its own wants and beliefs and not merely the wants and beliefs of certain powerful individuals of that conglomerate collective.

An interesting analysis of collective intentionality has been presented by Michael Bratman (1993). According to Bratman’s view shared intentions are complexes of appropriately constrained individual intentions interrelated in appropriate ways. The intentions of individuals form a web, and that web is the collective intention. His account is attractive because it captures the widespread intuition that collective intentions should rely on the intentions of individuals, but it also shows that collective intentions are different from collections of individual intentions. I think that the account of shared intention developed by Bratman is fundamentally an account of substantive collective intention.

3. Collective Responsibility and the Values of Individual Freedom

The second controversy concentrates on the consequences of ascribing collective responsibility in practice. Is it possible, participants in this controversy ask, for individual members of a group to be collectively responsible for group-based harms in cases where they did not directly cause it? In other words, the question is whether the whole community can be held responsible for the harms produced by particular group members. Does responsibility attach to all members of the perpetrator’s group, no matter what their actual role?

Individualists claim; it is impossible, for it is unfair, whether we are considering a result produced by more than one person's action or by a single person, to blame a person for a result that he did not intend to produce. The very idea of collective responsibility appears to deny the traditional link between responsibility and control, since one's group membership is often beyond one's power of choice, e.g., membership in racial, ethnic or national communities. It also seems to violate the liberal ethics of the “separateness of persons”. Rawls (1971) has used this phrase to communicate the idea that each individual has a separate, non-fungible value. According to Rawls, in
assigning responsibility, one must consider each person separately and weigh his moral accountability on the basis of his own actions. The notion of collective responsibility threatens to assign to each member of a group the worst crimes of his fellows, thereby ignoring the differences among, and independent dignity of, individuals. The potential for abuse is significant. In treating the issue of collective responsibility, we must be mindful of this danger. Then, individualists try to establish general criteria for distributing collective responsibility among group members.

Suppose that an interview with a man who survived a massacre committed on his family by some members of a rival ethnic group. He said; I know that not all of them were in on it. Not all of them have blood on their hands. But, you know, not one of them has come up to me and said, “I’m sorry for what my people did”. He felt that even those among his neighbors who did not actually harm him or his family owed him something-an apology. It seems that it is difficult to justify this man’s claim, but intuitively his reaction is perfectly understandable.

Imagine a person (X) who is a member of an ethnic group that has committed atrocities against another group. X himself, however, has not directly caused any harm. He didn’t himself kill, beat, or rape, though many of his group did. Might X still share the responsibility foe the crimes? It seems most people would say that he might. For instance, there are many ways in which X might have been an indirect or partial cause of the atrocities. If X publicly advocated violence against the rival ethnic group, he shares a lot of the blame for the resulting harm. If he voted for a political regime that he believed capable of these sorts of acts, and if that regime later went on to commit these acts, X is certainly partly responsible.

Another way in which individuals can come to share in the responsibility for the wrongs committed by their groups is through acts of omission. Suppose that X was in a position to prevent some of the crimes committed by his group without subjecting himself to any unreasonable risks, and that he knew this, but still he did nothing. X would surely share in the responsibility for the outcome. Here, the link between responsibility and causation is broken. In omitting to act, X did not cause the harms in any straightforward sense. Still, the link between responsibility and what is in the agent’s control remains. X could have acted but did not. His omission was a choice. Since choice and control are still relevant to the attribution of responsibility, the separateness of persons principle is respected. However, the lack of a causal link does have an effect on our understanding of the sort of responsibility involved. Depending on the details of the case, “shame” may be a more appropriate response to acts of omission than blame.
The reason comes from a consideration of the following situation. A student is beaten in full sight of classmates (X, Y, and Z) assembled around him. And, if someone tried to subdue the attacker, things would settle down (all classmates have ability to subdue). Let's suppose that X subdue the attacker. Now, can X's action release Y and Z from their own obligations to subdue, or can Y and Z be blamed for failing in their obligations? In this case, it is obvious that someone (X or Y or Z or X&Y or X&Z or Y&Z or X&Y&Z) ought to subdue the attacker. But, for example, we cannot go as far as to argue that “Z” ought to subdue the attacker. Therefore, we cannot claim that Y and Z can be blamed for failing in their obligations.

In my opinion, above-mentioned “ought to” (“someone ought to subdue the attacker”) is proper for the supererogation and is not relevant with the notion of obligation. So, even if X or Y or Z does not try to subdue the attacker, they cannot be blamed for it.

I have discussed several ways in which people can share in the responsibility when their fellows cause harms. Larry May (1992) says “just as our relationships with others are multitudinous and wide-ranging, so is the domain of responsibility”. It is very difficult to give a fully unified account of collective responsibility. But, if we establish the conception of collective responsibility, we will be more likely to act in ways that can reconcile groups in the aftermath of harms and avoid future harms.

Reference
Aim of this talk

- I examine several theses in Searle (1995) and point out some difficulties in his description.
- I clarify some notions and show how to overcome these shortcomings.


Searle’s notion of institutional facts

- According to Searle (1995), institutional facts can be explained by means of three notions, i.e.
  
  - collective intentionality
  - assignment of function
  - constitutive rule.

Problems with Searle’s notion of collective intentionality

- Searle claims collective intentionality is primitive and biologically founded.
  - His insistence on the primitiveness of collective intentionality makes it difficult to analyze interactions of intentional states among agents.

Problems with Searle’s notions

- Searle's schema of assignments of function and schema of constitutive rules are too inflexible, so that the application of these schemata becomes unnecessarily restricted.
Collective Intentionality

Collective Intentionalit

Primitiveness of collective intentionality (Searle (1995) p. 25f)
– It is indeed the case that all my mental life is inside my brain, and all your mental life is inside your brain, and so on for everybody else. But it does not follow from that all my mental life must be expressed in the form of a singular noun phrase referring to me.

The form that my collective intentionality can take is simply “we intend,” “we are doing so-and-so,” and the like. In such cases I intend only as part of our intending. The intentionality that exists in each individual head has the form “we intend.”

Problem with this description
A problem with this description is that it is unclear what is stated by “I intend only as part of our intending”.
• A possible interpretation:
  – We, i.e. group G, intend that \( p \) if and only if every member of G believes that we intend that \( p \).
  \[ I_G p = \forall x \in G \left( B_I x G p \right) \]

Problems with Searle’s notions

Collective Intentionality multi-agent BDI logic
• In this presentation, I use multi-agent BDI logic explained in Meyer and Veltman (2007).
• I slightly modify notations of some modal operators.
  – I write \( ^G \text{C} \) instead of \( \text{C} \).
  – \( ^G \text{C} \) means: Group G collectively believes that \( p \).

Multi-agent BDI formulas
• \( B_i p \): i believes that \( p \).
• \( ^E_G p \): everybody in G believes that \( p \).
• \( ^G \text{C} p \): Group G collectively believes that \( p \).
• \( I_i p \): i intends that \( p \).
• \( ^E_G p \): everybody in G intends that \( p \).
• \( ^M_G p \): all members in G mutually intend that \( p \).
• \( ^G \text{C} p \): Group G collectively intends that \( p \).
Models for n-agent BDI logic

- Models for n-agent BDI logic are Kripke structures of the form
  \[ M = (W, V, R_1, \ldots, R_n, R_{F,G}, R_{D,G}, S_{1}, \ldots, S_{n}, S_{F,G}, S_{D,G}) \]
  where
  - \( W \) is a non-empty set of states (or worlds);
  - \( V \) is a truth assignment function per state;
  - \( G \) is a subset of \( \{1, \ldots, n\} \);
  - The \( R_i \) are accessibility relations on \( W \) for interpreting the modal operations \( B_i \), assumed to be serial, transitive and euclidean relations, while the \( S_i \) are accessibility relations on \( W \) for interpreting the modal operations \( I_i \), assumed to be serial relations.
  - \( R_{F,G} = \cup_{i \in G} R_i \) and \( S_{F,G} = \cup_{i \in G} S_i \);
  - \( R_{D,G} = R_{D,G}^* \) and \( S_{D,G} = S_{D,G}^* \), the transitive closure of \( R_{D,G} \) and \( S_{D,G} \) respectively.

Interpretation of multi-agent BDI formulas

- \( M, w \models \rho \) iff \( V(w)(\rho) = true \), for \( \rho \in \mathcal{P} \);
- The logical connectives are interpreted as usual;
- \( M, w \models B_i \rho \) iff \( M, w \models \rho \) for all \( w' \) with \( R_i(w, w') \);
- \( M, w \models E_i \rho \) iff \( M, w \models \rho \) for all \( w' \) with \( R_{E_i,G}(w, w') \);
- \( M, w \models E_{i,G} \rho \) iff \( M, w \models \rho \) for all \( w' \) with \( S_{E_i,G}(w, w') \);
- \( M, w \models I_i \rho \) iff \( M, w \models \rho \) for all \( w' \) with \( S_{I_i,G}(w, w') \);
- \( M, w \models M_i \rho \) iff \( M, w \models \rho \) for all \( w' \) with \( S_{M_i,G}(w, w') \);

Valid formulas of multi-agent BDI logic (1)

- \( \mathcal{E}_{G,D} \equiv \forall x \in G \, (\mathcal{I}_x \rho) \)
- \( \mathcal{M}_{G,D}(\rho \rightarrow q) \rightarrow (\mathcal{B}_{G,D} \rho \rightarrow \mathcal{M}_{G,D} q) \)
- \( \mathcal{M}_{G,D} \rho \rightarrow \mathcal{E}_{G,D} \rho \)
- \( \mathcal{M}_{G,D} \rho \rightarrow \mathcal{E}_{G,D} \mathcal{M}_{G,D} \rho \)
- \( \mathcal{M}_{G,D}(\rho \rightarrow \mathcal{E}_{G,D} \rho) \rightarrow (\mathcal{E}_{G,D} \rightarrow \mathcal{M}_{G,D} \rho) \)
- Definition of collective intention
  - \( \mathcal{C}_{G,D} \rho = \mathcal{I}_G \rho \land \mathcal{B}_G \mathcal{M}_{G,D} \rho \)

Valid formulas of multi-agent BDI logic (2)

- \( \mathcal{E}_{G,D} \equiv \forall x \in G \ (I_x \rho) \)
- \( \mathcal{M}_{G,D}(\rho \rightarrow q) \rightarrow (\mathcal{B}_{G,D} \rho \rightarrow \mathcal{M}_{G,D} q) \)
- \( \mathcal{M}_{G,D} \rho \rightarrow \mathcal{E}_{G,D} \rho \)
- \( \mathcal{M}_{G,D} \rho \rightarrow \mathcal{E}_{G,D} \mathcal{M}_{G,D} \rho \)
- \( \mathcal{M}_{G,D}(\rho \rightarrow \mathcal{E}_{G,D} \rho) \rightarrow (\mathcal{E}_{G,D} \rightarrow \mathcal{M}_{G,D} \rho) \)

Clarification

- Multi-agent BDI logic clarify interactions among agents in a group.
- I introduce index \( G \) to clarify problems for a person who belongs to different groups.
- If a person \( A \) belongs to both \( G_1 \) and \( G_2 \), then the union of the collective belief set of \( G_1 \) and that of \( G_2 \) must be consistent because of the validity of
  - \( \mathcal{B}_{G_1,D} \rho \rightarrow \mathcal{E}_{G_1,D} \rho \) and \( \mathcal{B}_{G_2,D} \rho \rightarrow \mathcal{E}_{G_2,D} \rho \)
- Otherwise, \( A \)'s belief becomes inconsistent.
Description of a Problem

- If a person $A$ belongs to both $G_1$ and $G_2$, then the collective belief set of $G_1$ and that of $G_2$ must be consistent because of the validity of $\neg C_{G_1,p} \rightarrow \neg E_{G_1,p}$ and $\neg C_{G_2,p} \rightarrow \neg E_{G_2,p}$.
- In reality, this is sometimes not the case. There can be a person who cannot decide which group he wants to join, where the union of the collective beliefs of the two groups are inconsistent.
- This is a real problem and it should be also describable within Searle’s framework. I doubt if this is possible.

Conditional intention and self-restriction

- A conditional intention “$i$ intends to do $A$, if $i$ realizes that $p$” can be expressed by $B_i(\rho \rightarrow I_i do(i, A))$.
- A conditional self-restriction “$i$ does not intend to do $A$, if $i$ realizes that $p$” can be expressed by $B_i(\rho \rightarrow \neg I_i do(i, A))$.
- I presuppose here: $B_i(I_i do(i, A)) \rightarrow I_i do(i, A)$.

Conditional obligation and prohibition

- A conditional obligation “Do $A$, if $p$” is accepted by $x$, if $x$ has a conditional intention expressed by $B_x(\rho \rightarrow I_x do(x, A))$.
  - A categorical obligation can be expressed by “Do $A$, if $T$”, where $T$ expresses a tautology.
- A conditional prohibition “Do not $A$, if $p$” is accepted by $x$, if $x$ has a conditional intention expressed by $B_x(\rho \rightarrow \neg I_x do(x, A))$.
  - A categorical prohibition can be expressed by “Do not $A$, if $T$”, where $T$ expresses a tautology.

Searle’s constitutive rules

- A constitutive rule has the form “$X$ counts as $Y$ in context $C$”.
- The structure of collective rules can be iterated.
- Institutional facts exist only in systems of constitutive rules. (1995, p. 28)

Why can rules create facts?

- Rules work properly only in a rule system.
- An action changes a former state.
- A rule describes what kind of actions are allowed and what kind of actions are prohibited.
- A rule system contains descriptive sentences and rules.
- A rule system introduces a new language based on the given language.
- A rule system provides a new interpretation strategy for actions.
Example: Rules of chess

- Chess is a two-player game that is played on an 8-by-8 chessboard, with thirty-two pieces (sixteen for each player) of six types; each type of piece moves in a distinct way.
- The goal of the game is to protect the most valuable piece, the king, and trap (checkmate) the opposing king.

From Wikipedia

Rules of chess: Game play

- The player controlling the white pieces moves first.
- After the initial move by white, players alternate moves.
- Play continues until a draw is called, a player resigns or a king is trapped by means of a checkmate.

Rules of chess: Movement

- Each piece moves in a different way. Generally, a piece cannot pass through squares occupied by other pieces, but it can move to a square occupied by an opposing piece, which is then "captured" (removed from the board). Only one piece can occupy a given square.
  - The rook moves orthogonally to the players (forward, backward, left or right) any number of squares.
  - The knight moves in an "L"-shape (two spaces in one direction and one space orthogonally to it). It is the only piece that can jump over other pieces.
  - The bishop moves diagonally any number of squares and always stays on one of the two chequered colours.
  - The queen moves orthogonally or diagonally, any number of squares.
  - The king moves orthogonally or diagonally only one square at a time.
  - The pawn moves one space straight forward (away from the player). On its first move it can optionally move two spaces forward. If there is an enemy piece diagonally (either left or right) one space in front of the pawn, the pawn may move diagonally to capture that piece. A pawn cannot capture or jump over a piece directly in front of it.

Rules of chess constitute a rule system

- The rule system of chess enables interpretations of actions in chess.
- These interpretations are shared by the players.
  - Example for interpretation of actions in chess
    - Description of an action (in the standard English):
      - Peter took one of white figures on a board and placed it on a different square of the board.
    - Interpretation of this action (described in the chess-language):
      - Peter moved a rook two spaces forward.

Interpretation of Searle’s form for constitutive rules

- Searle’s form for constitutive rules has the form "\textit{X counts as Y in context C}.
- I interpret this form as follows:
  - \textit{C} is a rule system.
  - \textit{X} is described by a language that does not presuppose the rule system.
  - \textit{Y} describes \textit{X} from the viewpoint of the rule system.

My interpretation of "\textit{X counts as Y in context C}:

- My interpretation.
  - \textit{C} is a rule system.
  - \textit{X} is described by a language that does not presuppose the rule system.
  - \textit{Y} describes \textit{X} from the viewpoint of the rule system.
- Example for interpretation of actions in chess
  - Description of an action (in the standard English):
    - Peter took one of white figures on a board and placed it on a different square of the board.
  - Interpretation of this action (described in the chess-language):
    - Peter moved a rook two spaces forward.
A rule system provides an interpretation system for real actions

• Question:
  – Why can a rule system create facts?

Answer to:
Why can a rule system create facts?

- A rule system provides a new interpretation system for real actions.
- This interpretation system is collectively believed by the players and they interpret actions (in the game) in the same way.
- A descriptive part in a rule system introduces new notions and describes constraints for these notions.
- In this way, we can explain why a rule system can create facts.
  – This explanation is deeper than Searle's one.

Assignment of Function

Form of the assignment of function in Searle (1995)

• The function of $A$ is to $X$.
• Example (p. 14)
  That river is good to swim in.
• Objects with a function (p. 14)
  – chairs, tables, houses, cars, lecture halls, pictures, streets, gardens


• Whenever the function of $X$ is to $Y$, $X$ and $Y$ are parts of a system where the system is in part defined by purposes, goals, and values generally.
• Whenever the function of $X$ is to $Y$, then $X$ is supposed to cause or otherwise result in $Y$.

Goal-directed Actions (My view)

- We can use things in order to achieve a certain goal.
- In some rule systems, there are things whose use in it is restricted by its rules.
  – Example in chess:
    • Chess pieces such as king, rook, queen, pawn, knight, and bishop.
- The assignment of function can be explained, when we explain what a rule system is.
Acceptance of a rule system

• A rule system is collectively accepted by \( G \) iff all sentences in it are collectively believed or accepted by \( G \).

• A normative proposition \( p \) is collectively accepted by \( G \) iff it is collectively believed by \( G \) that proposition \( p \) is accepted by all members of \( G \).

G-institutional-facts

• The truth conditions of a G-institutional-statement depend on G-collective-beliefs.

• A G-institutional-fact is expressed by a corresponding true G-institutional-statement.

• G-institutional-facts are socially constructed, whereas physical facts are not.

Relations among Institutional Facts

• There are three cases for two groups \( G_1 \) and \( G_2 \):
  1. \( G_1 \subseteq G_2 \) (or \( G_2 \subseteq G_1 \));
  2. \( G_1 \cap G_2 = \emptyset \);
  3. \( G_1 \cap G_2 = \emptyset \).

Questions

– \( G_1 \subseteq G_2 \) : Are all \( G_2 \)-institutional facts also \( G_1 \)-institutional facts?
– \( G_1 \cap G_2 = \emptyset \) : Are all \( G_1 \)-institutional facts consistent with \( G_2 \)-institutional facts?
G-declaration according to Nakayama (2004)

- A G-declaration is addressed to G-members and it expresses a speaker's desire for a G-collective-belief.
- A G-declaration can be successful only if it is stated by a person whose authority concerning the declared content is accepted by G.
- A G-declaration can create a G-institutional fact, because it can create G-collective-beliefs.

Importance of G-declaration

- A G-declaration can create G-institutional facts.
  - Introduction of legal systems
  - Nomination of ministers, a chairman, …
  - Control of meetings

Social Organizations

A thesis about the ontological status of social organizations

- Social organizations, such as states, companies, and universities, are four-dimensional fusions of individual objects.


1. A social organization G has a structure.
2. This structure is so formed that it enables the continuation of the existence of the social organization.
3. Any member of G knows that he belongs to G.
4. G collectively believes that G exists.

Acceptance based on a (expert) group
The requirement of collective belief is sometimes too strong

- Conditions for G-collective-belief are sometimes too strong, so that they are normally not satisfied by real situations.
- We need a more flexible notion for collectivity.
  - G-acceptance based on a (expert) group E

A recursive characterization of G-acceptance based on a (expert) group E

a) A proposition $p$ is accepted by $G$ based on group $E$, if proposition $p$ is collectively believed by $E$ and it is collectively believed by $G$ that $E$ is the expert group for the subject expressed by $p$.
b) A proposition $p$ is accepted by $G$ based on group $E$, if there is a group $F$ such that proposition $p$ is accepted by $E$ based on group $F$.

Putnam’s division of linguistic labor

- Putnam’s division of linguistic labor corresponds to an acceptance based on an expert group.

Putnam’s hypothesis of the division of linguistic labor (Putnam (1975) p. 228)

- Every linguistic community exemplifies the sort of division of linguistic labor just described:
  - that is, possesses at least some terms whose associated ‘criteria’ are known only to a subset of the speakers who acquire the terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets.
  - The subset of the speaker corresponds to an expert group in my description.

Consistency of G-acceptance

- Let $S(G)$ be the set of G-accepted sentences. Then, $S(G)$ should be consistent.
  - Otherwise, any sentence can be inferred from $S(G)$.
  - Suppose $G_1$ and $G_2$ are expert groups in $G$ ($G_1$ $\subseteq$ $G$ and $G_2$ $\subseteq$ $G$). If $G_1$ and $G_2$ have different views, then there is no G-accepted sentence with respect to the controversial issues.

Consistency of G-law-systems

- Law systems that are applicable to the organization $G$ should be consistent each other or it must be written in them how to resolve contradictions.
- In a modern society $G$, the correctness of the official procedure is often crucial for G-acceptance.
Meta-rules for G-acceptance

• In our modern society G, we have meta-rules that define when a proposal of rules is G-accepted.
  – Example for meta-rules for Japan-acceptance
    • Some meta-rules for Japan-acceptance is written in the Japanese constitution.
    • A Japanese law is Japan-accepted, if it is Diet-accepted.

Political Decision Group

• In a democratic system, it is often decided by a G-election, who belongs to a political decision group for G.
  – For example, members of the Diet.

Conclusion

Conclusion 1

• I pointed out some problems of Searle (1995) with respect to description of the social reality.
• I proposed the reason why a rule system can create new facts.
• I proposed a new interpretation of (G-)declaration and explained why a successful G-declaration can create institutional facts.
• I pointed out that we need an appropriate notion of G-acceptance-based-on-a-group, in order to explain the structure of social reality.

Conclusion 2

• I pointed out that it is crucial for an explanation of social reality to make it explicit who is involved in particular institutional facts and in a particular collective intentionality.
• The structure of social reality is far more complex than described in Searle (1995).
Substructuralized modal logics applied to the two wise girls puzzle

Hisashi KITAMURA Koji NAKATOGAWA Yohei FUKAYAMA

1 Introduction

In this paper, we will discuss the relation of substructural logics to a logical puzzle called “the two wise girls puzzle” that is presented in Yasugi and Oda [8]. In our ordinary life, we often expand our own beliefs by hearing the utterances expressing some beliefs, and make some inference by use of them. This puzzle focuses on the situation in which the beliefs of several agents influence one another in this way. Using an inference system, Yasugi et al. formulate the inference involving the beliefs held by the agents of the puzzle in question, and discuss the solvability of the puzzle. To begin with, we will review the puzzle and an outline of Yasugi et al’s approach to it. Next, we will introduce a system of substructural logics and show that there is a doubtful point to the solvability of the puzzle in the system. Moreover, we will say that there are several candidates for the expression of the modal axiom. We will assume that the readers are familiar with the sequent calculus LK.

2 “The two wise girls puzzle”

We will explain the basic setting of the puzzle in question. There are three persons, and we call two of them “agent1” and “agent2.” Suppose that agent1 and agent2 are aligned in the same direction and that agent2 is located in front of agent1. Therefore, agent1 can see agent2 and agent2 cannot see agent1, but agent2 can listen to the agent1’s voice if agent1 says something. Suppose further that these two agents wear their hats. In this case, neither
knows the color of the hat of agent2, but the location of these two agents enables agent1 to know the color of her hat. Agent2 wears a white hat but the color of the agent1’s hat is not questioned. The observer, who is different from agent1 and agent2, tells the agents that at least one agent wears a white hat. The observer says to agent1, “Do you know if your hat is white?” Then agent1 answers “No! I don’t know.” After that, the observer asks the agent2 the same question, and then agent2 answers “Yes, I know.” If agent2 reaches the correct conclusion, that is, is able to know that she herself wears a white hat, what inference makes her obtain the conclusion in question?

Yasugi and Oda formulated this situation by use of an inference system and explained the question of why agent2 reached the correct conclusion. Their system is the sequent calculus based on a modal propositional logic and is called KD4^2. In addition to the usual logical symbols of propositional logics, this contains the modal operators B_1 and B_2, which are used to express the belief of agent1 and agent2, respectively. Moreover, KD4^2 contains the propositional constant 1W which means the proposition that agent1 wears a white hat and the propositional constant 2W which means the proposition that agent2 wears a white hat.

The axioms are not different from the ones of the sequent calculus LK. As for inference rules, the following two inferences rules (B_i → B_i) and (B_2 → B_2) are added to the ones of the inference system LK^1.

\[ \frac{\Gamma, B_i \Gamma' \rightarrow \Delta}{B_i \Gamma, B_i \Gamma' \rightarrow B_i \Delta (B_i \rightarrow B_i)} \quad (i = 1, 2) \]

---

^1 Yasugi and Oda’s system does not contain the rules of exchange and contraction in an explicit way. However, since they define the sequent not as the form “the list of formulae → the list of formulae” but as the form “the set of formulae → the set of formulae” [8, p. 148], it would be appropriate to regard the rules as implicitly presupposed. In this paper, when we introduce the system CFL_\text{KD4}^2, we define the sequent as the form “the sequence of formulae → the sequence of formulae.”
Here each of the symbols \( \Gamma, \Gamma' \) and \( \Delta \) denotes the set of logical formulae, and \( \Delta \) has at most one element. In addition, \( B_i \Gamma = \{ B_i(A) | A \in \Gamma \} (i = 1, 2) \).

These rules are introduced as alternatives of the axioms involving the modal operators \( K, D \) and \( 4 \). When one defines the sequent calculus of modal logics, these three axioms are often introduced as the following form of inference rules:

\[
\frac{\Gamma \rightarrow A}{B_i \Gamma \rightarrow B_i(A)} (B_i-K) \quad \frac{\Gamma \rightarrow (B_i-D)}{B_i \Gamma \rightarrow (B_i-D)} \quad \frac{B_i \Gamma \rightarrow A}{B_i \Gamma \rightarrow B_i(A)} (B_i-4) \quad (i = 1, 2)
\]

These correspond to the axioms \( K, D \) and \( 4 \), respectively\(^2\). In fact, when \( \Gamma' = \emptyset \) and \( \Delta \neq \emptyset \), \( (B_i-K) \) is \( (B_i-K) \), and when \( \Gamma' = \Delta = \emptyset \), \( (B_i-D) \) is \( (B_i-D) \), and when \( \Gamma = \emptyset \) and \( \Delta \neq \emptyset \), \( (B_i-D) \) is \( (B_i-4) \). Conversely we can show that \( (B_i-D) \) holds by use of \( (B_i-D) \) and \( (B_i-4) \) when \( \Delta = \emptyset \), and by use of \( (B_i-K) \) and \( (B_i-4) \) when \( \Delta \neq \emptyset \).

By the use of the modal operators \( B_1 \) and \( B_2 \), they described the sets of sentences expressing the beliefs held by agent1 and agent2. They described \( \Gamma_1 \) the set of sentences denoting beliefs held by agent1 as follows:

\[
\Gamma_1 = \{ B_1(1W \lor 2W), B_1(2W) \}
\]

Here \( B_1(1W \lor 2W) \) is the belief held by agent1 when she hears the observer’s utterance that at least one agent wears a white hat while \( B_1(2W) \) is the belief held by agent1 when she looks at the back of agent2. In contrast, since the agent2’s belief changes as agent2 hears the agent1’s utterance, they described the situations before and after this change. They described \( \Gamma_2 \) the set of sentences denoting beliefs held by agent2 before she hears the agent1’s utterance as follows:

\[
\Gamma_2 = \begin{cases} 
  B_2(1W \lor 2W), B_2(B_1(1W \lor 2W)), & \\
  B_2(2W \supset B_1(2W)), B_2(\neg 2W \supset B_1(\neg 2W)) 
\end{cases}
\]

\( B_2(2W \supset B_1(2W)) \) and \( B_2(\neg 2W \supset B_1(\neg 2W)) \) are the beliefs obtained by agent2 when she considers the location of agent1 and herself. Moreover, they described \( \Gamma'_2 \) the set of sentences denoting beliefs held by agent2 after she hears the agent1’s utterance as follows:

\[
\Gamma'_2 = \Gamma_2 \cup \{ B_2(\neg B_1(2W)) \}
\]

\(^2\)Strictly speaking, \( (B_i-K) \) corresponds to what is brought by combining \( B_i(A \supset B) \supset (B_i(A) \supset B_i(B)) \) with the necessitation \( A \supset B_i(A) \). Using \( (B_i-K) \), we can show that \( (B_i-D) \) corresponds to the axiom \( D \), that is \( B_i(A) \supset \neg B_i(\neg A) \). In the same way, using \( (B_i-K) \), we can show that \( (B_i-4) \) corresponds to the axiom \( 4 \), that is \( B_i(A) \supset B_i(B_i(A)) \).
Γ₂’ is an extension of Γ₂. \( B₂(\neg B₁(2W)) \) is the belief obtained by agent2 when she hears the agent1’s answer to the observer, “No, I don’t know.”

In the following we note the results obtained on the basis of what we have prepared.

(a) \( \Gamma₁ \vdash B₁(1W) \)
(b) \( \Gamma₁ \vdash B₁(\neg 1W) \)
(c) \( \Gamma₂' \vdash B₂(2W) \)
(d) \( \Gamma₂ \nvdash B₂(2W) \)

(a) and (b) show that agent2 does not know whether or not she wears a white hat. This justifies the agent1’s answer to the observer, “No, I don’t know.” (c) and (d) show that agent2 knows that she wears a white hat after she hears the agent1’s answer, and that agent2 does not know this until she hears the agent1’s answer. The original problem is to show that (c) holds, that is, to show the inference process in which agent2 reaches the correct conclusion. In order to show that (c) holds, Yasugi and Oda presented a procedure to describe the proof figure of the sequent \( \Gamma₂' \rightarrow B₂(2W) \). In the later section, we will present a more well-organized proof figure than the one obtained on the basis of this procedure, and use the former for further discussion.

3 Our substructural logic-based approach

Yasugi and Oda’s approach was to describe the agent2’s inference, using the proof system of modal logics on the basis of the sequent calculus LK. In contrast, we will take up a somewhat weaker system than Yasugi and Oda’s one, and examine whether or not their result carries over to our adopted system.

The general method for weakening the inference system of modal logics is to change or eliminate the axioms or the inference rules concerning the modal operators in question. Since Yasugi and Oda’s system is a variant of the modal logic KD₄, it would be possible to examine whether or not we obtain the same result, by use of weaker systems KD and K₄ than this system or of the systems between each of these two and KD₄. However, such a restriction would make the connection between the behavior of modal operators and the notion of belief loose. The first reason is that the axiom D which asserts that no one holds the contradictory belief can be understood as a postulate to construct a logic for belief. The second reason is that it is natural to posit the axiom 4 which asserts that each one has the ability in which when
one consider something, one can think of the relevant consideration itself. Rather, we will not control the inference rules governing the behavior of modal operators, but we will obtain a weaker proof system by restricting the inference rules concerning the structure.

### 3.1 Substructural modal logic $CFL_e KD4^2$

The system we adopt is $CFL_e KD4^2$, the Classical Full Lambek calculus with exchange rules and the modal axioms $K$, $D$ and $4$ about 2 modal operators. We obtain this system by eliminating the weakening and contraction rules from the set of the structural inference rules of the sequent calculus $LK$, and instead by introducing the propositional constants and the logical connectives and adding the axioms and the inference rules governing these. Here we will explain only the difference from $LK$ briefly. For the details, see Watari et al [7] . (Our present paper follows the notation adopted in that paper.) The comprehensive description of substructural logics is given in the textbook [4] by Restall.

The substantial difference of our system from $LK$ is the presence of binary connectives $*$ and $+$. In linear logics, these are called “multiplicative conjunctive” and “multiplicative disjunctive,” respectively. In contrast, the conjunctive connective and the disjunctive connective in $LK$ are called “additive conjunctive” and “additive disjunctive,” respectively. The notational correspondence between our system and linear logics is in the following.

<table>
<thead>
<tr>
<th></th>
<th>Conjunctive</th>
<th>Disjunctive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multiplicative</td>
<td>Additive</td>
</tr>
<tr>
<td>Linear logics</td>
<td>$\otimes$</td>
<td>$&amp;$</td>
</tr>
<tr>
<td>Our system</td>
<td>$*$</td>
<td>$\land$</td>
</tr>
</tbody>
</table>

In the system where there are weakening rules and contraction rules, $*$ and $+$ coincide with $\land$ and $\lor$, respectively with respect to the behavior. In this sense, $*$ and $+$ are one kind of conjunctive connective and disjunctive connective, respectively, which are brought by the subdivision of conjunction and disjunction by the restriction of the inference rules. The function

---

3This name shows that our system is what is brought by combining the sequent calculi $CFL_e$ with the modal logics $KD4^2$. Following this way of naming, we may call Yasugi and Oda’s system $LKKD4^2$. Moreover, since $LK$ accords with the system $CFL_{ecw}$ which is brought by adding the rules of weakening and contraction to $CFL_e$, we may write the system as $CFL_{ecw} KD4^2$.

4However, as for the modal operators, in order to compare our system with Yasugi and Oda’s one easily, we change the relevant notation into $B$. 

5
of the propositional constants that is introduced newly can be understood in relation to these connectives. We will introduce new logical symbols of $\text{CFL}_e \text{KD}4^2$ which contains these. Then we will state the axioms and the inference rules governing the behavior of these. Moreover, we will refer to the intuitive meaning carried by these new connectives and finally apply these to the puzzle in question.

The logical symbols added to the system $\text{LK}$ are the propositional constants $t, f, 1W$ and $2W$, the modal operators $B_1$ and $B_2$ and the binary connectives $\ast$ and $\plus$. The propositional constants $1W$ and $2W$ are not different from the ones explained in the previous section. Modal operators $B_1$ and $B_2$ are used again, but the inference rules for them is changed in accordance with the description in Watari et al.’s paper. In our system, the rules $(B_i - \text{K}), (B_i - \text{D})$ and $(B_i - \text{4})$ are adopted in place of $B_i \rightarrow B_i$ and we have seen in section 2 that they are equivalent.

The symbols $t$ and $f$ play a role as the unit elements such that we regard the symbols $\ast$ and $\plus$ as binary operations on the set of formulae. That is to say, using the axioms and the inference rules described later, we can show the following. For a formula $A$,

$$
\begin{align*}
&t \ast A \rightarrow A & A \rightarrow t \ast A & A \ast t \rightarrow A & A \rightarrow A \ast t \\
&f + A \rightarrow A & A \rightarrow f + A & A + f \rightarrow A & A \rightarrow A + f
\end{align*}
$$

The same thing holds between $\land$ and $\top$, and between $\lor$ and $\bot$. That is to say, $\top$ and $\bot$ are the unit elements of $\land$ and $\lor$. As in the case of $\ast$ and $\plus$, the rules of weakening and contraction cannot differentiate the pair $t$ and $\top$ and the pair $f$ and $\bot$. Since in $\text{LK}$, $\top$ and $\bot$ can be understood as a true proposition and a false one, respectively, we can think that $t$ and $f$ that are introduced newly are one kind of a true proposition and a false one, respectively, which are brought by the restriction of the inference rules.

The axioms and the inference rules of $\text{CFL}_e \text{KD}4^2$ that we will use in this paper are in the following. However, here we describe only the axioms and the inference rules added to the system $\text{LK}$. As we said in the beginning of this section, the rules of weakening and contraction are eliminated from the inference rules concerning the structure. Therefore we can use only exchange rules and cut rules as the structural inference rules.

- **Axioms**

  $$
  \rightarrow t \quad f \rightarrow \quad \Gamma \rightarrow \Delta, \top \quad \Gamma, \bot \rightarrow \Delta
  $$

- **Inference rules**

  $$
  \frac{\Gamma \rightarrow \Delta}{\Gamma, t \rightarrow \Delta} \quad (\text{tw}) \quad \frac{\Gamma \rightarrow \Delta}{\Gamma \rightarrow \Delta, f} \quad (\text{fw})
  $$
Here the symbols $A$ and $B$ denote formulae, and each of the symbols $\Gamma$, $\Gamma'$, $\Delta$ and $\Delta'$ denotes the sequence of formulae, and when $\Gamma = A, B, \ldots$, $\Gamma' = A, B, \ldots$. Basically, this system depends on Watari et al.’s system, but they differ in that the inference rules concerning modal operators are relativized to each agent in order to apply them to the puzzle in question.

### 3.2 Resource-sensitivity

For further discussion, we describe the feature of the inference rules concerning the symbol $\,$. As we have already said, since the symbol $\,$ is one kind of disjunctive connective, it is useful to compare with this the inference rules concerning $\,$ that is also a disjunctive connective. The difference between these becomes clear, in the left rule concerning $\,$, especially in the case where $\Gamma = \Gamma'$ and $\Delta = \Delta'$.

\[
\begin{align*}
\Gamma, A, B & \rightarrow \Delta \quad \text{(* left)} & \Gamma' \rightarrow A, \Delta & \Gamma' \rightarrow B, \Delta' \\
\Gamma, A \ast B & \rightarrow \Delta \\
\Gamma, A & \rightarrow \Delta & \Gamma', B & \rightarrow \Delta' \\
\Gamma, \Gamma', A + B & \rightarrow \Delta, \Delta' \quad \text{(+ left)} & \Gamma & \rightarrow \Delta, A + B \\
\Gamma & \rightarrow A & \Gamma & \rightarrow (B, \Delta, \Delta') \quad \text{(+ right)} \\
\end{align*}
\]

In these two inferences, the premise of the inference coincides while the conclusion of the inference does not. In the left rule concerning $\,$, two $\Gamma$s that are in the premise of the inference appear in the conclusion as they are, while in the left rule concerning $\,$, two $\Gamma$s and two $\Delta$s are each altered into one symbol. That is to say, we can say that both differ in that one preserves the resources used in the inference while the other does not. The sensitivity of the newly introduced disjunctive connectives — of even this entire proof system — to the increase or decrease of the resources used in the inference is called “resource-sensitivity” or “resource-consciousness.” That $\,$ and $\,$ are given as separate disjunctive connectives is a realization of the fact that our system has a resource-sensitive feature.

Such a resource-sensitivity can be seen when we compare the right rule concerning $\,$ with the one concerning $\,$. The right rule concerning $\,$ has the following two forms:

\[
\begin{align*}
\Gamma & \rightarrow \Delta, A & \Gamma & \rightarrow \Delta, A \ast B \\
\Gamma & \rightarrow \Delta, B & \Gamma & \rightarrow \Delta, A \ast B \\
\end{align*}
\]

\[i = 1, 2\]
In the rule of the latter form, we consider the case where $\Delta = \Delta', A$.

\[
\frac{\Gamma \rightarrow \Delta', A, B}{\Gamma \rightarrow \Delta', A \lor B, B} (\lor \text{ right})
\]

\[
\frac{\Gamma \rightarrow \Delta', A, B}{\Gamma \rightarrow \Delta', A + B} (+ \text{ right})
\]

Again the premises in the two inferences coincide while the conclusions there do not. When we use the right rule concerning $\lor$, the new symbol $B$ is introduced as a disjunct in the disjunction that appears newly in the conclusion. In contrast, when we use the right rule concerning $+$, the symbol $B$ contained in the premise is used as a disjunct in the disjunction that appears newly in the conclusion. In consequence, when we use the right rule concerning $\lor$, the number of $B$s which are prepared as the resources of the inference increases, and when we use the right rule concerning $+$, such an increase does not happen. This fact again shows that the symbol $+$ has the resource-sensitivity.

Since the duality that is seen between $\land$ and $\lor$ is also seen in between $*$ and $+$, the symbol $*$ has the resource-sensitivity again. Since $*$ and $+$ are a conjunctive connective and a disjunctive one that are separated from $\land$ and $\lor$ by restricting the structural inference rules, $*$ and $+$ can be understood as a resource-sensitive aspect potentially existing in $\land$ and $\lor$. By making the notions of conjunction and disjunction more fine-grained in this way, our system obtains a certain “weakness” that will be described later.

For further discussion, we would like to give the symbol $+$ a certain meaning in a naive sense. The characterization by the inference rules certainly is one way of giving it some meaning, but we can give it a naive meaning in a different way from this. One way is to utilize the logical formula which is equivalent to $A + B$ but does not contain $+^5$. Because we develop our discussion within the proof-theoretic framework, it is appropriate to use the mutual deducibility between two logical formulae as the equivalency. For example, we use the following facts:

\[
\vdash A + B \rightarrow (\neg A) \supset B \quad \vdash (\neg A) \supset B \rightarrow A + B
\]

That is to say, $A + B$ and $(\neg A) \supset B$ involve the relation of mutual deducibility. Thus, since the meaning of $(\neg A) \supset B$ is “if $A$ does not hold, $B$ holds,” we can adopt this as the meaning of $A + B$. In contrast, the meaning of the proposition $A \lor B$ containing the disjunctive connectives that is not $+$, is “$A$ holds or $B$ holds.” However, in this case, we do not know whether $A$

---

$^5$It is possible to give a semantics by use of an appropriate algebra [7, pp. 432–434], but since we intend to apply the logical puzzle described in the natural language to a logical system, we cannot depend on such a semantics which is quite different from the natural language.
holds or $B$ holds [6, p. 23]. The inference rule concerning $\lor$ reflects this consideration well. For example, if we want to deduce $C$ from $A \lor B$, we must show that $C$ is deduced from $A$ and that $B$ is deduced from $A$. That is because we do not know whether $A$ holds or $B$ holds. This is what the left rule concerning $\lor$ requires, as shown in the following.

$$\frac{A \to C \quad B \to C}{A \lor B \to C} \quad (\lor \text{ left})$$

### 3.3 Strength of our system

Since Yasugi and Oda’s system differs from our system with respect to the logical symbol, it is impossible to compare the sizes of the set of theorems easily. However, the claim that our system is a weak one in a sense can be understood from the fact that we cannot prove representative tautologies as shown in the following. For formulae $A$, $B$ and $C$,

- $A \supset (B \supset A)$
- $A \supset (B \supset C) \supset ((A \supset B) \supset (A \supset C))$
- $(\neg A \supset B) \supset ((\neg A \supset \neg B) \supset A)$

For example, when we draw the proof figure by going back from the conclusion, under the situation in which we take $A$, $B$ and $C$ as formulae, we cannot accomplish the proof figure as in the following figure.

$$\frac{A,B \to A}{A \to B \supset A} \quad \frac{A \supset (B \supset A)}{\to A \supset (B \supset A)}$$

In this case, the impossibility of the use of the rule of weakening prevents us from completing the proof figure. In addition, as for the other two logical formulae, the proof figure cannot be completed, since we do not have the rule of contraction. In contrast, our system is not too weak. As in the case of $(A \land B) \supset (A \lor B)$, in our system, we can accomplish the proof which does not require the use of the rules of weakening and contraction in $\text{LK}$. Moreover, we can prove some logical formulae which we cannot prove in $\text{LK}$ without the use of the rules of weakening and contraction, by replacing the conjunctive connective and the disjunctive connective occurring here with $*$ and $+$, respectively. For example, in our system we cannot prove the mutual deducibility between $A \supset B$ and $(\neg A) \lor B$ since it requires the use of the rules of contraction and weakening. However, the mutual
deducibility holds between $A \supset B$ and the proposition such that we replace the disjunctive connective appearing in $(\neg A) \lor B$ with +. In addition, the law of contradiction and the law of excluded middle are its examples.

4 Applying $CFL_eKD^2$ to the puzzle

Based on the preparation developed above, we will examine whether or not Yasugi and Oda’s result holds in the system $CFL_eKD^2$. In particular, our examination consists in the question of whether or not we can prove that $B_2(2W)$ is deduced from $\Gamma'_2$ the set of sentences describing the agent2’s belief after she hears the agent1’s answer to the observer. That is to say, what matters is whether or not we can draw the proof figure leading to the sequent $\Gamma'_2 \rightarrow B_2(2W)$. We construct the proof figure by going back from the conclusion. The following figure is an example of the ones constructed in this way.

\[
\begin{array}{c}
? \\
1W \rightarrow 2W, 1W & 2W \rightarrow 2W, 1W \\
1W, \neg 2W \rightarrow 1W & 2W, \neg 2W \rightarrow 1W \\
1W \lor 2W, \neg 2W \rightarrow 1W & \neg 2W, 2W \\
B_1(1W \lor 2W), B_1(\neg 2W) \rightarrow B_1(1W) & B_1(1W \lor 2W), \neg B_1(1W), B_1(\neg 2W) \rightarrow B_1(1W) \\
\neg 2W, 2W & B_1(1W \lor 2W), \neg B_1(1W), \neg 2W \rightarrow B_1(\neg 2W) \rightarrow 2W \\
B_1B_1(1W \lor 2W), B_1\neg B_1(1W), B_1(\neg 2W \rightarrow B_1(\neg 2W)) \rightarrow B_12W \\
\end{array}
\]

Yasugi and Oda [8, p. 155] pay attention to the fact that the restricted elements of $\Gamma'_2$ are sufficient in order to obtain the required figure. Accordingly, we set the necessary elements in the bottom of the figure. As seen in the top of the proof, since we do not have the rule of weakening, we cannot complete this proof figure. Of course, because there are infinitely many ways of drawing the proof figure, we must discuss this matter by preparing the appropriate semantics in order to verify that we cannot prove the relevant matter without the use of the rule of weakening, but we will leave this to further study. Rather, we note the following fact. That is to say, if the symbol $\lor$ appearing in the formula $B_2(B_1(1W \lor 2W))$ contained in the bottom of the figure is the symbol +, we can accomplish the proof.
We have so far described the two kinds of disjunctive connectives, but we have said little things about the question of how we make these two disjunctive connectives distinguished with respect to the use in relation to the ordinary language. Since $\Gamma_2'$ is what we take over from Yasugi and Oda’s previous study, with respect to this element there may be a portion which should be re-described by use of the logical symbol of substructural logics. Now is it correct to replace the disjunctive connectives appearing in the formula $B_2(B_1(1W \lor 2W))$ that is an element of $\Gamma_2'$? Do we obtain a certain method for the examination of the correctness in question?

What we encounter here is the problem concerning the correspondence between the formal language and the natural language. Ideally, when we are given a sentence containing a disjunction in the ordinary language, it would be good to have a method for determining whether we should write the disjunction as $\_\$ or as $\_\_$. We gave naive meanings to sentences containing $\_\$ or $\_\_$. That is to say, we showed the possibility that we read $A \_\$ B as “$A$ holds or $B$ holds (but we do not know whether either holds.)” and read $A \_\_ B$ as “If $A$ does not hold, $B$ holds.” However, this translation seems to be rather weak as a method for the classification of the disjunctive sentences. That is because it seems difficult to avoid the criticism that these two meaning characterization eventually says the same thing. Some people cannot recognize the difference between these two meanings of the disjunctions, by inferring “$A$ holds or $B$ holds, and if $A$ does not hold, it is natural that $B$ holds” and “suppose that if $A$ does not hold, then $B$ holds; $A$ holds or $A$ does not hold; if $A$ holds, we can say that $A$ hold or $B$ holds; even if $A$ does not hold, because we can deduce that $B$ holds from the supposition, again we can say that $A$ holds or $B$ holds.” Since there is no strange point in the inference of this person, the problem consists in the meaning given in English to the disjunctive connectives. We need further examination on this point.

5 On the axiom D

In this section, we will give our own opinion to the modal axiom D. We have added the inference rules corresponding to the axioms $K$, $D$ and $4$ to the sequent calculus $CFL_e$. The axiom $D$ corresponding to the rule $B_1\cdot D$ is
$B_i(A) \supset \neg B_i(\neg A)$ where $A$ is a formula. There are other ways for describing the axiom $D$. For example, one of the ways is to state $\neg B_i(A \land \neg A)$. The axiom $D$ is originally what is introduced in order to express the deontic modality, but because we use $B_i$ as an operator for the expression of the belief, the axiom of the latter form means “It is not believed that $A$ and not $A$,” that is, “No contradiction is believed.” We call the former $D_1$ and the latter $D_2$. In the system $\text{LK}$ with the rule $B_i\text{-K}$, using the rules of contraction and weakening, we can prove that $D_1$ and $D_2$ involve the relation of mutual deducibility. However, in $\text{CFL}_e$ with $B_i\text{-K}$, these two deductions fail. Therefore, we cannot identify $D_1$ with $D_2$.

In $\text{CFL}_e$, the notion of disjunction is divided into two ones, so that two operations “$\lor$” and “$+$” occur. In addition, “$\bot$” and “$f$” are introduced as each unit element of these two operations, respectively. In the system containing the rules of weakening and contraction, as in $\text{LK}$, these two can be identified and we can regard these as propositional constants expressing any inconsistent proposition. Moreover, In $\text{CFL}_e$, the notion of conjunction is also divided into two ones, so that two operations “$\land$” and “$*$” occur. Therefore if we take $\text{CFL}_e$ as the basis of a proof system, this means that we have much more ways for the expressions of inconsistent propositions. If the proposition “No contradiction is believed” is a meaning of the axiom $D$, the following will work as candidates of the expressions.

- $D_1 : B_i(A) \supset \neg B_i(\neg A)$
- $D_2 : \neg B_i(A \land \neg A)$
- $D_3 : \neg B_i(\bot)$
- $D_4 : \neg B_i(f)$
- $D_5 : \neg B_i(A \ast \neg A)$

The first two are $D_1$ and $D_2$ which has already mentioned. $D_3$, $D_4$ and $D_5$ are what can be obtained by replacing an inconsistent proposition in $D_2$ with other expressions. We cannot identify any two of these. That is because the deducibility holding between these in the system $\text{CFL}_e$ with the rule $B_i\text{-K}$ is given in the following.
In $\text{LK}$ with $B_i-K$, all of these involve the relation of mutual deducibility, but in $\text{CFL}_e$ we would be able to choose any one of these expressions. We have adopted $D_1$, which corresponds to $B_i-D$ well. We have no idea of which expression can denote the content of the sentence “No contradiction is believed” in the best way. Since the axiom $D$ is introduced as one of requirements satisfied by the belief or the obligation, we may say that the fact that such diverse formulations are possible means that the use of substructurized logics works as help for understanding the notions of belief and obligation. Here what becomes important again is the correspondence between the natural language and the formal language that has been mentioned in the previous section. Do we refer to the multiplicative conjunction or the additive conjunction of a sentence and its denial by the word “contradiction”? The problem of such a correspondence exists in the foundation of our inquiry, and it is this that makes the expressions of the axioms diverse.

6 Summary

We have recapitulated “the two wise girls puzzle” and Yasugi and Oda’s approach to it. Next, we have introduced the system of modal substructurized logics involving the resource-sensitive property, and using this system, we have shown that there is a doubtful point as to the solvability of the puzzle established by Yasugi and Oda$^6$. Finally, we have pointed out that we can solve the puzzle in the introduced system of substructurized logics, depending on how we establish the correspondence between the connectives inherent to such logics and the counterparts of the ordinary language. In consequence, we have confirmed that the correspondence between substructurized logics and the ordinary language is important. Such an importance as regards the

$^6$The epistemic system introduced in Sadrzadeh[5] to analyze the muddy children paradox will provide different solutions to the wise girls puzzle of Yasugi and Oda.
correspondence has also been clear from the fact that this importance even
affects the theoretical system characterizing the expressions of the axioms.

References


dissertation, Université du Québec à Montréal, 2006.


[7] Osamu Watari, Takeshi Ueno, Koji Nakatogawa, Mayuka F. Kawaguchi,
and Masaaki Miyakoshi. Sequent systems for classical and intuitionisti-
tic substructural modal logic. In *Proc. of the 7th and 8th Asian Logic
Conferences*, pp. 423–442. Singapore University Press & World Scientific,
2003.

Treating Speech Acts as Acts

If we are to take the notion of speech act seriously, we must be able to treat speech acts as acts.

By characterizing speech acts in terms of dynamic changes they bring about, it becomes possible to treat them within a general theory of action.

Perlocutionary Acts as Acts

Perlocutionary acts are acts that really produce "real effects" upon the attitudes and the actions of addressees (Austin 1975, pp. 101-3).

Locutionary act: He said to me "Shoot her!" meaning by "shoot" shoot and referring by "her" to her.

Illocutionary act: He urged (advised, ordered, etc.) me to shoot her.

Perlocutionary act (a): He persuaded me to shoot her.
Perlocutionary act (b): He got me to shoot her.

Illocutionary Acts: a Problem

Illocutionary acts do not directly affect brute facts, except for those physical conditions involved in the production of sounds and written symbols.

Nor do they directly affect attitudes and actions of addressees.

A Gap

For instance, intuitively, a command "See to it that φ!"

makes worlds where φ holds preferred over those where it does not – at least, if we accept the preference induced by the issuer of the command. (Van Benthem & Liu, to appear)

Illocutionary Acts as Acts

Illocutionary acts affect institutional facts.

How is it possible to capture changes in institutional facts?
Example 1: On A Hot Day (1)

Suppose you are reading an article on logic in the office you share with your boss and a few other colleagues. While you are reading, the temperature of the room rises, and it is now above 30 degrees Celsius.

Example 1: On A Hot Day (2)

Then, suddenly, you hear your boss's voice. She said, "Open the window!"

She commanded you to open the window.

What effects does her command have on the current situation?

Example 1: On A Hot Day (3)

Her act of commanding did not affect the state of the window directly.

Nor did it affect the number of alternatives you have. It is still possible for you to turn on the air conditioner, to ignore the heat, or to open the window.

Example 1: On A Hot Day (4)

But it has now become impossible for you to choose alternatives other than that of opening the window without going against your obligation.

It is now obligatory upon you to open the window, although it was not so before.

Acts of Commanding

Acts of commanding seem to affect deontic status of possible courses of actions.

We have considered how we can capture the kind of changes acts of commanding bring about in Yamada (2007a, b).

The Idea of Command Logic

We model changes acts of commanding bring about in terms of a new update logic. We combine a multi-agent variant of the language of monadic deontic logic with a dynamic language to talk about the situations before and after the issuance of commands, and the commands that link those situations.
The results

Although the resulting language inherits various inadequacies from the language of monadic deontic logic, some interesting principles are captured and seen to be valid nonetheless.

Related Paradigms

- Speech Act Theory (Austin, Searle)
- Propositional Dynamic Logic
- Update Semantics (Veltman)
- Logic of Public Announcements (Plaza, Groeneveld & Gerbrandy, Baltag, Moss & Solecki, Kooi & van Bentham, etc.)
- Deontic Logic

Logics of Public Announcements

The Static Base: Epistemic Logics

Adding Announcement Modalities

The Dynamic Extension: Logics of Public Announcements

The Same Strategy Works for Acts of Commanding

The Static Base
The static language: $\mathcal{L}_{MDL}^+$
Multi-agent deontic logic: MDL$^+$

Adding Command Modalities

The Dynamic Extension
The language of Command Logic: $\mathcal{L}_{CL}$
Eliminative Command Logic: ECL

The Language of Deontic Logic

$Op: \text{It is obligatory that } p.$

$Pp: \text{It is permitted that } p.$

$Fp: \text{It is forbidden that } p.$

Relativisation

$Op_i p: \text{It is obligatory upon } i \text{ that } p.$

$Pp_i p: \text{It is permitted of } i \text{ that } p.$

$Fp_i p: \text{It is forbidden of } i \text{ that } p.$

where $i$ is an agent in a given set $I$ of agents.
The Static Language $L_{MDL^+}$

Definition 1. Take a countably infinite set $A_{prop}$ of proposition letters and a finite set $I$ of agents, $p$ ranging over $A_{prop}$ and $i$ over $I$. The multi-agent monadic deontic language with an alethic operator $L_{MDL^+}$ is given by:

$$\phi ::= \top | p | \neg \phi | \phi \land \psi | \Box \phi | O \phi$$

$L_{MDL^+}$ -models (1/2)

Definition 2. By an $L_{MDL^+}$-model, I mean a quadruple $M = \langle W^M, R^M_i, R^K_i, V^M \rangle$ where:

(i) $W^M$ is a non-empty set
(ii) $R^K_i \subseteq W^M \times W^M$

$L_{MDL^+}$ -models (2/2)

(iii) $R^K_i$ is a function that assigns a subset $R^K_i(i)$ of $R^K_i$ to each $i \in I$
(iv) $V^M$ is a function that assigns a subset $V^M(p)$ of $W^M$ to each $p \in A_{prop}$.

We usually abbreviate $R^K_i(i)$ as $R^K_i$.

Truth definition for $L_{MDL^+}$ (1/2)

Definition 3. Let $M$ be an $L_{MDL^+}$-model and $w \in W^M$. If $p \in A_{prop}$, $i \in I$, and $\phi, \psi$ are sentences of $L_{MDL^+}$, then

(a) $M,w \models p$ iff $w \in V^M(p)$
(b) $M,w \models \top$
(c) $M,w \models \neg \phi$ iff $M,w \not\models \phi$

Truth definition for $L_{MDL^+}$ (2/2)

(d) $M,w \models (\phi \land \psi)$ iff $M,w \models \phi$ and $M,w \models \psi$
(e) $M,w \models \Box \phi$ iff for every $v$ such that $\langle w, v \rangle \in R^K_i$, $M,v \models \phi$
(f) $M,w \models O \phi$ iff for every $v$ such that $\langle w, v \rangle \in R^K_i$, $M,v \models \phi$

Completeness of MDL^+

A sound and complete proof system for MDL^+ is given in Yamada (2007a). Thus:

Theorem 1: MDL is strongly complete with respect to the class of $L_{MDL^+}$-models.
Example 1 in $\mathcal{L}_{\text{MDL}}$

The situations you were in before and after the issuance of your boss’s command in Example 1 can be represented by two related $\mathcal{L}_{\text{MDL}}$-models $M$ and $N$ respectively. For example, we may say:

1. $M, s \models \neg O_a p \land \neg O_a \neg p$
2. $N, s \models O_a p$

What we cannot say in $\mathcal{L}_{\text{MDL}}$

Sentences of $\mathcal{L}_{\text{MDL}}$ can be used to describe the situations before and after the issuance of your boss’s command.

But they cannot be used to talk about the act of commanding that links these situations.

The Language of the Logic of the Act of Commanding (Command Logic)

$O_i \xi$ : It is obligatory upon the agent $i$ to see to it that $\xi$.

$[i, \psi] \phi$ : After every act of commanding $i$ (the addressee) to see to it that $\psi$, $\phi$ holds.

$[i, \psi] O_i \xi$ : After every act of commanding $i$ to see to it that $\psi$, it is obligatory upon $i$ to see to it that $\xi$.

The Language of the Logic of Public Announcements

$K_i \psi$ : Agent $i$ knows that $\psi$.

$[A!] \phi$ : After every truthful public announcement that $A$, $\phi$ holds.

$[A!]K_i \psi$ : After every truthful public announcement that $A$, $i$ knows that $\psi$.

The Dynamified Language $\mathcal{L}_{\text{CL}}$

Definition 5. Take the same countably infinite set $A_{\text{prop}}$ of proposition letters and the same finite set $I$ of agents, $p$ ranging over $A_{\text{prop}}$ and $i$ over $I$ as before. The language of Command Logic $\mathcal{L}_{\text{CL}}$ is given by:

$\phi := \top \mid p \mid \neg \phi \mid \phi \land \psi \mid \Diamond \phi \mid O_i \phi \mid [\pi] \phi$

$\pi := i, \phi$
Truth definition for $\mathcal{L}_{\text{CL}}$ (1/3)

Definition 4  Let $M$ be an $\mathcal{L}_{\text{MDL}^+}$-model and $w \in W^M$. If $p \in \text{Aprop}$, $i \in I$, and $\phi, \psi$ are sentences of $\mathcal{L}_{\text{CL}}$, then

(a) $M,w \models \phi$ iff $w \models V^i(p)$
(b) $M,w \models \top$
(c) $M,w \models \neg \phi$ iff $M,w \not\models \phi$

Truth definition for $\mathcal{L}_{\text{CL}}$ (2/3)

(d) $M,w \models \text{ECL} (\phi \land \psi)$ iff $M,w \models \text{ECL} \phi$ and $M,w \models \text{ECL} \psi$
(e) $M,w \models \text{ECL} \Box \phi$ iff for every $v$ such that $\langle w,v \rangle \in R_v^M$, $M,v \models \text{ECL} \phi$
(f) $M,w \models \text{ECL} O_i \phi$ iff for every $v$ such that $\langle w,v \rangle \in R_v^M$, $M,v \models \text{ECL} \phi$.

On this truth definition

We abbreviate $\{ (x,y) \in R_v^M(i) \mid M,y \models \text{ECL} \phi \}$ as $R_v^M \upharpoonright \phi$.

As we always have $R_v^M \upharpoonright d \subseteq R_v^M$, we also have $R_v^M \upharpoonright \phi \subseteq R_v^M$. Hence $M_{1,x}$ is guaranteed to be an $\mathcal{L}_{\text{MDL}^+}$-model.

Boss’s Act of Commanding in MDL$^+$

$$\neg O_s p \land \neg O_s \neg p$$

$$\neg O_s p \land \neg O_s \neg p$$

$$\neg O_s p \land \neg O_s \neg p$$

$$\neg O_s p \land \neg O_s \neg p$$

Boss’s Act of Commanding in ECL (Eliminative Command Logic)

$$\neg O_s p \land \neg O_s \neg p$$

$$\neg O_s p \land \neg O_s \neg p$$

$$\neg O_s p \land \neg O_s \neg p$$

$$\neg O_s p \land \neg O_s \neg p$$
Boss’s Act of Commanding in ECL (2)

CUGO Principle

If neither $O_i$’s nor $i$’s occur in $\phi$, we have

$$\boxdot_i O_i \phi.$$  

Usually, you should do what your boss commands you to do. In other word, commands usually generate obligations.

Why not CGO but CUGO

The following formula is not valid:

$$[i, P, q] O_i P, q.$$ 

If $\psi$ involves deontic operators or command operator indexed with $i$, the truth of $\psi$ at $w$ in $M$ does not guarantee the truth of $\psi$ at $w$ in $M_i, \phi$.

An Open Question

Let $S_{CGO}$ be the set of sentences $\phi$ such that we have $\models_{ECL} [i, \phi] O_i \phi$. Let $S_{\text{free}}$ be the set of $E_{\text{MDL}}$ -sentences in which no deontic operators for $i$ occur.

As $O_i \phi \rightarrow O_i \phi \in S_{CGO}$, we have $S_{\text{free}} \subseteq S_{CGO}$.

But exactly how large $S_{CGO}$ is is an open question.

Reduction Axioms

1. $[i, \phi] p \leftrightarrow p \text{ where } p \in \text{Aprop}$
2. $[i, \phi] \top \leftrightarrow \top$
3. $[i, \phi] \rightarrow \psi \leftrightarrow \neg[i, \phi] \neg \psi$
4. $[i, \phi] (\psi \land \chi) \leftrightarrow ([i, \phi] \psi \land [i, \phi] \chi)$
5. $[i, \phi] \square \psi \leftrightarrow \square[i, \phi] \psi$
6. $[i, \phi] O_i \psi \leftrightarrow O_i ([i, \phi] \psi)$
7. $[i, \phi] O_i \psi \leftrightarrow O_i [i, \phi] \psi$ \text{ where } i \neq j.

Translation function $t$ (1/2)

- $t(p) = p$
- $t(\top) = \top$
- $t(\neg \phi) = \neg t(\phi)$
- $t(\phi \land \psi) = t(\phi) \land t(\psi)$
- $t(\square \phi) = \square t(\phi)$
- $t(O_i \phi) = O_i t(\phi)$
Translation function $t$ (2/2)

\[

t(\lbrack i \rbrack \rho) = \rho \\
t(\lbrack i \rbrack \top) = \top \\
t(\lbrack i \rbrack \lnot \psi) = \lnot t(\lbrack i \rbrack \psi) \\
t(\lbrack i \rbrack (\psi \land x)) = t(\lbrack i \rbrack \psi) \land t(\lbrack i \rbrack x) \\
t(\lbrack i \rbrack \psi) = t(\lbrack i \rbrack \psi) \\
t(\lbrack i \rbrack \lnot \psi) = \lnot t(\lbrack i \rbrack \psi) \\
t(\lbrack i \rbrack \psi) = O_i t(\lbrack i \rbrack \psi) \\
t(\lbrack i \rbrack [\psi] x) = t(\lbrack i \rbrack \psi) t(\lbrack i \rbrack x) \\

\]

Proof System for ECL

The proof system for ECL (Eliminative Command Logic) contains all the axioms and all the rules of the proof system for MDL*, and in addition all the reduction axioms listed above and the following rule:

\[
\begin{array}{c}
\Psi \\
\hline
[i, \phi] \psi
\end{array}
\]

Completeness

The completeness of ECL is derived from the completeness of MDL*.

Theorem 2

ECL is strongly complete with respect to the class of $L_{\text{MDL}^*}$-models.

Example 2

Suppose you are in a combat troop and now waiting for your captain's command to fire. Then you hear the command, and it has become obligatory upon you to fire. But before that, you were not permitted to fire. This forbiddance is now no longer in force. Thus it seems that after his command, you are permitted to fire, at least in the sense of lack of forbiddance.

Built-in assumptions 1

Commands are assumed to be always eliminative.

Proposition 2

We have:

\[
\models_{\text{ECL}} [i, \phi] P_i \phi \rightarrow P_i \phi.
\]

Built-in assumptions 2

Commands of the form $\lbrack i \rbrack \rho$ is assumed to have no effects on deontic accessibility relations for any agent other than $i$.

Corollary 2

If $\psi \in S_{\text{free}}$, then:

\[
M, w \models_{\text{ECL}} \psi \text{ iff } M_{[i] \rho}, w \models_{\text{ECL}} \psi.
\]
Built-in assumptions 3

Commands are assumed to have no preconditions.

Compare:

PAL: \[ K(\phi \rightarrow K[\phi] \psi) \]
ECL: \[ O(\phi \rightarrow [\psi, \phi]) \]

Dead end principle

We have:

\[ \models_{ECL} [! (p \land \neg p)] O, \phi. \]

Since \( R^{\psi_1}_i (p \land \neg p) \) is empty, no world is \( R^{\psi_1}_i (p \land \neg p) \)-accessible for \( i \). Absurd commands lead to an obligational dead end.

D axiom cannot be added

We have:

\[ M \models_{ECL} O(p \rightarrow P, \phi) \land \neg P, \neg \phi. \]

Thus \( O, \phi \rightarrow P, \phi \) cannot be added to ECL.

Restricted sequential conjunction principle

If \( \phi, \psi \in S_{\text{free}} \), we have:

\[ \models_{ECL} [! !], (\phi \land \psi) \land \neg (\phi \land \psi). \]

Since \( R^{\psi_1}_i (\phi \land \psi) \) can be distinct from \( (R^{\psi_1}_i \phi) \land \psi \), unrestricted form of sequential conjunction principle is not valid.

Restricted order invariance principle

If \( \phi, \psi \in S_{\text{free}} \), we have:

\[ \models_{ECL} [! !]\psi x \rightarrow [! !], (\phi \land \psi) x. \]

Since \( (R^{\psi_1}_i \phi) \land \psi \) can be distinct from \( (R^{\psi_1}_i \phi) \land \psi \), unrestricted form of order invariance principle is not valid.

Another Interesting Validity

We have:

\[ \models_{ECL} [! !], (\neg p) O, \phi. \]

Everything comes to be obligatory if a pair of contradictory commands is issued. Thus you will be in an obligational dead end.
**An Extension: ECL II**

We may distinguish command issuers:

\[ [\! [i/j] \! ] p ] ( [\! [i/k] \! ] (\neg p) \land \! [O_{i/k} \! ] (\neg p) ) . \]

If an authority \( k \) commands you to see to it that \( \neg p \) after another authority \( j \) commands you to see to it that \( p \), then you will be in an obligational dilemma.

**Why an obligational dilemma**

Although no worlds are both \( R_{i/j} \)-accessible and \( R_{i/k} \)-accessible, there may be an \( R_{i/j} \)-accessible world and an \( R_{i/k} \)-accessible world in such a situation.

**Example 3**

**A Contingent Dilemma (1/2)**

\[ [\! [i/j] \! ] p ] ( [\! [i/k] \! ] q ) (\neg p) \land \! [O_{i/k} \! ] q ) . \]

\( p \): you attend the conference in Hakodate on May 8 2007.

\( q \): you join the demonstration in São Paulo on May 8 2007.

You will be in an obligational dilemma, if \( p \) and \( q \) happen to be incompatible.

**A Contingent Dilemma (2/2)**

\[ [\! [i/k] \! ] p ] ( [\! [i/k] \! ] q ) (\neg p) \land \! [O_{i/k} \! ] q ) . \]

If some very rapid means of transportation were available, it would be possible for you to obey both commands. But unfortunately, no such means of transportation happens to be available in the real world.

**Logical Dynamics of Multi-agent Language Games**

(1) Acts of promising can also be considered as deontic updaters, and acts of asserting can be considered as updaters of propositional commitments.

(2) Perlocutionary Acts can also be considered as updaters of systems of knowledge, belief and preference.

**Treating Illocutionary Acts and Perlocutionary Acts Together**

ECL II can be combined with DEUL (Dynamic Epistemic Upgrade Logic) of van Benthem & Liu (to appear).

DEUL can be interpreted as dealing with perlocutionary acts of getting addressees to prefer something.
The Language of Epistemic Preference Logic

*The Language of Dynamic Epistemic Upgrade Logic*

1. **The Language of Epistemic Preference Logic**
   
   - $K_i \phi$: the agent $i$ knows that $\phi$
   - $[\text{pref}]_i \phi$: in every world the agent $i$ considers at least as good as the current one $\phi$ holds
   - $U \phi$: in every world $\phi$ holds

2. **The Language of Dynamic Epistemic Upgrade Logic**
   
   - $[\phi!]_i \psi$: after every truthful public announcement that $\phi$, $\psi$ holds
   - $[\phi!]_i K_i \psi$: after every truthful public announcement that $\phi$, $i$ knows that $\psi$
   - $[\# \phi]_i \psi$: after every act of publicly suggesting $\phi$, $\psi$ holds
   - $[\# \text{pref}]_i \psi$: after every act of publicly suggesting $\phi$, in every world $i$ considers at least as good as the current one, $\psi$ holds

3. **The Need for Reinterpretation**

   As an act of suggesting $\phi$ is an illocutionary act, it works as a trigger of preference upgrade only if we accept the preference induced by its issuer.

   Preferences relativized to individual agents can be considered as propositional attitudes. So acts of getting addressees to prefer something can be considered as perlocutionary acts.

4. **The Language of Dynamic Deontic Epistemic Preference Logic (DDEPL)**

   **Definition** Take a countably infinite set $\text{Aprop}$ of proposition letters and a finite set $I$ of agents, with $p$ ranging over $\text{Aprop}$ and $i, j$ over $I$.

   The language of DDEPL is given by:

   - $\phi ::= \top | \neg \phi | \phi \land \psi | U \phi | O \psi | [\text{pref}]_i \phi | [\text{Com}]_{ij} \phi | [\text{G-pref}]_{ij} \phi | \text{Ann} \phi$
   - $K_i [\text{pref}]_i \phi$: the agent $i$ knows that $\phi$
   - $[\text{pref}]_i \phi$: in every world the agent $i$ considers at least as good as the current one, $\phi$ holds
   - $U \phi$: in every world $\phi$ holds
   - $O \psi \phi$: it is obligatory upon the agent $i$ with respect to $j$ to see to it that $\phi$

5. **The Language of Dynamic Deontic Epistemic Preference Logic (DDEPL)**

   - $[\text{Com}]_{ij} \psi$: after every successful act of commanding the agent $i$ by $j$ to see to it that $\phi$, $\psi$ holds
   - $[\text{G-pref}]_{ij} \psi$: after every successful $j$'s act of getting the agent $i$ to consider every world where $\phi$ holds at least as good as the current one, $\psi$ holds
What we Can Say in the Language of Dynamic Deontic Epistemic Preference Logic

\[ \text{O}_{ij} \phi \land \langle \text{pref} \rangle_i \neg \phi \]

It is obligatory upon an agent \( i \) with respect to \( j \) to see to it that \( \phi \), but \( i \) find some none- \( \phi \)-world at least as good as the current one.

Deontic Epistemic Preference Model

A deontic epistemic preference model is a tuple \( M = \langle W, \sim, \{ \tau_i | i \in I \}, \{ \rho_i | i \in I \}, R, V \rangle \), where

- \( W \) is a non-empty set of possible worlds
- \( \sim \) is a usual equivalence relation of epistemic accessibility for agent \( i \)
- \( \tau_i \) is a reflexive transitive relation of preference ordering for agent \( i \)
- \( R \) is a deontic accessibility for agent \( i \) with respect to \( j \)
- \( V \) is a valuation for proposition letters

Thank you for your attention!

The End
Institutions in Channel Theory

Stijn De Saeger
Japan Advanced Institute of Science and Technology
stijn@jaist.ac.jp
July 31, 2007

Abstract

In this talk, we propose an account of institutions based on channel theory, and show that the key elements in the theory (classifications, local logics) offer an expressive framework for the formalization of institutional action.

1 Introduction

Multiagent systems have been characterized as technological extensions of human society in [4]. To the extent that many of the problems encountered in formalizing the behaviour of software agents do indeed have close analogues in the “real” world, many aspects of agent societies have been modeled after their real world counterparts, drawing on a wealth of available research in such fields as epistemic logic, game theory, belief revision, and so on.

In order to formalize communication between (artificial) agents, a number of different agent communication languages (ACLs) have been proposed in the artificial intelligence community. Despite individual differences, they mostly seem to agree that an appropriate formalization of the theory of speech acts, pioneered by Searle [5] and Austin [1], is the way to go. This is hardly surprising: the concept of speech acts has a solid philosophical foundation, and it allows to cash out an account of agent communication in terms of rational action, making it an extension of an already well-studied problem in AI.

When it comes to giving a semantics for these ACLs, the dominant approach has been a mentalistic or intention-based one, in the tradition of the Gricean intention recognition model of cognition (see work by Cohen & Levesque, Bratman). A prime example is the definition of “inform” speech acts in terms of the feasibility preconditions (FP) and postconditions (RE or “rational effect”) on agents’ belief states in the FIPA specification of agent communication.  

\[ \text{http://www.fipa.org/repository/aclspecs.html} \]
\[ \langle \text{inform}(i, j, \varphi) \rangle \]

\[
\begin{align*}
\text{FP} & : B_i \varphi \land \neg B_i (B_j \varphi \lor B_j \neg \varphi) \\
\text{RE} & : B_j \varphi
\end{align*}
\] (1)

The preconditions say that \(i\) should believe \(\varphi\), and should also believe that \(j\) does not already have an opinion on whether or not \(\varphi\). In those situations, the effect of an inform act is that \(j\) comes to believe \(\varphi\) as well. As some researchers have claimed, this approach faces some issues. Besides long-standing problems with the formalization of intensional concepts like belief (e.g. logical omniscience), there is a tension between the essentially public nature of communication and the private nature of agent beliefs. Communication is by its nature a social activity, and defining its meaning in terms of subjective mental states that are in principle not open for inspection (either by other agents or by some system-level observer) seems to miss an important conceptual point: Belief updates fail to capture the social updates triggered by speech acts. If \(i\)’s inform action goes uncontested by \(j\), \(i\) is from that point on entitled to treat \(j\) “as if” she believed \(\varphi\), according to the implicit rules of the dialogue game.

In response to the problems faced by mentalistic theories of agent communication, a number of researchers (Colombetti, [3] and Singh, [6]) have advocated to ground a theory of speech acts in the social updates they effect, rather than the epistemic ones — essentially defining an act by agent \(i\) of informing agent \(j\) that \(\varphi\) as committing \(i\) to the truth of \(\varphi\), vis-a-vis \(j\). Such a social semantics for ACLs is appealing for a number of reasons, essentially replacing the thorny issue of talking about the private beliefs of agents by the more transparent notion of adherence to social norms and commitments. At the same time though, it places speech act semantics firmly in the realm of institutions:

\[
\ldots \text{“institutions” are systems of constitutive rules. Every institutional fact is underlain by a (system of) rule(s) of the form “} X \text{ counts as } Y \text{ in context } C\text{“}. (J. Searle, [5] p.51)
\]

As argued convincingly in [3], institutional action and natural action have profound differences, which poses a new set of problems for traditional AI theories of rational action. AI theories like BDI logics are traditionally defined in terms of agent intentions and processes of physical causation between events: agent \(i\) forms an intention to update agent \(j\)’s belief state with \(\varphi\), and this intention itself sets in motion a causal chain of events leading to its fulfillment. However, the success (or otherwise) of an institutional action depends on parameters beyond the control of the agents involved — institutions crucially involve some kind of collective intensionality that transcends any single instance of such event. For instance, asking for a beer counts as a commitment to purchase in

\[2\text{ Not to be confused with the mentalistic notion of commitment as a kind of “persistent intention.”}\]
the context of pubs because the institute of commerce says that it does. As Searle argues, these bodies of social convention are not normative rules, which are concerned with the distribution of obligations and permissions over a society of agents. Instead they are in a very real sense constitutive of what we recognize as successful social action, through the distribution of institutional powers among agents. In order to develop a proper social semantics of speech acts, we need a formal account of institutions and their logical properties.

2 An Account of Institutions

How are we to make sense formally of a statement like “X counts as Y in context C” (and the ways it might fail to hold). As a first approximation, we say a given event e supports an institutional fact Y in a context C when:

i. e has a physical property X, such that

ii. X is a proxy for Y by virtue of some institutional context C, in which

iii. “X \Rightarrow_c Y” is a constitutive rule of C.

Channel Theory  We propose to use channel theory (Barwise & Seligman, [2]) as a platform for formalizing institutional action. The main objects in channel theory are domain classifications, morphisms between these classifications, and the logics they give rise to. Let \( C_P \) be a classification representing physical reality. Formally, \( C_P \) is a triple \( \langle S_P, \Sigma_P, \|=P \rangle \) classifying so-called “brute facts” in \( S_P \) (i.e. a set of situation or event tokens having a spatio-temporal extension\(^3\)) according to natural event types in \( \Sigma_P \), which is expressed in the binary classification relation \( \|=P \subseteq S_P \times \Sigma_P \). Thus for \( s \in S_P \), \( raiseHand_i \in \Sigma_P \), we write

\[
s \|=P raiseHand_i
\]  

if s is an event of agent i raising her hand.

On the other hand there is the realm of social reality, which we may represent similarly using a classification \( C_S = \langle S_S, \Sigma_S, \|=S \rangle \), classifying situations \( S_S \) according to their social meanings \( \Sigma_S \) in a relation \( \|=S \subseteq S_S \times \Sigma_S \). For instance, the following instance of this relation says that \( s' \) is a situation in which agent i makes a bid.

\[
s' \|=S makesBid_i
\]  

Institutions are systems of constitutive rules, i.e. constraints linking physical actions to their social implications. They account for the fact that the event \( s \) of i raising her hand in \( C_P \) may count as an event \( s' \) of making a bid in the context of auctions, but also as an event of, say, volunteering to solve a problem in math class.

\(^3\) For our purposes, nothing substantial hangs on the distinction between situations and events, so we treat them as one.
Classifications give rise to a hierarchy of contexts or local logics. Let a pair of sets \( \Gamma, \Delta \subseteq \Sigma_P \) be a constraint on a situation \( c \) (written “\( \Gamma \vdash_c \Delta \)” iff when \( c \models X \) for all \( X \in \Gamma \) then \( c \models Y \) for some \( Y \in \Delta \). By extension, \( \Gamma \vdash_C \Delta \) iff \( \Gamma \vdash_c \Delta \) for all \( c \in C \subseteq S \). For example, “\( \text{scratchHead} \uparrow_S \text{raiseHand} \)” is a constraint holding for all situations in \( S_P \). Similarly, constraints on the social classification \( C_S \) (like for instance “\( \text{makeBid}_i \uparrow_S \text{mustPurchase}_i \)” represent normative rules. A context \( \langle C, \vdash, N \rangle \) on a classification \( C \) then consists of a set of constraints \( \vdash \) (closed under logical consequence) and a set of situations \( N \subseteq S \) that are considered “normal” for this context, meaning that they satisfy all constraints in \( \vdash \).

Institutions then are theory-like objects stipulating how a piece of information in \( C_P \) relates to \( C_S \) in a principled way. This suggests a binary channel construction \(^4\) (Fig. 1) around an institutional classification \( C_I = \langle S_I, \Sigma_I, \models_I \rangle \), where \( S_I \) is the Cartesian product of \( S_P \) and \( S_S \), i.e. a set of pairs \( \langle s_1, s_2 \rangle \) (for \( s_1 \in S_P, s_2 \in S_S \)) \(^5\) and \( \Sigma_I \) is the disjoint union of \( \Sigma_P \) and \( \Sigma_S \). Let \( \langle f^\uparrow, f^\downarrow \rangle \) and \( \langle g^\uparrow, g^\downarrow \rangle \) be the usual infomorphisms \(^6\), such that \( \langle s_0, s_1 \rangle \models_I X \) iff \( X = f^\uparrow(X') \) and \( s_0 \models_P X' \) (resp. \( X = g^\uparrow(X') \) and \( s_1 \models_S X' \)). We call a set of constraints \( \{ \Gamma, \Delta \mid \Gamma, \Delta \subseteq \Sigma_I \} \) on \( C_I \) closed under identity, weakening and cut a body of constitutive rules, and a context \( \langle C_I, \vdash_I, N_I \rangle \) on \( C_I \) an institution.

Then institutions act as theories on the alignment of \( C_P \) and \( C_S \).

We can now give formal substance to the claim that \( i \)’s raising a hand counts as making a bid in the social context \( A \in \text{CXT}(C_S) \) of auctions, namely as a constraint of the corresponding institution:

\[
f^\uparrow(\text{raiseHand}_i) \vdash_{g[A]} g^\downarrow(\text{makeBid}_i)
\]

In this picture, any instance \( s \) of a “hand raising” event counts as an act of “bidding” in a social context \( A \) if every interpretation \( \langle s, s' \rangle \) of \( s \) in a social situation \( s' \) of context \( A \) is normal with respect to the constraint \( f^\uparrow(\text{raiseHand}_i) \vdash_{g[A]} g^\downarrow(\text{makeBid}_i) \).

---

4 See [2].

5 Note that a pair \( \langle s_1, s_2 \rangle \) does not necessarily denote two distinct situations. They could be different perspectives (i.e. “physical” and “social”) on the same event.

6 Given classifications \( C_A \) and \( C_B \), an infomorphism \( f : C_A \cong C_B \) from \( C_A \) to \( C_B \) is a pair of contravariant functions \( \langle f^\uparrow, f^\downarrow \rangle \) satisfying that \( \forall s \in S_B, \sigma \in \Sigma_A : f^\uparrow(s) \models_A \sigma \iff s \models_B f^\downarrow(\sigma) \). (Again see [2]).
Count-as Conditionals and Nonmonotonicity  Modeling count-as conditionals as the constraints of some local logic allows us to account for the essential context-dependence of constitutive norms. For example, $f(r\text{aiseHand}_i) \vdash_{g[A]} g(A) \cap (\text{makeBid}_i)$ may be a valid constitutive rule in the institutional context $g[A]$ of auctions, while still allowing the raising of a hand to count as some other act in another social context (say, “$f(r\text{aiseHand}_i) \vdash_{g[V]} g(V) \cap (\text{vote}_i)$” in the context $V$ of voting). One well-known logical property of count-as conditionals is their nonmonotonicity. Constitutive rules like $f(r\text{aiseHand}_i) \vdash \text{makeBid}_i$ generally do not admit left strengthening (eg. $r\text{aiseHand}_i, \text{scrathHead}_i \nvdash \text{makeBid}_i$) or right weakening (eg. $r\text{aiseHand}_i, \nvdash \text{makeBid}_i, \text{disownsChildren}_i$).

In channel theory, nonmonotonicity is dealt with at the level of contexts, rather than rules. That is, inside a given context all inferences are monotonic, by default. Those instances of left strengthening or right weakening that are problematic are cases where a strengthened (respectively weakened) constraint turns out to be in conflict with some situations $N' \subseteq N$ assumed normal with respect to this context, and would thus warrant a shift to some stronger (weaker) institutional context. 7

References


7 As shown in [2], the set of contexts of a classification $C$ forms a complete lattice under context subsumption ($\subseteq$), where $(C, \vdash_1, N_1) \subseteq (C, \vdash_2, N_2)$ iff $\vdash_1 \subseteq \vdash_2$ and $N_1 \supseteq N_2$. A context strengthening or weakening then corresponds to moving to some weaker or stronger context in the $\subseteq$ hierarchy.
Adaptive Bases of Human Rationality

Nobuyuki Takahashi
Department of Behavioral Science
The Center for the Study of Cultural and Ecological Foundations of the Mind
Hokkaido University

Part 1
General framework of CEFOM/21
In studying human behavior, rationality has been one of the most central concepts in behavioral and social sciences, but …

<table>
<thead>
<tr>
<th>Rationality and Adaptationist Perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>In reality people often behave in an irrational manner.</td>
</tr>
</tbody>
</table>

No predictable power at the macro level

Impractical desktop theory

This situation has begun to change since 1990s.

- The rise of experimental/behavioral economics
- The wide acceptance of evolutionary perspectives in psychology
- Interdisciplinary collaborations using evolutionary game theory as a common language

The myth of "tabula rasa" has been almost completely refuted by empirical evidence.

**The human mind is not, in fact, infinitely malleable but instead works under certain constraints.**
Rationality and Adaptationist Perspective

Why do certain constraints exist?

• We are endowed with psychological mechanisms that enable us to behave in certain ways as adaptive tools for the social environment.

• These psychological mechanisms have allowed humans to build and maintain societies.

✓ The few research centers (e.g., George Mason University, Zurich University, the Max Planck Institute, the Santa Fe Institute, UCLA and UCSB) share the basic perspective.

✓ CEFOM/21 (Center for the Study of Cultural and Ecological Foundations of the Mind, a 21st Century Center of Excellence) has been one of the few research centers around the world to study the adaptive bases of human rationality both theoretically and empirically since its establishment in 2002.

Part 2

From empirical findings to theory
Why do people cooperate in a one-shot PDG?

**Puzzle**

• Prisoner’s dilemma game (PDG) is the most well-known game in research on human cooperation.

**The most significant finding in empirical studies is that people actually cooperate very often even in a one-shot PDG where cooperation is irrational.**

---

**Prisoner’s Dilemma Game (PDG)**

• There are two players, A and B.
• Each player has two behavioral choices: Cooperation or Defection.
• Each player’s payoff is determined by a joint decision.
• Player A’s payoff is indicated above and to the right of the diagonal, and player B’s payoff is represented below and to the left.

**Dominant choice: Defection**

Defection produces an individually better outcome no matter what the choice of the partner is.

<table>
<thead>
<tr>
<th>Player B’s Choice</th>
<th>Player A’s Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cooperation</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
</tr>
</tbody>
</table>
Why do people cooperate in a one-shot PDG?

So many people actually choose cooperation.

• A rational player should choose defection.
• However, even in a one-shot PD played between unrelated players under complete anonymity, many people actually choose cooperation.
• This is one of the most basic findings in experimental research during the past few decades.

Are ordinary people irrational?

Player B’s Choice | Player A’s Choice
---|---
Cooperation | Defection
C | 2 2 0 3
D | 3 0 1 1

Why do people cooperate in a one-shot PDG?

Choosing cooperation is not a simple logical error.

Traditional answer: Cooperation is a result of confusion (c.f., Andreoni, 1995).

People fail to understand the incentive structure of the game that makes defection a dominant choice.

Our answer: Social Exchange Heuristic (SEH) (Kiyonari, Tanida, & Yamagishi, 2000)

It is triggered by the construal of the situation as one involving social exchange, and once triggered, subjectively transforms the nature of the exchange, resulting in the perception of the situation as an assurance game. With the transformed perception of the situation, people come to intuitively believe in the desirability of mutual cooperation.
Why do people cooperate in a one-shot PDG?

Social exchange is the foundation of society.

• Everyday, everybody is engaged in social exchange.
  Social exchange – Giving / receiving resources
  Resources can be anything from economic resources (e.g., money) to psychological/sociological resources (e.g., love, respect).

• Social exchange has been the most important activity throughout the history of human beings.

• Social exchange has a characteristic of PD. Although there is a possibility of being exploited, successful exchange brings mutual benefit.
Why do people cooperate in a one-shot PDG?

**PDG as social exchange**

In a social exchange relation, each party pays or does not pay a cost, $c$, to provide the other partner a benefit, $b$, where $b > c$.

<table>
<thead>
<tr>
<th>Social exchange</th>
<th>PDG</th>
<th>A PDG matrix representing four outcomes of a social exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paying the cost ($c$)</td>
<td>Cooperation</td>
<td>Player A’s Choice</td>
</tr>
<tr>
<td>Not paying the cost</td>
<td>Defection</td>
<td>Player B’s Choice</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Cooperation</th>
<th>Defection</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>$b-c$</td>
<td>$b$</td>
</tr>
<tr>
<td>D</td>
<td>$b$</td>
<td>$-c$</td>
</tr>
</tbody>
</table>

Temptation payoff: $b > b-c$  
Reward payoff: $b-c > 0$  
Punishment payoff: $0 > -c$  
Sucker payoff: $-c$

Why do people cooperate in a one-shot PDG?

**Assurance Game (AG) and Prisoner’s Dilemma Game (PDG)**

**Defection is the dominant choice in PDG.**

**There is no dominant choice in AG.**

• When the partner defects, defection produces an individually better outcome.

• However, when the partner cooperates, cooperation produces an individually better outcome.

Assurance Game

<table>
<thead>
<tr>
<th>Player B’s Choice</th>
<th>Player A’s Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation</td>
<td>Defection</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
</tbody>
</table>

Prisoner’s Dilemma Game

<table>
<thead>
<tr>
<th>Player B’s Choice</th>
<th>Player A’s Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation</td>
<td>Defection</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
</tr>
</tbody>
</table>
Why do people cooperate in a one-shot PDG?

*Heuristic as a rule of thumb*

Heuristics are simple, efficient decision-making rules that work well under most circumstances, but in certain cases lead to systematic cognitive biases.

---

**Confusion**

Cooperation is irrational. People choose cooperation *because they are confused* (e.g., they are not smart enough, the instructions during the experiment are too vague). Therefore, *when they are not confused, they would choose defection*.

**Social Exchange Heuristic (SEH)**

It is an important design feature of human cognitive functioning that makes mutual cooperation possible in social exchange. *People choose cooperation when SEH is triggered*. When it is triggered, *it subjectively transforms PD into AG*, and motivates people to seek mutual cooperation.

In order to distinguish which explanation is better, a series of experiments was conducted.
### Why do people cooperate in a one-shot PDG?

**Experimental Design**

- **Reality (high, middle, low)**
- **Partner’s cooperation (known, unknown)**

<table>
<thead>
<tr>
<th>A) Realistic sense of exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confusion – Realistic sense of exchange decreases confusion.</td>
</tr>
<tr>
<td>→ Realistic sense of exchange <strong>decreases cooperation.</strong></td>
</tr>
<tr>
<td>SEH – Realistic sense of exchange triggers SEH.</td>
</tr>
<tr>
<td>→ Realistic sense of exchange <strong>increases cooperation.</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B) Partner’s cooperation</th>
</tr>
</thead>
<tbody>
<tr>
<td>If players know that their partners decided to cooperate…</td>
</tr>
<tr>
<td>Confusion – Players face a less complex decision task than those who make decisions simultaneously.</td>
</tr>
<tr>
<td>→ Players <strong>cooperate less.</strong></td>
</tr>
<tr>
<td>SEH – Sequential nature of the game promotes the sense of exchange, which triggers SEH.</td>
</tr>
<tr>
<td>→ Players <strong>cooperate more.</strong></td>
</tr>
</tbody>
</table>

### Why do people cooperate in a one-shot PDG?

**Manipulation of Reality**

- **Full experiment** – The participants played a PDG once and were paid exactly the amount specified in the payoff matrix.

  - **Reality – High**

- **Vignette experiment** – The participants were told to imagine that they had been participating in the experiment described in the vignette, and to decide whether they would have cooperated or defected if they had been in the experiment.

  - **Vignette with money**
    - Participants were shown the payoff matrix expressed in terms of monetary value (“yen”) as in the original, full experiment.

  - **Reality – Middle**

  - **Vignette with score**
    - Participants were shown the payoff matrix expressed in terms of scores (“points”). There was no mention of money.

  - **Reality – Low**
Why do people cooperate in a one-shot PDG?

*Manipulation of partner’s cooperation*

What if players know that their partners decided to cooperate?

- **Simultaneous-game condition**
  - Participants and their partners make their decisions simultaneously.
  - Players don’t know what their partners would do.

- **2nd-player condition**
  - Participants were told that their partners would make the decision first. *After informed that their partners had chosen C, they made their own decisions.*
  - Players know that their partners have already decided to cooperate.

---

*Experimental Design*

Reality (high, middle, low) ▷ Partner’s cooperation (known, unknown)

Payoff matrix used in the experiment

<table>
<thead>
<tr>
<th>Participant’s choice</th>
<th>Partner’s choice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>1200</td>
</tr>
<tr>
<td>D</td>
<td>1800</td>
</tr>
</tbody>
</table>
### Why do people cooperate in a one-shot PDG?

#### Hypothesis regarding reality

<table>
<thead>
<tr>
<th>Confusion – Realistic sense of exchange decreases confusion.</th>
<th>→ Realistic sense of exchange <strong>decreases cooperation.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation rate …</td>
<td></td>
</tr>
<tr>
<td>Full &lt; Vignette with money &lt; Vignette with score</td>
<td></td>
</tr>
<tr>
<td>Reality (High)</td>
<td>(Middle)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEH – Realistic sense of exchange triggers SEH.</th>
<th>→ Realistic sense of exchange <strong>increases cooperation.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation rate …</td>
<td></td>
</tr>
<tr>
<td>Full &gt; Vignette with money &gt; Vignette with score</td>
<td></td>
</tr>
<tr>
<td>Reality (High)</td>
<td>(Middle)</td>
</tr>
</tbody>
</table>

### Why do people cooperate in a one-shot PDG?

#### Hypothesis regarding partner’s cooperation

If players know that their partners decided to cooperate…

<table>
<thead>
<tr>
<th>Confusion – Players face a less complex decision task than those who make decisions simultaneously.</th>
<th>→ Players <strong>cooperate less.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation rate …</td>
<td></td>
</tr>
<tr>
<td>Simultaneous &gt; 2(^{nd})-player</td>
<td></td>
</tr>
<tr>
<td>(unknown)</td>
<td>(known)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEH – Sequential nature of the game promotes the sense of exchange, which triggers SEH.</th>
<th>→ Players <strong>cooperate more.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation rate …</td>
<td></td>
</tr>
<tr>
<td>Simultaneous &lt; 2(^{nd})-player</td>
<td></td>
</tr>
<tr>
<td>(unknown)</td>
<td>(known)</td>
</tr>
</tbody>
</table>
Why do people cooperate in a one-shot PDG?

Result (1)

Coordination rate …

Full > Vignette with money > Vignette with score

Realistic sense of exchange promoted cooperation.

Confusion was refuted, and SEH was supported.

Why do people cooperate in a one-shot PDG?

Result (2)

In the full experiment and the vignette with money

Simultaneous < 2nd-player

Confusion was refuted, and SEH was supported.

In the vignette experiment with score,

Simultaneous > 2nd-player

Only when the outcome of the game was truly trivial, participants acted rationally.
Why do people cooperate in a one-shot PDG?

Is SEH Adaptive?

It is clear that people have SEH. However, by definition, SEH produces a logical error.

How can it be adaptive?

Error management as an adaptationist explanation for heuristics

- Decision-making under uncertainty often results in erroneous inferences, but some errors are more costly than others.
- SEH helps people reduce the likelihood of one type of error while increasing the likelihood of another type of error in social exchanges.

<table>
<thead>
<tr>
<th>Player B’s Choice</th>
<th>Player A’s Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation</td>
<td>Defection</td>
</tr>
<tr>
<td>C</td>
<td>b-c</td>
</tr>
<tr>
<td></td>
<td>b-c</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Uncertain nature of social exchange

If this exchange opportunity is truly a one-shot game with no possibility of reputation, \( D \) is the dominant choice.

However .... two possibilities of receiving sanctions

- Possibility a) This can be the beginning of a long-term relationship between you and your partner.
  
  Choosing \( C \) is adaptive as TFT dictates (e.g., Axelrod, 1984).

- Possibility b) Information about your behavior (reputation) could potentially spread to other members of the community.
  
  Choosing \( C \) is adaptive in order to avoid ostracism from a community (Davis and Holt, 1993).

Managing our behaviors under such uncertainty is one of the most important adaptive tasks humans face.
### Why do people cooperate in a one-shot PDG?

**Our decisions can produce two types of errors**

The consequences of 4 possible states of the social exchange

<table>
<thead>
<tr>
<th>Inferences</th>
<th>True nature of exchanges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defection is sanctioned</td>
<td>Defection is sanctioned</td>
</tr>
<tr>
<td>(punishment/ostracism)</td>
<td>Correct inference</td>
</tr>
<tr>
<td></td>
<td>Choosing C</td>
</tr>
<tr>
<td></td>
<td>Outcome: Gains from mutual cooperation</td>
</tr>
<tr>
<td>Defection is not sanctioned</td>
<td>Type I Error</td>
</tr>
<tr>
<td></td>
<td>Choosing C</td>
</tr>
<tr>
<td></td>
<td>Outcome: Failure in saving the cost of cooperation</td>
</tr>
<tr>
<td></td>
<td>Correct inference</td>
</tr>
<tr>
<td></td>
<td>Choosing D</td>
</tr>
<tr>
<td></td>
<td>Outcome: Savings in the cost of cooperation</td>
</tr>
</tbody>
</table>

The cost of Type II error is huge in human societies.

**Having SEH (i.e., perceiving an exchange situation under uncertainty as one in which defection is sanctioned) is adaptive.**

**SEH is a systematic bias in favor of committing Type II errors less often despite the cost of the increased probability of committing Type I errors.**

---

### Why do people cooperate in a one-shot PDG?

**Summary of Part 2**

Cooperation in a one-shot PD is a logical error. However, many people actually cooperate.

Experimental results showed that …

- Realistic sense of exchange promoted cooperation.
- People cooperated more when the partner has already cooperated in the sequential game than in the simultaneous game.

Humans have Social Exchange Heuristic.

Making logical errors is more adaptive than making the logically correct decisions in the “real” environment.
Part 3

From theory to empirical findings

What kind of altruism can be adaptive?

Altruism as unilateral resource giving

- Altruistic behavior  - Altruistic motivation

A

Help!

B
What kind of altruism can be adaptive?

*Altruism as unilateral resource giving*

- Altruistic behavior
- Altruistic motivation

A

Give

B

B loses his resource. Giving is costly. How can B’s unilateral resource giving be rational?
What kind of altruism can be adaptive?

**Indirect reciprocity**
Reciprocity not by the recipient but by the third party

![Diagram of indirect reciprocity](image)

If B’s giving is reciprocated not by A but by another person (C), B’s giving can be rational.

However, it has been shown that not indiscriminate giving but discriminate giving is necessary for the emergence of indirect reciprocity.

What kind of discriminate altruism can maintain indirect reciprocity?

---

What kind of altruism can be adaptive?

1. **Theory** – Mathematical analysis, computer simulation
2. **Empirical findings**
What kind of altruism can be adaptive?

Mathematical analysis and computer simulation using evolutionary game theory

Framework of giving game

• A pair consisting of a donor and recipient is chosen randomly from a population.
• The donor decides whether to give his resource to his recipient with a cost of c. When a donor gives, the recipient receives the benefit b (b>c).
• Each player has a reputation score S which has two values: “Good” or “Bad.”
  - The donor gives if he thinks the recipient’s score is Good.
  - The donor doesn’t give if he thinks the recipient’s score is Bad.

<table>
<thead>
<tr>
<th>Four types of recipient</th>
<th>(a) Current recipient’s previous behavior</th>
<th>(b) Previous recipient’s S</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
<td>Bad</td>
</tr>
<tr>
<td>Gave</td>
<td>Good or Bad</td>
<td>Good or Bad</td>
</tr>
<tr>
<td>Didn’t give</td>
<td>Good or Bad</td>
<td>Good or Bad</td>
</tr>
</tbody>
</table>

What kind of altruism can be adaptive?

Mathematical analysis and computer simulation using evolutionary game theory

Framework of giving game

• The donor decides recipient’s score by using two types of information based on recipient’s previous behavior as a donor.
  (a) Their previous behavior (gave or didn’t give)
  (b) Previous recipient’s reputation score (Good or Bad)

• Each player has 4 rules that determine whether each type of recipients is considered good or bad.
• Strategies: sets of 4 rules – 16 strategies are possible.

Strategies dictate what kind of recipients are regarded as good.
What kind of altruism can be adaptive?

Previously proposed solutions (1)

**Image Scoring Strategy** (Nowak and Sigmund, 1998a, b)

Round $m - 1$

X

Y

Recipient

Donor

Round $m$

Y

Z(IS)

Is Y good?

Y

Z(IS)

Recipient

Donor

Round $m$

Y

Z(IS)
What kind of altruism can be adaptive?

Previously proposed solutions (1)

Image Scoring Strategy (Nowak and Sigmund, 1998a, b)

- IS gives the recipient who gave to his/her recipient, and does not give to the recipient who did not give in the previous round.
- IS is a variant of Tit-For-Tat.
### What kind of altruism can be adaptive?

**Previously proposed solutions (1)**

<table>
<thead>
<tr>
<th>(a) Current recipient’s previous behavior</th>
<th>(b) Previous recipient’s S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Bad</td>
</tr>
<tr>
<td>Gave</td>
<td>Good</td>
</tr>
<tr>
<td>Didn’t give</td>
<td>Bad</td>
</tr>
</tbody>
</table>

However…

Image scoring strategy (GGBB) was rejected by later studies (Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003).

### What kind of altruism can be adaptive?

**Previously proposed solutions (2)**

**Standing Strategy**

(Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003)

- Like IS, **standing** gives to the recipient who gave to his/her recipient in the previous round.
- Unlike IS, **standing** uses 2nd order information.
  - **Standing** does not always consider not-giving “bad.”
  - **Standing** distinguishes between justifiable not-giving and unjustifiable not-giving.
What kind of altruism can be adaptive?

Previously proposed solutions (2)

Justifiable not-giving

If the recipient did not give to a bad person, standing gives to the recipient.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z(Standing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did not give</td>
<td>1st order info</td>
<td>Round m - 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>Z(Standing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>Round m</td>
</tr>
</tbody>
</table>
What kind of altruism can be adaptive?

*Previously proposed solutions (2)*

**Justifiable not-giving**

If the recipient did not give to a *bad* person, *standing* gives to the recipient.

<table>
<thead>
<tr>
<th>W</th>
<th>X</th>
<th>2nd order info</th>
<th>Round m - 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>Did not give</th>
<th>Y</th>
<th>1st order info</th>
<th>Round m - 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td></td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>Z(Standing)</th>
<th>Round m</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What kind of altruism can be adaptive?

*Previously proposed solutions (2)*

**Justifiable not-giving**

If the recipient did not give to a *bad* person, *standing* gives to the recipient.

<table>
<thead>
<tr>
<th>W</th>
<th>X</th>
<th>2nd order info</th>
<th>Round m - 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did not give</td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>Did not give</th>
<th>Y</th>
<th>1st order info</th>
<th>Round m - 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad</td>
<td></td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y</th>
<th>Z(Standing)</th>
<th>Round m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>?</th>
<th></th>
<th></th>
</tr>
</thead>
</table>
What kind of altruism can be adaptive?

Previously proposed solutions (2)

**Unjustifiable not-giving**

If the recipient did not give to a **Good** person, **standing** does not give to the recipient.

![Diagram showing the process of giving and not-giving between two individuals, W and X, and later recipient Y, with the conditions of giving and not-giving and the information flow between rounds.]

- **Round m - 2**: Y (Standing) does not give.
- **Round m - 1**: X (Good) did not give to Y.
- **Round m**: W (Gave) to X.

**Previously proposed solutions (2)**

However…

Standing strategy (GGBG) was rejected by later studies (Takahashi and Mashima, 2006).

<table>
<thead>
<tr>
<th>(a) Current recipient’s previous behavior</th>
<th>(b) Previous recipient’s S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good, Gave</td>
<td>Good, Good</td>
</tr>
<tr>
<td>Bad, Didn’t give</td>
<td>Bad, Good</td>
</tr>
</tbody>
</table>
What kind of altruism can be adaptive?

Common characteristics of rejected strategies

It does not matter whom one gives to.
Giving is always regarded as Good.

<table>
<thead>
<tr>
<th>Image scoring strategy (GGBB)</th>
<th>Standing strategy (GGBG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Current recipient’s previous behavior</td>
<td>(b) Previous recipient’s S</td>
</tr>
<tr>
<td>Gave</td>
<td>Good</td>
</tr>
<tr>
<td>Didn’t give</td>
<td>Bad</td>
</tr>
</tbody>
</table>

| (a) Current recipient’s previous behavior | (b) Previous recipient’s S |
| Gave | Good | Good |
| Didn’t give | Bad | Good |

What kind of altruism can be adaptive?

New solutions (Takahashi and Mashima, 2006)

Strict discriminator strategy (SDISC) and Extra standing strategy (ES) (Takahashi and Mashima, 2003)

| (a) Current recipient’s previous behavior | (b) Previous recipient’s S |
| Gave | Good | Bad |
| Didn’t give | Bad | Bad |

| (a) Current recipient’s previous behavior | (b) Previous recipient’s S |
| Gave | Good | Bad |
| Didn’t give | Bad | Good |

• Unlike IS and Standing, SDISC and ES does not always consider giving “good.”

They distinguishes between “good” givers and “bad” givers.
What kind of altruism can be adaptive?

*Strict discriminator strategy and extra standing strategy*

**Helping a “good” person is “good”!**

If the recipient gave to a *good* recipient, *SDISC* or *ES* gives to the recipient.

- **Round m - 2**
  - W gave X
  - Good

- **Round m - 1**
  - X gave Y
  - 1st order info

- **Round m**
  - Y gives Z (*SDISC* or *ES*)

What kind of altruism can be adaptive?

*Strict discriminator strategy and extra standing strategy*

**Helping a “good” person is “good”!**

If the recipient gave to a *bad* recipient, *SDISC* or *ES* does not give to the recipient.

- **Round m - 2**
  - W did not give X
  - Bad

- **Round m - 1**
  - X gave Y
  - 1st order info

- **Round m**
  - Y does not give Z (*SDISC* or *ES*)
What kind of altruism can be adaptive?

Common characteristics of new solutions

- Unlike IS and Standing, SDISC and ES do not give to previous givers who gave to a “bad” recipient.

They regard giving to Bad as Bad.

SDISC and ES regards “saints” (indiscriminate altruists) as “traitors.”

<table>
<thead>
<tr>
<th>Strict discriminator strategy (GBBB)</th>
<th>Extra standing strategy (GBBG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Current recipient’s previous behavior</td>
<td>(b) Previous recipient’s S</td>
</tr>
<tr>
<td>Gave</td>
<td>Good</td>
</tr>
<tr>
<td>Didn’t give</td>
<td>Bad</td>
</tr>
</tbody>
</table>

Three keys to make indirect reciprocity possible

(a) To regard a person who didn’t give despite having had a chance to give to a Good person as “Bad” (i.e., to exclude a free rider)

(b) To regard a person who gave to a Good person as “Good” (i.e., to help a conditional giver)

(c) To regard a person who gave to a Bad person as “Bad” (i.e., to exclude an unconditional giver).

The most important characteristic for the emergence of indirect reciprocity
What kind of altruism can be adaptive?

1. Theory – Mathematical analysis, computer simulation

2. Empirical findings

Do people actually regard those who gave to bad as bad and exclude them?

Experiment
### Experimental setting

**8-person Giving game**
- There were 17 rounds.
- In each round, participants were endowed 50 yen and asked to decide whether to
  - give it to one of the other 7 participants (The endowment they gave were doubled and given to their recipients: recipients received 100 yen) or
  - keep it for themselves (They received 50 yen).
- After each round, they received the feedback information (=how much they received in the round).
- When they made their decisions, they could see
  - 1st-order information
  - 2nd-order information about all members.
What kind of altruism can be adaptive?

Displayed information

This person…

Gave in the last round

to the person who had not given two rounds before

1) 1st-order information: whether each target gave or didn’t give in the last round
What kind of altruism can be adaptive?

Displayed information

This person…

1) 1st-order information: whether each target gave or didn’t give in the last round

2) 2nd-order information: (If 1st-order information was “gave”), whether the person had given to a giver or a non-giver in the last round

What kind of altruism can be adaptive?

Dependent variable (DV)

There were always three types of targets.

1) Giver to Giver (GtoG) 2) Giver to Non-giver (GtoNG) 3) Non-giver (NG)

1st order info

Gave in the last round

2nd order info

toward a person who had given 2 rounds before

toward a person who hadn’t given 2 rounds before

Didn’t give in the last round

DV: The preference score for each type of target

This DV indicates to what extent each participant gave to (or avoided) the target of each type selectively rather than randomly.
What kind of altruism can be adaptive?

Result

What types of target did participants choose as their recipients?
### What kind of altruism can be adaptive?

#### Result

**Average of preference of each type (SD)**

<table>
<thead>
<tr>
<th></th>
<th>1) Giver to Giver</th>
<th>2) Giver to Non-giver</th>
<th>3) Non-giver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.14 (0.23)</td>
<td>0.00 (0.18)</td>
<td>-0.14 (0.13)</td>
</tr>
</tbody>
</table>

Results of a comparison between 3 types (ANOVA) showed that the effect of type of target is statistically significant ($F(2,62)=13.23, p<.0001$).

**Giver to Giver > Giver to Non-giver > Non-giver**

Participants gave more to previous givers than to previous non-givers. Participants gave more to “Giver to Giver” than to “Giver to Non-giver”.

Two predictions were supported.

---

### What kind of altruism can be adaptive?

#### Conclusion

*Actual behavioral patterns are consistent with what theoretical studies have argued to be adaptive.*

To **exclude not only free riders but also unconditional cooperators who benefit free riders** is needed.
Last remarks

When we study human rationality, we need to consider its adaptive bases.

Adaptationist approach argues that many of our psychological traits have been shaped by adaptive tasks we (or our ancestors) faced.

The mechanism that has shaped our psychological traits is still unspecified. It can be natural selection, sexual selection, individual learning, cultural evolution (i.e., imitation of successful strategies), and so on.

Although CEFOM/21 will end this month, we will continue our effort to be at the cutting-edge of research on the fundamental sociality of the human mind.

Thank you for your attention

For details, please read the following articles.


Moral Conflicts between Groups of Agents

Allard Tamminga*

SOCREAL 2007
Hokkaido University
March 10, 2007

1 Introduction

Two groups of agents, \( G_1 \) and \( G_2 \), face a \textit{moral conflict} if \( G_1 \) has a moral obligation and \( G_2 \) has a moral obligation, such that these obligations cannot both be fulfilled.

Before we consider what actions are good or bad, right or wrong, it is proper to consider first what is meant by, and what not, (..) the expression ‘doing an action’ or ‘doing something’ (Austin 1957, p. 178)

In our analysis of moral conflicts between groups of agents, we adopt Austin’s suggestion. On the basis of the well established \textit{stit} logics of agency developed by Nuel Belnap and others,\(^1\) we present a consequentialist system of multi-agent deontic logic, which is a generalization of John Horty’s utilitarian deontic logic.\(^2\)

\[ \diamond \phi \text{ 'It is possible that } \phi \text{' } \]
\[ [G_1] \phi \text{ 'Group } G \text{ of agents sees to it that } \phi \text{' } \]
\[ \circlearrowleft_{G_2} \phi \text{ 'In the interest of group } G \text{ of agents, group } G \text{ of agents ought to see to it that } \phi \text{' } \]

Two groups of agents, \( G_1 \) and \( G_2 \), face a \textit{basic moral conflict} if and only if there is a formula \( \phi \), such that both \( \circlearrowleft_{G_1} \phi \) and \( \circlearrowleft_{G_2} \neg \phi \) are true.

\(^*\)Department of Theoretical Philosophy, Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands. E-mail: A.M.Tamminga@rug.nl. Joint work with Barteld Kooi.


2 Language and Semantics

We use a modal language $L$ built from a countable set $P = \{p_1, p_2, \ldots\}$ of atomic propositions and a finite set $A = \{a_1, \ldots, a_n\}$ of individual agents. $L$ is the smallest set (in terms of set-theoretical inclusion) satisfying the conditions (i) through (v):

(i) $P \subseteq L$
(ii) If $\phi \in L$ and $\psi \in L$, then $(\phi \land \psi) \in L$ and $(\phi \rightarrow \psi) \in L$
(iii) If $\phi \in L$, then $\lnot \phi \in L$ and $\Diamond \phi \in L$
(iv) If $\phi \in L$ and $G \subseteq A$, then $[G] \phi \in L$
(v) If $\phi \in L$ and $F \subseteq A$ and $G \subseteq A$, then $\Box_F G \phi \in L$.

2.1 Consequentialist Models

Definition 1 A consequentialist model $\mathcal{M}$ is an ordered pair $\langle \mathcal{S}, \mathcal{I} \rangle$, where $\mathcal{S}$ is a choice structure and $\mathcal{I}$ an interpretation.

2.2 Choice Structures

Definition 2 A choice structure $\mathcal{S}$ is a triple $\langle W, A, \text{Choice} \rangle$, where $W$ is a non-empty set of possible worlds, $A$ a finite set of agents, and $\text{Choice}$ a choice function.

2.2.1 Choice functions

Given a non-empty set $W$ of possible worlds and a finite set $A$ of individual agents, we define choice sets of individual agents by a choice function from individual agents to sets of sets of possible worlds, i.e., $\text{Choice} : A \rightarrow \mathcal{P}(\mathcal{P}(W))$, meeting the conditions that (1) for each individual agent $a$ in $A$ it holds that $\text{Choice}(a)$ is a partition of $W$, and (2) for each selection function $s$ assigning to each individual agent $a$ in $A$ a set of possible worlds $s(a)$ such that $s(a) \in \text{Choice}(a)$ it holds that $\bigcap_{a \in A} s(a)$ is non-empty.

Given a choice function $\text{Choice}$ from individual agents to sets of sets of possible worlds and given the corresponding set $\text{Select}$ of selection functions $s$ assigning to each individual agent $a$ in $A$ an option $s(a)$ in $\text{Choice}(a)$, we define

$$\text{Choice}(G) = \{ \bigcap_{a \in G} s(a) : s \in \text{Select} \},$$

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

2
2.2.2 \( G \)-choice equivalence of worlds

In choosing an option \( K \) from \( \text{Choice}(G) \), the group \( G \) of agents restricts the total set of possible worlds to the possible worlds in the set \( K \). A formula of the form \([G] \phi\), informally interpreted as ‘Group \( G \) of agents sees to it that \( \phi \)’, is true in a world \( w \) if and only if \( \phi \) is true in all possible worlds that are elements of the option of \( G \) that contains \( w \). Or, equivalently, if and only if for all possible worlds \( w' \) that are \( G \)-choice equivalent to world \( w \) it holds that \( \phi \) is true in world \( w' \).

**Definition 3 (\( G \)-Choice Equivalence)** Let \( \mathcal{S}(= \langle W, A, \text{Choice} \rangle) \) be a choice structure. Let \( G \subseteq A \). Let \( w, w' \in W \). Then \( w \sim_G w' \) (\( w \) and \( w' \) are \( G \)-choice equivalent) is defined to be:

\[
\text{iff } \quad \text{for all } K \in \text{Choice}(G) \text{ with } w \in K \text{ it holds that } w' \in K.
\]

2.3 Interpretations

**Definition 4** An interpretation \( I \) is an ordered pair \( \langle \text{Utility}, V \rangle \), where \( \text{Utility} \) is a utility function and \( V \) a valuation function.

2.3.1 Utility functions

We assume that individual utilities are given by a utility function from ordered pairs consisting of an individual agent and a possible world to the real numbers between \(-5\) and \(5\), i.e., \( \text{Utility} : A \times W \rightarrow [-5, 5] \).

The group utility a group \( F \) of agents assigns to a possible world \( w \) is defined as the arithmetical mean of the individual utilities the individual agents in \( F \) assign to \( w \):

\[
\text{Utility}(F, w) = \frac{1}{|F|} \sum_{a \in F} \text{Utility}(a, w),
\]

if \( F \) is non-empty. Otherwise, \( \text{Utility}(F, w) = 0 \).

2.3.2 \( F \)-dominance between \( G \)’s options

Roughly, a formula of the form \( \circ_F^G \phi \), informally interpreted as ‘In the interest of group \( F \) of agents, group \( G \) of agents ought to see to it that \( \phi \)’, is true in a world \( w \) if and only if for all options \( K \) in \( \text{Choice}(G) \) that do not ensure \( \phi \) there is a strictly \( F \)-better option \( K' \) in \( \text{Choice}(G) \) such that (1) option \( K' \) ensures \( \phi \), and (2) all options \( K'' \) that are at least as \( F \)-good as \( K' \) also ensure \( \phi \). We interpret “\( F \)-betterness” decision-theoretically.
When a group $G$ performs a collective action by choosing an option $K$ from $\text{Choice}(G)$, it constrains the set $W$ of possible worlds to that set $K$ of possible worlds. It may be, however, that the agents who are not members of $G$ are the members of the group $A - G$ perform a collective action by choosing an option $S$ from $\text{Choice}(A - G)$, thereby constraining the set $K$ to the set of possible worlds $K \cap S$. Hence, $G$ usually will not be able to fully determine the outcome of its collective actions, since the final outcome also depends on the actions of agents in $A - G$. Nevertheless, we can define an $F$-dominance relation over $G$’s options. If $K$ and $K'$ both are in $\text{Choice}(G)$, then, intuitively, $K$ weakly $F$-dominates $K'$ if and only if option $K$ promotes the utility of group $F$ at least as well as option $K'$, regardless of the collective action of the agents in $A - G$.

**Definition 5 (F-Dominance)** Let $M(= (S, I))$ be a consequentialist model. Let $F, G \subseteq A$ and let $K, K' \in \text{Choice}(G)$. Then $K \succeq_{F} K'$ ($K$ weakly $F$-dominates $K'$ for $G$) is defined to be:

$$K \succeq_{F} K' \text{ iff for all } S \in \text{Choice}(A - G) \text{ and for all } w, w' \in W,$$

$$\text{it holds that if } w \in K \cap S \text{ and } w' \in K' \cap S, \text{ then Utility}(F, w) \geq \text{Utility}(F, w').$$

As usual, $K \succ_{F} K'$ ($K$ strongly $F$-dominates $K'$ for $G$) if and only if $K \succeq_{F} K'$ and $K' \not\succ_{F} K$.

### 2.4 Semantics

**Definition 6 (Semantical Rules)** Let $M(= (S, I))$ be a consequentialist model. Let $w \in W$ and let $\phi, \psi \in \mathcal{L}$. Then

(i) $M, w \models p$ iff $V(p, w) = \text{TRUE},$ if $p \in \mathcal{P}$

(ii) $M, w \models \neg \phi$ iff $M, w \not\models \phi$

(iii) $M, w \models \phi \land \psi$ iff $M, w \models \phi$ and $M, w \models \psi$

(iv) $M, w \models \phi \rightarrow \psi$ iff $M, w \not\models \phi$ and/or $M, w \models \psi$

(v) $M, w \models \lozenge \phi$ iff there is a $w'$ in $W$ such that $M, w' \models \phi$

(vi) $M, w \models [G] \phi$ iff for all $w'$ in $W$ with $w \sim_{G} w'$ it holds that $M, w' \models \phi$

(vii) $M, w \models \Box_{G} \phi$ iff for all $K$ in $\text{Choice}(G)$ with $K \subseteq \llbracket \phi \rrbracket_{M}$ there is a $K'$ in $\text{Choice}(G)$ with $K' \subseteq \llbracket \phi \rrbracket_{M}$ such that (1) $K' \succeq_{F} K$, and (2) for all $K''$ in $\text{Choice}(G)$ with $K'' \succeq_{F} K'$ it holds that $K'' \subseteq \llbracket \phi \rrbracket_{M}$.

We introduce the following notational conventions: Given a model $M$, we write $M \models \phi$, if for all worlds $w$ in $W$ it holds that $M, w \models \phi$. We write $\models \phi$, if for all
models $\mathcal{M}$ it holds that $\mathcal{M} \models \phi$. Given a choice structure $\mathcal{S}$, we write $\mathcal{S} \models \phi$, if for all interpretations $\mathcal{I}$ of $\mathcal{S}$ it holds that $(\mathcal{S}, \mathcal{I}) \models \phi$.

**Lemma 1** Let $\phi, \psi \in \mathcal{L}$. Then

(i) $\models \Box_G \phi \rightarrow \Diamond_G \phi$ (‘ought’ implies ‘can’)
(ii) If $\models \phi \iff \psi$, then $\models \Box_G \phi \iff \Box_G \psi$
(iii) If $\models \phi$, then $\models \Box_G \phi \wedge \Box_G \psi$
(iv) $\models \Box_G (\phi \wedge \psi) \rightarrow \Box_G (\phi \wedge \Box_G \psi)$ (deontic agglomeration)

2.4.1 An example: the Prisoner’s Dilemma

<table>
<thead>
<tr>
<th></th>
<th>Don’t confess</th>
<th>Confess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don’t confess</td>
<td>3, 3</td>
<td>0, 4</td>
</tr>
<tr>
<td>Confess</td>
<td>4, 0</td>
<td>1, 1</td>
</tr>
</tbody>
</table>

This payoff matrix can be translated into a consequentialist model $\mathcal{M} = (\mathcal{S}, \mathcal{I})$. The choice structure $\mathcal{S}$ is given by $W = \{w_1, w_2, w_3, w_4\}$, $A = \{a, b\}$, $\text{Choice}(a) = \{\{w_1, w_2\}, \{w_3, w_4\}\}$, and $\text{Choice}(b) = \{\{w_1, w_3\}, \{w_2, w_4\}\}$. The interpretation $\mathcal{I}$ is given by

- $\text{Utility}(a, w_1) = 3$
- $\text{Utility}(b, w_1) = 3$
- $\text{Utility}(a, w_2) = 0$
- $\text{Utility}(b, w_2) = 4$
- $\text{Utility}(a, w_3) = 4$
- $\text{Utility}(b, w_3) = 0$
- $\text{Utility}(a, w_4) = 1$
- $\text{Utility}(b, w_4) = 1$,

and $V(p, w) = \text{TRUE}$ if and only if $w \in \{w_3, w_4\}$, and $V(q, w) = \text{TRUE}$ if and only if $w \in \{w_2, w_4\}$. We read $p$ as ‘Agent $a$ confesses’ and $q$ as ‘Agent $b$ confesses’.

Given $\mathcal{M}$, it holds that

- $\mathcal{M} \models \Box_a^p \wedge \Box_a^{a,b} \neg p$ and $\mathcal{M} \models \Box_b^q \wedge \Box_b^{a,b} \neg q$
- $\mathcal{M} \models \Box_a^p \wedge \Box_a^{a,b} \neg p$ and $\mathcal{M} \models \Box_b^q \wedge \Box_a^{a,b} \neg q$
- $\mathcal{M} \models \Box_a^{a,b} (-p \wedge \neg q)$ and $\mathcal{M} \models \Box_b^{a,b} (-p \wedge \neg q)$. 
3 Two Characterizations of Moral Conflicts

Definition 7 (Characterization) Let $\mathcal{C}$ a class of choice structures and let $\phi \in \mathcal{L}$. Then $\phi$ characterizes $\mathcal{C}$, if for all choice structures $\mathcal{S}$ it holds that $\mathcal{S} \in \mathcal{C}$ if and only if $\mathcal{S} \models \phi$.

3.1 Moral Conflicts of Type $\bigcirc_{\mathcal{G}_1}^{\mathcal{F}_1} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}_2} \neg p$

A moral conflict of type $\bigcirc_{\mathcal{G}_1}^{\mathcal{F}_1} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}_2} \neg p$ might occur in a choice structure $\mathcal{S}$ if and only if there are groups $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G}_1$ of agents in $\mathcal{S}$ such that $\mathcal{F}_1$ is non-empty, $\mathcal{F}_2$ is non-empty, $\mathcal{F}_1$ and $\mathcal{F}_2$ are not identical, and $\mathcal{G}_1$ has at least two non-identical options for acting:

Theorem 1 Let $\mathcal{C}$ be the class of choice structures $\mathcal{S}$ such that for all $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G} \subseteq A$ it holds that $\mathcal{F}_1 = \emptyset$ or $\mathcal{F}_2 = \emptyset$ or $\mathcal{F}_1 = \mathcal{F}_2$ or $\text{Choice}(\mathcal{G}) = \{W\}$. Let $p \in \mathcal{P}$. Then

$$\bigwedge_{\mathcal{F}_1, \mathcal{F}_2, \mathcal{G} \subseteq A} \neg(\bigcirc_{\mathcal{G}_1}^{\mathcal{F}_1} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}_2} \neg p) \text{ characterizes } \mathcal{C}.$$ 

3.2 Moral Conflicts of Type $\bigcirc_{\mathcal{G}_1}^{\mathcal{F}} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}} \neg p$

A moral conflict of type $\bigcirc_{\mathcal{G}_1}^{\mathcal{F}} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}} \neg p$ might occur in a choice structure $\mathcal{S}$ if and only if there are groups $\mathcal{F}, \mathcal{G}_1, \mathcal{G}_2$ of agents in $\mathcal{S}$ such that $\mathcal{F}$ is non-empty, $\mathcal{G}_1 - \mathcal{G}_2$ has at least two non-identical options for acting, $\mathcal{G}_2 - \mathcal{G}_1$ has at least two non-identical options for acting, and $\mathcal{G}_1 \cap \mathcal{G}_2$ has at least two non-identical options for acting:

Theorem 2 Let $\mathcal{C}'$ be the class of choice structures $\mathcal{S}$ such that for all $\mathcal{F}, \mathcal{G}_1, \mathcal{G}_2 \subseteq A$ it holds that $\mathcal{F} = \emptyset$ or $\text{Choice}(\mathcal{G}_1 - \mathcal{G}_2) = \{W\}$ or $\text{Choice}(\mathcal{G}_2 - \mathcal{G}_1) = \{W\}$ or $\text{Choice}(\mathcal{G}_1 \cap \mathcal{G}_2) = \{W\}$. Let $p \in \mathcal{P}$. Then

$$\bigwedge_{\mathcal{F}, \mathcal{G}_1, \mathcal{G}_2 \subseteq A} \neg(\bigcirc_{\mathcal{G}_1}^{\mathcal{F}} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}} \neg p) \text{ characterizes } \mathcal{C'}.$$ 

Part of the proof: ($\Leftarrow$) Suppose $\mathcal{S} \notin \mathcal{C}'$. Then there must be $\mathcal{F}, \mathcal{G}_1, \mathcal{G}_2 \subseteq A$ such that $\mathcal{F} \neq \emptyset$ and $\text{Choice}(\mathcal{G}_1 - \mathcal{G}_2) \neq \{W\}$ and $\text{Choice}(\mathcal{G}_2 - \mathcal{G}_1) \neq \{W\}$ and $\text{Choice}(\mathcal{G}_1 \cap \mathcal{G}_2) \neq \{W\}$. To prove that $\mathcal{S} \not\models \bigwedge_{\mathcal{F}, \mathcal{G}_1, \mathcal{G}_2 \subseteq A} \neg(\bigcirc_{\mathcal{G}_1}^{\mathcal{F}} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}} \neg p)$, it suffices to construct a model $\mathfrak{M} = (\mathcal{S}, \mathcal{I})$ in which there is a $w$ such that $\mathfrak{M}, w \models \bigcirc_{\mathcal{G}_1}^{\mathcal{F}} p \land \bigcirc_{\mathcal{G}_2}^{\mathcal{F}} \neg p$. We conclude from the four properties that there are at least two non-identical options $K_1$ and $K_2$ in $\text{Choice}(\mathcal{G}_1 - \mathcal{G}_2)$, at least two non-identical options $L_1$ and $L_2$ in $\text{Choice}(\mathcal{G}_2 - \mathcal{G}_1)$, and at least two non-identical options
$M_1$ and $M_2$ in $\text{Choice}(G_1 \cap G_2)$, and that there is an agent $a$ in $F$. Note that if $K \in \text{Choice}(G_1 - G_2)$ and $M \in \text{Choice}(G_1 \cap G_2)$, then $K \cap M \neq \emptyset$ and $K \cap M \in \text{Choice}(G_1)$. Note that if $L \in \text{Choice}(G_2 - G_1)$ and $M \in \text{Choice}(G_1 \cap G_2)$, then $L \cap M \neq \emptyset$ and $L \cap M \in \text{Choice}(G_2)$. We now define a suitable interpretation $\mathcal{I} = \langle \text{Utility}, V \rangle$. First, $\text{Utility}$ is defined as follows:

$$\text{Utility}(a, w) = \begin{cases} 1, & \text{if } w \in K_1 \cap M_2 \text{ or } w \in L_2 \cap M_1 \\ 0, & \text{otherwise,} \end{cases}$$

and for all agents $b$ in $A - \{a\}$ and for all worlds $w$ in $W$, we fix $\text{Utility}(b, w) = 0$.

Second, we stipulate $V(p, w) = \text{TRUE}$ if and only if $w \in K_1 \cap M_2$.

Let $\mathfrak{M} = \langle \mathcal{S}, \mathcal{I} \rangle$ and let $w \in W$. Now it is easy to show that (1) for all $R \in \text{Choice}(G_1)$ with $R \neq K_1 \cap M_2$ it holds that $K_1 \cap M_2 \succ^{\mathfrak{M}}_G R$ and (2) for all $S \in \text{Choice}(G_2)$ with $S \neq L_2 \cap M_1$ it holds that $L_2 \cap M_1 \succ^{\mathfrak{M}}_G S$. Hence, $\mathfrak{M}, w \models \diamond^{\mathfrak{M}}_G p$ and $\mathfrak{M}, w \models \diamond^{\mathfrak{M}}_G \neg p$. Therefore, $\mathfrak{M}, w \models \diamond^{\mathfrak{M}}_G p \land \diamond^{\mathfrak{M}}_G \neg p$.  

Let us take a closer look at the countermodel to interpret it properly. The group $G_1 \cap G_2$ of agents cannot make a principled choice from $\text{Choice}(G_1 \cap G_2)$ to maximize the interest of group $F$. If $G_1 \cap G_2$ is taken to belong to group $G_1$, it has to choose option $M_2$ to maximize $F$’s interest. On the other hand, if $G_1 \cap G_2$ is seen as a subgroup of group $G_2$, it must rather choose option $M_1$ to maximize $F$’s interest. Obviously, $G_1 \cap G_2$ cannot choose both options. The group $G_1 \cap G_2$ is wearing two hats here.
Utilitarian deontic logic
and a sequel
Yuko Murakami
murakami@ni.ac.jp
National Institute of Informatics
March 2007, SOCREAL (Hokkaido University)

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

Ross' paradox
• Post the letter!
• Therefore, post or burn the letter!
• The paradox applies to all monotonic deontic logics, including SDL
  – "a modality □ is monotonic" means: when A □ B is a theorem, so is □ A □ □ B
• Many alternatives have been proposed
  – Most prospective is stit (see-to-it-that)

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

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

This talk
• Gives some ideas what is going on Horty
• Proposes a logic
• Some siderations
• An application in AI (not mine!)
• If time allows…
  – Comment on stit
  – An alternative proposal
Simplified semantics

- Set of possible worlds
- Partition
- Value function

- **Multi-S5** as action modalities
- With **Independence of agents**:
  - any combination of agents’ choices can be realized by a possible world

Example

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

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

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

Compare options: dominance

- None for both personae

Forget the non-logic option

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

Another example

- No dominant choice for A
- $b_2$ dominates $b_1$ for B
Simplified semantics

- Set of possible worlds
- Partition
- Value function

- **Multi-S5** as action modalities
- With the interaction assumption
  - *Independence of agents*: any combination of agents’ choices can be realized

Horty’s definitions: State

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

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

Df Preference

- Binary relation on any set of possible worlds \( K \) and \( K' \).

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

  \( K' \) is strongly preferred to \( K \), written \( K < K' \), iff \( K \leq K' \) and not \( K' \leq K \).

Df Dominance

- On choices for each agent \( a \):

  \( K' \) weakly dominates \( K \), written \( K \leq K' \), iff for each \( S \in \text{state}_a, K \cap S \leq K' \cap S \).

  \( K' \) strongly dominates \( K \), written \( K < K' \), iff \( K < K' \) and not \( K' < K \).

Compare options: dominance

- None for A
- For B, \( b_2 \) (strongly) dominates \( b_1 \).

Notes

- Unlike payoff function,
  - Value is assigned to PW
  - Value does not depend on agents

Language

- The language \( \mathcal{L} \) contains
  - propositional variables \( p_0, p_1, \ldots \)
  - terms for agents \( a_0, a_1, \ldots \)
  - an identity symbol \( = \) for agent terms
  - truth-functional operators: \( \land, \lor, \neg \)
  - and modal operators: intuitions—the first is universal operator, the second (with an agent index) is S5, and the last is defined later

- Usual abbreviations \( [\exists], [\Box], \text{and } [\text{some more abbreviations…}] \)
More abbreviations

(∃)A = df ~[∀]¬A.

α ≠ β = df ¬(α = β)

◊αA = df ◊αA.

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

Truth condition

M, v = (∃)A iff for each K ∈ choiceᵥ such that K ⊆ ||A||, there is a K′ ∈ choiceᵥ such that
(i) K ⊂ K′,
(ii) K′ ⊆ ||A||, and
(iii) K′ ⊆ ||A||" for every K′ ∈ choiceᵥ with K′ ⊂ K′′.

Example: infinite choice for a single agent

- W=N
- R=identity relation
- Value(n)=n

1. v(p)={x:x>2} ◊αp holds everywhere
2. v(p)={x:x is even} ◊αp holds nowhere

Axioms...

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

And more axioms and rule

• Independence of agents (scheme)
niff(β₀,…,βₖ)(◊α₀A₀ ∧ … ∧ ◊αₖAₖ) → (∃)◊α₀A₀ ∧ … ∧ (∃)◊αₖAₖ.
  - E.g. Independence of 2 agents
    α ≠ β ∧ (∃)◊αA ∧ (∃)◊βB → (∃)(◊αA ∧ ◊βB);
• Rule
  - Necessitation for universal operator from A to infer [∀]A
• Rules for other operators can be derived.

Results

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

- Axiomatizability: common for several classes of frames proposed by Horty and Thomason
- Logics sharing the same axiomatic system
  - w/ 0-1 value assignment
  - Agent-relative value assignment
- Horty’s proposal needs subtler investigations than presented here:
  - $S_5$ for action is too simple!

A sequel of utilitarian deontic logic: Ethical robots

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

Comparison of ethical codes

- $J$: harsh utilitarian code governing $R_1$
- $O$: common sense code governing $R_2$
- $J^*$: harsh code governing both
- $O^*$: common sense code governing both

$J$ if $J$ holds, then $R_1$ terminates life support.

$O$ if $O$ holds, $R_2$ should not delay pain med.

etc.

Input comments on outcomes

- $R_1$ terminates life support and $R_2$ does not delay pain med $\rightarrow (-!)$ [strongly negative]
- $R_1$ refrains from life support termination and $R_2$ delivers appropriate pain med $\rightarrow (+!!)$ [best]
- $R_1$ refrains from life support termination but $R_2$ withholds the med $\rightarrow (-)$ [bad]
- $R_1$ stop the support and $R_2$ withholds $\rightarrow (-!!)$ [worst]

Can be represented in formulas of Horty’s language

Key assumption also coded

- If either $R_1$ or $R_2$ is ever obligated to see to it that they are obligated to see to it that $P$ is carried out, they in fact deliver.

The theorem prover answers

- to the query

  Does each ethical code implies (+!!)?
  i.e. Does a code implies the best outcome?
  
- No to $J, O, J^*$
- Yes to $O^*$

  That is: The system picks up a ethical code which produces a desirable outcome.
Stit

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

Moral dilemma revisited

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

Adding the principle

- Ross’ paradox revives in Stit with an additional assumption

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

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

Proposal

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

Modal logic of partitions

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

Need philosophical examination

- Metaphysical status of possible worlds
- Information and action
- Moral theory
References: stit and other logics

• Belnap et al. (2001) *Facing the Future*. OUP.
• Hory (2001) *Agency and Deontic Logic*. OUP.

References: ethical robots

• *Toward Ethical Robots via Mechanized Deontic Logic* K Arkoudas, S Bringsjord, P Bello - *Machine Ethics: Papers from the AAAI Fall Symp, 2005*
• *Toward a General Logicist Methodology for Engineering Ethically Correct Robots* S Bringsjord, K Arkoudas, P Bello - IEEE Intelligent Systems, 2006
A Strategic View of Promising
Jun Miyoshi

1 Introduction

The aim of this paper is to explain why people make a promise and keep it. I try to achieve it from the strategic point of view but not the traditional. Traditionally, it is thought to be a rule (or code, contract, convention, and so on) that a promisor should keep her promise unless the promisee releases her.\(^1\) Theorists who agree to it say that, thus, people are following this rule when they keep a promise. However, my view is that a promisor is choosing the action which maximizes her interest when she keeps the promise and the action belongs to an equilibrium. Promising and fulfilling it is a kind of strategy to get the maximum in the situation.

The remainder of this paper is organized as follows. First, I bring forward a general argument about how to explain regularity of human social behavior. I pose some doubts about the rule-following model of social actions, which is very common in philosophy and sociology, and then defend a view based on the concept of equilibrium. Second, I introduce game theory and apply it to the problems of promising. It is shown that giving and keeping a promise is the best strategy and a part of a Nash equilibrium in some repeated games.

2 Rule-Following and Equilibrium

People often do the same actions as others do. For example, they say “Good morning” (or just “Morning”) seeing someone in the morning, and they keep the promises they made. Most of these actions do not appear to be determined causally like natural phenomena, because they are done intentionally. Showing why people do these similar actions requires different justifications from those in natural sciences.

Philosophy, sociology, and other social sciences have been trying to explain such non-causal regularity of human behavior. A common explanation is that there is a rule telling people to do the actions in question and they are following it. It is very persuasive in some cases since the subjects who do the actions admit it. For example, we, as participants in the regularity, know that saying “Good morning” is the correct way of offering a greeting in the morning. In other words, it is a constitutive rule known to English speakers that saying “Good morning” counts as a greeting in the

\(^1\) I refer to a promisor, speaker, or player 1 by “she” and to a promisee, hearer, or player 2 by “he”.

morning.²

However, I think that this rule-following model of social behavior has at least five weaknesses. First, we do not know who, when, where, and how decided such a rule. This aporia is of the same type as that of social contract theory. Second, even if there was a historical fact that someone had enacted legislation for it, it would not imply that we should obey it. For instance, suppose that some Mesopotamians set a rule for promising thousands of years ago. Does it have any practical meaning to us? Third, hypothesizing a rule behind a horde of similar actions raises other problems: why should people follow the rule, and why do they follow it regularly actually? Obviously, a rule in itself never forces people to do what it tells. Here, we must find another reason why people follow the rule and, moreover, we can not appeal to another rule, to prevent infinite regress. Fourth, when we deal with some illocutionary acts, they are so basic that it is impossible to make the rules of them. For example, as for proposing and agreeing, you can neither propose the rule of proposing nor agree to the rule of agreeing, if they are rule-following actions. Also, we do not know how to promise to follow the rule of promising. How can these rules be established without these actions? Lastly, the rule-following model is involved in not a few philosophical controversies.³ Resolving them looks much more difficult than getting around them.

Another way to deal with the regularity of human behavior is to analyze it as an equilibrium in the given circumstance. Roughly, “equilibrium” means that each of all agents is doing her or his best. If a set of actions is an equilibrium, it entails that any agent’s deviating from the set of actions will make her or his profit smaller. Accordingly, an equilibrium has a kind of self-enforcing power and stability, if it is achieved, for rational agents. In other words, it is reasonable to suppose that people do spontaneously the actions which belong to an equilibrium. Therefore, explanations based on the concept of equilibrium can avoid the above defects of the rule-following model.

Let me describe my basic scheme for the analysis. It has two hypotheses. The first one is that a person is rational or tries to maximize her or his profit. Here, “profit” does not mean money or pleasure. It is utility value given as a real number measured technically by the order of those things the person prefers.⁴ The second hypothesis is that a person is intelligent enough to calculate the profit which will be gained through an action in question depending on how much relevant information she or he has. On these hypotheses, we can formulate people’s social actions in a given

² For the idea of constitutive rule, see Searle, 1969, section 2.5 and Searle, 1995, 43-51.
³ One of the most famous ones will be the paradox of rule-following. See Wittgenstein, 1953, and Kripke, 1983. There have been many disputes about what Wittgenstein really thought on it.
⁴ Strictly speaking, this is so called von Neumann-Morgenstern utility. See von Neumann and Morgenstern, 1953, Appendix.
situation and work out its equilibrium by game theory. If the result supports the actual behavior, it should be recognized as a correct explanation of it.

Some might argue, against these hypotheses, that a human does not seem so rational since many experiments and observations have revealed surprising examples of human irrationality. Though it has some plausibility, but I think that humans usually pursue the maximization of their profits. Otherwise, we would have to say that they did something intentionally with no reason. In fact, when someone does not appear to be rational, we will commonly expect to find some causes of it such as causal interference and luck of relevant information in her or his decision process. The cases that people act unaccountably irrationally, if they are, are very rare and extraordinary. I think that we can let them go out of range of formal analysis. They will be a matter of causal studies.

The question why people keep their promises is a case to which the equilibrium-based view can be applied. According to the traditional view, it is a rule that people should fulfill their promises.\(^5\) Especially, some speech act theorists assert that this is a part of constitutive rules for the speech act of promising or their logical derivation.\(^6\) But supposing such a rule does not necessitate people’s observing promises. Actually, many philosophers have been considering why a promise is not or should not be broken even if they admit the existence of such a rule.\(^7\) I think that this problem can be solved with the ideas of rationality and equilibrium as I discuss in the following section. If this is right, we will have a more promising method of understanding human behavior, in some sense, and can avert at least some defects of the rule-following model.

### 3 Promising

#### 3. 1 Game Theory

In this subsection, I introduce game theory. My targets are Nash equilibrium and folk theorem. Readers who already know them can skip this subsection.

We use the theory of non-cooperative games. A non-cooperative game is such that two or more players compete with each other. Because we require just two persons, that is, a promisor and a promisee, I mention two-person games. Let us call the two persons player 1 and player 2. In a non-cooperative game, each player chooses one strategy independently and gains a utility value

---

\(^5\) Typically, Prichard, 1949.
\(^6\) Searle, 1969, chapter 8.
according to the both players’ choices. A strategy is a sequence of actions a player can do. If a player can do one action in a game, her or his strategy consists of one action.

Now, first, let us see Nash equilibrium. A Nash equilibrium is a pair of strategies which are the best responses to each other. For example, (A, B) is the Nash equilibrium of the game shown in Table 1. The table means that player 1 chooses doing A or not doing A and that player 2 doing B or not doing B. The number in the left side of each box shows the utility value player 1 gains and the right one player 2’s if the pair of actions indicated by the box is selected. For instance, when (A, B) is chosen, player 1 has 4 and player 2 also has 4. Clearly, doing A is better than not doing A for player 1 and doing B is better than not doing B for player 2. Therefore, the set of strategies (A, B) is composed of the best responses to each other. A Nash equilibrium is understood as a stable point such that no player will change the choice.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>not B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4, 4</td>
<td>3, 2</td>
</tr>
<tr>
<td>not A</td>
<td>2, 3</td>
<td>1, 1</td>
</tr>
</tbody>
</table>

Table 1

A Nash equilibrium is not always the optimal. In other words, it may not be the pair of strategies that gives both players the highest utility values among the possible ones. An example is shown in Table 2. The Nash equilibrium is (not A, not B) since, for player 1, not doing A is better than doing A whichever player 2 chooses, and, for player 2, not doing B is better for the similar reason. But (A, B) gives the higher values to both A and B than (not A, not B). Thus, in this game, the Nash equilibrium is not the optimal. This type of game is called “Prisoner’s Dilemma.”

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>not B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3, 3</td>
<td>1, 4</td>
</tr>
<tr>
<td>not A</td>
<td>4, 1</td>
<td>2, 2</td>
</tr>
</tbody>
</table>

Table 2

Next, we see folk theorem. Is (A, B) in Table 2 impossible in any way? No. It is possible in repeated games. Suppose that players 1 and 2 play the game infinitely many times. In the repeated game,

---

8 For the formal definition of Nash equilibrium, see Nash, 1950 and 1951.
9 I limit the values to being from 1 to 4, for simplicity. Fixing the range of utility values does not harm the generality of the discussion. See Von Neumann and Morgenstern, 1953, 24, 25, and 627.
10 Compared to a repeated game, an original game of it is called “stage (game)”.


each player has an uncountable set of strategies. For instance, player 1’s strategy might be \{A, A, A \ldots\}, \{A, not A, A, not A \ldots\}, or \{not A, not A, not A \ldots\}. Among these, she can choose the trigger strategy such that she does A as far as player 2 does B but, once he does not do B at any stage, then she keeps on not doing A from the following stage. Player 2 also can take his trigger strategy in the same way. Then, it is provable that the pair of these trigger strategies is a Nash equilibrium of the repeated game under certain conditions. Consequently, \( (A, B) \) is chosen repeatedly. This is called folk theorem.

**Proof.**

Let \( \delta_1 \) and \( \delta_2 \) be the discount factors of players 1 and 2 respectively and \( 0 < \delta_1 < 1, 0 < \delta_2 < 1 \).

Suppose that player 2 uses a trigger strategy described above.

On the one hand, if player 1 keeps doing A, her expected utility value is
\[
3 + 3\delta_1 + 3\delta_1^2 + 3\delta_1^3 + \ldots = \frac{3}{1-\delta_1}.
\]

On the other hand, if player 1 does not do A from a stage, her expected utility value is
\[
4 + 2\delta_1 + 2\delta_1^2 + 2\delta_1^3 + \ldots = 2 + \frac{2}{1-\delta_1}.
\]

Then, we have
\[
\frac{3}{1-\delta_1} \geq 2 + \frac{2}{1-\delta_1} \text{ if and only if } \delta_1 \geq \frac{1}{2}.
\]

This means that doing A is the better for player 1 as far as player 2 chooses a trigger strategy and her discount rate is not less than 1/2.

In the same way, it can be shown that doing B is the better for player 2 as far as player 1 chooses a trigger strategy and his discount rate is not less than 1/2.

These entail that the pair of trigger strategies is a Nash equilibrium of the repeated game.

Folk theorem has great importance for understanding social actions. It supplies rational foundations of social cooperation in the long run. That is, socially cooperative attitude can be the best strategy for individuals in the repeated game of lifetime. If this is right, mutual altruistic actions can be interpreted as rational actions in the sense that they lead their agents to the maximization of profit.

3.2 Analysis of Promising

In this subsection, we extend folk theorem and apply it to promising. The idea is that many promises are made and kept in give-and-take relationships. While all players take helpful actions at one turn in the game in Table 2, we often see that a person helps another earlier and the latter helps back the
former later. This can be described as that each player helps the other in turn in a repeated game.

First, we extend folk theorem. Suppose that two players play a repeated game. But the games are mixture of the two kinds of stage game. One is in Table 3-1 and the other is in Table 3-2. These two stage games are randomly repeated infinitely. Strictly, the repeated game is such that each stage game is either one in Table 4-1 or one in Table 4-2 with the same probabilities of 1/2.\textsuperscript{11}

\begin{center}
\begin{tabular}{|c|c|c|}
\hline
 & B & not B \\
\hline
A & 1, 4 & 1, 2 \\
not A & 2, 2 & 2, 2 \\
\hline
\end{tabular}
\end{center}

Table 3-1

\begin{center}
\begin{tabular}{|c|c|c|}
\hline
 & B & not B \\
\hline
A & 4, 1 & 2, 2 \\
not A & 2, 1 & 2, 2 \\
\hline
\end{tabular}
\end{center}

Table 3-2

Table 3-1 means that player 1 can altruistically help player 2 by doing A and he can reject her help by not doing B, and Table 3-2 means that player 2 can altruistically help player 1 by doing B and she can reject his help by not doing A.

Trigger strategies can be devised in these games too. Player1’s trigger strategy is such that she begins by doing A and, if player 2 does not do B at any stage, then she does not do A at the next stage and keeps it later on. Similarly, player 2’s trigger strategy is such that he begins by doing B and, if player 1 does not do A, then he does not do B and keeps it later on.

Then, it can be proved that, under a certain condition, a Nash equilibrium of the repeated game is the pair of these trigger strategies. In this equilibrium, the repetition of (A, B) or helping each other continually is rationally realized.

\textit{Proof.}

Since the games are symmetrical for the two players, just considering the choice of player 1 at a stage in Table 3-1, which gives the only chance for her to get more by not doing A, is enough.

\textsuperscript{11} This is not a game with incomplete information in the proper sense. Each player knows her or his and the other’s types at every turn in this repeated game. They just can not predict which types they have at later turns.
Suppose that player 1 is playing a stage game in Table 3-1 and player 2 uses a trigger strategy. Note that the next stage game can be one in Table 3-1 or one in Table 3-2 with the probability of 1/2. Thus, if player 1 does A at a stage, the utility value she will gain at the next game is

\[ 1 \times \frac{1}{2} + 4 \times \frac{1}{2} = \frac{5}{2}. \]

On the one hand, if she does A, her expected utility is

\[ 1 + \frac{5}{2} \delta_1 + \frac{5}{2} \delta_1^2 + \frac{5}{2} \delta_1^3 + ... = \frac{3}{2} + \frac{5/2}{1 - \delta_1}. \]

On the other hand, if she does not do A, her expected utility is

\[ 2 + 2\delta_1 + 2\delta_1^2 + 2\delta_1^3 + ... = \frac{2}{1 - \delta_1}. \]

Then, we have

\[ \frac{3}{2} + \frac{5/2}{1 - \delta_1} \geq \frac{2}{1 - \delta_1} \text{ if and only if } \delta_1 \geq \frac{2}{3}. \]

Therefore, under the condition that each player's discount factor is not less than 2/3, the pair of trigger strategies is a Nash equilibrium of the repeated game.

Now, the general result of these games can be shown. The generalized form of the games is given in the following tables. The assumptions are

\[ x_1 \leq z_1, \ y_1 \leq w_1, \ x_2 \geq z_2, \ x' \leq y'_2, \ z' \leq w'_2, \ x' \geq y'_1, \ (x_1 + x'_1)/2 \geq (w_1 + w'_1)/2, \text{ and } (x_2 + x'_2)/2 \geq (w_2 + w'_2)/2. \]

<table>
<thead>
<tr>
<th>B</th>
<th>not B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x_1, x_2</td>
</tr>
<tr>
<td>not A</td>
<td>z_1, z_2</td>
</tr>
</tbody>
</table>

Table 3'-1

<table>
<thead>
<tr>
<th>B</th>
<th>not B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>x'_1, x'_2</td>
</tr>
<tr>
<td>not A</td>
<td>z'_1, z'_2</td>
</tr>
</tbody>
</table>

Table 3'-2

In the repeated game, the pair of trigger strategies is a Nash equilibrium if and only if

\[ \delta_1 \geq \frac{z_1 - x_1}{z_1 - x_1 + \left( \frac{x_1 + x'_1}{2} - \frac{w_1 + w'_1}{2} \right)} \]

and
Proof.
Consider player 1’s choice at a stage in Table 3’-1 while player 2 uses a trigger strategy.
On the one hand, if she does A, her expected utility value is
\[ x_1 + \frac{x_1 + x'_1}{2} \delta_1 + \frac{x_1 + x'_1}{2} \delta_1^2 + \ldots = x_1 - \frac{x_1 + x'_1}{2} + \frac{(x_1 + x'_1)/2}{1 - \delta_1}. \]
On the other hand, if she does not do A, her expected utility value is
\[ z_1 + \frac{w_i + w'_i}{2} \delta_1 + \frac{w_i + w'_i}{2} \delta_1^2 + \ldots = z_1 - \frac{w_i + w'_i}{2} + \frac{(w_i + w'_i)/2}{1 - \delta_1}. \]
Then, we have
\[ x_1 - \frac{x_1 + x'_1}{2} + \frac{(x_1 + x'_1)/2}{1 - \delta_1} \geq z_1 - \frac{w_i + w'_i}{2} + \frac{(w_i + w'_i)/2}{1 - \delta_1} \]
if and only if \( \delta_2 \geq \frac{z_1 - x_i}{z_1 - x_i + \left( \frac{x_i + x'_i}{2} - \frac{w_i + w'_i}{2} \right)} \).
In the same way, we have the result about \( \delta_2 \).

From these results, we can say that people are cooperative generally, even in a probable sequence of different situations, because being cooperative makes their profits larger than not being cooperative. Of course, the above propositions are not perfectly general and natural, but I do not think it is very unreasonable to say that they describe some important features of social behavior. In addition, this seems consistent with our ordinary experience. We tend to be kind to those who are kind to us and unkind to those unkind to us. These suggest that our life is like repeated games and our attitudes are close to trigger strategies.

Next, we apply this to promising. I take the following conversation between players 1 and 2 as a standard example of promising.

Example 1
1 “Would you lend me the book? I promise to give it back next Monday.”
2 “OK. Here you are.”
1 “Thanks.”

The situation where this conversation occurs is as follows: player 2 owns the book, and player 1 does
not have another copy of it but has to read it for, say, an examination, but he will be troubled if he loses the book. For player 1, the best is that player 2 lends her the book and she does not give it back to him. The second is that he lends it and she gives it back. The worst is that he does not lend it. For player 2, the best is that he does not lend the book to player 1. The second is that he lends it to her and she gives it back. The worst is that he lends it and she does not return it. These preferences are expressed in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>lending</th>
<th>not lending</th>
</tr>
</thead>
<tbody>
<tr>
<td>giving back</td>
<td>3, 3</td>
<td>1, 4</td>
</tr>
<tr>
<td>not giving back</td>
<td>4, 1</td>
<td>1, 4</td>
</tr>
</tbody>
</table>

Table 4

This game just as shown in Table 4 does not imply that (giving back, lending) is chosen. The Nash equilibria are (giving back, not lending) and (not giving back, not lending).\textsuperscript{12} This means that player 2 will not pass the book to player 1 in the short run.

Yet, considering it as a part of the repeated game given in Table 4'-1 and Table 4'-2 can make his lending the book rationally possible.

<table>
<thead>
<tr>
<th></th>
<th>cooperating</th>
<th>not cooperating</th>
</tr>
</thead>
<tbody>
<tr>
<td>cooperating</td>
<td>3, 3</td>
<td>1, 4</td>
</tr>
<tr>
<td>not cooperating</td>
<td>4, 1</td>
<td>1, 4</td>
</tr>
</tbody>
</table>

Table 4'-1

<table>
<thead>
<tr>
<th></th>
<th>cooperating</th>
<th>not cooperating</th>
</tr>
</thead>
<tbody>
<tr>
<td>cooperating</td>
<td>3, 3</td>
<td>1, 4</td>
</tr>
<tr>
<td>not cooperating</td>
<td>4, 1</td>
<td>4, 1</td>
</tr>
</tbody>
</table>

Table 4'-2

Suppose that the games in Table 4'-1 and in Table 4'-2 are repeated infinitely. Let the probability that either game is played at a stage be 1/2. In the same way as the above repeated game (Tables 3-1 and 3-2), we can show that the pair of trigger strategies is a Nash equilibrium and (cooperating, cooperating) can be realized in these games.

\textsuperscript{12} This game has two Nash equilibria, but it has just one subgame perfect equilibrium (not giving back, not lending). The latter equilibrium is more refined than the former.
Proof.
Since the games are symmetrical for the two players, just considering the choice of player 1 is enough.
According to the general result, the pair of trigger strategies is a Nash equilibrium if and only if
\[
\delta_1 \geq \frac{z_1 - x_1}{z_1 - x_1 + \left(\frac{x_1 + x_1'}{2} - \frac{w_1 + w_1'}{2}\right)}.
\]
In this repeated game, it means
\[
\delta_1 \geq \frac{4 - 3}{4 - 3 + \left(\frac{3 + \frac{3}{2} - 1 + \frac{4}{2}}{2}\right)} = \frac{1}{1 + 1/2} = \frac{2}{3}.
\]
Therefore, under the condition that each player’s discount factor is not less than 2/3, the pair of trigger strategies is a Nash equilibrium of this repeated game.

It is very plausible that the players’ particular actions in the situation, say, lending the book and giving it back, are cooperative actions. This makes the game shown in Table 4 one stage of the repeated game in Tables 4'-1 and 4'-2. Therefore, (lending, giving back) in Table 4 can be seen as a part of a Nash equilibrium of this repeated game. Thus, it is justified that player 2 lends his book to player 1 and she gives it back to him, i.e. she keeps her promise. In short, keeping a promise is doing a cooperative action. Not doing it makes the other player pull the trigger. I think that this example is common enough for us to say that the scheme generally explains why people keep their promises.

Now, the remaining problem is why people make a promise. The above example suggests that it is because they need to persuade others to do something difficult to do in the short run. Actually, (lending, giving back) is not a Nash equilibrium of the one-shot game in Table 4. Arguably, promising, in such a circumstance, works as showing clearly the promisor’s recognition of the promisee’s action as a cooperative action in the long relationship between them. It places the actions the players are considering in the context of a repeated game.

This view will be supported by the following two facts. First, it also explains why just an intention-communicating is not as effective as a promise. An intention-communicating is telling what the speaker intends to do.⁷ See the example below.

Example 2
1 “Would you lend me the book? I will give it back next Monday.”

---

⁷ Owens, 2006. “Communicating” seems to include stating, describing, saying, etc.
2 “Do you promise?”
1 “Yes. I promise.”
2 “OK. Here you are.”
1 “Thanks.”

The utterance “I will give it back next Monday” does not unequivocally mean a promise. It may be an intention-communicating. Asking “Do you promise?” is trying to make it clear which it is. Thus, it is reasonable to infer that player 2 would not decide to lend the book if player 1 did not promise to put it back. Why was not her intention enough for him?

The game-theoretic view can supply an answer. If player 1 does not promise but only tells her intention, it implies that she thinks that what she hopes player 2 to do is a part of a Nash equilibrium of the situation as one shot game. It follows from this that her evaluation of his having the book for him may be lower than player 2’s himself, that is, she may think that if she takes the book away he will not be troubled as much as he will actually be. If so, she will not behave as he expects rationally on the basis of his own evaluation of it. Therefore, a mere intention-communicating is not enough. Player 2 must confirm how much player 1 evaluates the utility value of that thing for him. Her saying that she will promise informs him of her good evaluation of it.

We now can see an important difference between promising and intention-communicating. It is the value of the hearer’s action to be done in exchange of the speaker’s action she promises or says she intends to do. If this value is high enough, then she should promise to do something for it. If it is not, she should not promise but just intention-communicate to do something for it. This difference comes from the fact that promising puts the exchange in some repeated game but intention-communicating in one-shot game.

Second, it also explains when and why a promise does not work well. See the example below.

Example 3
1 “Would you lend me one million dollars? I promise to give it back next Monday.”
2 “No way.”

Suppose that player 2 can afford to pay one million dollars if he collects all his money. But it is very natural that he hesitates to lend so much money even when player 2 promises to pay it back. This is a typical case that promising is not effective.
Why is not a promise effective in this case? An answer is given by the scheme of repeated game. Ordinarily, people do not give and take one million dollars. It means that they are not dealing with so much money in their repeated game. In other words, player 2 will not have adequately many chances to borrow one million dollars or more from player 1. It entails that, for her, the cost caused by his trigger strategy will be smaller than the profit of her taking away his one million dollars. Therefore, the promise to pay back one million dollars is easy to breach and does not have practical effect. Lending one million dollars or more is too much to be secured only by the long term relationships. This points the other limit of promising. If the value of the action for the promisee to do is too high, then the promisor should do something more than promising, such as giving a mortgage.

4 Conclusions

I have posed the two questions: why do people make a promise and why do they keep it? I have argued that it is because they try to maximize their profits and put their interactions in the context of a repeated game which has a new equilibrium.

In conclusion, I mention the role of a rule for promising. I know that many people think that it is a rule that you should keep your promise and that they actually keep their promises being conscious of the rule. I do not think it is completely false. In my view, the rule plays the role of heuristics or a guide to the best for you to do after you made a promise. Many elements of a real situation where you are interacting with others are often so obscure and complicated that you can not calculate a Nash equilibrium of the situation. In that case, it is very reasonable to decide to keep your promise following the rule, which is a kind of lesson learned from many people’s experience. Possibly, it will reflect the general fact that cooperating with others is more valuable than you are inclined to feel. But this does not mean that it is an inviolable rule. It may be broken when trigger strategies are known not to be strong enough in the circumstance where a promise is made. For example, suppose that you promised to play golf with your friend today, but just when you were about to go, you found that your child was very sick. Clearly, you should break the rule, and take the child to the doctor. It is since taking care of your kid is more important than what your friend will do in revenge for your omitting playing golf with her or him.14

14 Actually, people seem to be very flexible toward promise-breaking. At least, they hardly pull the trigger against only one or two violations. It will be because they know that the future is unpredictable and that individuals often must respond to unexpected accidents. I think that the social mechanism of mutually admitting promise-breaking can be analyzed from the strategic point of view.
Bibliography


Moral discussion using any common language suffers from ambiguities and vagueness of expression, imprecision of idiomatic expressions and the thousand thousand vagaries any common language subjects its topics of analysis. Common language, by nature metaphoric, impedes as much as facilitates communication. The expression “Words fail” comes to mind, when trying to articulate oneself, particularly about moral decision-making, since so many variables and hidden (oops, metaphor!) assumptions are lurking (oops, metaphor!) beneath the surface (oops, yet another metaphor!) of expression. One should also add the potential for laziness of linguistic execution, even within the inherent limitations of normal language. British novelist and essayist George Orwell, in his essay “Politics and the English Language,” rails against what he saw as a reciprocal correlation between lazy use of language and a degeneration of thought resulting from the lazy use of language. For these reasons, one appreciates immensely the discipline and rigor with which formal language must subject its object of investigation. Enter, meta-language and formalization where common language fails us.

My intent is not to denigrate any attempts to create discrete languages, meta-languages, hierarchies (e.g. organizational) to analyze more precisely moral relationships (e.g. obligation, commitment, promise-keeping), moral concepts (e.g. moral agent, choice, consequence). It is certainly of interest ‘to see how far one can go’ with formal and meta-linguistic techniques. To be sure, more secure concepts, fewer ambiguities and vague expressions can only help. One can even speak of beauty in formal language, measured by its level of precision, its economy of expression, simplicity or even its elegance. The lure of abstraction is so great that is useful to remember Plato’s thinly veiled message in his allegory of the cave. The mind, after struggling hard to rise above the common and rough world of sensible and social reality, may want to linger and inhabit its mansions of thought, beautiful and pleasing as they are. Plato, as we know, has those liberated from the dross of existence, those who have been forced free of the familiar to return to the rough and imperfect world people actually inhabit. We should heed his advice, however implicit it may be in meta-ethical discussions.
The value of an artificial or meta-ethical language is contingent partly upon its applicability to this rough world of ours. Here, in the shadows hide vast and varied repositories of emotions, sub-conscious drives, shifting power structures, cultural expressions and practices, historical elements. An artificial or meta-ethical language can extract essential connections between concepts from the mountains of detail, can categorize and list all possible choices of a moral agent in a given situation, devoid of the detritus of historical traditions and of lived time within messy social contexts.

Logical analysis of moral decision-making provides some but not all of the most decisive elements that guide a person to the needed steps along the corridors of moral decision-making. Logic as method can show the steps, but not why one should at that moment, under these conditions, act. Premises added or premises deleted will lead to different conclusions. Enter premise a, b, and c, x is derived; delete a, b, and c and replace with d, e, and f, another conclusion, y, is derived. Combine them in one of several ways, yet other conclusions follow. This is to say that formal logical method (discrete symbolic formalization, creation of meta-language) can be ably manipulated to provide a desired and predictable result. Conclusion x is desired; conclusion y is not. This view makes logic a handmaiden to the will, at least when it comes to moral questions.

Albert Camus noted that “There is but one truly serious philosophical problem and that is suicide” (The Myth of Sisyphus). Let’s follow his lead, then, and use it as a place to start. Considering whether and under which conditions one would like to remain alive is both a perennial question with a long history of answers as well as a topical question, given all the novel ways a human being can be kept alive and supplemented. A simplified argument follows:

If I can no longer attend to myself, then I am debilitated.
If I am debilitated, then I am no longer want to live.
Therefore, if I can no longer attend to myself, then I no longer want to live.

\[ \neg A \Rightarrow D \]
\[ D \Rightarrow \neg L \]
\[ \neg A \Rightarrow \neg L \]

The logic is clear; the hypothetical syllogism admits of an easy test of validity. An analysis of the premises admits less. The first premise seems less prone to error. Of course, definitions of ‘attend to myself’ and ‘debilitated’ need to be defined. The second premise is much more likely to strain, while trying to establish the link between not being able to attend to oneself and no longer wanting to live. That is, statements such as these can only be verified by the person herself or himself. There is no universal truth to such statements; they are personalized, reflective, phrased with one person in mind. Camus overcomes all negations of life with a supreme sense of scorn. Sometimes, insufficient scorn may be available for the moment.
One could just as well phrase a similar argument, slightly altered, to produce the desired result:

If I can no longer attend to myself, I am debilitated
If I am debilitated, then I still want to live, provided I can hire someone to care for me and provided I can enjoy at least part of the day.
Therefore, if I can no longer attend to myself, I still want to live, provided that I can hire someone to care for me.

Translated to notation, the argument takes the same hypothetical syllogistic form:

\[
\neg A \supset D \\
D \supset [L \supset (C \cdot E)] \\
\neg A \supset [L \supset (C \cdot E)]
\]

In effect, we produce the argument whose conclusion appeals most to us. Logic by itself can recommend or convince, but only along certain official corridors of thought. To continue the metaphor, we can envision many doors, if you will, off these corridors; each door opened identifies variables, which when linked, will lead to several logically acceptable conclusions, when closed yet other conclusions. However, why one opens these doors in the first place is not always a matter of logic, but of preference or other motivations.

These doors could also represent more detail. For instance, \( C = \) ‘provided I can hire someone to care for me.’ What level of care? Private nurse, 24/7? Care at home by a close relative? What will she be allowed to do? Will she be able to drink a coffee and smoke a cigarette on the balcony? Will she be able to tend to flowers? Will she be able to watch and listen to the birds? What is her definition of “acceptable”

Let us consider the example of suicide further. Suicide as an established category must be defined. Many lead lives that at first appeal to the same established category as traditional notions of suicide would yield, such as bungee cord jumpers, mountain-climbers, cigarette-smokers, drivers on the Autobahn. Though many of us engage in risky behaviors, even debilitating over the long-term, as those described, they typically are not added to the list of suicides, though that would need extended argument in another paper to show why they could or could not be included. In deciding to live or die, one may open the door to statistical evidence for illumination. According to the World Health Organization, in 2000, a person committed suicide every 40 seconds or 2,160 per day or 788,400 for the year. For the same year, it was the third leading cause of death between the ages of 15-44, regardless of sex. Generally, over 90 percent of suicides occur during times of depression, resulting from loss of job, of love, or of honor (WHO).
Despite recent disturbing increases among the young and the growing understanding of correlations between drug use and depression with the choice of suicide, one finds of course that most never commit suicide. One therefore should not commit suicide, since most do not and find good reasons to continue on till the last.

Most people do not commit suicide.
Therefore, I will not commit suicide.

Though the conclusion is admirable, the evidence is very weak, since what many people do is no reliable indication of how one ought to act. Even if the statistical argument is too weak, one may open other doors in the corridor of thought that introduce other variables, either in the form of principles or more evidence which leads them to the sought-after conclusion. Here is an example:

Most people do not commit suicide.
Most find either that life is appealing or that life must be lived to its end.
If life is appealing, then I should persist. And if life must be lived to its end, then I must persist.
Therefore, I should continue to live or I must continue to live.

Translated to notation, the argument takes this form:

\[ \neg S \]
\[ A \lor M \]
\[ (A \supset P) \land (M \supset P) \]
\[ P \lor P \]

The constructive dilemma is clearly visible and easy to test for validity. The initial premise – that most people do not commit suicide - does not come into play, other than as a general statement surrounding the argument. Escaping through the proverbial horns of a dilemma, we could add another component to the disjunctive statement. Or, by taking the argument by the horns, we could disprove one of the conjuncts.

Despite the clarity of the argument, we have to contend with the effects of actual suicides. There have been spates of celebrity suicides, one of most recent hitting the Korean pop music scene with the suicide deaths of Eun-joo (March 2005), Lee, and Yuni (21 January 2007), and most recently, Jeong Da-Bin (10 February 2007) (Jackson, BBC). There is fear of copy-cat cases (Park, Chan-Kyong). Some young Korean women now apparently feel that if these celebrities were not strong enough to withstand the vagaries of career or cruelty from others, despite their beauty and talent, how could they?
Tests for validity and soundness help us see the problem more transparently, help us appreciate the relationships of variables. Yet there is a hidden foundation beneath the smooth workings of inference.

Premise
Premise <---------Unstated assumptions

Conclusion

These unstated assumptions are non-rational in nature; they may comprise multiple human emotions, social expectations, even chance. A poet’s observation, such as e.e. cummings, gives resonance to this thought when he says “love is a deeper season than reason;…” (100 selected poems, #83) The poet’s reference claims that an emotion, at least one, is ultimately more vital than reason. He does not exclude reasoning altogether, but suggests that reasoning is not as profound a justification for action as love, that at least logic abstracted from our rough social reality, cannot adequately capture or express moral issues.

Russian novelist Leo Tolstoy illustrates cummings’ point about the limits of reason, however transparent. Tolstoy’s short story “The Death of Ivan Illych” chronicles the physical decline and ultimate transformation of a middle-class St. Petersburg court attorney by the same name, as he experiences a protracted terminal illness. At one point, as the poor fellow is still able to reflect and in turn reject the course of his ailments, Illych speaks to himself:

The syllogism he had learnt from Kiezewetter’s Logic: ‘Caius is a man, men are mortal, therefore Caius is mortal,’ had always seemed to him correct as applied to Caius, but certainly not as applied to himself.

The logic is impeccable and simple. Rephrased, the argument would be

If Caius is a man, then Caius is mortal.
Caius is a man.
Therefore, Caius is mortal.

Where propositional logic fails to represent the famous syllogism, predicate logic can:

1. (x)(Cx ⊃ Mx)
2. Cx
3. Mx
There is no way out for poor Caius, nor for anyone else, for that matter. Yet Ivan Ilych revolts against this corridor of logic and the door its premises must open:

That Caius, man in the abstract, was mortal, was perfectly correct, but he was not Caius, not an abstract man, but a creature quite, quite separate from all others. He had been little Vanya, with a mamma and a papa, with Mitya and Volodya, with the toys, a coachman and a nurse, afterward with Katenka and with all the joys, griefs, and delights of childhood, boyhood, and youth. What did Caius know of the smell of that striped leather ball Vanya had been so fond of? Had Caius kissed his mother's hand like that, and did the silk of her dress rustle so for Caius? Had he rioted like that at school when the pastry was bad? Had Caius been in love like that? Could Caius preside at a session as he did? ‘Caius really was mortal, and it was right for him to die; but for me, little Vanya, Ivan Ilych, with all my thoughts and emotions, it's altogether a different matter. It cannot be that I ought to die. (The Death of Ivan Ilych, 496)

There is no doubt that Ilych’s wishes and delusions to be precluded from the inimicable conclusion will fail. Brute finalities confront all wishes and win in the end. We can see, however, the profusion of detail, the profundity of memories from one man’s life contrasted sharply with a simple valid argument. This intensely personal approach cannot be achieved through abstraction. In fact, abstraction is transparent precisely because the ruddy details have been removed, leaving the corridor of reasoning uncluttered. Sans memory, sans feelings, sans the collective weight of a person’s lived life, How many thousand thousand doors must be opened in the corridors of reasoning to illuminate or darken the decision-making process of how to act or be? Somehow, after months of denial, after having lost all his fair-weather friends, having alienated his beautiful and healthy wife, Ivan Ilych, found his answer at the touch of his young son’s hand and tears, in the last moments of his life.

Suddenly, he knew what he must do. This illumination, this sudden insight, had less to do with logic than with love. So too is it with some of our moral acts and orientations. As we walk down the corridors of reasoning, we must keep in mind that that decision process as to which doors we open to illuminate the dark are often hidden in impenetrable recesses, beyond the reach of reason.

Logic as vehicle to decision-making will take us where we want to go, dependent on our desire to do so, our willingness to act. It takes an act of will, however, to be logical. The frugal or elegant houses of thought we build to extract the essentials from the non-essentials can take on a life of their own, almost forfeiting Plato’s advice to return to the rough world of social reality. Literature
gives other insight into human reality, but it too fails for the very reasons that logic succeeds.
WORKS CITED


