SOCREAL 2013
3rd International Workshop on Philosophy and Ethics of Social Reality

25 - 27 October 2013
Hokkaido University, Sapporo, JAPAN

Abstracts

Edited by Tomoyuki Yamada

Under the Auspices of
Center for Applied Ethics and Philosophy (CAEP)
Graduate School of Letters, Hokkaido University
and
Grant-in-Aid for Scientific Research on Innovative Areas:
Prediction and Decision Making (23120002)
The Ministry of Education,
Culture, Sports, Science and Technology (MEXT), Japan

Preface

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

The first edition of SOCREAL Workshop was held on 9 - 10 March 2007, and the second edition was held on 27 - 28 March 2010. Building upon the success of SOCREAL 2007 and 2010, its third edition, SOCREAL 2013, will be held on 25 - 27 October 2013 under the Auspices of Center for Applied Ethics and Philosophy (CAEP), Graduate School of Letters, Hokkaido University, and Grant-in-Aid for Scientific Research on Innovative Areas: Prediction and Decision Making (23120002) of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

This volume includes the abstracts of the invited lectures and the accepted lectures to be given at SOCREAL 2013. Accepted lectures are selected by peer reviewing of the abstracts by the members of the program committee. We have selected 10 papers out of 24 submissions. We thank all the researchers who submitted their papers for their interest in SOCREAL 2013 and the members of program committee for their reviewing work.

WORKSHOP CO-CHAIRS
Thomas Ågotnes
Tomoyuki Yamada

Committees

PROGRAM COMMITTEE

Thomas Ågotnes (University of Bergen, Norway, and Southwest University, China)
Johan van Benthem (University of Amsterdam, The Netherlands, and Stanford University, USA)
Jose Carmo (Universidade da Madeira, Portugal)
Mamoru Kaneko (Waseda University, Japan)
Fenrong Liu (Tsinghua University, China)
Yuko Murakami (Tohoku University, Japan)
Yasuo Nakayama (Osaka University, Japan)
Mitsuhiro Okada (Keio University, Japan)
Manuel Rebuschi (Nancy University, France)
Nobuyuki Takahashi (Hokkaido University, Japan)
Allard Tamminga (Rijksuniversiteit Groningen, The Netherlands)
Tomoyuki Yamada (Hokkaido University, Japan)
Berislav Žarnić (University of Split, Croatia)

LOCAL ORGANIZING COMMITTEE

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

The Program and the Table of Contents

DAY 1: 25 OCT Friday

True Lies

Thomas Ågotnes

University of Bergen, Norway, and Southwest University, China

A lie is a statement that is false, or at least believed to be false, when it is announced. But the world after the lie is not the same as the world before the lie, so is the statement necessarily still false after the lie is announced - is the lie still a lie? This talk is about true lies. These are ”self-fulfilling” lies that become true when they are made. The analysis is based on formal modal epistemic logic. True lies are conceptually related to Moore sentences, sentences that are true but become false when they are announced, but the exact relationship between the two types of sentences is not trivial. I will also discuss impossible lies (lies that stay false when announced) as well as the relationships to successful formulas (truths that stay truths when announced) and self-refuting truths (truths that become false when announced). The talk is based on joint work with Hans van Ditmarsch (Nancy) and Yanjing Wang (Beijing).

Toward sheaf semantics for a multi-agent substructural modal logic

Yohei Fukayama[1] Prof. & Dr. Koji Nakatogawa[2] Dr. Hisashi Kitamura[3]
Department of Philosophy, Hokkaido University, Japan

Kitamura, Nakatogawa and Fukayama (2007) chose a certain substructural logic, as an initial and basic tool to analyze the two wise girls puzzle discussed in Yasugi and Oda (2002). This substructural logic is named as $\mathbf{CFL_eKD4^2}$ by Fukayama, and is based on the two systems, $\mathbf{CFL_eKD}$ and $\mathbf{CFL_eKT4}$, which are introduced in Watari, Nakatogawa and Ueno (1999). $\mathbf{CFL_eKD4^2}$ is a Classical Full Lambek with exchange rules and the axioms K, D and 4 about two modal operators. The 2007 paper contains a detailed logical analysis, due to the effort of Fukayama, of the possibility for a solution of that puzzle, by replacing the connectives in the ordinary sentential logic with the ones in a substructural logic. In this study, we will offer an overview of several semantics relevant to that study, before we spell out possible world semantics to the system in question. In particular, we will focus on a development of the semantics specified on the basis of the notion of a sheaf. Watari, Ueno and Nakatogawa (1999) supply some algebraic semantics to various substructural modal logics, and they contain the semantics to the sequent calculi $\mathbf{CFL_eKD}$ and $\mathbf{CFL_eKT4}$ close to our system[4]. As an attempt to develop some possible world semantics to the sequent calculus in substructural logics, Ono and Komori (1985) define the so-called Kripke model via some algebraic semantics. In contrast to this, Restall (2000:239-248) employs ternary relations as accessibility relations, and gives a certain type of semantics which can still be regarded as a kind of possible world semantics. These attempts are significant, but we would like to obtain a semantics which is more general than previous ones, by giving a natural extension of the semantic notions specified on the basis of the Kripke structure (which we will state precisely later).

Shehtman and Skvortsov (1990) and Awodey and Kishida (2008) describe a Kripke structure by using the notion of a sheaf over a topological boolean algebra. One can give a more general consideration to this concept, by introducing a presheaf as a set-valued contravariant functor stated in category-theoretic terms. In particular, it is important to take into account a presheaf on W, that is, a contravariant functor from the set W of all possible worlds with accessibility relations to the category of sets, in order

[1] fukayama@let.hokudai.ac.jp
[2] koji@logic.let.hokudai.ac.jp
[3] h.kitamura@airedale-xing.com
[4] The modal logic **KT4** accords with the one **S4**.

to obtain a natural connection between a Kripke structure and a sheaf. Those who construct the semantics mentioned above intend to give semantics to predicate modal logics, but one can obtain semantics of propositional modal logics by a partial simplification of the mentioned semantics (Moerdijk and van Oosten (2007:15)).

In what follows, we will give an overview of a Kripke structure, a presheaf and an interpretation of sentences by a presheaf. In this case, we presuppose that the reader is familiar with some knowledge of set theory and category theory. By the term *Kripke structure*, we mean a triple $< W, R, V >$ consisting of a set W, a binary relation R on W, the family $V = \{V_w\}_{w \in W}$ of functions V_w assigning the truth values T or F to each atomic sentence in the language of modal logics with respect to each element of W. The elements of W are called *possible worlds*. The relation R is called an *accessibility relation*. Each function V_w is called an *assignment function*. Each function V_w is extended to the assignment of the values T and F to the complex sentences consisting of the sentences σ and τ in the following way:

- $V_w(\sigma \wedge \tau) = T \Leftrightarrow V_w(\sigma) = T$ and $V_w(\tau) = T$,
- $V_w(\sigma \vee \tau) = T \Leftrightarrow V_w(\sigma) = T$ or $V_w(\tau) = T$,
- $V_w(\sigma \supset \tau) = T \Leftrightarrow$ if $V_w(\sigma) = F$ then $V_w(\tau) = T$,
- $V_w(\neg \sigma) = T \Leftrightarrow V_w(\sigma) = F$,
- $V_w(\Box \sigma) = T \Leftrightarrow$ if wRv , then $V_v(\sigma) = T$, for any $v \in W$.

Consider the case in which we have a Kripke structure $M = < W, R, V >$ and a sentence A. M is called a *model* of A if $V_w(A)=T$ for each w in W. Based on this notion, we can define the relation of logical implication and the notion of validity.

A binary relation on a set is called a *preorder* if it is reflexive and transitive. We can regard a set W with a preorder R as a category. An object of the category is an element of W, and an arrow is a pair $<a, b>$ of elements of W such that aRb holds. The composition is defined in terms of the transitivity of R, and the existence of identity arrows is shown by the reflexivity of R.

Let **C** be a small category, i.e., neither the class of its objects nor the class of its arrows is a proper class. A *presheaf* (of sets) on **C** is a contravariant functor from a small category to the category of sets. The presheaves on **C** constitute a category by taking a natural transformation between them as its arrow. Moreover, this category satisfies the axioms of elementary topos. Many studies, including Mac Lane and Moerdijk (1992), Goldblatt (1984/2006) and Bell (1988/2008), deal with the relation between topos and logic. In what follows, we describe an overview Moerdijk and van Oosten (2007) has given to the semantics on a sentential logic under the notion of presheaf. The interpretation of sentences by the notion of presheaf, which we will in-

troduce below, relies on many points made in Moerdijk and van Oosten (2007). That is because they often provide concise notions.

The notion of Kripke structure and the one of presheaf are connected via the *Yoneda embedding functor*. It is a functor form a set *W* with a preorder *R* as a category to the category of presheaves on *W*, which assigns to each element *w* of *W* a contravariant functor Hom(-, *w*) from *W* to the category of sets (i.e., it is a presheaf on *W*.) Hom(-, *w*) assigns to each element *v* of *W* the set of arrows from *v* to *w*. We can identify it with the set $\downarrow$(w) of the elements *v* of *W* satisfying *vRw*, which is named the *down set* of *w*.

Consider the set of arrows whose codomain is *w* with the property that this set is closed under the composition from the right. This is called a *cosieve* on *w*. In particular, the set of arrows whose codomain are *w* are called the *maximal cosieve* on *w*. Consider a functor Ω assigning to each element *w* of *W* the set of all cosieves on *w*. Further consider a family $t = \{t_w\}_{w \in W}$. Each t_w is a function from 1(w) to Ω(w), where 1 is the terminal object in the category of presheaves on *W* and 1(*w*) is the value of 1 as a functor at *w*. For each *w* in *W*, 1(*w*) is a singleton, so we write it as {*}. t_w assigns to * the maximal cosieve on *w*. Then *t* is a natural transformation from 1 to Ω. Ω and *t* together constitute the subobject classifier in the category of presheaves on *W*. Moreover, a one-to-one correspondence exists between the set of all downward-closed subsets of $\downarrow$(*w*) (If *y* belongs to it and *xRy* holds, then *x* also belongs to it.) and Ω(*w*) (Moerdijk and van Oosten (2007:6)).

An atomic sentence is interpreted as a subobject of 1 in the presheaves on W. The interpretation of the atomic sentence p is written as [p]. In addition, the arrow from 1 to Ω classifying [p] is written as {p}, whose existence is shown by using the property of the subobject classifier. Since {p} is a natural transformation, it has the function $\{p\}_w$ from {*} to Ω(*w*) as one of its components. Under these notion, the situation in which the atomic sentence p is true in the world *w* is defined by $id_w \in \{p\}_w(*)$, where id_w is the identity arrow on *w* in *W*. Moreover, we obtain the counterpart of the evaluation of the mentioned complex sentences as a theorem by adding the interpretations of logical connectives and a modal operator.

The notion of (pre-)sheaf has been widely used in algebraic geometry. In mathematics, a sheaf of modules or of rings rather than a sheaf of sets is employed in order to obtain from algebraic structures a topological space relevant to it. Our approach is under Kripke structure and sheaf, and it will generate a geometry-relevant understanding concerning the perpetually changing knowledge state of agents[5].

[5] In July in 2011, a symposium with the title “Set within Geometry” was held in Nancy in France

References

Awodey, S. & Kishida, K. (2008). Topology and modality: The topological interpretation of first-order modal logic. *The Review of Symbolic Logic*, **1**, 146-166.

Bell, J. L. (2008). *Toposes and local set theories: An introduction*. New York: Dover. (Original work published 1988)

Goldblatt, R. (2006). *Topoi: The categorial analysis of logic.* New York: Dover. (Original work published 1984)

Kitamura, H, Nakatogawa, K., & Fukayama, Y. (2007). Substructuralized modal logics applied to the two wise girls puzzle. In *SOCREAL 2007: International workshop on philosophy and ethics of social reality 2007* (pp. 40-53). Graduate Program in Applied Ethics (GPAE), Graduate School of Letters, Hokkaido University. Retrieved from http://hdl.handle.net/2115/29932

Mac Lane, S. & Moerdijk, I. (1992). *Sheaves in geometry and logic: A first introduction to topos theory.* New York: Springer.

Moerdijk, I. & van Oosten, J. (2007). Topos theory. Retrieved from http://www.staff.science.uu.nl/~ooste110/syllabi/toposmoeder.pdf

Ono, H., & Komori, K. (1985). Logics without the contraction rule. *The Journal of Symbolic Logic*, **50**, 169-201.

Restall, G. (2000). *An introduction to substructural logics.* London: Routledge.

Shehtman, V. & Skvortsov, D. (1990). Semantics of non-classical first-order predicate logics. In P. P. Petkov (ed.), *Mathematical logic* (pp. 105-116). New York: Plenum Press.

Watari, O., Ueno, T., & Nakatogawa, K. (1999). Sequent systems for classical and intuitionistic substructural modal logics. In R. Downey, D. Decheng, S. P. Tung, Y. H.

---

as a satellite meeting in an international conference concerning scientific methodology that is supposed to be given once every four years. F. W. Lawvere gave an invited lecture at the symposium. The leading idea of this satellite meeting is expressed in the question: "how far the subject matter of set theory can be viewed as art of geometry." Most of the participants are concerned about the issue: how various mathematical theories developed from 19 century onward including set theory and topology is to be located in a context of historical development of geometry. The main topic of that meeting seemed to be how it is possible to understand a historical development of the subject matter of set theory within a much larger and major historical context of the geometry from 19 century onward. The beginning of this historical context is Gauss's Theorema Egregium, and it has opened the new field of geometry, later called "manifold" in a broad sense. Not only Riemann and Einstein, but also E. Cartan, Grothendieck and F. W. Lawvere can be regarded as mathematicians who made contributions to this historical context of the development of "manifold." The main concern of the participants was the methodological and philosophical considerations of the historical development of mathematics. See below:
http://www.archmathsci.org/conferences-and-workshops/symposium-sets-within-geometry-nancy-france-26-29-july-2011-2/

Qiu, & M. Yasugi (Eds.), *Proceedings of the 7th & 8th Asian Logic Conferences* (pp. 423-442). Singapore: World Scientific.

Yasugi, M., & Oda, S. (2002). A note on the wise girls puzzle. *Economic Theory,* **19**, 145-156.

Logical normativity in communication ethics

Berislav Žarnić

The language as a normative source can be viewed from two perspectives. On the one side, there are regulative requirements like sincerity and trust. In particular the explicit pronunciation of the requirement of non-deceptive use of language is likely to be found in any system of general ethics. On the other side, there are more specific, discourse dependent, constitutive requirements of coherence that arise from the logical nature of language. E.g. the two speech acts, one of which is insincere for not expressing speaker's intentional state while the other is incoherent for refusing entailments of speaker's previous discourse, both can have the same deontic status of being forbidden. Nevertheless the origin of their deontic status is not the same: the source of regulative requirements, such as non-deceptiveness, comes from the purpose of language to enable reaching understanding while the source of constitutive requirements, such as coherence, lies in the nature of language—in its logical structure. The difference of origins is revealed by the effects of requirement violation. The violation of the first type of language-use requirements changes the character of communication, e.g. from cooperative to non-cooperative communication. The violation of the second type of language use requirements destroys communication: language-mediated interaction ceases to be possible.

For the description of the diverse character of language requirements one needs: (i) a discriminative ontology suitable for (ii) a comprehensive theory on relations between language and types of worlds together with (iii) an expressively rich formal language adequate for the theory. A discriminative ontology has been given in [3] and it can be briefly summarized as in Table 1.

Objective world	Social world	Subjective world
physical facts	norms	mental facts
external world		*internal world*

Table 1: Habermas [3] ontology.

It has been claimed [5] that there are four main language-world relations: (i) with respect to objective world there is the relation of representation, (ii) with respect to subjective worlds there is the relation of expression of the speaker's intentional state and the relation of alteration of intentional state of the hearer, (iii) with respect to social world there is the relation of assigning deontic status to acts, e.g. by norm promulgation or by requesting, and the relation of modification of linguistic commitments by language use. The relations are depicted in Figure 1

The formal language of dynamic epistemic logic by van Benthem and others, systematically investigated in [4], can be applied for the description of diversity of language-based relations. For the purpose of defining the syntax of a

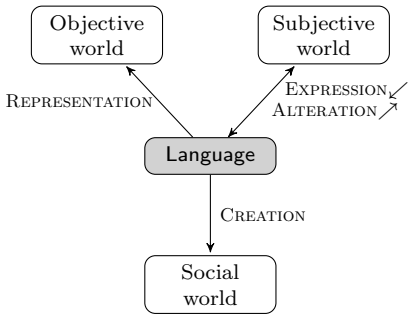

Figure 1: Diversity of language relations.

simplified formal language let us adopt the following notational conventions: $\mathsf{i}, \mathsf{j}, \ldots$ for actors from a communication group; $p, q, \ldots$ for propositional letters; $\circledast_\mathsf{i} \in \{\mathrm{B}_\mathsf{i}, \mathrm{D}_\mathsf{i}\}$ for generic intentionality operator that stands in place of 'i believes that ...' and 'i desires that ...'; $\mathsf{i}\ \mathrm{stit}\colon$ for modal operator of action 'i sees to it that ...'; $\odot_\mathsf{i} \in \{\mathbf{P}_\mathsf{i}, \mathbf{F}_\mathsf{i}, \mathbf{O}_\mathsf{i}\}$ for generic deontic operator that stands in place of 'it is permitted for i that ...', 'it is forbidden for i that ...' and 'it is obligatory for i that ...'.

$$
\begin{array}{ll}
\mathcal{L}_{\text{reality}} & \varphi ::= p \mid \circledast_\mathsf{i}\varphi \mid \mathsf{i}\ \mathrm{stit}\colon \varphi \mid \odot_\mathsf{i}\varphi \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \chi \\
\mathcal{L}_{\text{utterance}} & \xi ::= !\mathsf{i}\ \mathrm{stit}\colon \varphi \mid \cdot\varphi \mid !\mathsf{i}\ \mathrm{stit}\colon \varphi \to \cdot\varphi \\
\mathcal{L}_{\text{locution}} & \chi ::= \mathsf{i}\colon \underline{\xi} \\
\mathcal{L}_{\text{effect}} & \epsilon ::= \varphi \mid [\chi]\epsilon \mid \neg\epsilon \mid (\epsilon \wedge \epsilon)
\end{array}
$$

The typical sentential form $[\chi]\varphi \in \mathcal{L}_{\text{effect}}$ describes an effect of a locution χ in terms of description $\varphi \in \mathcal{L}_{\text{reality}}$ of resulting states in subjective worlds and in the social world. For example, the social effect of imperative locution can be described by the formula

$$
[\mathsf{i}\colon\underline{!\mathsf{j}\ \mathrm{stit}\colon \varphi}]\mathbf{O}_\mathsf{i}(\mathsf{j}\ \mathrm{stit}\colon \varphi \vee \mathsf{j}\colon\underline{\neg\mathsf{j}\ \mathrm{stit}\colon \varphi}) \tag{1}
$$

which states that when i asks of j to see to it that φ an obligation is created for j either to perform the requested act or to announce refusal. The expressive relation can be captured by the formula which shows that in any case after i utters imperative to the effect that j sees to it that φ no new information will be added if i further says that she desires that j sees to it that φ:

$$
[\mathsf{i}\colon\underline{!\mathsf{j}\ \mathrm{stit}\colon \varphi}]\psi \leftrightarrow [\mathsf{i}\colon\underline{!\mathsf{j}\ \mathrm{stit}\colon \varphi}][\mathsf{i}\colon\underline{\cdot\mathrm{D}_\mathsf{i}\ \mathsf{j}\ \mathrm{stit}\colon \varphi}]\psi \quad \text{for any} \quad \psi \in \mathcal{L}_{\text{reality}} \tag{2}
$$

From a logical point of view, the generation of the speaker's linguistic commitments by her own discourse (sequence of locutions) is a most interesting phenomenon in communication ethics. Linguistic commitments of a monologic discourse mirror logical relations between performed and unperformed utterances. We propose the following definition: actor i is committed to ξ_n after i's discourse $\xi_0 \ldots \xi_{n-1}$ iff $([\mathsf{i}\colon\underline{\xi_0}] \ldots [\mathsf{i}\colon\underline{\xi_{n-1}}]\mathbf{P}_\mathsf{i}\ \mathsf{i}\colon\underline{\xi_n}$, and $[\mathsf{i}\colon\underline{\xi_0}] \ldots [\mathsf{i}\colon\underline{\xi_{n-1}}]\mathbf{F}_\mathsf{i}\ \mathsf{i}\colon\underline{\xi'}$ for all utterances ξ' such that ξ_n and ξ' are incompatible, and $[\mathsf{i}\colon\underline{\xi_0}] \ldots [\mathsf{i}\colon\underline{\xi_{n-1}}]\mathbf{O}_\mathsf{i}(\chi \to \mathsf{i}\colon\underline{\xi_n})$ for some locution χ).

The distinction between regulative and constitutive requirements in communication ethics can be expressed in the formal language $\mathcal{L}_{\text{effect}}$ as a difference in

their effects. E.g. Grice's maxim of quality [2] *Don't say what you believe to be false* is a regulative requirement which translates to (3) and whose violation is communicatively coherent (4).

$$\mathrm{B_i}\neg\varphi \to \mathbf{F}_{\mathsf{i}}\ \mathsf{i}{:}\underline{\cdot\varphi} \tag{3}$$

$$\mathrm{B_i}\neg\varphi \wedge \neg[\mathsf{i}{:}\underline{\cdot\varphi}]\bot \tag{4}$$

On the other hand, the deontic reading of Moore-type sentence *Don't deny the sincerity conditions of your speech-acts* yields a constitutive requirement. Formula (5) shows one among many linguistic commitments created by the assertive locution. Proposition (6) shows communicative incoherence of the denial of sincerity conditions of an assertion. Proposition (7) gives a general form for any locution type where function Ψ delivers sincerity conditions for an utterance.

$$[\mathsf{i}{:}\underline{\cdot\varphi}]\mathbf{F}_{\mathsf{i}}\ \mathsf{i}{:}\underline{\cdot\mathrm{B_i}\neg\varphi} \tag{5}$$

$$[\mathsf{i}{:}\underline{\cdot\varphi}][\mathsf{i}{:}\underline{\cdot\mathrm{B_i}\neg\varphi}]\bot \tag{6}$$

$$\text{If } \circledast_{\mathsf{i}}\,\varphi \in \Psi(\xi), \text{ then } [\mathsf{i}{:}\underline{\xi}]\ \mathbf{F}_{\mathsf{i}}\ \mathsf{i}{:}\underline{\cdot\neg\circledast_{\mathsf{i}}\,\varphi}. \tag{7}$$

If Broome's [1] theory of requirements is applied to communication ethics, language turns out to be a normative source. If viewed in this manner, it exhibits an unique trait. For other normative sources it is possible that their codes (sets of requirements) violate the logic of language in which they are expressed. This kind of imperfection is not possible in the code of language-use. Logical requirements or linguistic commitments are constitutive requirements and they mirror the logical structure of language. Therefore, language user has no option but to satisfy her linguistic commitments. We either comply with the logical requirements of communication ethics or we fail in our attempt to use the language.

References

[1] John Broome. *Rationality Through Reasoning.* The Blackwell / Brown Lectures in Philosophy. Wiley-Blackwell, 2013.

[2] Paul Grice. *Studies in the Way of Words.* Harvard University Press, 1991.

[3] Jürgen Habermas. *The Theory of Communicative Action: Reason and the Rationalization of Society.* Beacon Press, Boston, 1984 [1981].

[4] Johan van Benthem. *Logical Dynamics of Information and Interaction.* Cambridge University Press, 2011.

[5] Berislav Žarnić. Logical root of linguistic commitments. In Anna Brożek, Jacek Jadacki, and Berislav Žarnić, editors, *Theory of Imperatives from Different Points of View, vol. 2*, pp. 7–25. Wydawnictwo Naukowe Semper, Warsaw, 2011.

Why there was no success in resolving Jörgensen's dilemma

Gleb V. Karpov
Saint Petersburg State University, Russia
glebsight@gmail.com

In the period from 40th to 60th of XX century a series of informal ways of how to solve Jörgensen's dilemma were introduced. (The situation when nondescriptive sentences are premises of inferences that people put into practice regardless of the fact that there is no good logical foundation for such inferences was originally discovered by Jörgen Jörgensen and later was named Jörgensen's dilemma [Jörgensen, 1937]. The approaches suggested clearly fall into two trends according to the way they treated the imperatives – kind of nondescriptive sentences associated in most cases with commands.

One of them – let me call it *parallelism* – reduces imperatives to declarative sentences, the other one interprets imperatives and imperative inference in terms, which are different from truth and false.

Parallelism tendency includes *moderate* and *marginal* branches. Its *marginal* branch represented by R. Hare papers [Hare, 1949] denies that there is a need to construct a special theory in order to explain the idea of imperative inference. The background of such denial is the belief that logical properties of imperatives and declaratives are identical. Followers of *moderate* branch of parallelism (W. Dubislav, O. Weinberger) argued that while declaratives can be treated as true or false in a quite natural manner, imperatives can not. However, imperatives can be evaluated as successfully or not successfully performed utterances. One of the suggested attempts to overcome the gap between truth and success, and thus between truthfunctional logic and that of imperatives, is to do so by means of "Dubislav convention" [Dubislav, 1938], according to which the logical value of each successfully performed imperative is seen as the corresponding truth value of the declarative sentence. The definition of imperative inference is then the following: imperative !A implies imperative !B if the corresponding declarative sentence A implies the declarative sentence B.

This definition is developed further in the *logic of satisfaction* for imperatives suggested by A. Hofstadter and J. C. C. McKinsey [Hofstadter, McKinsey, 1939] who introduce laws for imperatives as parallel counterparts to basic postulates of propositional logic (primarily to the rules of introduction and elimination of propositional connectives). In most cases this parallelism gives disappointing results: on one hand, in many aspects one arrives at quite trivial results, and gets a number of obvious paradoxes, on the other. This makes the logic of satisfaction one of the main objects for criticism from those who hold the idea of creating a special logic for imperatives which would be non-isomorphic to truth-functional logical theories.

The second trend in solving Jörgensen's dilemma is represented by two logical projects: one of them is the unification of the *logic of subjective validity* and the *logic of satisfaction*; the other is the logic of *satisfactoriness*. The unified logic of subjective

validity and satisfaction has been forwarded by A. Ross [Ross, 1944]. Ross assumes that in accordance with some conventional procedures imperatives indicate a certain speaker's psychological state which is specific to performing them and he thinks such psychological state can function as a basis for defining of imperative inference. Then it is possible to draw an imperative !B from an imperative !A if it is not the case that speaker can perform subjectively valid premise-imperative !A without performing subjectively valid consequence-imperative !B. However, Ross succeeded in developing only one inference rule for negation, and even this rule rests on our linguistic intuition and not on deductive reasons.

In his logic of satisfactoriness A. Kenny [Kenny, 1966] suggests an idea to take the notions of speaker's intentions and goals succession as the correlate for imperative inference definition. His definition says that an imperative !B can be inferred from an imperative !A if the case when !A satisfies some set of speaker's intentions and goals , is also the case the imperative !B satisfies the same set relative to the same speaker. An imperative is said to satisfy some set of speaker's intentions if its successful performance makes true the (set of) sentences describing these intentions. Such somewhat sophisticated definitions introduced by Kenny imply the consequence that the inference rules of logic of satisfactoriness in fact mirror those of classical logic: "!A implies !B" is the case if "B implies A" is true. This approach eliminates Ross's paradox, but leads to other paradoxes, e.g. from imperative !A one can infer imperative !(A$\wedge$B) in accordance with the satisfactoriness inference rules.

None of imperative logic projects does provide considerable results. I believe that the reason for this is wrong way of attacking Jörgensen's dilemma rather than the idea that in fact there are no imperative inferences. Despite all the differences among the inference definitions in the imperative logics advanced so far, they have much in common. I regard some of these commonplaces to be ***methodological mistakes*** which are the reasons of failure of many earlier attempts to resolve Jörgensen's dilemma.

First of all I should mention that every imperative logic project in the period rests upon the analogy between classical and imperative logic. The essence of this analogy is the questionable idea that standard truthfunctional logical connectives constitute sound molecular units out of atomic ones with regards to imperatives in the same way as they do with regards to declaratives. However, since there is no full correspondence between logical connectives and words of natural languages we use to link imperatives into reasonings, there is no reliable foundation that could enable us to speak about molecular imperatives like !A$\wedge$!B, !A$\vee$!B, !A$\rightarrow$!B.

Another wrong idea underlying the failure to give a sound definition of imperative inference goes hand in hand with the previous one. The idea that it is possible to construct molecular imperatives out of atomic ones like !A, !B, etc. with the help of standard logical connectives, does not provide us with a sound basis for further formalization. Looking back to the attempts made so far I conclude that in the earlier versions of imperative logic no such basis have been advanced. This is also a reason why

it is extremely hard to find the examples in natural language for such pairs of formalisms as:

!A∧!B and !(A∧B),

!A∨!B and !(A∨B),

!A→!B, !(A→B) and A→!B,

which would be manifest of the formal aspects of the distinction between the first and the second member of the pairs

The third methodological mistake (which is also mentioned by Jörg Hansen [Hansen, 2008]) is that no clear account of logical status of conclusion in imperative inference have been formulated so far. There is a bulk of questions concerning both the nature and the identification of it. What does it mean in logical terms when we say that one imperative implies another? Who is the author of such an implication and should the authorship for imperatives in fact be considered? And if we consider speaker to be the one who puts into practice implication of that kind, how then two different operations of commanding and making conclusion are combined in one action? These questions still have no answer and therefore it is very difficult to advance in imperative inference.

There is a ***possibility to overcome all methodological mistakes*** mentioned here, that originates in changing the way we look at imperatives. ***First of all***, instead of dividing each imperative into propositional content and something that puts this propositional content into action ("*dictor*" in early Hare's terminology or "*force*" in Frege's terminology) it is more productive to consider imperative in *monistic* (*atomic*) non-dividable way, because this prevents us from those paradoxes which arise due to "Dubislav convention". Such monistic approach is based on Wittgenstein's idea that we picture facts to ourselves in accordance with the conviction that things are related to one another in the same way as the elements of the picture [Wittgenstein, 1974, prop. 2.1 and 2.15].

Dualistic approach says that two sentences can have the same descriptive part and differ only in their dictors. But when we try to explicate the descriptive part from that of dictor's part in a pair of sentences where one sentence describes a fact but the other prescribes an action as, e.g. in

"Peter plays the piano" (1)

And

"Peter, play the piano, please!" (2)

we see that even the subjects of corresponding propositions are different: "the man who plays the piano" (the subject of the first proposition, as given in the sentence (1) is not equivalent to the subject of the second proposition, as given in the sentence (2), because in the latter we have "the man, who's playing piano I want to hear". The intensionals of these subjects is not the same and if it is so we cannot establish the corresponding relation between the parts of imperatives and the parts of indicatives, because their parts differ. And if so it is much better to consider imperatives as having no parts at all.

Thus if we refrain from dealing with imperatives interpreted *dualistically*, we immediately get a situation that is free from paradoxes that have occurred in the domains of both isomorphic and non-isomorphic approaches and vague formalisms discussed above. E.g. the combination of two imperatives !A and !B, even if their imperative moods are perfectly the same, gives only an aggregation !A*!B, and never gives something that can be expressed like !(A∧B), because it is impossible to think about co-called propositional content "A" that can be separated from imperative operator "!" and still be the *same* propositional content as in "!A". Even if I say that it is raining, I use special assertive mood and this fact is dropped out of a descriptive sentence and it cannot be expressed as simply "A".

The next step is to regard imperatives as actions which are resistant to combinations into complexes by means of logical connectives. Anyone who observes his personal social linguistic practice can easily notice that imperatives, if they are treated as actions can be: performed and executed or not executed, supported by other imperatives or conflict with them, set out in chains or be independent one from another; they can be repeated, be equal, force somebody to some specific act or provide a choice to its addressee; finally they can stipulate some other acts or be stipulated by some other acts.

Thanks to these two steps we can leave fruitless analogy between classical and imperative logic and consider: something that is called by the followers of dualistic approach an "imperative negation" of !A simply as a denial to do A; "conjunction of two imperatives" as a sequential performance of !A and !B; "disjunction of two imperatives" as a choice that is given to the addressee; "implication from one imperative to another" as an imperative enthymeme – the way to say something by means of saying something different; and finally we can consider "imperative equivalence" of dualistic approach as the fact of possible interchangeability of !A and !B.

All these considerations lead us to formulating some rules which govern the usage of imperatives connected with each other with the help of *language* and not logical connectives.

The rule for the sequence of imperatives: when performing several different imperatives take care that the addressee knows the order you wish him to execute these imperatives.

The rule for the situation when one imperative stipulates the other: always try to perform all your imperatives explicitly; if you perform implicit imperative !B by means of imperative !A which is explicit, take care that the addressee is able to proceed correctly from !A to !B.

The rule for interchangeability of imperatives: if you mean !B while performing !A (or vice versa) take care that your addressee is informed as much as you that the communicative function of !A and that !B in that context is the same.

Hence we can draw a conclusions from imperative premises not in accordance with some logical inference rules, governing the relations between some properties of imperatives that are different from truthfulness, but in accordance with some social conventional procedure, the nature and the properties of which needs to be investigated.

References

[Hare, 1949] Hare R. M., Imperative Sentences, Mind, New Series, Vol. 58, No. 229 (Jan., 1949), pp. 21-39.

[Dubislav, 1938] Dubislav, W., Zur Unbegrundbarkeit der Forderungssatze Theoria, 3, 1938, pp. 330-342.

[Hofstadter, McKinsey, 1939] Hofstadter, A. McKinsey, J. C. C. On the Logic of Imperatives, Philosophy of Science, Vol. 6, No. 4 (Oct., 1939), pp. 446-457.

[Kenny, 1966] Kenny, A. J. Practical Inference, Analysis, Vol. 26, No. 3 (Jan., 1966), pp. 65-75.

[Ross, 1944] Ross, A. Imperatives and Logic, Philosophy of Science, Vol. 11, No. 1 (Jan., 1944), pp. 30-46.

[Hansen, 2008] Hansen, J. Is there a Logic of Imperatives? Stable URL: http://icr.uni.lu/leonvandertorre/papers/esslli08.pdf.

[Wittgenstein, 1974] Wittgenstein, L. Tractatus Logico-Philosophicus. English translation by D. F. Pears and B. F. McGuinness.

[Jörgensen, 1937] Jörgensen, J., Imperatives and Logic, Erkenntnis 7 (1937/8), pp. 288-296.

Norm dynamics: institutional facts, social rules and practice

Alessio Antonini, Cecilia Blengino, Guido Boella, Leendert van der Torre

1 Introduction

Norms are created by social agents in a complex environment. Norms are adopted, implemented and used according to the result of the dialectic between the "normative message" of norms and the receiving contexts. Law scholars refer with "social rules" to what was actually adopted by the population and with "legal practice" to the pragmatic rules adopted by institutional agents. Local cultures, communities, language affect are components of norms context affecting the very meaning of norms. Further, norms are vaguer than language, there is a calculated incompleteness or incomprehensibility of the legal text [1].

Assuming a perspective focused on communication, any rule appears substantially as a message: a communication act with special "prescriptive" properties [2][p. 159]. As messages, norms iterate within a discursive space. The interaction does not take place in an empty space, but there is always the presence of "other speakers" [2][p. 163]. This dialectic affects norms, up to change also substantially their original meaning. Characters and interests of the audience inevitably influence the message itself [3]: the same legal discourse, while inverting the rules, it appears different depending on the social group to which messages are directed ("law", "legal discourse", "doctrinal discourse", etc.).

Alessio Antonini
Dipartimento di Informatica, Universit degli Studi di Torino, e-mail: antonini@di.unito.it

Guido Boella
Dipartimento di Informatica, Universit degli Studi di Torino, e-mail: boella@di.unito.it

Cecilia Blengino
Dipartimento di Giurisprudenza, Universit degli Studi di Torino,
e-mail: ceciliapiera.blengino@unito.it

Leendert van der Torre
CSC, University of Luxembourg, e-mail: leon.vandertorre@unilu.lu

In this contribution we take in account the norm dynamics exposing its structure as a multi-phase, continuous, complex, cyclic social process. In particular, we answer to the following questions:

1. how to expose the components of norm dynamics?
2. how to generalize norms life-cycle?
3. how to deal with the different agents' roles in norm dynamics?

Starting from a representation of norms as a network of social objects [4], we introduce the role of agents in norms creation [5] and the role of agents in the effectiveness of norms [6]. We connect agents to social objects through actions, powers and obligations (roles), as described by Broersen et al. (2001) [7] or Boella and van der Torre [8]. This leads to a representation of norms with agents' roles and phases of dynamics of norms. Using real cases and representations, we expose a scenario much closer to real life and a general outline of an equivalent model for norms. We aim to an incremental representation of norms to collect the traces of agents actions and make the hidden relations emerge.

The rest of this contribution is structure as follows. In section 2 we show two real scenarios to introduce a representation for norm dynamics. In section 3 we address our methodology.

2 Anatomy of norm dynamics

Now we consider norm creation in the social delegation cycle works as defined by Boella and van der Torre [5]. The social delegation cycle starts with a set of individual agent desires and goals. The first step leads to group or social goals via merging, the second step leads to norms and sanctions via planning, and the crucial third step of acceptance checks that the norms lead to satisfaction of the individual desires and goals the cycle started with. This is meant to be a logical model, that is, these logical relations exist between individual and group goals, and norms. It is not a protocol, that is, we do not have to go through these steps one by one. For example, in determining the group goals, the fact whether there is a norm which can implement it, may play a role in its adoption. How do we go from these abstract relations to a more refined model of norms? How do we represent the relations between norms, other objects and agents?

A norm n changes over time, for instance following the meaning shift of legal concepts or changes in legal texts [4]. An ontological shift is caused by agents' actions. Roles define agents involvement in society. Also, an agent can hold different roles, so they follow a personal mediation among all their roles, norms and expectations. In figure 1 we represent with nodes G_n^0,G_n^1 and G_n^2 the status of a norm graph G_n at time t_0, t_1 and t_2. Each shift is consequence of different set of agents, grouped by roles.

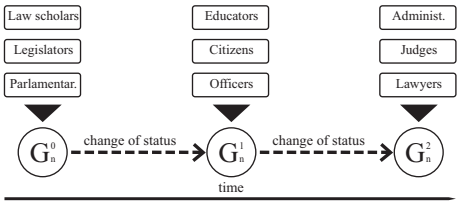

Fig. 1: Roles can change aspect of a norm *n* over time.

2.1 Role interpretation and local cultures in norms

Now, we'll discuss to limit cases to expose the general phases of norm dynamics and the features of a general representation of it.

Example 1 (1st part). Italian Constitution rules the Prosecution with two fundamental principles: the "mandatory prosecution" and the "reasonable duration of the action". A norm stets the Prosecutors to be a magistrates and a managers, they must pursue any crime and solve them quickly with their limited resources; this obliges them to order their work. These norms underline both the relevance of leadership styles and the relevance of the "local legal cultures" [9] on the output of each Prosecutor office. If a formalistic legal culture conducts not to comply to the mandatory prosecution norm, by the other side a managerial legal culture requires the Prosecutor to choose which crime give priority.

1) the way they deal with role conflicts is not a priori but a set of small choices they make and revision time to time.

Example 2 (2nd part). Prosecution is strongly dependent on the organization of each local judicial office. Each prosecutor office is linked by specific relationships with institutional actors (local court and advocacy) and with not institutional actors (politicians, social services, health services, trade associations, victim association, neighbourhood committees, etc).

2) We should expect behaviors that are agents' interpretations of roles.

Example 3 (3rd part). The analysis of the prosecutor offices of Turin and Bari shows how the ways every Prosecutor connects external inputs to the outputs of the Office create different organizational choices and, finally, different judiciary policies [10]. Inevitably prosecutors decide their job schedule taking in account what they think is more important to their local community. Prosecution is largely influenced by not juridical inputs, as local claim for justice.

3) The locality of norms arise from local cultures and local procedures that are implemented to deal with the close environment.

Example 4 (4th part). The influence of the perception about the social alarm produced by certain crimes is evident in the Prosecutors decision to create specialized

groups for specific types of crime. The perception of the social alarm also influences the choices about how much money and how many people destine to each activity and each specialized group. For instance, the fear for organized crime lead the prosecutor office of Bari to destine two- thirds of its magistrates to pursue this crime and consider less important crimes like thefts and muggings, while the office of Turin created a specific work group aimed to combat street crime, mostly to meet the demands of citizens committees.

4) The locality of norm transposition drive to different local dynamics.

Example 5 (5th part). Crime perception influences the definition of the internal proceedings with the construction - in each office - of different ways to treat "notitiae criminis" of different type. If a crime is considered more serious than others it will be assigned to a "specialized group" and it will be pursued carefully. Considering the limits of court resources and the prescription times of crimes, in each office some crimes will not even considered despite the reporting of authorities. Furthermore, to comply to the norm that oblige to speed trials, the office of Turin has defined automatic procedures for the "notitiae criminis" that Prosecutor believes are easily solved. In this way, prosecutors sometimes involuntarily pursue with greater hardness less serious crimes.

5) Local dynamics result on local norms.

In those examples agents' choices are not arbitrary, they mediate between their roles conflicts (magistrates and managers) and society requests (social expectations). Agents cannot just follow rules but interpret the underlying general principles.

2.2 Emerging norms

In the following example we show also that legal texts are part of complex dynamics.

Example 6. Currently, citizens' right to withdraw the cure is not recognized in Italy. It is missing a norm that defines the role of the anticipatory declaration, "living will", about what actions should be taken for their health if they are no longer able to make decisions due to illness or incapacity. In Italy, the only normative source is the article 32 of the Constitution that states the non-mandatory of medical treatment. Despite the lack of norms, in many judgments are recognized the validity of living will.

6) Norms are not just created, they can also spontaneously emerge from the current system.

In example 6 shows a norm that emerges from the normative system and the social expectations: living wills are normed even without a legal text that defines what they are and how they should be collected.

2.3 Structure of norm dynamics

Looking at agents' actions, a norm is the result of a continuous process that redefines it. In figure 2, we show three main phases, each of them present their own dynamics:

a. only a set of roles can directly participate to it,
b. local culture or domain knowledge required,
c. scope of agents' actions (local interest for their traces).

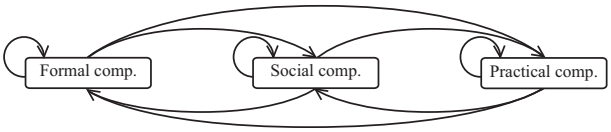

Fig. 2: Institutional-creation process of norms, transposition of norms in society and use of norms in law practice and regulations.

However, each component is strongly connected to others making them a single system. Agents bridge norm phases:

a. agents can play different roles (they can act in different phases),
b. agents do not always act according to their roles,
c. agents have access to almost any information,
d. agents share (common) knowledge and beliefs despite their role,
e. relations among agents are not bounded to their role.

In figure 3 we represent an instance of norm n with its three components G_{n_0}, G_{n_1} and G_{n_2}. Each component is made by an explicit structure, like T_{law} legal text and C_{law} law concepts, but there are also many connection between them (dashed lines) emerging from agents actions.

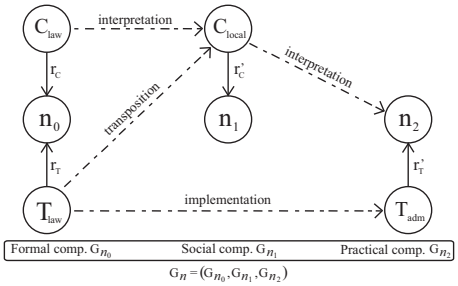

Fig. 3: Explicit and emerging structure of a norm n.

Figure 3 represent a norm G_{n_0} made by a legal text and legal concepts[1]. G_{n_1} represent what is called social norm and G_{n_2} legal practice. We indicate with dashed

[1] It is an extreme simplification of norm content, for further discussion see "Beyond the rules representation of norms: norms as social objects" [4].

edges some emerging relations that we expect from agents'actions (not limited by their roles). We labeled them with the name of the social phenomenon: *interpretation of norms and social expectations, implementations of norms in local procedures and regulations, and norm transposition in society*.

3 Conclusions

We introduce a complex representation of norms that takes in account the institutional, social and practical aspects of norms. Our representation is based on a methodology for building social objects from agents' actions. We extended an early representation of norms with phases and agents' roles involved in norm dynamics, figure 1. Using some examples, we exposed the characteristic of norms dynamics, figure 2. After, we showed in figure 3 an instance of representation that takes in account the complexity of norm dynamics.

We start to expose important features of norms dynamics and requirements for norms models.

1. *How to represent norms taking into account the institutional, social and practical aspects of norms?* We show how to represent norms is necessary to include the representations of agents' actions: norm creation, norm use and norm practice, like in figure 3.
2. *How to generalize the dynamics of norms with a process that involves agents roles?* We exposed the multiple level of norm dynamics: phase dynamics (impact of agents' actions and beliefs on norms), figure 2, and inter-phase dynamics.
3. *hot to represent the role of agents in norm dynamics?* Finally, we showed norm dynamics phases dependencies: agents' groups, agents' roles and local cultures. In particular, each context give birth to a new phase with its own local dynamics.

Commonly, norms are represented as rules. That solution has great advantages and few disadvantages related to the readability of big set of rules. However, in legal practice, the norm is not exhausted by rules extracted from the legislative text, but it is something that emerges from all the legislative system, the interpretations, the judgments and in general from the whole social system around norms. So, using directly rules to represent norms brings out several problems due to the arbitrary and rigidity of the rule implementation.

The idea of "social object" suggest the use of an object oriented modeling technique. Also, social objects are instances of models (or other social objects) suggesting the use of classes. However, it is impossible to pre-determine agents' behavior. So, we describe open models - what components an instance should contain - that will result in graphs. We provide an additional level or representation (a network of social objects) that connects agents with concepts representations like ontologies or rules. We use those representations to gain insight into the nature of norms and the relations with agents' mind. A more detailed model is expected to arise in future studies.

References

1. D. Liebwald, "Law's capacity for vagueness," *International Journal for the Semiotics of Law*, vol. 26(2), pp. 391–423, 2013.
2. V. Ferrari, *Lineamenti di sociologia del diritto*. Laterza, 1997.
3. B. S. Jackson, *Semiotics and legal theory*. Henley, 1985.
4. A. Antonini, G. Boella, and L. van der Torre, "Beyond the rules representation of norms: norms as social objects," in *Proceedings of Rules 2013 Conference*, 2013.
5. G. Boella and L. van der Torre, "Δ: The social delegation cycle," in *Proceedings of the 7th International Workshop on Deontic Logic in Computer Science (DEON 2004)*, pp. 29–42, 2004.
6. G. Boella and L. van der Torre, "Substantive and procedural norms in normative multiagent systems," *Journal of Applied Logic*, vol. 6(2), pp. 152–171, 2008.
7. J. Broersen, M. Dastani, H. J., Z. Huang, and L. van der Torre, "The boid architecture: conflicts between beliefs, obligations, intentions and desires," in *Proceedings of the fifth international conference on Autonomous agents (AGENTS '01)*, pp. 9–16, 2001.
8. G. Boella and L. van der Torre, "The ontological properties of social roles in multi-agent systems: Definitional dependence, powers and roles playing roles," *Artificial Intelligence and Law Journal (AILaw)*, 2007.
9. R. Cotterrell, *The concept of legal culture*, pp. 12–31. Darmouth: Aldershot, 1997.
10. C. Blengino, *Esercizio azione penale e processi organizzativi*, pp. 117–226. Giuffrè Editore, 2007.

Some perspectives on ceteris paribus preference

Fenrong Liu

Tsinghua University, China

The feature of ceteris paribus or ”everything else being equal” is central to the notion of preference. After the study of Von Wright, it has been widely studied in logic and AI, and the latest take-up is van Benthem, Girard and Roy (JPL paper). In this talk I will first provide a review of the previous works with a focus on how ceteris parabus preference is understood, then I will introduce more recent ideas, both conceptually and technically.

Dynamic Normative Logic and Information Update

Yasuo Nakayama
Graduate School of Human Sciences, Osaka University
1-2 Yamada-oka, Suita, Osaka, 565-0871 JAPAN
nakayama@hus.osaka-u.ac.jp

In Nakayama (2010), *Logic for Normative Systems* (LNS) was proposed. In this paper, I show how to deal with information update within LNS. I call LNS with information update device *Dynamic Normative Logic* (DNL). Recently, the dynamic epistemic logic (DEL) has been established as a framework for logical description of social interactions.[1] DNL can be considered as an alternative framework for the same purpose. DNL can explicitly express conditions for social behaviors and describe interactions between social actions and normative inference in detail.

1. Logic for Normative Systems and Dynamic Normative Logic

The following is a modification of LNS in Nakayama (2010).[2]

Let T and OB be a set of sentences in *First-Order Logic* (FOL) and q be a sentence of FOL.

(1a) A pair $\langle T, OB\rangle$ consisting of *belief base* T and *obligation base* OB is called a *normative system* ($NS = \langle T, OB\rangle$).

(1b) q belongs to the *belief set* of normative system NS (abbreviated as $\mathbf{B}_{NS}\, q$) $\Leftrightarrow$ q follows from T.

(1c) q belongs to the *obligation set* of NS (abbreviated as $\mathbf{O}_{NS}\, q$) $\Leftrightarrow$ $T\cup OB$ is consistent & q follows from $T\cup OB$ & q does not follow from T.

(1d) q belongs to the *prohibition set* of NS (abbreviated as $\mathbf{F}_{NS}\, q$) $\Leftrightarrow$ $\mathbf{O}_{NS}\, \neg q$.

(1e) q *belongs to the permission set of* NS (abbreviated as $\mathbf{P}_{NS}\, q$) $\Leftrightarrow$ $T\cup OB\cup\{q\}$ is consistent & q does not follow from T.

(1f) A normative system $\langle T, OB\rangle$ is consistent $\Leftrightarrow$ $T\cup OB$ is consistent.

(1g) In this paper, we interpret that NS represents a normative system accepted by a person or by a group at a particular time. Thus, we insert *what a person* (or *a group*) *believes to be true* into the *belief base* and *what he believes that it ought to be done* into the *obligation base*.

Based on the above definition, we can easily prove the following main theorems that characterize

[1] For the development of the dynamic epistemic logic, you nay consult van Benthem (2011). His description is restricted on various kinds of (dynamic) extension of *propositional* modal logics.

[2] We use &, $\Rightarrow$, and $\Leftrightarrow$ as meta-semantic abbreviation for *and*, *if … then*, and *if and only if*.

LNS, where $NS = \langle T, OB \rangle$.

(2a) $(\mathbf{B}_{NS}\,(p \to q)\ \&\ \mathbf{B}_{NS}\,p) \Rightarrow \mathbf{B}_{NS}\,q$.

(2b1) $(\mathbf{O}_{NS}\,(p \to q)\ \&\ \mathbf{O}_{NS}\,p) \Rightarrow \mathbf{O}_{NS}\,q$.

(2b2) $\mathbf{O}_{NS}\,p \Rightarrow \mathbf{P}_{NS}\,p$.

(2b3) $\mathbf{F}_{NS}\,p \Rightarrow$ not $\mathbf{P}_{NS}\,p$.

(2c1) $\mathbf{P}_{NS}\,p \Rightarrow$ not $\mathbf{B}_{NS}\,p$.

(2c2) $\mathbf{B}_{NS}\,p \Rightarrow$ (not $\mathbf{O}_{NS}\,p$ & not $\mathbf{F}_{NS}\,p$ & not $\mathbf{P}_{NS}$ p).

(2d1) $(\mathbf{O}_{NS}\,(p \to q)\ \&\ \mathbf{B}_{NS}\,p) \Rightarrow \mathbf{O}_{NS}\,q$.

(2d2) $(\mathbf{O}_{NS}\,(p \wedge q)$ & not $\mathbf{B}_{NS}\,p) \Rightarrow \mathbf{O}_{NS}\,p$.

(2d3) $(\mathbf{O}_{NS}\,(p \wedge q)\ \&\ \mathbf{B}_{NS}\,p) \Rightarrow \mathbf{O}_{NS}\,q$.

(2d4) $(\mathbf{O}_{NS}\,(p \vee q)\ \&\ \mathbf{B}_{NS}\,\neg p) \Rightarrow \mathbf{O}_{NS}\,q$.

(2d5) $(\mathbf{O}_{NS}\,(p \vee q)\ \&\ \mathbf{F}_{NS}\,p) \Rightarrow \mathbf{O}_{NS}\,q$.

(2d6) $(\mathbf{O}_{NS}\,p$ & not $\mathbf{B}_{NS}\,q) \Rightarrow \mathbf{O}_{NS}\,(p \vee q)$.

(2d7) $(\mathbf{B}_{NS}\,(p \to q)\ \&\ \mathbf{O}_{NS}\,p\ \&\ \mathbf{P}_{NS}\,q) \Rightarrow \mathbf{O}_{NS}\,q$.

(2e1) $(\mathbf{O}_{NS}\,\forall x_{1\ldots}\,\forall x_n\,(P(x_1,\ldots,x_n) \to Q(x_1,\ldots,x_n))\ \&\ \mathbf{B}_{NS}\,P(a_1,\ldots,a_n)$ & not $\mathbf{B}_{NS}\,Q(a_1,\ldots,a_n))$ $\Rightarrow \mathbf{O}_{NS}\,Q(a_1,\ldots,a_n)$. [This means: If $\forall x_{1\ldots}\,\forall x_n\,(P(x_1,\ldots,x_n) \to Q(x_1,\ldots,x_n))$ is an obligation and you believe $P(a_1,\ldots,a_n)$, then $Q(a_1,\ldots,a_n)$ is an obligation unless you believe that it was already done.]

(2e2) $(\mathbf{F}_{NS}\,\exists x_{1\ldots}\,\exists x_n\,(P(x_1,\ldots,x_n) \wedge Q(x_1,\ldots,x_n))\ \&\ \mathbf{B}_{NS}\,P(a_1,\ldots,a_n)$ & not $\mathbf{B}_{NS}\,\neg Q(a_1,\ldots,a_n))$ $\Rightarrow \mathbf{F}_{NS}\,Q(a_1,\ldots,a_n)$.

We update normative system $\langle T, OB \rangle$ through extending T or OB with new information p (i.e. $T \cup p$ or $OB \cup p$). In this paper, we call sometimes a normative system a *normative state*. As we see in the next section, a normative state of a person can be dependent on that of other person. To emphasize aspects of information update, we call LNS with information update device *Dynamic Normative Logic* (DNL).

2. An Application of DNL

To clarify update processes, we divide belief base T into two parts, namely elementary theory ET and a set of facts $FACT$. Thus, it holds, $T = ET \cup FACT$ & $ET \cap FACT = \varnothing$. In the example in this section, only $FACT$ is updated.

As an example, we consider a simple scene in a restaurant described by (van Benthem 2011: 4):

> In a restaurant, your Father has ordered Fish, your Mother ordered Vegetarian, and you have Meat. Out of the kitchen comes some new person carrying the three plates. What will happen?

We assume, here, that the asked person is a boy. The following list describes possible developments of the scene and translations of the described sentences into formula of FOL.

The waiter asks, 'Who has the Meat?	*ask* (*w*, *Family*, *ιj ordered* (*j*, *meat*), 0)
The boy says 'Me'.	*answer* (*b*, *ιj ordered* (*j*, *meat*), *b*, 0)
The waiter serves him with the meat plate.	s*erve* (*w*, *b*, *meat*, 0)
The waiter asks, 'Who has the Fish?'	*ask* (*w*, *Family*, *ιj ordered* (*j*, *fish*), 1)
The father says 'Me'.	*answer* (*f*, *ιj ordered* (*j*, *fish*), *f*, 1)
The waiter serves him with the fish plate.	s*erve* (*w*, *f*, *fish*, 1)
The waiter serves the mother with the vegetarian plate without asking.	s*erve* (*w*, *m*, *v*, 2)

To describe this scene within DNL, we need to make explicit each component of *NS* in this story.

Elementary Theory for the group (i.e. the family members and the waiter): ET^G = {(3a), (3b), (3c), (3d)}.

(3a) [Set Theoretical Principles] $\forall G_1 \forall G_2 (G_1 = G_2 \leftrightarrow \forall x (x \in G_1 \leftrightarrow x \in G_2)) \wedge \forall x \forall G_1 \forall G_2 (x \in G_1 \cup G_2 \leftrightarrow (x \in G_1 \vee x \in G_2)) \wedge \forall x \forall G_1 \forall G_2 (x \in G_1 - G_2 \leftrightarrow (x \in G_1 \wedge \neg x \in G_2))$.

(3b) $\forall i \exists^{=1} x$ *ordered* $(i, x) \wedge \forall i \forall j \forall x \forall y$ (*ordered* $(i, x) \wedge$ *ordered* $(j, y) \wedge i \neq j \rightarrow x \neq y$).
(Each family member ordered exactly one plate.)

(3c) $\forall i \forall x (\exists n$ *answer* $(i, \iota j$ *ordered* $(j, x), i, n) \rightarrow i = \iota j$ *ordered* $(j, x))$, where ιj *ordered* (j, x) refers to the person who ordered x. This use of ι-operator is justified by (3b).
(If someone answers that he ordered x, then he is the person who ordered x.)

(3d) $\forall i \forall G_1 \forall n$ (*served* $(G_1, n) \wedge \exists x$ *serve* $(w, i, x, n) \rightarrow$ *served* $(G_1 \cup \{i\}, n+1)$).
(At stage n where G_1 is already served, if the waiter serves person i with a plate, then $G_1 \cup \{i\}$ is served at stage n+1.)

Elementary Theory for the waiter: ET^w = {(3e)}

(3e) $\forall x \forall D \forall n$ (*have-plate* $(*, D, n) \wedge x \in D \wedge \exists i$ *serve* $(*, i, x, n) \rightarrow$ *have-plate* $(*, D - \{x\}, n+1)$), where sign '*' indicates that this belief is a *de se belief* (i.e. belief about himself).
(The waiter believes: At stage n where he has plates D, if he serves someone with plate x, then he has plates $D - \{x\}$ at stage n+1.)

Obligation Base for the group: OB^G = {(4a)}

(4a) $\forall i \forall x$ (*ordered* $(i, x) \rightarrow \forall n$ (*ask* $(w,$ *Family*, ιj *ordered* $(j, x), n) \rightarrow$ *answer* $(i, \iota j$ *ordered* $(j, x), i, n)))$.
(If the waiter asks the family 'Who ordered x?', then the person who ordered x should answer that he (or she) did. This rule expresses a social norm for guests in a restaurant.)[3]

[3] Here, the speech act of asking is interpreted as a request for an answer from a person who has sufficient

Obligation Base for the waiter: OB^w = {(4b)}

(4b) $\forall i\ \forall x$ (*ordered* (i, x) $\rightarrow \forall D\ \forall n$ (*have-plate* (*, D, n) $\wedge\ x \in D \rightarrow$ *serve* (*, i, x, n))), where sign '*' indicates that this obligation is a *de se norm* (i.e. norm about himself).
(The waiter should serve a guest with the meal that he (or she) ordered.)

(5a) Initial State:
$FACT^G_0$ = {*Family* = {b, f, m}, *Plate* = {*meat*, *fish*, v}, *served* ($\varnothing$, 0)}.
$FACT^b_0$ = {*ordered* (*, *meat*), $* \in$ *Family*}. The content of $FACT^b_0$ means 'I ordered *meat* and I belong to the *Family*', where 'I' refers to the boy.
$FACT^f_0$= {*ordered* (*, *fish*), $* \in$ *Family*}.
$FACT^m_0$ = {*ordered* (*, v), $* \in$ *Family*}.
$FACT^w_0$ = {*have-plate* (*, *Plate*, 0)}.

(5b) Normative systems on state n (In this story, $FACT^G_n$ is updated along the development of the situation.)
$G(n) = \langle T^G_n, OB^G \rangle$, where $T^G_n = ET^G \cup FACT^G_n$.
$boy(n) = \langle T^G_n \cup FACT^b_0, OB^G \rangle$,
$father(n) = \langle T^G_n \cup FACT^f_0, OB^G \rangle$,
$mother(n) = \langle T^G_n \cup FACT^m_0, OB^G \rangle$,
$waiter(n) = \langle T^G_n \cup ET^w \cup FACT^w_0, OB^G \cup OB^w \rangle$.

Based on (5b), we can easily show that $\mathbf{B}_{G(n)}$ expresses a shared belief among four people in the story, namely it holds: $\mathbf{B}_{G(n)}p \Rightarrow (\mathbf{B}_{waiter(n)}p\ \&\ \mathbf{B}_{boy(n)}p\ \&\ \mathbf{B}_{father(n)}p\ \&\ \mathbf{B}_{mother(n)}p)$.

I propose to interpret the restaurant story as a game played by the waiter and three guests who are cooperative with the waiter. We assume here that each of players obeys and performs any obligation that is required in each situation. It is the goal of this game that the waiter correctly distributes all plates he had at the initial state. Now, we can describe the development with help of DNL as follows.

(6a) By constructing a finite model, we can prove: $\mathbf{P}_{waiter(0)}$ *ask* (*, *Family*, *ιj ordered* (*j*, *meat*), 0).
Thus, the waiter asks, 'Who has the Meat?': $FACT^G_1 = FACT^G_0 \cup$ {*ask* (*w*, *Family*, *ιj ordered* (*j*, *meat*), 0)}.
Then, because of (4a) and (5a): $\mathbf{O}_{boy(1)}$ *answer* (*, *ιj ordered* (*j*, *meat*), *, 0). Following this obligation, the boy says 'Me': $FACT^G_2 = FACT^G_1 \cup$ {*answer* (*b*, *ιj ordered* (*j*, *meat*), *b*, 0)}.
Now, because of (3c) and (5b): $\mathbf{B}_{waiter(2)}$ *ordered* (*b*, *meat*), which means that the waiter realizes that the boy ordered meat. Then, from (4b) follows: $\mathbf{O}_{waiter(2)}$ *serve* (*, *b*, *meat*, 0).

---

information.

Following this obligation, the waiter serves the boy with the meat plate: $FACT^G_3 = FACT^G_2 \cup \{serve\ (w, b, meat, 0)\}$. Now, from (3d) and (3e) follows: $\mathbf{B}_{G(3)}$ *served* $(\{b\}, 1)$ & $\mathbf{B}_{waiter(3)}$ *have-plate* $(*, \{fish, v\}, 1)$.

(6 b) Similarly as (6a), we obtain the following updates and attitude changes:

$\mathbf{P}_{waiter(3)}$ *ask* $(*, Family, \imath j\ ordered\ (j, fish), 1)$.

$FACT^G_4 = FACT^G_3 \cup \{ask\ (w, Family, \imath j\ ordered\ (j, fish), 1)\}$.

$\mathbf{O}_{father(4)}$ *answer* $(*, \imath j\ ordered\ (j, fish), *, 1)$.

$FACT^G_5 = FACT^G_4 \cup \{answer\ (f, \imath j\ ordered\ (j, fish), f, 1)\}$.

$\mathbf{B}_{waiter(5)}$ *ordered* $(f, fish)$ & $\mathbf{O}_{waiter(5)}$ *serve* $(*, f, fish, 1)$.

$FACT^G_6 = FACT^G_5 \cup \{serve\ (w, f, fish, 1)\}$.

$\mathbf{B}_{G(6)}$ *served* $(\{b, f\}, 2)$ & $\mathbf{B}_{waiter(6)}$ *have-plate* $(*, \{v\}, 2)$.

(6c) In the third stage, the waiter infers who ordered the vegetarian plate without asking. Because of (3b): $\mathbf{B}_{waiter(6)}$ *ordered* (m, v). Thus, $\mathbf{O}_{waiter(6)}$ *serve* $(*, m, v, 2)$. Following this obligation, the waiter serves the mother with the vegetarian: $FACT^G_7 = FACT^G_6 \cup \{serve\ (w, m, v, 2)\}$. Then, we obtain: $\mathbf{B}_{G(7)}$ *served* $(\{b, f, m\}, 3)$ & $\mathbf{B}_{waiter(7)}$ *have-plate* $(*, \varnothing, 3)$. This shows that the waiter realized that he had accomplished his current task.

Now, you may recognize that this interaction in the restaurant is similar to many language games described in Wittgenstein (1953). Actually, simple language games can be described within DNL. Furthermore, other puzzles like *The Cards* and *The Muddy Children* (cf. van Benthem 2011: 8, 12) can be solved within DNL.

3. Concluding Remarks

In this paper, we extended LNS and defined DNL. Then, we have shown how to describe information update within DNL and applied DNL to a logical elucidation of social interactions in a restaurant scene. The method used in this paper is applicable to descriptions of social interactions among multiple agents, especially when these interactions involve belief update that affects normative attitudes.

References

van Benthem, J. (2011) *Logical Dynamics of Information and Interaction*, Cambridge University Press.

Nakayama, Y. (2010) "Logical Framework for Normative Systems," *SOCREAL 2010: Proceedings of the 2nd International Workshop On Philosophy and Ethics of Social Reality*, 27–28 March 2010, Sapporo: 19-24.

Wittgenstein, L. (1953) *Philosophical Investigations*.

"To be or not to be"≠ "to kill or not to kill" a logic on action negation

Xin Sun

Individual and Collective Reasoning Group, University of Luxembourg
xin.sun@uni.lu

Abstract. This paper defines a new action negation operator such that it is more dilemma-free than the existed treatment. A dynamic deontic logic is built on top of this new theory of action. Such new logic satisfies the free choice axiom and avoids all the implausible results often arise with the validation of free choice axiom.

Key words: action negation, dynamic deontic logic

1 Introduction

Research on deontic logic can be divided into two main groups: the ought-to-be group and the ought-to-do group. The ought-to-do group originates from the famous Finnish philosopher von Wright's pioneering paper [11]. Belong to this branch there are dynamic deontic logic [8, 7], and deontic action logic [9, 3, 10].

One issue of dynamic deontic logic is to characterize the negation of action. In the dynamic logic literature [6], the negation of action is usually interpreted as set theoretical complement with respect to the universal relation. [1] and [2] point out that such treatment is not the best choice when dynamic logic is applied to deontic setting. Several new versions of action negation are defined in [1, 2].

In this paper, we will define another action negation operator which is intuitively natural and technically dilemma-free. In Section 2 we recall dynamic logic. In section 3 we define a new operator for action negation therefore give arise to a new dynamic logic. In Section 4 we apply the new logic to the deontic setting. We conclude this paper in Section 5.

2 Dynamic logic

In this section we recall the definitions of dynamic logic. Let $\mathbb{P}$ be a countable set of propositional letters and $\mathbb{A}$ a countable set of symbols of action generators. The language of dynamic logic can be defined by the following BNF:

Definition 1 (Language of dynamic logic). *For $a \in \mathbb{A}$ and $p \in \mathbb{P}$,*

- $\alpha := a|\alpha \cup \alpha|\alpha \cap \alpha|\alpha;\alpha|\alpha^*|\overline{\alpha}$
- $\phi := p|\top|\neg\phi|\phi \wedge \phi|[\alpha]\phi$

Here symbols of the form α are action terms and ϕ are formulas. Formulas are interpreted by the relational model, which can be defined as follows.

Definition 2 (Relational model). *A relational model $\mathbb{M} = (S, R^{\mathbb{A}}, V)$ is a triple:*

- *S is a nonempty set of possible states.*
- *$R^{\mathbb{A}} : \mathbb{A} \to 2^{S \times S}$ is an action interpretation function, assigning a binary relation over $S \times S$ to each action generator $a \in \mathbb{A}$.*
- *V is the valuation function for propositional letters.*

The action interpretation function $R^{\mathbb{A}}$ can be extend to a new function R to interpret arbitrary actions as follows:

- $R(a) = R^{\mathbb{A}}(a)$ for $a \in \mathbb{A}$.
- $R(\alpha \cup \beta) = R(\alpha) \cup R(\beta)$
- $R(\alpha \cap \beta) = R(\alpha) \cap R(\beta)$
- $R(\alpha;\beta) = R(\alpha) \circ R(\beta)$
- $R(\alpha^*) = (R(\alpha))^*$

Here $\circ$ is the composition operator for relations and * is the reflexive transitive closure operator of relations. We leave the case for $R(\overline{\alpha})$ to the next section because that is the theme of this paper. With the function R in hand, we can define the semantics for formulas of dynamic logic use relational model as following:

Definition 3 (Semantics of dynamic logic). *Let $M =< W, R^{\mathfrak{A}}, V >$ be a relational model. Let $w \in W$.*

- *$M, w \models p$ iff $w \in V(p)$*
- *$M, w \models \neg\phi$ iff not $M, w \models \phi$*
- *$M, w \models \phi \wedge \phi$ iff $M, w \models \phi$ and $M, w \models \phi$*
- *$M, w \models [\alpha]\phi$ iff for all v, if $(w, v) \in R(\alpha)$ then $M, v \models \phi$*

3 A New Treatment of Action

In this section we first review the known treatment for dynamic logic on action negation, then we define a new alternative.

3.1 Action negation in the literature

The traditional interpretation of action negation [6] is to let $R(\overline{\alpha}) = W \times W - R(\alpha)$, *i.e.* the set theoretical complement with respect to the universal relation. Jan Broersen [1] and [2] argues that the universal relation is not the ideal background for complement when dynamic logic is applied to normative reasoning, or deontic logic. Instead we should restrict the universal relation such that

those worlds which are unreachable by any action are out of concern. With this intuition Jan Broersen suggest we replace the universal relation $W \times W$ in the interpretation of $R(\overline{\alpha})$ by relations like $\bigcup_{\alpha \in \mathbb{A}} R(\alpha)$, $(\bigcup_{\alpha \in \mathbb{A}} R(\alpha))^+$, $(\bigcup_{\alpha \in \mathbb{A}} R(\alpha))^*$ etc.

Jan Broersen's approach is more natural than its traditional counterpart in the deontic setting. But there is a shortcoming both of them can not overcome. For an illustration, first note that for any two action α and β, the action $\alpha \cup \overline{\alpha}$ and $\beta \cup \overline{\beta}$ are identical in both traditional and Broersen's approach. Now suppose Hamlet receives the following authorization: "you are permitted either to be or not to be" and James Bond receives the following authorization:"you are permitted either to kill or not". Intuitively, the first permission offers Hamlet a free choice between to live and to dead and the second offers 007 the license to kill or not. These two permissions convey very different information and should be distinguished. But they are identical in both the traditional and Jan Broersen's approach. Therefore those two approach both lead us to a dilemma.

A dilemma needs a solution. In the following section we will develop a new interpretation of action negation such that the above dilemma is solved.

3.2 A new approach of action negation

We first make a classification about actions. Since actions are interpreted by relations and the simplest relation is a set contains one ordered pair of states, we can naturally call an action α particle if $R(\alpha)$ contains exactly one ordered pair. For a particle action α, we call the first component of $R(\alpha)$ the pre-condition of α. Formally, if $R(\alpha) = (s_1, s_2)$, then $pre(\alpha) = \{s_1\}$. And we call the second component of $R(\alpha)$ the post-condition of α, formally $post(\alpha) = \{s_2\}$. Intuitively, a particle action is a deterministic change from one state to another.

Based on particle action, we build atomic action as a union of particle actions which share the same pre-condition. For instance, for two particle actions α_1 and α_2 with $R(\alpha_1) = \{(s_1, s_2)\}$, $R(\alpha_2) = \{(s_1, s_3)\}$, the action α_3 such that $R(\alpha_3) = \{(s_1, s_2), (s_1, s_3)\}$ is an atomic action.

For an atomic action α, its pre-condition is the same as its consisted particle actions. The post condition of α is the union of the post conditions of its consisted particle actions. Therefore $post(\alpha_3) = \{s_2, s_3\}$. Intuitively, an atomic action is a nondeterministic change from a specific state to other states. For example, if we let s_1 represents "China", s_2 represents "USA", and s_3 represents "Canada", then α_3 means "go to north America from China".

A normal action is a union of simple actions which possibly bears different pre-conditions. For example, let s_1 be "Italy" and s_2 be "Luxembourg", then the action α with $R(\alpha) = \{(s_1, s_2), (s_2, s_1)\}$ is a normal action which can be read as "go to Luxembourg from Italy, or go to Italy from Luxembourg". For a normal action α, we defined its pre-condition as the union of the pre-conditions of its consisted simple actions. Formally, $pre(\alpha) = \{s \in S | (s, t) \in R(\alpha)\}$.

Now we have a classification of actions and the pre/post condition of action has been defined. It is the time to grasp what the negation of an action is. For an atomic action α, say $R(\alpha) = \{(s, s), (s, t)\}$ and $W = \{s, t, u\}$, we tend to

define $R(\overline{\alpha}) = \{(s, u)\}$. The intuition is, we understand α as a non-deterministic movement from s to either t or s itself, then the negation of α can be understood as "go to those states other than α goes". More formally, we define $R(\overline{\alpha}) = pre(\alpha) \times (W - post(\alpha))$ for a simple action α.

For a normal action, we can calculate $R(\overline{\alpha})$ via following steps:

1. We decompose α to atomic actions $\alpha_1, \ldots, \alpha_n$ such that $R(\alpha) = R(\alpha_1) \cup \ldots \cup R(\alpha_n)$ and for every i, j, $pre(\alpha_i) \neq pre(\alpha_j)$. It can be verified such decomposition is unique and each α_i is a maximal sub-atomic-action of α in the sense that for every atomic action β, if $R(\beta) \subseteq R(\alpha)$ then there exist a unique a_i in the decomposition such that $R(\beta) \subseteq R(\alpha_i)$
2. For each $i \in \{1, \ldots, n\}$, we calculate $R(\overline{\alpha_i})$. Since α_i is an atomic action, we have $R(\overline{\alpha_i}) = pre(\alpha_i) \times (W - post(\alpha_i))$.
3. We take the union of these $R(\overline{\alpha_i})$ to form $R(\overline{\alpha}) = R(\overline{\alpha_1}) \cup \ldots \cup R(\overline{\alpha_n})$.

Equivalent to the procedure above, we can define the negation of action in a more concise manner as follows:

Definition 4. $R(\overline{\alpha}) = Pre(\alpha) \times S - R(\alpha)$.

It is not hard to verify that for two action α and β, as long as $pre(\alpha) \neq pre(\beta)$, we have $R(\alpha \cup \overline{\alpha}) \neq R(\beta \cup \overline{\beta})$. Hence the dilemma from subsection 3.1 is solved.

4 From action to deontic logic

There are several possible approaches to develop deontic logic based on the logic of action above. One is as in [4], for every action we choose a subset of its post-condition to be the ideal outcomes, then use those ideal outcomes to form a neighborhood, served as the source of normativity. A second approach is to build deontic logic via deontic-dynamic reduction as in [2]. In this section we focus on the second approach.

The methodology is to introduce 'normative constant' for obligation, permission and prohibition respectively. Let c_F be the constant of a prohibition, c_O for obligation and c_P be for permission. Intuitively, $V(c_F)$ is the set of states which are morally forbidden and $V(c_O)$ the set of morally required states. V_P is the morally permissive states. For each valuation V of a relational model $M =< W, R^{\mathfrak{A}}, V >$, $V(c) \subseteq W$, for $c \in \{c_F, c_O, c_P\}$. We moreover require $V(c_F) \cap V(c_P) = \emptyset$ and $V(c_O) \subseteq V(c_P)$. For those states which are not in $V(c_F) \cup V(c_P)$, we consider them as morally neutral.

We define normative operators based on normative constant as follows:

- $P(\alpha) := [\alpha]c_P$
- $F(\alpha) := [\alpha]c_F$
- $O(\alpha) := [\overline{\alpha}]\neg c_O$

According to the above definition, an action is permitted iff the post-condition of its execution will always belong to the morally permissive states. An action is

forbidden iff all the post-condition of its execution will lead to morally forbidden states. An action is obligatory iff as long as we execute its negation, the outcome will not be morally required.

It can be verified that the above deontic operators satisfies the following logical properties:

- $\models P(\alpha \cup \beta) \rightarrow P(\alpha) \wedge P(\beta)$
- $\not\models O(\alpha) \leftrightarrow \neg P(\overline{\alpha})$
- $\not\models O(\alpha) \rightarrow O(\alpha \cup \beta)$
- $\not\models P(\alpha) \rightarrow P((\alpha \cap \beta) \cup (a \cap \overline{\beta}))$

Those properties are essential to verify the free choice axiom meanwhile block all the potential implausible results arises with the validation of free choice axiom [4],[5].

5 Conclusion

This paper defines a new action negation operator such that it is more dilemma-free than the existed treatment of the action negation. A dynamic deontic logic is build on top of this new logic. Such new logic satisfies the free choice axiom and avoids all the implausible results often arise with the validation of free choice axiom.

References

1. J. Broersen. Relativized action negation for dynamic logics. In P. Balbiani, N-Y. Suzuki, F. Wolter, and M. Zakharyaschev, editors, *Advances in Modal Logic*, volume 4, pages 51–70, 2003.
2. J. Broersen. Action negation and alternative reductions dynamic deontic logics. *Journal of applied logic*, 2004.
3. P. Castro and T. Maibaum. Deontic action logic, atomic boolean algebras and fault-tolerance. *Journal of Applied Logic*, 2009.
4. D. Gabbay, X. Sun, and L. Gammaitoni. The paradoxes of permission, an action based solution. *to appear in Journal of Applied Logic*, 2013.
5. S. Hansson. The varieties of permissons. In *Handbook of deontic logic and normative systems*. College Publication, 2013.
6. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic logic*. The MIT Press, 2000.
7. Ron Van Der Meyden. The dynamic logic of permission. *Journal of Logic and Computation*, 6:465–479, 1996.
8. J. J. Meyer. A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, pages 109–136, 1988.
9. K. Segerberg. A deontic logic of action. *Studia Logica*, 1982.
10. R. Trypuz and R. Kulicki. A systematics of deontic action logics based on boolean algebra. *Logic and Logical Philosophy*, 2009.
11. G. H. von Wright. Deontic logic. *Mind*, pages 1–15, 1951.

Measurement-Theoretic Foundations of Preference Aggregation Logic for Weighted Utilitarianism (Extended Abstract)

Satoru Suzuki

Komazawa University,
bxs05253@nifty.com

1 Motivation

Harsanyi [3, 4] develops expected utility theory of von Neumann and Morgenstern [5] to provide *two axiomatizations* of *utilitarianism*. Weymark [7, 8] refers to these results as Harsanyi's *Aggregation* and *Impartial Observer Theorems*.[1] Sen [6] argues that von Neumann-Morgenstern expected utility theory is an *ordinal* theory and, therefore, *any increasing* transform of an expected utility function is a satisfactory representation of an individual's preference relation. However, utilitarianism requires a *cardinal* theory of utility and so Harsanyi is not justified in giving his theorems utilitarian interpretations. Sen's informal discussion of these issues is formalized by Weymark [7]. Broome [1] calls this argument the "*Standard Objection*" to Harsanyi's theorems. The aims of this talk are as follows:

1. We show that *Domotor's exact reformulation [2]* of Harsanyi's Aggregation Theorem in terms of *measurement theory* can result in dodging the *Standard Objection* to Harsanyi's theorems.[2]
2. We propose a new version of complete logic for preference aggregation represented by a weighted utilitarian rule—Preference Aggregation Logic for Weighted Utilitarianism (PALU) by means of *measurement theory*.

2 Standard Objection and Domotor's Theorems

We define a prospect and an ordered mixture space as follows:

Definition 1 (Prospect and Ordered Mixture Space). *Let $\mathscr{A}$ be a nonempty set of alternatives, $\mathscr{J}$ a real unit interval $\{x \in \mathbb{R} : 0 \leq x \leq 1\}$, $[\ \ ] : \mathscr{A} \times \mathscr{J} \times \mathscr{A} \to \mathscr{A}$ a mixture operation such that $[\mathbf{a}, 1, \mathbf{b}] = \mathbf{a}, [\mathbf{a}, \alpha, \mathbf{b}] = [\mathbf{b}, 1-\alpha, \mathbf{a}]$ and $[[\mathbf{a}, \beta, \mathbf{b}], \alpha, \mathbf{b}] =$*

[1] In this talk, we are not concerned with Impartial Observer Theorem because the model of Preference Aggregation Logic for Weighted Utilitarianism (PALU) is based only on Aggregation Theorem.

[2] As far as the author knows, no one shows that Domotor's exact reformulation of Harsanyi's Aggregation Theorem can result in dodging the Standard Objection to Harsanyi's theorems. So in the author's opinion, the first aim of this talk is necessary since the model of Preference Aggregation Logic for Weighted Utilitarianism (PALU) is based on Aggregation Theorem.

$[\mathbf{a}, \alpha\beta, \mathbf{b}]$, $\mathscr{S} := \{1, \ldots, n\}$ *a* ***society set****,* $\precsim_i$ *a* $[\quad]$*-monotonic ordering of an agent* $i \in \mathscr{S}$ *on* $\mathscr{A}$*, and* $\precsim$ *a* $[\quad]$*-monotonic ordering of* $\mathscr{S}$ *on* $\mathscr{A}$*, where* $\mathbf{a}, \mathbf{b} \in \mathscr{A}$ *and* $\alpha, \beta \in \mathscr{J}$*. We call* $[\mathbf{a}, \alpha, \mathbf{b}]$ *a* ***prospect****,* $(\mathscr{A}, [\quad], \precsim_i)$ *an* ***ordered mixture space of*** *i, and* $(\mathscr{A}, [\quad], \precsim)$ *an* ***ordered mixture space of*** $\mathscr{S}$*.*

We define a weak order, etc. as follows:

Definition 2 (Weak Order, etc.).

1. Weak Order
$\precsim$ is a weak order (connected and transitive) on $\mathscr{A}$. (Similarly for $\precsim_i$.)
2. Continuity
For any $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathscr{A}$ for which $\mathbf{c} \prec \mathbf{b} \prec \mathbf{a}$, there exists $\alpha \in \mathscr{J}$ such that $[\mathbf{a}, \alpha, \mathbf{c}] \sim \mathbf{b}$, where $\mathbf{a} \prec \mathbf{b} := \mathbf{b} \not\precsim \mathbf{a}$ and $\mathbf{a} \sim \mathbf{b} := \mathbf{a} \precsim \mathbf{b}$ and $\mathbf{b} \precsim \mathbf{a}$. (Similarly for $\prec_i$ and $\sim_i$.)
3. Independence
For any $\mathbf{a}, \mathbf{b} \in \mathscr{A}$ and any $\alpha \in \mathscr{J}$, if $\mathbf{a} \sim \mathbf{b}$, then $\mathbf{a} \sim [\mathbf{a}, \alpha, \mathbf{b}]$.
For any distinct $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathscr{A}$ and any $\alpha \in \mathscr{J}$, if $\mathbf{a} \sim \mathbf{b}$, then $[\mathbf{a}, \alpha, \mathbf{c}] \sim [\mathbf{b}, \alpha, \mathbf{c}]$. (Similarly for $\sim_i$.)
4. Pareto Indifference
For any $a, b \in \mathscr{A}$, if, for any $i \in \mathscr{S}$, ($\mathbf{a} \sim_i \mathbf{b}$), then $\mathbf{a} \sim \mathbf{b}$.

Weymark [8] formulates Harsanyi's Aggregation Theorem as follows:

Theorem 1 (Harsanyi's Aggregation Theorem). *Suppose that $\precsim_i$ ($i \in \mathscr{S}$) and $\precsim$ are binary relation on $\mathscr{A}$ that satisfies Weak Order, Continuity and Independence and also suppose that Pareto Indifference is satisfied these relations. Let U_i be an expected utility representation of $\precsim_i$, and U be an expected utility representation of $\precsim$. Then, there exist $\alpha_i (i \in \mathscr{S}), \beta \in \mathbb{R}$ such that, for any $\mathbf{a} \in \mathscr{A}$,*

$$(1)\ U(\mathbf{a}) = \sum_{i=1}^{n} \alpha_i U_i(\mathbf{a}) + \beta.$$

An implication of (1) is that, for any $\mathbf{a}, \mathbf{b} \in \mathbf{A}$,

$$(2)\ U(\mathbf{a}) \leq U(\mathbf{b}) \textit{ iff } \sum_{i=1}^{n} \alpha_i U_i(\mathbf{a}) \leq \sum_{i=1}^{n} \alpha_i U_i(\mathbf{b}).$$

Remark 1. The conclusion Harsanyi intends to draw is that alternatives are socially ranked using an *weighted utilitarian* rule.

We define an increasing transform, etc. as follows:

Definition 3 (Increasing Transform, etc.).

- *$f : \mathbb{R} \to \mathbb{R}$ is an* ***increasing transform*** *if for any $x, y \in \mathbb{R}$, $f(x) \leq f(y)$ iff $x \leq y$.*
- *A utility function that is unique up to an increasing transform is said to be* ***ordinal****.*
- *$f : \mathbb{R} \to \mathbb{R}$ is a* ***positive linear transform*** *if there exist $\alpha, \beta \in \mathbb{R}$ such that $f(x) = \alpha x + \beta$ for any $x \in \mathbb{R}$.*
- *A utility function that is unique up to a positive linear transform is said to be* ***cardinal****.*

– *The n-tuple of positive linear transforms $F = (f_1, \ldots, f_n)$ is called* ***co-cardinal****.*

According to Weymark [8], the underlying reason for the problems identified by Sen is that in order for utilitarianism to be meaningful, it must be possible to compare *utility differences* (gains and losses) both intrapersonally and interpersonally. The need of difference comparability can be seen most clearly rewriting (2) as follows:

$$(3)\ U(\mathbf{a}) \leq U(\mathbf{b}) \text{ iff } \sum_{i=1}^{n} \alpha_i (U_i(\mathbf{a}) - U_i(\mathbf{b})) \leq 0.$$

The utility difference sum in (3) does not change if the utility profile $\mathbf{U}$ is replaced by the profile $\mathbf{V} = F \circ \mathbf{U} := (f_1 \circ U_1, \ldots, f_n \circ U_n)$ for some *co-cardinal* n-tuple of transform F. Let $\mathscr{U}^C$ denote the set of such profile of utility functions. The profiles in $\mathscr{U}^C$ are the *only* profiles that preserve the utility difference sum in (3). However, nothing in the version of expected utility theory that Harsanyi employed in his theorems rules out the use of *non-linear* increasing transform of U_i. So, because the set of admissible profiles is not always a subset of $\mathscr{U}^C$, $\lesssim$ is *not always weighted utilitarian*. On the other hand, there are two main problems in *measurement theory*:

1. the representation problem: justifying the assignment of numbers to objects,
2. the uniqueness problem: specifying the transformation up to which this assignment is unique.

A solution to the former can be furnished by a *representation theorem*, which establishes that the specified conditions on a qualitative relational system are (necessary and) sufficient for the assignment of numbers to objects that represents (or preserves) all the relations in the system. A solution to the latter can be furnished by a *uniqueness theorem*, which specifies the transformation up to which this assignment is unique. Domotor [2] proves the following theorem.

Theorem 2 (Strict Positive Moment). *Let $(\mathscr{V}, \prec)$ and $(\mathscr{W}, \prec')$ be two real linear finite dimensional strict partially ordered vector spaces with strict partial ordering relations $\prec$ and $\prec'$. Furthermore, let $f : M \to \mathscr{V}$ and $g : M \to \mathscr{W}$ be two mappings from a nonempty set M into the spaces $\mathscr{V}$ and $\mathscr{W}$ such that $0 \prec f(c)$ and $0 \prec g(c)$ for some $c \in M$. Then for there to exist a* ***strictly positive linear operator*** *$\mathscr{F} : \mathscr{V} \to \mathscr{W}$ such that $\mathscr{F}(f(x)) = g(x)$ for $x \in M$ and*

$$\textit{If } 0 \sim v, \textit{then } 0 \sim' \mathscr{F}(v). \quad \textit{If } 0 \prec v, \textit{then } 0 \prec' \mathscr{F}(v).$$

where $v \in \mathscr{V}$, it is necessary and sufficient that

$$\textit{If } 0 \sim \sum_{i \leq m} \alpha_i f(x_i), \textit{then } 0 \sim' \sum_{i \leq m} \alpha_i g(x_i). \quad \textit{If } 0 \prec \sum_{i \leq m} \alpha_i f(x_i), \textit{then } 0 \prec' \sum_{i \leq m} \alpha_i g(x_i).$$

hold for any $x_i \in M$ and $\alpha_i \in \mathbb{R} (1 \leq i \leq m)$.

We define a strictly positive social utility structure as follows:

Definition 4 (Strictly Positive Social Utility Structure). *A finite collection of relational structures $\mathbb{M}(\mathscr{S})$ is called* ***strictly positive social utility structure*** *of the society $\mathscr{S}$ if the following conditions are met:*

1. Weak Order, Continuity, Independence, and Pareto Indifference
$\precsim_i$ and $\precsim$ satisfy Weak Order, Continuity, Independence, and Pareto Indifference.

2. Strong Pareto
For any $\mathbf{a},\mathbf{b} \in \mathscr{A}$, if, for any $i \neq j \in \mathscr{S}$ and some j, ($\mathbf{a} \precsim_i \mathbf{b}$ and $\mathbf{a} \prec_j \mathbf{b}$), then $\mathbf{a} \prec \mathbf{b}$.

By showing that a strictly positive social utility structure of Definition 4 establishes the existence of the strictly positive linear operator $\mathscr{F}$ of Theorem 2, Domotor proves the following theorems:

Theorem 3 (Representation Theorem). *Let $\mathbb{M}(\mathscr{S})$ is a strictly positive social utility structure iff there exist utility functions $U_i, U : \mathscr{A} \to \mathbb{R}$ and positive reals γ_i such that for any $\mathbf{a},\mathbf{b} \in \mathscr{A}$, $\alpha \in \mathscr{J}$ and $i \in \mathscr{S}$,*

1. *$\mathbf{a} \precsim_i \mathbf{b}$ iff $U_i(\mathbf{a}) \leq U_i(\mathbf{b})$,*
2. *$U_i([\mathbf{a},\alpha,\mathbf{b})]) = \alpha U_i(\mathbf{a}) + (1-\alpha)U_i(\mathbf{b})$,*
3. *$U'(\mathbf{a}) = \sum_{i \in \mathscr{S}} \gamma_i U_i(\mathbf{a})$, where $U' = g \circ U$ for some $g \in \mathscr{G}$ ($\mathscr{G}$:= the group of **positive linear transforms**).*

Theorem 4 (Uniqueness). *The constants γ_i are given uniquely by the choice of U_i in a fixed scale $g_i \in \mathscr{G} (i \in \mathscr{S})$, where g_i is a transform of U_i.*

Remark 2. When a strictly positive social utility structure is given, the only permissible transform of U_i is a *positive linear* one, that is, the n-tuple of transforms $(g_1,\ldots,g_n)$ is *co-cardinal*. So at least in a strictly positive social utility structure, Domotor's representation and uniqueness theorem results in *dodging the Standard Objection to Harsanyi's theorems*.

3 Preference Aggregation Logic for Weighted Utilitarianism PALU

Next we will construct a preference aggregation logic for weighted utilitarianism PALU on the basis of Domotor's representation and uniqueness theorems.

3.1 Language of PALU

We define the language $\mathscr{L}_{\mathsf{PALU}}$ of PALU as follows:

Definition 5 (Language of PALU).

- *Let $\mathscr{S}$ denote a nonempty **society** set of **agents**, $\mathscr{V}$ a set of individual variables, $\mathscr{C}$ a set of individual constants, $\mathbf{WP}_i$ a **weak preference relation symbol** of i, $\mathbf{WP}$ a **social weak preference relation symbol**.*
- *The language $\mathscr{L}_{\mathsf{PALU}}$ of PALU is given by the following BNF grammar:*

$$\varphi ::= t_1 = t_2 \mid \mathbf{WP}_i(t_1,t_2) \mid \mathbf{WP}(t_1,t_2) \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi \mid \forall x\varphi$$
$$t ::= x \mid a,$$

  where $x \in \mathscr{V}$ and $a \in \mathscr{C}$.
- *$\bot, \vee, \to, \leftrightarrow$ and $\exists$ are introduced by the standard definitions.*

- $\mathbf{WP}_i(t_1,t_2)$ *means that an agent i does not prefer* t_1 *to* t_2.
- $\mathbf{WP}(t_1,t_2)$ *means that* t_1 *is not socially preferable to* t_2 *in terms of weighted utilitarianism.*
- *We define a strict preference relation symbol* $\mathbf{SP}_i$ *and an indifference relation symbol* $\mathbf{ID}_i$ *as follows (Similarly for* **SP** *and* **ID***):*

$$\mathbf{SP}_i(t_1,t_2) := \neg\mathbf{WP}_1(t_2,t_1),$$
$$\mathbf{ID}_i(t_1,t_2) := \mathbf{WP}_i(t_1,t_2) \text{ and } \mathbf{WP}_i(t_2,t_1).$$

- *The set of all well-formed formulae of* $\mathscr{L}_{\mathsf{PALU}}$ *is denoted by* $\Phi_{\mathscr{L}_{\mathsf{PALU}}}$.

3.2 Semantics of PALU

Model of $\mathscr{L}_{\mathsf{PALU}}$ We define a model $\mathfrak{M}$ of $\mathscr{L}_{\mathsf{PALU}}$ as follows:

Definition 6 (Model $\mathfrak{M}$ of $\mathscr{L}_{\mathsf{PALU}}$). *$\mathfrak{M}$ is a tuple* $(\mathscr{S},\mathscr{A},a^{\mathfrak{M}},b^{\mathfrak{M}},\ldots,[\quad],\precsim_i,\precsim)$, *where:*

- $\mathscr{S} := \{1,\ldots,n\}$ *is a society set,*
- $\mathscr{A}$ *is a nonempty set of alternatives,*
- $a^{\mathfrak{M}},b^{\mathfrak{M}},\ldots \in \mathscr{A}$,
- $\{(\mathscr{A},[\quad],\precsim_i,\precsim) : i \in \mathscr{S}\}$ *is a* ***strictly positive social utility structure*** *of* $\mathscr{S}$ *of Definition 4.*

Truth in PALU We provide $\mathscr{L}_{\mathsf{PALU}}$ with the following satisfaction definition relative to $\mathfrak{M}$:

Definition 7 (Satisfaction, Truth and Validity). *When an (extended) assignment function* $s(\tilde{s})$ *is given, what it means for* $\mathfrak{M}$ *to satisfy* $\varphi \in \Phi_{\mathscr{L}_{\mathsf{PALU}}}$ *with s, in symbols* $\mathfrak{M} \models_{\mathscr{L}_{\mathsf{PALU}}} \varphi[s]$ *is inductively defined as follows:*

- *The satisfaction clauses of* $=,\top,\neg,\wedge$ *and* $\forall$ *are standard ones,*
- $\mathfrak{M} \models_{\mathscr{L}_{\mathsf{PALU}}} \mathbf{WP}_i(t_1,t_2)[s]$ *iff* $\tilde{s}(t_1) \precsim_i \tilde{s}(t_2)$,
- $\mathfrak{M} \models_{\mathscr{L}_{\mathsf{PALU}}} \mathbf{WP}(t_1,t_2)[s]$ *iff* $\tilde{s}(t_1) \precsim \tilde{s}(t_2)$.

If $\mathfrak{M} \models_{\mathscr{L}_{\mathsf{PALU}}} \varphi[s]$ *for all s, we write* $\mathfrak{M} \models_{\mathscr{L}_{\mathsf{PALU}}} \varphi$ *and say that* φ *is true in* $\mathfrak{M}$. *If* φ *is true in all models of* $\mathscr{L}_{\mathsf{PALU}}$, *we write* $\models_{\mathscr{L}_{\mathsf{PALU}}} \varphi$ *and say that* φ *is valid.*

The next important corollary follows from Theorem 3 and Definitions 6 and 7.

Corollary 1 (Weighted Utilitarian Rule). *In* $\mathfrak{M}$ *of* $\mathscr{L}_{\mathsf{PALU}}$, *there exist utility functions* $U_i : \mathscr{A} \to \mathbb{R}$ *such that for any* $\tilde{s}(t_1),\tilde{s}(t_2) \in \mathscr{A}$,

$$\mathfrak{M} \models_{\mathscr{L}_{\mathsf{PALU}}} \mathbf{WP}(t_1,t_2)[s] \quad \textit{iff} \sum_{i=1}^{n} \alpha_i U_i(\tilde{s}(t_1)) \leq \sum_{i=1}^{n} \alpha_i U_i(\tilde{s}(t_2)).$$

Remark 3. This corollary indicates that we can reason about preference aggregation represented by an *weighted utilitarian rule* in terms of PALU.

3.3 Syntax of PALU

Proof System of PALU We extend a proof system of *first-order logic with an equality symbol* in such a way as to add the syntactic counterparts of the *Connectedness* of $\precsim_i$ and $\precsim$, the *Transitivity* of $\precsim_i$ and $\precsim$, *Pareto Indifference*, and *Strong Pareto*:

Definition 8 (Proof System of PALU).

- *all valid formulae of first-order logic with an equality symbol,*
- $\forall x \forall y(\mathbf{WP}_i(x,y) \vee \mathbf{WP}_i(y,x))$
  (Syntactic Counterpart of Connectedness of $\precsim_i$),
- $\forall x \forall y(\mathbf{WP}(x,y) \vee \mathbf{WP}(y,x))$
  (Syntactic Counterpart of Connectedness of $\precsim$),
- $\forall x \forall y \forall z((\mathbf{WP}_i(x,y) \wedge \mathbf{WP}_i(y,z)) \rightarrow \mathbf{WP}_i(x,z))$
  (Syntactic Counterpart of Transitivity of $\precsim_i$),
- $\forall x \forall y \forall z((\mathbf{WP}(x,y) \wedge \mathbf{WP}(y,z)) \rightarrow \mathbf{WP}(x,z))$
  (Syntactic Counterpart of Transitivity of $\precsim$),
- $\forall x \forall y((\mathbf{ID}_1(x,y) \wedge \cdots \wedge \mathbf{ID}_n(x,y)) \rightarrow \mathbf{ID}(x,y))$
  (Syntactic Counterpart of Pareto Indifference),
- $\forall x \forall y(((\mathbf{SP}_1(x,y) \wedge \mathbf{WP}_2(x,y) \wedge \cdots \wedge \mathbf{WP}_n(x,y)) \vee \cdots \vee (\mathbf{WP}_1(x,y) \wedge \cdots \wedge \mathbf{WP}_{n-1}(x,y) \wedge \mathbf{SP}_n(x,y))) \rightarrow \mathbf{SP}(x,y))$,
  (Syntactic Counterpart of Strong Pareto),
- *Modus Ponens,*
- *Generalization.*

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

Remark 4. The proof system of PALU has neither a syntactic counterpart of **Continuity** nor that of **Independence** both of which are satisfied in $\mathfrak{M}$ of $\mathscr{L}_{\mathsf{PALU}}$ of Definition 6. For $\mathscr{L}_{\mathsf{PALU}}$ is not so fine-grained as to express them. However, it is not a defect of PALU. For PALU is designed to capture *behavior* about preference aggregation represented by a weighted utilitarian rule, whereas both Continuity and Independence are the mere *structural* properties required for the existence of a weighted utilitarian in a mixture space.

Theorems That Are Characteristic of Preference Aggregation Represented by Weighted Utilitarian Rule We can prove the following theorems that are characteristic of preference aggregation represented by a weighted utilitarian rule:

Proposition 1 (Theorems That Are Characteristic of Preference Aggregation Represented by Weighted Utilitarian Rule).

$\vdash_{\mathsf{PALU}} \forall x\forall y(((\mathbf{WP}_2(x,y) \wedge \cdots \wedge \mathbf{WP}_n(x,y) \wedge \mathbf{WP}(y,x)) \rightarrow \mathbf{WP}_1(y,x)) \vee \cdots \vee ((\mathbf{WP}_1(x,y) \wedge \cdots \wedge \mathbf{WP}_{n-1}(x,y) \wedge \mathbf{WP}(y,x)) \rightarrow \mathbf{WP}_n(y,x)))$.

$\vdash_{\mathsf{PALU}} \forall x\forall y(((\mathbf{ID}_2(x,y) \wedge \cdots \wedge \mathbf{ID}_n(x,y)) \rightarrow (\mathbf{WP}_1(x,y) \leftrightarrow \mathbf{WP}(x,y))) \vee \cdots \vee ((\mathbf{ID}_1 \wedge \cdots \wedge \mathbf{ID}_{n-1}(x,y)) \rightarrow (\mathbf{WP}_n(x,y) \leftrightarrow \mathbf{WP}(x,y))))$.

$\vdash_{\mathsf{PALU}} \forall x\forall y(((\mathbf{WP}_1(x,y) \wedge \cdots \wedge \mathbf{WP}_n(x,y)) \wedge (\mathbf{SP}_1(x,y) \vee \cdots \vee \mathbf{SP}_n(x,y))) \rightarrow \mathbf{SP}(x,y))$.

3.4 Metalogic of PALU

We touch upon the metatheorems of PALU. It is easy to prove the soundness of PALU.

Theorem 5 (Soundness). *For any* $\varphi \in \Phi_{\mathscr{L}_{\mathsf{PALU}}}$, *if* $\vdash_{\mathsf{PALU}} \varphi$, *then* $\models_{\mathsf{PALU}} \varphi$.

We can also prove the completeness of PALU by means of Definition 6 and Definition 8.

Theorem 6 (Completeness). *For any* $\varphi \in \Phi_{\mathscr{L}_{\mathsf{PALU}}}$, *if* $\models_{\mathsf{PALU}} \varphi$, *then* $\vdash_{\mathsf{PALU}} \varphi$.

References

1. Broome, J.: Can there be a preference-based utilitarianism? In: Fleurbaey, M., et al. (eds.) Justice, Political Liberalism, and Utilitarianism: Themes from Harsanyi and Rawls, pp. 221–238. Cambridge University Press, Cambridge (2008)
2. Domotor, Z.: Ordered sum and tensor product of linear utility structures. Theory and Decision 11, 375–399 (1979)
3. Harsanyi, J.C.: Cardinal utility in welfare economics and in the theory of risk-taking. Journal of Political Economy 61, 434–435 (1953)
4. Harsanyi, J.C.: Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy 63, 309–321 (1955)
5. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (1944)
6. Sen, A.: Welfare inequalities and rawlsian axiomatics. Theory and Decision 7, 243–262 (1976)
7. Weymark, J.A.: A reconsideration of the harsanyi-sen debate on utilitarianism. In: Elster, J., E., R.J. (eds.) Interpersonal Comparisons of Well-Being, pp. 255–320. Cambridge University Press, Cambridge (1991)
8. Weymark, J.A.: Measurement theory and the foundations of utilitarianism. Social Choice and Welfare 25, 527–555 (2005)

Preconditions, Common Sense Reasoning, and Context Shifts

Tomoyuki Yamada

Hokkaido University

In doing things in everyday life, we rely on various regularities that hold normally. For example, by getting the switch of a flashlight on, you can light its bulb. The relevant regularity here may be expressed as follows (Barwise and Seligman, 1997):

(1) The switch being on entails that the bulb is lit.

What will happen, however, if the battery is dead. By applying the inference rule called weakening, we could derive the following:

(2) The switch being on and the battery being dead entails that the bulb is lit.

Since this conclusion is unacceptable, we might wish to revise (1) and say:

(3) The switch being on and the battery being live entails that the bulb is lit.

What will happen, however, if the bulb is gone. By weakening again, we would have:

(4) The switch being on, the battery being live, and the bulb being gone entails that the bulb is lit.

Barwise and Seligman (1997) proposes an interesting treatment of the regularities of this kind and their exceptions in terms of channel theory. We will examine how a similar treatment can be developed for the regularities on which we rely in doing things with words and their exceptions in the form of infelicitous speech acts.

Bibliography

Austin, J. L. *How to do Things with Words*, The Williams James Lectures, Harvard University, 1955. In: J. O. Urmson, and M. Sbisà, (eds.), *How to do Things with Words*, 2nd. ed., Harvard University Press, 1975.

Barwise, J. and Seligman, J. *Information Flow: The Logic of Distributed Systems*, Cambridge University Press, 1977.

van Ditmarsch, H., van der Hoek, W., and Kooi, B. *Dynamic Epistemic Logic*, Springer, 2007.

Yamada, T. 'Logical dynamics of some speech acts that affect obligations and preferences." *Snthese*, 165, Springer, 2008, pp. 295–315.

Title: A naturalistic approach to freedom and responsibility

Takuya Niikawa (Hokkaido University, PhD student), Riichiro Hira (National Institute of Basic Biology, Assistant Professor), Toshihiro Kotani (Kisarazu National College of Technology, Lecturer)

Contact details:
Takuya Niikawa: niitaku11@yahoo.co.jp
Riichiro Hira: wmtmt2@gmail.com
Toshihiro Kotani: tskotani@gmail.com

Abstract

What are the conditions to act freely? What are the conditions to attribute responsibility of an act to a subject? Cognitive neuroscience studies have suggested that various brain areas involve in decision making, and patients with lesions of some neural structures show impairment in rational decision making. These scientific findings suggest that in many cases of crimes, a psychiatric disorder causes the criminal act. This has led to problems in the applicability of naïve conceptions of responsibility and freedom. Suppose that a man killed a girl as a result of strong sexual desire. In the case where it has been found that there is a severe lesion in his brain area which is supposed to control sexual urges, should we judge that he is to be held responsible for the crime? Was he free when he committed the crime? How should we deal with cases in which kleptomania (inability to refrain from desire to steal items for reasons other than personal use or financial gain) or schizophrenia is associated with a crime? It seems that we do not have any robust intuition about how to answer such questions. In order to address such questions in an objective and convincing manner, it is likely necessary to form the new, sophisticated conceptions of freedom and responsibility. It seems to us that armchair considerations of freedom and responsibility are insufficient to provide such a conception. Instead, it can be constructed only through cooperative research by philosophy, neurophysiology, psychology, psychiatry, jurisprudence and other related fields. Our research project (it is called CORE-PhiB: "Cooperative Research on the Concept of Responsibility by Philosophy and Brain Science") aims to form the very conceptions of freedom and responsibility based on which we can provide an objective and convincing standard by appealing to which we can deal with difficult cases such as those mentioned above.

As the first step of this project, we try to naturalize the concept of free will (or freedom) in such a way to characterize the necessary-sufficient condition of a subject being free with respect to his/her action only in terms of cognitive-scientifically admissible properties such as neural properties. If such a characterization is successfully made, then we can arguably provide a scientifically-supported objective standard to determine whether a subject's action is done freely or not. However, one might wonder if the existence of free will is incompatible with the scientific view of the world. Others might wonder if cognitive neuroscience is essentially irrelevant to the philosophical disputes over free will. If either is the case, it seems in principle impossible to characterize the concept of free will in a naturalist way. The question to be asked here is, is the concept of free will such that we can naturalize it in the above sense? If the answer is in the affirmative, then the next question is this: how should we characterize it? This paper aims to address these questions.

There is a well-known philosophical problem called "the problem of free will" by Peter van Inwagen (2000; 2008). In order to understand how our attempt to naturalize the concept of free will is related to traditional philosophical disputes, it is helpful to start from the problem. Inwagen formulates the problem as follows:

> There are seemingly unanswerable arguments that (if they are indeed unanswerable) demonstrate that free will is incompatible with determinism. And there are seemingly unanswerable arguments that (if indeed …) demonstrate that free will is incompatible with indeterminism. But if free will is incompatible both with determinism and indeterminism, the concept "free will" is incoherent, and the *thing* free will does not exist. There are, moreover, seemingly unanswerable arguments that, if they are correct, demonstrate that the existence of moral responsibility entails the existence of free will, and, therefore, if free will does not exist, moral responsibility does not exist either. It is, however, evident that moral responsibility does not exist: if there were no such thing as moral responsibility nothing would be anyone's fault, and it is evident that there are states of affairs to which one can point and say, correctly, to certain people: "That is your fault." (Inwagen 2008, p.328)

He concludes that the point of the problem of free will is "to find out which of these seemingly unanswerable arguments is fallacious, and to enable us to identify the fallacy"

(2008, p.328). He confesses that he hasn't found out any satisfactory solution to this problem.

Manuel Vargas (2011; forthcoming) proposes an interesting solution to the problem. He insists that the fact that the concept of free will, as such, is incoherent does not mean that free will does not exist. It does not even mean that the concept of free will should be discarded. It just means that our commonsense thinking of free will contains some crucial errors. There is a possibility to *revise* (therefore not to discard) the concept of free will so as to remove the errors. As Inwagen pointed out, ordinary moral practices such as blaming or praising seem to presuppose the existence of free will. Such practices may be essential for our social life. Given this, it is inappropriate to easily dismiss the concept of free will. Thus, we should pursue the possibility of revising the concept of free will in a way to secure two points: (1) to keep the concept workable in relevant practices and (2) not to change the subject-matter from that which we intend to pick out by the original concept of free will. Roughly speaking, this view is what Vargas calls "revisionalist" and himself endorses.

His own revisionalist proposal is that free will is the capacity we have to recognize and respond to moral considerations. Based on this conception, Vargas (forthcoming) claims that the deepest problem that threatens our having free will is not a matter of high metaphysics, but rather the contexts in which we exercise our agency and the political challenges of structuring our environments to better support responsible agency. On the revisonalist view, we can see that free will is neither incompatible with determinism and indeterminism, because having a recognitional or responsive capacity is obviously irrelevant to both of the theses. In light of this, regardless of which thesis the scientific view of world favors, there seems no reason to think that free will is incompatible with the scientific view of world. Moreover, recognition of moral properties is one of the topics in cognitive neuroscience. Hence, it is plausible to think that the investigation of free will should be connected with cognitive neuroscience.

There are, however, at least two problems in Vargas's proposal. First, he does not consider how "the capacity we have to recognize and respond to moral considerations" can be realized in our brain. It is unclear how we can determine whether or not the capacity properly works in a given context. Second, more importantly, he does not sufficiently justify the idea that the concept of free will is essentially connected to moral practices. One may plausibly argue that it is begging the question to presuppose the

idea. For example, Inwagen holds that a subject has free will with respect to an act if and only if "we simultaneously have both the following abilities: the ability to perform that act and the ability to refrain from performing that act" (2008, p.329). In this definition, there is no reference to moral considerations. If this is the minimal concept of free will, the subject-matter changes by characterizing the concept of free will in terms of moral considerations. Therefore, it seems that revisionalists have to argue that the revised concept inherits the essential component of the original concept. What Vargas says is at best that some philosophers have the intuition that the concept of free will should be connected to moral practices. This is not enough to justify his revisionalist proposal.

We basically accept the revisionalist view that the concept of free will should be re-characterized. Moreover, we agree to his proposal in that free will should be identified with certain cognitive capacities or states. Nevertheless, there is an important methodological difference: we focus on the phenomenology of free will as a starting point of its conceptual characterization (Bayne 2008). We define the phenomenology of free will as what it is like to act freely. In typical cases of free action, we undergo a certain common distinctive phenomenology. It is plausible (or we argue so) that the presence of the phenomenology is essential to the concept of free will. Given this, it seems that we can grasp the concept of free will (or its essential parts) by considering when and by what mechanism we undergo the phenomenology. This claim is not begging the question, for this indicates *how we can grasp the essential component* of the concept of free will, rather than *what the component is*.

It is advantageous to naturalization of the concept of free will to begin with phenomenology since the phenomenology of mental states is generally regarded as supervening on brain states (content or mental state itself is more controversial). Here, we make the following assumptions: (1) the presence or absence of the distinctive phenomenology of free action is indirectly detectable via observation and introspective report, and (2) the neural basis of the phenomenology can be specified by neuroimaging technologies. Given these assumptions, we establish a working hypothesis that free will is nothing other than the cognitive capacities which is associated with the neural activities responsible for the phenomenology of free action. According to this hypothesis, a subject is free with respect to an action if and only if the cognitive capacities are properly exercised. This is the very naturalist characterization of free will. This hypothesis is evaluated by considering whether or not the cognitive capacities have the

role which free will is supposed to have in our various social practices. Although we plausibly predict that the capacities play the supposed role in moral practices, this is not a priori truth.

Of course, we need to engage in cognitive scientific research in order to specify the neural basis of the phenomenology of free will and then to clarify what cognitive capacities are associated with the neural activities. On our approach, it is indispensable to make experiments on not only animals but also *human beings* because they alone can report their phenomenal states. However, there are many methodological and ethical constraints on such experiments. How to direct such an experimental research is under consideration.

Reference list

Bayne, T. (2008) "The Phenomenology of Agency", *Philosophy Compass* 3 (1):182-202.
van Inwagen, P. (2000) "Free Will Remains a Mystery", *Philosophical Perspectives* 14:1-20.
van Inwagen, P. (2008) "How to Think About the Problem of Free Will", *Journal of Ethics* 12 (3/4):327 - 341.
Vargas, M. (2011) "Revisionist Accounts of Free Will: Origins, Varieties, and Challenges" In Robert Kane (ed.), *Oxford Handbook on Free Will, 2nd Edition*, Oxford UP.
Vargas, M. (forthcoming) "How to Solve the Problem of Free Will", In Paul Russell & Oisin Deery (eds.), *The Philosophy of Free Will*, Oxford UP.

Contemplating counterfactuals: On the connection between agency and metaphysical possibility

Sjur K. Dyrkolbotn [*1], Ragnhild H. Jordahl [†2], and Hannah A. Hansen [‡3]

[1]Durham Law School, Durham University, UK
[2]Department of Philosophy, University of Bergen, Norway
[3]Department of Information Science and Media Studies, University of Bergen, Norway

Extended Abstract

We consider the connection between the metaphysics of modality and agency, focusing on how it can be captured in logics for reasoning about multi-agent systems. We argue that philosophical insights can be gained from looking to these formalisms, and that they tend to come with implicit philosophical assumptions that we may consider to gain a better understand their broader meaning.

Indeed, social structures that have been designed with the aid of formal tools have become increasingly relevant to social reality, for both real and artificial agents.[1] Hence, philosophical assessment of logical tools appear especially relevant in this context. In addition, philosophy may offer interesting directions to pursue when developing these tools further.

In the full paper, we first argue that the connection between metaphysical modality and agency needs to be taken into account in order to arrive at a proper understanding of these notions. We observe, in particular, that agency appears to feature crucially in important metaphysical arguments concerning possibility, while metaphysical possibility seem to be at play in important arguments concerning agency.

Following up on this, we compare logics of agency with formal approaches to metaphysical theories of modality. We focus attention on branching time temporal logics, particularly variants of *alternating time temporal logic* (ATL) [Alur et al., 2002], and we relate these to the recently proposed *dispositional* account of modality, see [Borghini and Williams, 2008, Vetter, 2011]. The dispositional account makes the connection between possibility, causation and agency clearer at the philosophical level, so providing a formal interpretation of this theory seems like a particularly interesting research challenge.

Proposals have already emerged, giving a formal or semi-formal account of the dispositional theory. We note that these formalisms bear close resemblance to many logics considered in the theory of multi-agent systems and in the philosophy of agency, and we give a brief account of existing work, particularly [Jacobs, 2010, Vetter, 2010]. We go on to analyze the relationship between these approaches and related formalisms

[*]s.k.dyrkolbotn@durham.ac.uk
[†]ragnhild.jordahl@gmail.com
[‡]hannaha.hansen@gmail.com

[1]The growing importance of the social web over the last 10-15 years serves as an obvious example of this development.

from artificial intelligence. Moreover, by making the connection between metaphysical modality and agency explicit, we hope to shed new light on a number of well-known issues, both from philosophy and the theory of multi-agent systems.

Agency in the metaphysics of possibility

One of the main controversies in contemporary work on metaphysical modality arises from the tension between the theories of Lewis and Kripke respectively [Kripke, 1981, Kripke, 2005, Lewis, 1986, Lewis, 1971]. Both Lewis and Kripke build on the account given by Leibniz [Leibniz, 1998], who held that something is possible if and only it is true in some possible world, and necessary if and only it is true in all of them.

Lewis' theory relies on an ontology which posits the existence of concretely existing possible worlds, completely separated from our own. Kripke's theory, on the other hand, is based on an *actualistic* understanding of possible worlds; what actually exists is taken to be that which is part of our world, and all that is possible must, in principle, originate from this actuality.

It is commonly accepted that a powerful argument can be made against Lewis' theory by considering *identity* and *de re* modal claims, that is, claims about what is possible for a particular existing object, the identity of which we know in the actual world. How can it be, for instance, that something which is possible for *me* is witnessed by the existence of some other world, all the while I myself am part of this one? Lewis answers by saying that what is possible for me is witnessed by something which obtains in some possible world for someone who is not me, but is very much like me, namely my *counterpart* [Lewis, 1971].

This answer is held by many to be an affront to our intuitive understanding of modality. In a famous thought experiment [Kripke, 1981], Kripke makes this point by considering the possibility that Humphrey won the 1968 US presidential election. Why exactly would Humphrey care if someone very much like him won the election? Surely, when contemplating the possibility of victory, Humphrey is thinking about *himself*?

Kripke's argument, and the question of identity across possible worlds more generally, seems to owe much of its significance from considerations rooted in agency. For instance, we observe that modal agency, involving an agent contemplating the possible, is at the core of the Humphrey thought experiment. More generally, whenever a modal claim arises in real life this is invariably due to some agent engaging in modal reflection.[2] Moreover, when doing so, the agent is invariably embedded in structures that are present in physical and social reality, and his thoughts may in turn give rise to actions that can *change* these structures. It seems to us that a metaphysical theory of possibility had better take this into account.

The dispositional theory of metaphysical possibility

On the dispositional account, the possible is determined by dispositions found in the actual world; it remains rooted in this world, and we may describe modality as something that is *present* and *real* (i.e. not a phenomenon arising simply from the way we tend to use our language for example). To say that something is possible means

[2]That is not to say that modal agency subsumes or is constitutive of metaphysical possibility; this would involve excluding many possibilities that are often included in a metaphysical account, such as the possibility of a world with no agents (some may want their metaphysical theory to exclude this, but we prefer to remain agnostic about it). We are not, in particular, suggesting any kind of fictionalism about metaphysical possibilities, and the point we are making is not subsumed by previous work in this vein, as that of [Rosen, 1990, Rosen, 1995]. While agency should also be considered by such theories, their primary concern is with how possible worlds are to be made sense of, and how they come to be. This is not our topic; our argument is that *regardless* of what possible states of affairs are, it appears that how we *interact* with these in our social lives is relevant, also to the formulation of an appropriate metaphysical theory of possibility.

that there is some actual disposition for which this possibility — this possible state of affairs — is its manifestation. The (possible) manifestations can serve to characterize and individuate dispositions, but as dispositions themselves are actual, *they* determine what is metaphysically possible – what could possibly manifest – not the other way around. Hence we need not rely on possible worlds (real or metaphorical) as a primitive philosophical notion. Possible states of affairs can certainly be modeled formally as points in a directed graph – a powerful tool in modal logics – but according to the dispositional account this does not require us to commit to any particular position regarding possible worlds, not even their existence. Rather, possible states of affairs can be *traced back* to their origin in actuality, and while they have rich internal structure, this structure arises from how they could have come about, so that the discourse of possible worlds can remain entirely metaphorical without challenging the actual existence of metaphysical modalities.[3]

It is important to emphasize that dispositions always trace back to properties of objects present in the world here and now. New dispositions do not spontaneously appear along any (counterfactual) future time-lines, and all possibilities result from the possible manifestations of existing dispositions. Still, *higher-order* dispositions might need to be considered, i.e., dispositions that are merely possible, but which arise from manifestations of dispositions that are always closer – in a chain of possible manifestations – to dispositions existing in the actual world, see [Borghini and Williams, 2008].

The manifestations of dispositions might not come about, and objects tend to have many dispositions that will never materialize. Think of the glass that has the dispositional property of being fragile — this means that the glass will break if struck with sufficient force, but this disposition to break might very well never become actual. But even if the dispositions are never manifested, the existence of dispositional properties is enough to account for the possibility that the glass *might* break or that it *could have been* broken.

The connection between agency and dispositions can be elucidated by considering the term *powers*. It is used in the philosophy of causation, often as a synonym for dispositions [Mumford and Anjum, 2011b], but also in the philosophy of agency, where it has a different, but related, meaning [van Inwagen, 1983]. Roughly speaking, a power can be seen as a disposition involving agency by way of pointing to an ability that an agent has to bring about an outcome. In the example above, one might say of the glass that it is disposed to break, but one might also say of an agent that he has the power to break it. It seems wrong, however, to say that he is disposed to do so, simply because he can.

The distinction could be useful for a dispositional theory of possibility. If someone claims "it is possible for me to break the glass", it seems that the disposition of the glass to break if he hits it is no longer a sufficient truthmaker for this claim. What if we consider a world where this person does not exist, or he is necessarily prevented from hitting the glass for some other reason? Then it seems natural to also make reference to his power to hit the glass, not only the dispositional fact that it might break if he does so.

The interrelated nature of powers and dispositions is further underlined by the observation that mathematically speaking, the formal frameworks used in [Jacobs, 2010, Vetter, 2010] to study objects and their dispositions are strikingly similar to logics used to study agents and their actions in the theory of multi-agent systems. This observation is the starting point for our technical project, which aims to give an account of the dispositional theory, as well as the connection to agency, by means of multi-agent logics.

[3] We point to [Vetter, 2011] for a survey of recent work on dispositions and possibility

Agency and metaphysical possibility in formal logics

There is a vast landscape of formal logics that involve agency and possibility, and increasingly, these notions are also considered together, especially in logics for modeling interaction in a multi-agent system, see [Wooldridge, 2009, van Benthem, 2011].

In the Humphrey thought experiment, Humphrey knew he lost the election in 1968, but he was still free to contemplate the possibility of a different outcome. Moreover, by contemplating this possibility he was engaging in a form of metaphysical agency that does not in general appear reducible to other forms.[4] Towards a formal representation, we suggest turning to *multi-modal* logics, allowing us to study interactions between a modality representing metaphysical possibility, and another, distinct modality, which can be used for talking about agency involving reflection concerning such possibilities.[5] In this paper, we will focus on multi-modal logics that are based on a branching time notion of possibility. Such logics have attracted much interest, both in philosophy and AI, and they are particularly interesting because they have been extended in various ways by adding modal operators specifically directed at modeling agency. We point to [Belnap and Perloff, 1988, Horty and Belnap, 1995, van der Hoek and Wooldridge, 2003, Ågotnes et al., 2009, Broersen, 2011b] for a collection of work on such formalisms that is relevant to the points we are making in this paper.

We consider reinterpretations of the branching time formalisms, viewing transitions between states as resulting from the (possibly counterfactual) manifestations of dispositions. The temporal dimension can then be understood as modeling higher order manifestations of dispositions.[6]

In the full paper, we take this point of view further, suggesting that the study of such systems and the connections between them has the potential to shed light on a number of different, but related, questions, such as the relationship between free will and determinism [List, 2013, Strawson, 1962], the workings of higher order dispositions [Borghini and Williams, 2008], the applicability of notions involving moral responsibility [Frankfurt, 1969, Broersen, 2011a], the nature of necessity and the question of whether or not dispositional possibility is a distinct form of modality [Mumford and Anjum, 2011, Fine, 1994, Fine, 1995], and the distinction between knowing that it is possible to do something, and actually knowing *how* to do it [Jamroga and van der Hoek, 2004, Jamroga and Ågotnes, 2006].

The primary aim is to make a methodological point: since all of these questions involve the relationship between agency and metaphysical possibility, more work should be devoted to studying them in this light. By suggesting a formal interpretation of the dispositional theory we hope to make a convincing argument for the soundness of this research project.

[4] For instance, while such agency might be intimately related to future and past actions (and attitudes towards such actions), it need not be directed at any specific ones. Hence it does not seem possible, in general, to account for it in terms of causal decision theory, see [Joyce, 1999] for a presentation of this theory. Also note that agents may contemplate far fetched scenarios leading to highly concrete effects in the actual world. Consider, for instance, the agent who considers the possibility of cloning dinosaurs, and then forms the goal of going to see Jurassic Park at the cinema. Such agency, in particular, appears to be genuinely interacting with a metaphysical notion of possibility.

[5] Multi-modal logics is a rich topic which is being studied from many different angles and it attracts much technical interest, see [Kurucz et al., 2003].

[6] We mention that a related development, that also argues for the metaphysical importance of branching time possibility is presented in [Müller, 2012]. Here, however, the suggestion is made that branching time possibility is in itself metaphysically basic, in that it gives rise to the *real* notion of metaphysical possibility, which, albeit not as wide as that usually considered, is still wide enough to cover the interesting cases, including those that deserve primary attention in metaphysics.

References

[Ågotnes et al., 2009] Ågotnes, T., van der Hoek, W., and Wooldridge, M. (2009). Robust normative systems and a logic of norm compliance. *Logic Journal of the IGPL*, 18(1):4–30.

[Alur et al., 2002] Alur, R., Henzinger, T., and Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM (JACM)*, 49(5):672–713.

[Belnap and Perloff, 1988] Belnap, N. and Perloff, M. (1988). Seeing to it that: A canonical form for agentives. *Theoria*, 54(3):175–199.

[Borghini and Williams, 2008] Borghini, A. and Williams, N. E. (2008). A dispositional theory of possibility. *Dialectica*, 62(1):21–8211.

[Broersen, 2011a] Broersen, J. (2011a). Deontic epistemic stit logic distinguishing modes of mens rea. *J. Applied Logic*, 9(2):137–152.

[Broersen, 2011b] Broersen, J. (2011b). Making a start with the stit logic analysis of intentional action. *Journal of Philosophical Logic*, 40(4):499–530.

[Fine, 1994] Fine, K. (1994). Essence and modality. *Philosophical Perspectives*, 8:1–16.

[Fine, 1995] Fine, K. (1995). The logic of essence. *Journal of Philosophical Logic*, 24(3):241–273.

[Frankfurt, 1969] Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(3):829–39.

[Horty and Belnap, 1995] Horty, J. F. and Belnap, N. (1995). The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644.

[Jacobs, 2010] Jacobs, J. D. (2010). A powers theory of modality: or, how I learned to stop worrying and reject possible worlds. *Philosophical Studies*, 151:227–248.

[Jamroga and Ågotnes, 2006] Jamroga, W. and Ågotnes, T. (2006). What agents can achieve under incomplete information. In Stone, P. and Weiss, G., editors, *Proc. of the Fifth Intern. Joint Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 232–234. ACM Press.

[Jamroga and van der Hoek, 2004] Jamroga, W. and van der Hoek, W. (2004). Agents that know how to play. *Fundamenta Informaticae*, 63:185–219.

[Joyce, 1999] Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

[Kripke, 1981] Kripke, S. (1981). *Naming and Necessity*. Blackwell Publishing.

[Kripke, 2005] Kripke, S. (2005). Identity and necessity. In Loux, M. J., editor, *Metaphysics - Contemporary Readings*. Routledge.

[Kurucz et al., 2003] Kurucz, A., Wolter, F., Zakharyaschev, M., and Gabbay, D. M. (2003). *Many-Dimensional Modal Logics: Theory and Applications*, volume 148 of *Studies in Logic and the Foundations of Mathematics*. Elsevier.

[Leibniz, 1998] Leibniz, G. (1998). *Theodicy*. Open Court. First published in 1709.

[Lewis, 1971] Lewis, D. (1971). Counterparts of persons and their bodies. *Journal of Philosophy*, 68(7).

[Lewis, 1986] Lewis, D. (1986). *On the plurality of worlds*. Oxford University Press.

[List, 2013] List, C. (2013). Free will, determinism, and the possibility of doing otherwise. *Noûs*, 47(2).

[Müller, 2012] Müller, T. (2012). Branching in the landscape of possibilities. *Synthese*, 188:41–65.

[Mumford and Anjum, 2011] Mumford, S. and Anjum, R. L. (2011). Dispositional modality. In *Lebenswelt und Wissenschaft, Deutsches Jahrbuch Philosophie 2*. Meiner Verlag.

[Rosen, 1990] Rosen, G. (1990). Modal fictionalism. *Mind*, 99(395):327–354.

[Rosen, 1995] Rosen, G. (1995). Modal fictionalism fixed. *Analysis*, 55(2):67–73.

[Strawson, 1962] Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48:1–25.

[van Benthem, 2011] van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

[van der Hoek and Wooldridge, 2003] van der Hoek, W. and Wooldridge, M. (2003). Cooperation, knowledge and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75:125–157.

[Vetter, 2010] Vetter, B. (2010). *Potentiality and possiblity*. University of Oxford. PhD thesis.

[Vetter, 2011] Vetter, B. (2011). Recent work: Modality without possible worlds. *Analysis*, 71(4):742–754.

[Wooldridge, 2009] Wooldridge, M. (2009). *An Introduction to Multiagent Systems*. John Wiley & Sons, Inc., 2 edition.

Interactive Incompleteness for Prediction/Decision Making in Games

Tai-Wei Hu and Mamoru Kaneko

Abstract: Logical inference is an engine for human thinking, especially for decision making in an interactive situation with more than one person. We study the possibility of prediction/decision-making following Nash-Johansen noncooperative solution theory in a fixed-point extension of epistemic logic KD^2 to capture infinite regresses arising from prediction-decision making. This extension is shown to be Kripke-complete. Our main results show a sharp contrast between determinacy of final decisions and predictions in solvable games and the incompleteness result for unsolvable games, stating that a player can neither reach a decision nor disprove it as a decision. Behind this incompleteness result, we find the conflict between the reciprocity of prediction/decision making and independence of the two players' minds. We also show that the incompleteness result leads to a no-formula theorem for unsolvable games.

Tai-Wei Hu: Kellogg School of Management, Northwestern University, t-hu@kellogg.northwestern.edu
Mamoru Kaneko: Faculty of Political Sciences and Economics, Waseda University, mkanekoepi@waseda.jp

Compliance Games (extended abstract)

Piotr Kaźmierczak

Dept. of Computing, Mathematics and Physics
Bergen University College, Norway
phk@hib.no

Normative systems or *social laws*[1] have been studied in the multi-agent systems community as a framework for coordinating players [7,3,2,4,8]. The idea is that given a Kripke structure (which is simply a directed graph, sometimes labelled with extra elements, such as players), we list the edges that are *blacklisted*, i.e. deemed illegal by the system designer. The set of these edges is called a *normative system* or *social law*.

One of the key problems in the literature is that of *compliance* with a given social law. Ågotnes et al. [3] presented a logical approach to the problem by designing a logic of norm compliance which allows to reason about agents' goal achieving capabilities depending on whether they comply with given laws or not.

Here we present a new approach where agents are presented with a cooperative game in which successful coalitions are those that comply with the norm, and those not complying are punished (by means of null payoffs). We adapt a specific type of non-transferable utility game called Qualitative Coalitional Game to model compliance game scenarios, and we show how known decision problems for this class of games can be used to reason about compliance.

We begin by concisely presenting all the formal background for our work. First we describe the logical framework for Social Laws, which is based on Kripke semantics for modal logics. Following Ågotnes et al. [3], we define our models as agent-labelled Kripke structures in the following way:

Definition 0.1 (Agent-labelled Kripke Structure). *An* agent-labelled Kripke structure *(henceforth referred to simply as* structure*) K is a tuple $\langle S, R, V, \Phi, A, \alpha\rangle$ where:*

- *S is the non-empty, finite set of* states*,*
- *$R \subseteq S \times S$ is the total*[2] *relation between elements of S that captures* transitions *between states,*
- *Φ is a non-empty, finite set of propositional symbols,*
- *$V : S \to 2^{\Phi}$ is a labelling function which assigns propositions to states in which they are satisfied,*

[1] In the multi-agent systems literature normative systems and social laws stand for the same. However, in this paper we will always use the notion of a social law, since calling our restrictions "normative systems" can perhaps be confusing for readers with a background in deontic logic.

[2] That is, $\forall s \exists t\ (s,t) \in R$. This kind of relation is also sometimes called serial.

– *A is a non-empty finite set of agents, and*
– $\alpha : R \to A$ *a function that labels edges with agents.*

A *path* π over a relation R is an infinite sequence of states $s_0, s_1, s_2, \ldots$ such that $\forall u \in \mathbb{N} : (s_u, s_{u+1}) \in R$. $\pi[0]$ denotes the first element of the sequence, $\pi[1]$ the second, and so on. An *s-path* is a path π such that $\pi[0] = s$. $\Pi_R(s)$ is the set of s-paths over R, and we write $\Pi(s)$, if R is clear from the context.

Objectives are specified using the language of *Computation Tree Logic* (CTL), a popular branching-time temporal logic. We use an adequate fragment of the language defined by the following grammar:

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathsf{E}\bigcirc \varphi \mid \mathsf{E}(\varphi \mathcal{U} \varphi) \mid \mathsf{A}(\varphi \mathcal{U} \varphi)$$

where p is a propositional symbol. The standard derived propositional connectives are used, in addition to standard derived CTL connectives such as $\mathsf{A}\bigcirc \varphi$ for $\neg\mathsf{E}\bigcirc \neg\varphi$ (see [6] for details). Satisfaction of a formula φ in a state s of a structure K, $K, s \models \varphi$, is defined as follows:

$$\begin{aligned}
&K, s \models \top;\\
&K, s \models p \text{ iff } p \in V(s);\\
&K, s \models \neg\varphi \text{ iff not } K, s \models \varphi;\\
&K, s \models \varphi \vee \psi \text{ iff } K, s \models \varphi \text{ or } K, s \models \psi;\\
&K, s \models \mathsf{E}\bigcirc \varphi \text{ iff } \exists \pi \in \Pi(s) : K, \pi[1] \models \varphi;\\
&K, s \models \mathsf{E}(\varphi \mathcal{U} \psi) \text{ iff } \exists \pi \in \Pi(s), \exists u \in \mathbb{N}, \text{s.t. } K, \pi[u] \models \psi\\
&\qquad\qquad \text{and } \forall v, (0 \leq v < u) : K, \pi[v] \models \varphi;\\
&K, s \models \mathsf{A}(\varphi \mathcal{U} \psi) \text{ iff } \forall \pi \in \Pi(s), \exists u \in \mathbb{N}, \text{s.t. } K, \pi[u] \models \psi\\
&\qquad\qquad \text{and } \forall v, (0 \leq v < u) : K, \pi[v] \models \varphi.
\end{aligned}$$

A *social law* $\eta \subseteq R$ is a set of black-listed ("illegal") transitions, such that $R \setminus \eta$ remains total.[3] A set of all social laws over R is denoted as $N(R)$. We say that $K \dagger \eta$ is a structure with a social law η *implemented* on it, i.e. for $K = \langle S, R, \Phi, V, A, \alpha\rangle$ and η, $K \dagger \eta = K'$ iff $K' = \langle S, R', \Phi, V, A, \alpha'\rangle$ with $R' = R \setminus \eta$ and:

$$\alpha'(s, s') = \begin{cases} \alpha'(s, s') & \text{if } (s, s') \in R' \\ \text{undefined} & \text{otherwise} \end{cases}$$

Also, $\eta \restriction C = \{(v, v') | (v, v') \in \eta \ \&\ \alpha(v, v') \in C\}$ for any $C \subseteq A$ – that is to account for agents that do not necessarily comply with the social law (i.e. we can consider situation in which only those edges that are "owned" by members of C are blacklisted).

Example 0.1. We introduce a running example that illustrates modeling a very simple Kripke structure.

Figure 1 presents and example Kripke structure with:

[3] This is a so-called "reasonableness" constraint – we do not want social laws implementation of which results in systems with dead-end states.

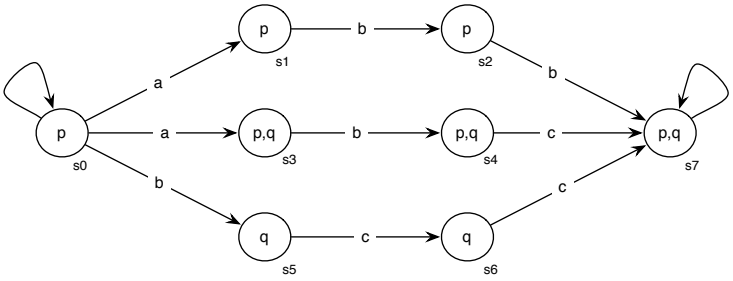

Fig. 1. Simple Kripke structure example.

- $S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$,
- $R = \{(s_0, s_0), (s_0, s_1), (s_1, s_2), (s_2, s_7), \ldots\}$,
- $\Phi = \{p, q\}$,
- $V(s_0) = \{p\}$, $V(s_1) = \{p\}$, $\ldots$, $V(s_7) = \{p, q\}$,
- $A = \{a, b, c\}$,
- $\alpha(s_0, s_1) = \alpha(s_0, s_3) = \{a\}$,
  $\alpha(s_1, s_2) = \alpha(s_3, s_4) = \alpha(s_0, s_5) = \alpha(s_2, s_7) = \{b\}$,
  $\alpha(s_4, s_7) = \alpha(s_5, s_6) = \alpha(s_6, s_7) = \{c\}$.

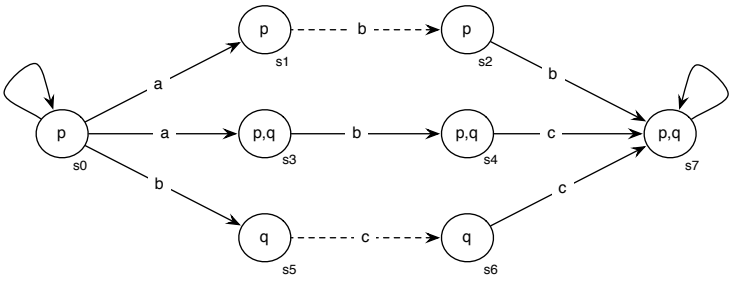

Fig. 2. Kripke structure with a social law implemented on it.

Figure 2 presents the same structure, but with a social law $\eta = \{(s_1, s_2), (s_5, s_6)\}$ implemented on it.

For the purpose of modelling compliance games we employ a formalism known as *Qualitative Coalitional Games* [9]. These are non-transferable utility cooperative games, which are particularly suitable for modelling goal-based scenarios. They are called "qualitative" because in contrast to most cooperative games, where a characteristic function assigns a numeric value (usually a real number) to each coalition, these games' characteristic functions assign a "good" or "bad" value to coalitions. Formal definitions follow below:

Definition 0.2 (Qualitative Coalitional Game). *A Qualitative Coalitional Game (abbreviated* QCG*) Γ is given by a tuple $\Gamma = \langle A, \Theta, \Theta_1, \ldots, \Theta_n, \nu \rangle$, where A is a finite set of players, $\Theta = \{\theta_1, \ldots, \theta_m\}$ is a set of* goals*, $\Theta_i \subseteq \Theta$ is a set of goals* per player*, and $\nu : 2^A \to 2^{2^\Theta}$ is a* characteristic function *of the game, assigning to each coalition $C \subseteq A$ a set of goals this coalition can achieve.*

We now use the above definition to formulate a particular kind of cooperative game, where agents are "rewarded" for complying to a social law (the characteristic function of the game pays coalitions whose members achieve at least one of their respective goals).

Definition 0.3 (Compliance Game). *A* compliance game *(abbreviated* CG*) is a* QCG *$\Gamma = \langle A, \Theta, \Theta_1, \ldots, \Theta_n, \nu \rangle$ induced by a Kripke structure $K = \langle S, R, \Phi, V, A, \alpha \rangle$ and defined as follows:*

- *A is the set of players from K,*
- *$\Theta = \{\varphi_1, \ldots, \varphi_m\}$ is a set of* goal formulas *expressed in the language of* CTL*,*
- $\nu(C) = \begin{cases} \{\varphi_1, \ldots, \varphi_m\} & \text{if } K \dagger (\eta \restriction C) \models \varphi_i \\ \emptyset & \text{otherwise.} \end{cases}$

Example 0.2. Take a look at Figure 1 first. It is implicitly a Kripke structure with an empty social law implemented on it (all transitions are legal), and then we could formulate the following CG based on it:

- $\Theta = \{\varphi_1, \varphi_2\}$,
- $\Theta_a = \Theta_b = \{\varphi_1\}$, $\Theta_c = \{\varphi_2\}$.

where $\varphi_1 = \mathsf{E}\Diamond p$ and $\varphi_2 = \mathsf{E}\Diamond q$.[4] In the structure from Figure 1, the following coalition are successful: $\{a, b\}$, $\{b, c\}$, $\{a, b, c\}$. However, in a structure presented in Figure 2, only the grand coalition is successful.

We now make some observations about what the power of Qualitative Coalitional Games can bring to the area of social laws. This is mostly based on the study of fourteen intuitive decision problems for those games studied by Wooldridge & Dunne [9], and the natural problems when reasoning about social laws are taken from the work of Ågotnes et al., especially [3].

The first problem is that of *C-sufficiency* – namely checking whether given a Kripke structure K, a social law η and a goal formula φ, a coalition C of agents in K are *sufficient* for achieving φ under η. Deciding *C-sufficiency* is proved to be co-NP-complete [3]. We can make an observation that given the framework defined as above, C-sufficiency can be represented as an instance of a game-theoretic decision problem known as *successful coalition* (SC) or *selfish successful coalition* (SSC). SC asks whether given a particular QCG and some coalition, is there a feasible choice available such that it will satisfy all members

[4] $\mathsf{E}\Diamond \equiv \mathsf{E}(\top \mathcal{U} \varphi)$, a standard abbreviation in CTL to denote the "at some point in the future" modality.

of said coalition. The problem of SSC is then to answer the question of whether it is the case that a coalition in question has a feasible choice that will satisfy a given agent only if he is part of this coalition. These two problems answer more general questions than C-sufficiency, but capture similar intuitions. Also, these two problems are shown to be NP-complete [9].

The second important problem is *C-necessity*, which, similarly to *C-sufficiency*, asks whether a given coalition is necessary to achieve a given goal under certain restrictions implemented upon a structure. This problem is also co-NP-complete [3], and it corresponds directly to the game-theoretic problem of finding a *minimal coalition* (MC), which happens to be co-NP-complete [9] as well.

We do not address problems of *C-sufficient feasibility* (checking whether there is a social law such that a given coalition C will be sufficient when this system is implemented) and *k-robustness* (checking whether a given social law is effective as long as number k of agents complies with it), but we can exploit game-theoretic nature of our formulation of social law compliance problems. Given the framework of QCGs we can answer many decision problems that are very interesting from the point of view of a social law designer. One of such problems is *core membership* and *core non-emptiness* – checking whether there are such outcomes of the game in which agents have no incentive to abandon a certain coalition structure. Knowing one has a non-empty core can be very useful, because then the system designer knows that a set of stable coalitions will emerge. And even if it does not, game theory literature offers some solutions to ensure a non-empty core, such as *cost of stability* [5] (paying subsidies to agents such that they remain in a coalition). Furthermore, with a game-theoretic framework defined as above, we can formulate questions about the nature of the game itself: e.g. is it *trivial* (every coalition is successful) or *empty* (every coalition fails)?, and about the nature of goals: are they realizable, are some of them necessary?

Finally, we are able to address one other problem mentioned in the literature, namely finding "influential" players and measuring their power by means of computing the Banzhaf index and Banzhaf measure, as it is done in [1].

In conclusion, we claim that the framework presented above is more powerful, more flexible, and more appropriate for the study of social laws than the work that has been presented in the literature until now. We plan to further study it, with an emphasis on computational properties of a number of decision problems sketched in the paragraph above.

References

1. T. Ågotnes, W. van der Hoek, M. Tennenholtz, and M. Wooldridge. Power in normative systems. In *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 145–152, 2009.
2. T. Ågotnes, W. van der Hoek, and M. Wooldridge. Normative system games. In *Proc. of 6th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2007)*, pages 1–8, 2007.

3. T. Ågotnes, W. van der Hoek, and M. Wooldridge. Robust normative systems and a logic of norm compliance. *Logic Journal of the IGPL*, 18(1):4–30, 2009.
4. T. Ågotnes, W. van der Hoek, and M. Wooldridge. Conservative social laws. In L. D. R. et al., editor, *Proceedings of ECAI 2012*, 2012.
5. Y. Bachrach, E. Elkind, R. Meir, D. Pasechnik, M. Zuckerman, J. Rothe, and J. S. Rosenschein. The cost of stability in coalitional games. In *Proc. of the 2nd Int. Symposium on Algorithmic Game Theory (SAGT '09)*, pages 122–134, Berlin, Heidelberg, 2009. Springer-Verlag.
6. E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science Volume B: Formal Models and Semantics*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1990.
7. Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73:231–252, 1995.
8. W. van der Hoek, M. Roberts, and M. Wooldridge. Social laws in alternating time: effectiveness, feasibility, and synthesis. *Synthese*, 156(1):1–19, Sept. 2006.
9. M. Wooldridge and P. E. Dunne. On the Computational Complexity of Qualitative Coalitional Games. *Artificial Intelligence*, 158, 2004.

Strategic Reasoning in Extensive Games with Short Sight

Chanjuan Liu[1], Fenrong Liu[2], and Kaile Su[1,3]

[1] School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2] Department of Philosophy, Tsinghua University, Beijing, China
[3] Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, Australia
chanjuan.pkucs@gmail.com, fenrong@tsinghua.edu.cn, kailepku@gmail.com

1 Introduction

To characterize the structures and reason about strategies of extensive games, much work has been done to provide the logical systems for such games. These logic systems focus on various perspectives of extensive games: (Harrenstein *et al.*, 2003) concentrated on describing equilibrium concepts and strategic reasoning. (van Benthem, 2002) used dynamic logic to describe games as well as strategies.

The assumption of common knowledge on game structures in traditional extensive games is sometimes too strong and unrealistic. For instance, in a game like chess, the actual game space is exponential in the size of the game configuration, and may have a computation path too long to be effectively handled by most existing computers. So we often seek sub-optimal solutions by considering only limited information or bounded steps foreseeable by a player that has relatively small amount of computation resources. Grossi and Turrini proposed the concept of *games with short sight* (Grossi and Turrini, 2012), in which players can only see part of the game tree. However, there is no work on the logical reasoning of the strategies in this game model.

Inspired by the previous logics for extensive games, this paper is devoted to the logical analysis of game-theoretical notions of the solutions concepts in games with short sight. The closely related work is (Harrenstein *et al.*, 2003), in which a logic was proposed for strategic reasoning and equilibrium concepts. In this work, however, we present a new logical system called LS for games with short sight. This logic introduces new additional modalities $[\lhd], [(\sigma_i)], [\mathring{\sigma}^s]$ to capture interesting features such as restricted sight and limited steps.

2 Preliminaries

In this section, we recall the definition of finite games in extensive form with perfect information and games with short sight proposed by (Grossi and Turrini, 2012).

Definition 1. (*Extensive game(with perfect information)*) *A finite extensive game (with perfect information) is a tuple* $G=(N, V, A, t, \Sigma_i, \succeq_i)$*, where* (V, A) *is a tree with* V*, a set of nodes or vertices including a root* v_0*, and* $A \subseteq V^2$ *a set of arcs.* N *is a non-empty set of the players, and* $\succeq_i$ *represents preference relation for each player* i*, which is a partial order over* V*. For any two nodes* v *and* v'*, if* $(v, v') \in A$*, we call* v' *a* successor *of* v*, thus* A *is also regarded as the successor relation. Leaves are the nodes that have no successors, denoted by* Z*.* t *is turn function assigning a member of* N *to each non-terminal node.* Σ_i *is a non-empty set of strategies. A strategy of player* i *is a function* $\sigma_i : \{v \in V \backslash Z |\ t(v) = i\} \to V$ *which assigns a successor of* v *to each non-terminal node when* $t(v) = i$*.*

As usual, $\sigma = (\sigma_i)_{i \in N}$ represents a strategy profile which is a combination of strategies of all players and Σ represents the set of all strategy profiles. For any $M \subseteq N$, σ_{-M} denotes the collection of strategies in σ excluding those for players in M. We define an outcome function $O : \Sigma \to Z$ assigning leaf nodes to strategy profiles, i.e., $O(\sigma)$ is the outcome if the strategy profile σ is followed by all players. $O(\sigma_{-M})$ is the set of outcomes players in M can enforce provided that the other players strictly follow σ. $O(\sigma'_i, \sigma_{-i})$ is the outcome if player i uses strategy σ' while all other players employ σ.

In games with short sight, players' available information is limited in the sense that they are not able to see the nodes in some branches of the game tree or have no access to some of the terminal nodes.

Definition 2. (*sight function*)*. Let* $G = (N, V, A, t, \Sigma_i, \succeq_i)$ *be an extensive game. A short sight function for* G *is a function* $s : V \backslash Z \to 2^{V|v} \backslash \emptyset$*, associating to each non-terminal node* v *a finite subset of all the available nodes at* v*, and satisfying:*

$v' \in s(v)$ *implies that* $v'' \in s(v)$ *for every* $v'' \lhd v'$ *with* $v'' \in V|_v$*, i.e. players' sight is closed under prefixes.*($\lhd$ *is the transitive closure of successor relation* A*.*)

Intuitively, function s associates any choice point with vertices that each player can see.

Definition 3. (*Extensive game with short sight*)*. An extensive game with short sight* (*Egss*) *is a tuple* $S = (G, s)$ *where* G *is a finite extensive game and* s *a sight function for* G*.*

Each game with short sight yields a family of finite extensive games, one for each non-terminal node $v \in V \setminus Z$:

Definition 4. (*sight-filtrated extensive game*) *Let* S *be an Egss given by* (G, s) *with* $G=(N, V, A, t, \Sigma_i, \succeq_i)$*. Given any non-terminal node* v*, a tuple* $S\lceil_v$ *is a finite extensive game by sight-filtration:* $S\lceil_v = (N\lceil_v, V\lceil_v, A\lceil_v, t\lceil_v, \Sigma_i\lceil_v), \succeq_i\lceil_v)$ *where*

- $N\lceil_v = N$*;*
- $V\lceil_v = s(v)$*, which is the set of nodes within the sight from node* v*. The terminal nodes in* $V\lceil_v$ *are the nodes in* $V\lceil_v$ *of maximal distance, denoted by* $Z\lceil_v$*;*
- $A\lceil_v = A \cap (V\lceil_v)^2$*;*
- $t\lceil_v = V\lceil_v \backslash Z\lceil_v \to N$ *so that* $t\lceil_v(v') = t(v')$*;*

– $\Sigma_i\lceil_v$ *is the set of strategies for each player available at* v *and restricted to* $s(v)$*. It consists of elements* $\sigma_i\lceil_v$ *such that* $\sigma_i\lceil_v(v') = \sigma_i(v')$ *for each* $v' \in V\lceil_v$ *with* $t\lceil_v(v') = i$*;*
– $\succeq_i\lceil_v = \succeq_i \cap (V\lceil_v)^2$.

Accordingly, we define the outcome function $O\lceil_v$: $\Sigma\lceil_v \to Z\lceil_v$ assigning leaf nodes of $S\lceil_v$ to strategy profiles.

3 A Logic of Extensive Games with Short Sight

3.1 $\mathcal{LS}$: Syntax and Semantics

Let P be the set of propositional variables, and Σ be the set of strategy profiles. The language $\mathcal{LS}$ is given by the following BNF:

$\varphi ::= p \mid \neg\varphi \mid \varphi_0 \wedge \varphi_1 \mid \langle\leq_i\rangle\varphi \mid \langle\mathring{\sigma}\rangle\varphi \mid \langle\mathring{\sigma}_{-i}\rangle\varphi \mid \langle\lhd\rangle\varphi \mid \langle\mathring{\sigma}^s\rangle\varphi \mid \langle\mathring{\sigma}^s_{-i}\rangle\varphi$

where $p \in P$, $\sigma \in \Sigma$. As usual, The dual of $\langle.\rangle\varphi$ is $[.]\varphi$.

Let $S = (N, V, A, t, \Sigma_i, \succeq_i, s)$ be an Egss. The tuple of $(V, R_{\leq_i}, R_{\mathring{\sigma}}, R_{\mathring{\sigma}_{-i}}, R_{\lhd}, R_{\mathring{\sigma}^s}, R_{\mathring{\sigma}^s_{-i}})$ is defined as the frame F_S for $\mathcal{LS}$, where for each player i, strategy profile σ, nodes v, v', the accessibility relations are given as follows.

$vR_{\leq_i}v'$ iff $v' \succeq_i v$
$vR_{\mathring{\sigma}}v'$ iff $v' = O|_v(\sigma|_v)$
$vR_{\mathring{\sigma}_{-i}}v'$ iff $v' \in O|_v(\sigma_{-i}|_v)$
$vR_{\lhd}v'$ iff $v' \in s_{t(v)}(v)$
$vR_{\mathring{\sigma}^s}v'$ iff $v' = O\lceil_v(\sigma\lceil_v)$
$vR_{\mathring{\sigma}^s_{-i}}v'$ iff $v' \in O\lceil_v(\sigma_{-i}\lceil_v)$

A model M for $\mathcal{LS}$ is a pair (F, π) where F is a frame for $\mathcal{L}$ and π a function assigning to each proposition p in P a subset of V, i.e., $\pi : P \to 2^V$. The interpretation for $\mathcal{LS}$ formulas in model M are defined as follows:

$M, v \models \langle\leq_i\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\leq_i}u$.
$M, v \models \langle\mathring{\sigma}\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\mathring{\sigma}}u$.
$M, v \models \langle\mathring{\sigma}_{-i}\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\mathring{\sigma}_{-i}}u$.
$M, v \models \langle\lhd\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\lhd}u$.
$M, v \models \langle\mathring{\sigma}^s\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\mathring{\sigma}^s}u$.
$M, v \models \langle\mathring{\sigma}^s_{-i}\rangle\varphi$ iff $M, u \models \varphi$ for some $u \in V$ with $vR_{\mathring{\sigma}^s_{-i}}u$.

The validities of a formula φ in models and frames are the same as the standard definitions (Blackburn *et al.*, 2001).

3.2 Axiom system

First, we have the following standard axioms.

(A_0) *Taut*, any classical tautology.

(A_1) K axiom for all modalities $[\leq_i], [\mathring{\sigma}], [\mathring{\sigma}_{-i}], [\lhd], [\mathring{\sigma}^s], [\mathring{\sigma}^s_{-i}]$.

Table 1 lists the other axioms of LS. The first column (N) is the *name* of the axiom. The second column denotes the *modalities* that each axiom is applied to.

N	Modality	Schema	Property
T	$[\leq_i]$ $[\lhd]$	$[\leq_i]\varphi \to \varphi$ $[\lhd]\varphi \to \varphi$	reflexivity
4	$[\leq_i]$	$[\leq_i]\varphi \to [\leq_i][\leq_i]\varphi$	transtivity
D	$[\mathring{\sigma}]$ $[\mathring{\sigma}^s]$	$[\mathring{\sigma}]\varphi \leftrightarrow \langle\mathring{\sigma}\rangle\varphi$ $[.]\varphi \leftrightarrow \langle.\rangle\varphi$	determinism
I	$([\mathring{\sigma}], [\mathring{\sigma}_{-i}])$ $([\mathring{\sigma}^s], [\mathring{\sigma}^s_{-i}])$	$[\mathring{\sigma}_{-i}]\varphi \to [\mathring{\sigma}]\varphi$ $[\mathring{\sigma}_{-i}]\varphi \to [\mathring{\sigma}]\varphi$	inclusiveness
M	$[\mathring{\sigma}]$ $[\mathring{\sigma}_{-i}]$	$[\mathring{\sigma}]([\mathring{\sigma}']\varphi \leftrightarrow \varphi)$ $[\mathring{\sigma}_{-i}]([\mathring{\sigma}'_{-i}]\varphi \leftrightarrow \varphi)$	terminating
Y	$([\lhd], [\mathring{\sigma}^s])$ $([\lhd], [\mathring{\sigma}^s_{-i}])$	$[\lhd]\varphi \to [\mathring{\sigma}^s]\varphi$ $[\lhd]\varphi \to [\mathring{\sigma}^s_{-i}]\varphi$	visibility

Table 1. Valid principles of LS

The third column shows the formula *schema*. The fourth column describes the *property* of the corresponding accessibility relation R.

K is used in all variants of the standard modal logic. T and 4 determine the preference of players to be *reflexive* and *transitive*. The sight of a player is reflexive. D ensures that a node reachable by a strategy profile σ from a node v is *determined*. I says that every outcome of strategy σ is *included* in the sets of outcomes by letting i free, and the other players following σ. M guarantees the final outcome vertices to be *terminated*. Y shows the *visibility* of all the nodes that can be reached from the current node v in sight-filtrated game $S\lceil_v$. D and I are the same as that for $[\mathring{\sigma}]$ and $[\mathring{\sigma}_{-i}]$.

The inference rules for LS are Modus Ponens (*MP*) and Necessitation (*Nec*).

Theorem 1. (*Soundness and Completeness Theorem*) *Logic of Extensive Games with Short sight LS is sound and complete w.r.t. all $\mathcal{LS}$-models.*

References

[Blackburn *et al.*, 2001] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal logic*. Cambridge University Press, 2001.

[Grossi and Turrini, 2012] Davide Grossi and Paolo Turrini. Short sight in extensive games. In *AAMAS*, pages 805–812, 2012.

[Harrenstein *et al.*, 2003] Paul Harrenstein, Wiebe van der Hoek, John-Jules Ch. Meyer, and Cees Witteveen. A modal characterization of nash equilibrium. *Fundamenta Informaticae*, 57(2-4):281–321, 2003.

[van Benthem, 2002] Johan van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11(3):289–313, June 2002.