LFDSN 2019

Logics for the Formation and Dynamics of Social Norm

4-5 May 2019, Hangzhou, China
Informal Proceedings

Edited By

Beishui Liao
Fenrong Liu
Huimin Dong

Tsinghua University & Zhejiang University

Beishui Liao Fenrong Liu Huimin Dong (Eds.)

LFDSN 2019

Logics for the Formation and Dynamics of Social Norm

Hangzhou, China, May 4-5, 2019

Proceedings

Editors' addresses:

Beishui Liao
Zhejiang University
Department of Philosophy
310028 Hangzhou, China

`baiseliao@zju.edu.cn`

Fenrong Liu
Tsinghua University
Department of Philosophy
100084 Beijing, China

`fenrong@tsinghua.edu.cn`

Huimin Dong
Zhejiang University
Department of Philosophy
310028 Hangzhou, China

`huimin.dong@xixilogic.org`

Preface

These are the proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN2019), which takes place in Hangzhou, China, 4-5 May, 2019. The workshop is funded jointly by the Convergence Research Project for Brain Research and Artificial Intelligence of Zhejiang University (titled "Logic, Cognition and Artificial Intelligence Research Team") led by Beishui Liao at Zhejiang University, and the National Project "Logics of information Flow in Social Networks" led by Fenrong Liu at Tsinghua University.

Dedicated to the two projects, the aim of this workshop is to look in new systematic ways at the contrast between individual values and social values. We go beyond traditional preference or judgment aggregation, placing a focus on agents' communication at a micro-level in social networks of a human society or human-machine integrated community. In that process, agents share new evidence or arguments and they change their preferences or values. Over time, norms then emerge, where some norms die out, while others stabilize. We want to understand this complex of phenomena, with an emphasis on the interplay between social structure and dynamic communication. Parts of these social processes have been studied in different paradigms such as evolutionary game theory, or simulation-based computer science. In this workshop our focus is on also bringing in methods from dynamic-epistemic logic and argumentation theory, broadly conceived, to understand how norms emerge and survive, tied to the agents' reasoning at each stage. We consider case studies, such as autonomous driving and the emergence of norms to resolve conflicts, but also general themes, such as the relation between legal rules, individual and social values.

An open call for papers was issued, inviting projects team members as well as other researchers in highly related areas to submit full papers or extended abstracts. Finally, we accepted 14 contributed papers. These papers nicely cover the main topics of this workshop, including collective attitudes and social interaction, social choice and decision making, formal argumentation and reasoning, evidences and games, probability and reasoning, reasons and beliefs, laws and ethics, etc. The overall mission of the workshop was to provide a platform for the diverse communities interested in this intersection.

In addition to the contributed talks, the conference program consists of four invited talks by leading international researchers in the fields of philosophy, logic, computer science and game theory. They are "Agreement and Disagreement in a Non-Classical World" by Adam Brandenburger, "Rethinking the Epistemology of Belief Dynamics" by Hanti Lin, "Towards Reasoning about Collective Memory" by R. Ramanujam, and "Deliberation, Single-Peaknedness, and Voting Cycles" by Olivier Roy.

We are grateful to the program committee members for their effort in helping us shaping the program. We thank our keynote speakers and all the authors who submitted their papers to our workshop for their contributions. Moreover, we would like to extend our gratitude to the local organizing committee for taking care of all the details that make this workshop possible. These proceedings appear in their dedicated CEUR at http://ceur-ws.org. Through the organization of the LFDSN workshop, we moreover enjoyed financial support from the Convergence Research Project for Brain Research and Artificial Intelligence of Zhejiang University and the National Project "Logics of information Flow in Social Networks" of Tsinghua University.

May 2019 Beishui Liao, Fenrong Liu, and Huimin Dong

Local Organizing Committee

Huimin Dong, Zhejiang University

Program Committee Chairs

Beishui Liao, Zhejiang University
Fenrong Liu, Tsinghua University & University of Amsterdam

Program Committee

Thomas Ågotnes, University of Bergen & Zhejiang University
Alexandru Baltag, University of Amsterdam
Johan van Benthem, University of Amsterdam & Stanford University & Tsinghua University
Adam Brandenburger, New York University
Hanti Lin, University of California, Davis
Xinwen Liu, Chinese Academy of Social Sciences
Piotr Kulicki, John Paul II Catholic University of Lublin
R. Ramanujam, Institute of Mathematical Sciences, Chennai
Olivier Roy, University of Bayreuth
Jeremy Seligman, University of Auckland & Tsinghua University
Leon van der Torre, University of Luxembourg & Zhejiang University
Yí N. Wàng, Zhejiang University

Contents

Chapter 1

Invited Talks

Agreement and Disagreement in a Non-Classical World
(Extended Abstract)

Adam Brandenburger [*] Patricia Contreras-Tejada [†]
Pierfrancesco La Mura [‡] Giannicola Scarpa [§] Kai Steverson [¶]

[*¶] New York University [†] Instituto de Ciencias Matemticas, Madrid
[‡] HHL Leipzig Graduate School of Management [§] Universidad Complutense de Madrid

In the domain of classical probability theory, Aumann (1976) proved the fundamental result that Bayesian agents cannot agree to disagree. Two agents Alice and Bob begin with a common prior probability distribution on a state space. Next, they each receive different private information about the true state and form their conditional (posterior) probabilities q_A and q_B of an underlying event of interest. Then, if these two values q_A and q_B are common knowledge between Alice and Bob, they must be equal: $q_A = q_B$. By "common knowledge" is meant that Alice knows Bob's probability is q_B, Bob knows Alice's probability is q_A, Alice knows Bob knows her probability is q_A, Bob knows Alice knows his probability is q_B, and so on indefinitely.

This result applies in the classical domain where classical probability theory applies. But in non-classical domains (such as the quantum world), classical probability theory does not apply, and so we cannot assume that the same facts about agreement and disagreement between Bayesian agents hold when they observe non-classical phenomena.

Inspired by their use in quantum mechanics, we employ signed probability measures ("quasi-probabilities") to investigate the epistemics of the non-classical world and ask, in particular: What conditions lead to agreement or allow for disagreement when agents may use signed probabilities?

We establish three results:

a. In a non-classical domain, and as in the classical domain, it cannot be common knowledge that two agents assign different probabilities to an event of interest.

b. In a non-classical domain, and unlike the classical domain, it can be common certainty that two agents assign different probabilities to an event of interest.

c. In a non-classical domain, it cannot be common certainty that two agents assign different probabilities to an event of interest, if communication of their common certainty is possible – even if communication does not take place.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

References

Aumann, R.J. (1976). Agreeing to Disagree. *Annals of Statistics*, 128(1):169–199.

Rethinking the Epistemology of Belief Dynamics

Hanti Lin

University of California, Davis

There are three major approaches to the epistemology of belief dynamics: 1. Belief revision theory (grouped with nonmonotonic logic and dynamic epistemic logic). 2. Bayesianism (broadly construed to include Bayesian statistics). 3. Learning theory (grouped with a significant part of frequentist statistics). Those three approaches may appear to compete with one another. For example, reliability is pursued by some but mostly ignored by others. Coherence is praised by some but set aside by some other. The qualitative modeling of belief is employed by some but banished by some other. Despite the difference among those three approaches, I argue that there is a way to make them work together as a unified whole. This is done by (i) identifying the core epistemological thesis and the moving parts of each of the three approaches, and (ii) reinterpreting some results that have been proved to connect or disconnect the three. The upshot is this: although almost each of us needs to slightly change the way how things get done (and this definitely applies to me), we are stronger together.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

Towards Reasoning about Collective Memory

R. Ramanujam

Institute of Mathematical Sciences, HBNI

Abstract

We offer a very simple model of how collective memory may form. Agents keep signalling within neighbourhoods, and depending on how many support each signal, some signals "win" in that neighbourhood. By agents interacting between different neighbourhoods, 'influence' spreads and sometimes, a collective signal emerges.

1 Background

> *Strictly speaking, there is no such thing as collective memory – part of the same family of spurious notions as collective guilt. But there is collective instruction . . . All memory is individual, unreproducible; it dies with each person. What is called collective memory is not a remembering but a* stipulating*: that this is important, and this is the story about how it happened, with the pictures that lock the story in our minds.*
>
> Susan Sontag (2003)

Any discussion on individual values and social values visits the question, *How are* **we** *to act?* at some point. The question, of course, is, who is this **we** referred to here ? Clearly this **we** is a social construction, one that depends on the very social norms and social values that we wish to reason about. Integral to such social construction of a collective is the memory ascribed to that collective. Group identity is constructed structurally by ascribing memory to the group, and in turn, such identity shapes its memory. Remembrance has a crucial impact on preferences and values, influences action.

It is here that Susan Sontag's quote above assumes significance. Sontag calls collective memory a process of stipulation. Somehow the collective ascribes importance to an item of memory, authenticates it and symbolizes it; then on, the symbolism "locks" the memory item, in Sontag's account.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

Note that this is a significant departure from the structural conception of memory, that visualises memory as a notebook, and remembering as looking it up. Wittgenstein strongly attacked such a conception of memory. Scholars like Sutton (2014) have discussed this at length. For Wittgenstein, social acts were important in shaping memory, and based on this, scholars like Rusu (2013) even talk of *social time*, and modern theories of connectionism and distributed memory build on many such notions.

For us, these remarks are relevant from two viewpoints. The 1950's saw the development of automata theory as a study of *memory structures*, and in theory of computation, automata provide a model of memory that Wittgenstein might have approved of. In this view, memory is not a table to be looked up, but is constituted by states of being of the automaton. Observations cause changes in state, some states remember (some of the past) and some forget. Thus, memory is built into system structure. Such a view is important for seeing memory and reasoning as *interdependent* rather than as separate (as psychologists used to consider). Logicians are used to equating automata and logics, as in the case of monadic second order logics of order. (Wittgenstein would have approved.)

The other viewpoint relates to distributed memory, where interacting agents rely on memory external to them. Computer science has evolved impressive models of highly flexible interaction and memory that has literally changed the everyday life of much of humanity in the last few decades.

In social theory, the notion of collective memory is influential. Maurice Halbwachs (1980) talked of how an individual's understanding of the past is strongly linked to a group consciousness, which in turn is a form of *group memory* that lives beyond the memories of individuals that form the group.

For the logician, these notions pose an interesting challenge: what are the logical properties of collective remembering ? What is the rationale followed by a group in ascribing / stipulating collective importance to events and their remembering ? Why is a particular idealisation chosen ? These are difficult questions to answer, but a more modest reformulation of such questions offers an approach to solutions. If the memory of an automaton is describable via logic, we can perhaps build a model of group and individual memory based on automata whose interactions lead to collective memory, which in turn influences behaviour of individual automata.

Why should one bother ? In (Parikh 2012) Rohit Parikh speaks of *cultural structures* providing an infrastructure to social algorithms (much as data structures do for computational algorithms). A queue is one such structure according to him, and we can see how it enables a specific kind of social behaviour. Epistemic reasoning is an essential component of social algorithms, as persuasively argued by Parikh. We can then see collective memory as an essential gradient of its infrastructure creating the 'common ground' in which social objectives and communications are interpreted.

What follows is a very simple, perhaps very simplistic, attempt at formalization of this notion, inspired by the study of *population protocols* in distributed computing (Aspnes and Ruppert 2009). We offer this formal model tentatively, as an initial step of a (hopefully) detailed research programme.

2 A model

Let N denote a fixed finite set of agent names. Let $\mathcal{C} \subseteq 2^N$ be a nonempty set of nonempty subsets of N, referred to as *neighbourhoods* over N.

For presenting the model we will make some simplifying assumptions. We fix an alphabet pair (Σ, Γ) where Σ is a finite input alphabet, Γ is a finite signal alphabet, both common to all agents.

Let $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$.

Let $I \in \mathcal{C}$ and $|I| = k$. A **distribution** over I is an m tuple of integers $\mathbf{y} = (y_1, \ldots, y_m)$ such that $y_j \geq 0$ and $\Sigma_{j=1}^{m} y_j = k, \;\; 1 \leq j \leq m$. That is, the jth component of $\mathbf{y}$ gives the number of agents in the neighbourhood I who give signal γ_j. Let $\mathbf{Y}[I]$ denote the set of all signal distributions of a neighbourhood I and let $\mathbf{Y} = \bigcup_{I \in \mathcal{C}} \mathbf{Y}[I]$.

The main idea is this. All agents initially receive an external input and assume some state. At each state, an agent produces a signal. Interactions occur in neighbourhoods nondeterministically, and an agent who is a member of many, could be interacting in different neighbourhoods. Each interaction induces a state transition that is determined only by the distribution of signals: it does not depend on who is signalling what, but how many are producing each signal. Such interactions keep occurring repeatedly until a stable configuration is reached.

Below, for $I \in \mathcal{C}$, we use the notation Γ^I for a vector of signals, one for each of the agents in I. Note that every such vector induces a distribution over Γ in $\mathbf{Y}[I]$.

Definition 1. *A group automaton over $\mathcal{C}$ is a pair $M = (\delta, \iota)$, where $\iota : \Sigma \rightarrow \Gamma$, and δ is a finite family of transition relations $\delta_I \subseteq (\mathbf{Y}[I] \times \Gamma^I)$, where $I \in \mathcal{C}$.*

A configuration χ is an element of Γ^N. The dynamics of M is given by a configuration graph G_M whose vertices are configurations and edges are labelled by neighbourhoods: an edge (χ, I, χ') is present when the signal distribution induced by χ causes a transition to Γ^I, and $\chi' = \chi \oplus \Gamma^I$ (with some abuse of notation). Note that ι induces an initial configuration χ_0. A history ρ is any finite or infinite path in G_M starting from χ_0. Let $\mathcal{H}_M$ denote the set of all maximal histories of M.

When we have stable configurations, we see them as formation of collective memory. For this it is of course essential that a history allows for signalling to spread across neighbourhoods; if some neighbourhoods never interact, then signalling can remain confined within pockets. So we may often consider only spanning histories where we impose the condition that every interaction that is infinitely often enabled (according to the transition rule) in an infinite history takes place infinitely often. This is a typical fairness condition used in the theory of computation, but weaker conditions may be sufficient for many systems. For instance, we might simply ask that the union of neighbourhoods in a history span all of N; this merely says that all the agents have interacted at least once.

As an example consider a system with two signals $\{g, b\}$, standing for "good" and "bad". Initially every agent perceives some global event as good or bad. The transition rule is simple: for any neighbourhood I, if more than half in I signal x, then all agents in I signal x in the new state. If they are exactly even, they continue evenly matched. Now if we ensure that δ offers a sufficiently rich interaction, we have either of the two perceptions becoming stable, depending on how many receive the input, as well as the structure of $\mathcal{C}$.

Theorem 1. *Given a system M, whether a stable configuration is reachable is decidable.*

Unfortunately, the size of the configuration graph can be very large (though finite) as also the description of δ, so decidability amounts to very little.

3 Discussion

We began with the intention of reasoning about collective memory. How do systems of signalling in neighbourhoods embody reasoning ?

Firstly, it should be clear that the history model is rich enough to interpret the formulas of a propositional temporal logic, and hence we can indeed talk in a logical language about the "remembrance of things past". Further, a natural equivalence relation $\sim_i$ can be defined, that equates histories for agent i, when the signal sequence projected to the i^{th} component in each of them is identical. Thus we can employ an epistemic temporal logic to discuss the dynamics.

Such a logical exercise is not sufficient in itself to uncover the process of *stipulation* mentioned by Sontag, or the *interdependence* between memory and reasoning demanded by Wittgenstein. However, in our opinion, the model holds considerable promise. For achieving the richness required, we hold two features to be essential: reinforcement of memory that comes through repeated interactions inside local neighbourhoods, but not confined to those neighbourhoods; complex social rules that determine influence in signalling. Both are present in this model.

What we have here is obviously a model in its infancy. Whether it can grow into a healthy tool for reasoning about collective memory is as yet a matter of hope.

References

Aspnes, J. and Ruppert, E. (2009). An introduction to population protocols, available from www.cs.yale.edu/homes/aspnes/papers/minema-survey.pdf.

Halbwachs, M. (1980). *The Collective Memory*. Harper & Row, New York.

Parikh, R. (2012). What Is Social Software?. *Games, Actions, and Social Software*, Jan van Eijck (Ed), LNCS 7010, 3–13.

Rusu, M. (2013). History and Collective Memory: The Succeeding Incarnations of an Evolving Relationship. *Philobiblon*, Vol. XVIII, No 2, 260–282.

Sontag, S. (2003). Regarding the pain of others. *Picador*.

Sutton, J. (2014). Remembering as Public Practice: Wittgenstein, memory, and distributed cognitive ecologies. *Proc* 36^{th} *Wittgenstein Symposium*, (Eds) D. Moyal-Sharrock, V.A. Munz, A. Coliva.

Deliberation, Single-Peaknedness, and Voting Cycles

Soroush Rafiee Rad Oliver Roy

University of Bayreuth

Deliberation in social networks can help avoiding voting cycles. This claim is usually made via the notion of "meta-agreements" which, it has been claimed, entails that the group members' preferences are single-peaked, which in turn guarantees the existence of a Condorcet winner. We provide evidence from simulations that this detour through meta-agreement, or even single-peaked preferences, is in general unnecessary. To the extend that deliberation induces a process of rational preference change, it almost completely avoids voting cycles. Whether it does so through the creation of single-peaked preferences depends on the whether the participants rank only strictly the alternatives or are allowed to be indifferent. In the former case deliberation does eliminate cycles by creating single-peaked preferences. In the latter case, however, single-peaked preferences play little to no role in the elimination of voting cycles. This, on the one hand, supports an optimistic view about the outcome of group deliberation. It does, on the other hand, puts important qualifications to the centrality of single-peaked preferences in that process.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

Chapter 2

Contributed Papers

Pluralism about Reason Accrual

Davide Fassio

Zhejiang University

Abstract

One of the most complex and mysterious phenomena discussed in the literature on reasons is the so-called 'accrual of reasons'. This phenomenon concerns the ways in which a set of reasons can interact and be combined to determine deliberative outcomes and overall obligations. Reasons accrue in variable and complicated ways. Philosophers have provided several models of this phenomenon. Unfortunately, none of them seems exempt from problems. The aim of this paper is to critically consider the various models advanced in the literature and show that they are all wanting because incomplete. While each model has considerable explanatory virtues, none of them alone can fully explain the many ways in which reasons accrue. My tentative solution is to assume that different mechanisms of reason accrual are operative in different contexts. This sort of 'accrual pluralism' could explain all the cases while avoiding problematic consequences of other approaches.

Normative reasons are considerations that count in favour of or against certain attitudes, actions or omissions. For example, that Tom promised to be on time at the party is a reason for her to do so, that the butler's fingerprints are on the murder weapon is a reason to believe that he is the culprit, and that a joke is funny is a reason to appreciate it. Many philosophers consider reasons as the most fundamental building blocks of normativity, to which other normative notions could be reducible, or in terms of which they could be explained.[1] Moreover, the notion of reason plays an

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

[1] e.g., (Alvarez 2010; Dancy 2000; Kolodny 2005; Parfit 2011; Pollock 1974; Raz 1999; Scanlon 1998, 2014; Schroeder 2007; Skorupski 2010). In the epistemic domain, see e.g., (Conee and Feldman 2004; Mitova 2015; Turri 2009; Kearns and Star 2008; Schroeder 2015; Sylvan 2014, 2016).

important role in our everyday normative reasoning. Reasons are considered the most important 'contributory' elements in normative discourse: reasons' combinations contribute to make better or worse cases for actions and attitudes, determining all things considered normative conditions, such as whether an agent is all things considered permitted, obliged or forbidden to perform a certain action or hold a certain attitude.[2]

One of the most complex and mysterious phenomena discussed in the literature on reasons is the so-called 'accrual of reasons'. This phenomenon concerns the ways in which a set of reasons can interact and be combined to determine deliberative outcomes and overall obligations. Reasons accrue in variable and complicated ways. In some circumstances, this accrual resembles a mere weight of different strengths of reasons, and the outcome an additive function of these weights. However, in other situations, the outcome is some more complex factor of the combination. For example, it may happen that the combination of two reasons for a certain act or attitude F results in a weaker normative support for F, or even in an obligation to refrain from F-ing. As Jonathan Dancy observed,

> "[reasons] can combine in peculiar and irregular ways [. . .]. There is no guarantee that the case for doing an action, already made to some extent by the presence of one reason, will be improved by adding a second reason to it. Reasons are like rats, at least to the extent that two rats that are supposedly on the same side may in fact turn and fight among themselves; similarly, the addition of the second reason may make things worse rather than better" (Dancy 2004, p.15).

Later we will consider several examples of this complex phenomenon.

Philosophers have provided several models of reason accrual. Unfortunately, none of them seems exempt from problems. Some accounts lead to implausible consequences. Others lack the resources to explain some of the relevant cases of accrual. The aim of this paper is to critically consider the various models advanced in the literature and show that they are all wanting because incomplete. While each model has considerable explanatory virtues, none of them alone can fully explain the many ways in which reasons accrue. I shall also tentatively argue that the solution to this problem consists in assuming that different mechanisms of reason accrual are operative in different contexts. This sort of 'pluralism' about reason accrual could explain all the cases while avoiding problematic consequences of other approaches.

This is the plan of the paper. In the next four sections (1-4) I shall consider four models of reasons' accrual: the 'weighing' account, the 'independent reason' account, the mixed account, and the holistic account. I shall argue that each of them is affected by several problems. These models seem either unable to fully explain the complexity of the phenomenon or doomed to implausible consequences. However, I shall also argue that these models are not radically misleading. Each of them seems particularly well fit to explain a restricted range of cases. In 5 I shall tentatively argue that the right approach to the accrual of reasons is pluralist. We should accept that there is not a single mechanism regulating the ways in which reasons combine to determine overall obligations. Rather, there is a plurality of mechanisms appropriate in different contexts.

Before proceeding further, let me add three clarificatory remarks: first, this paper is concerned with reason accrual in situations in which the strength of reasons is not derivative. I will set aside discussions of weighing mechanisms of what many call 'instrumental reasons', whose strength is derivative from source reasons. The peculiar mechanisms of weight and accrual for such type

[2] See, e.g., (Dancy 200; Lord and McGuire 2016).

of reasons would deserve a special treatment that unfortunately I cannot offer here.[3] Second, the weight of reasons and the way they accrue may be grounded in very different kinds of property. A short list includes prudence, authority (moral, aesthetic, social), convention, specificity, reliability, and second-order reasoning.[4] In this paper I will be exclusively concerned with the general mechanisms of reason accrual. My discussion will remain neutral on the underlying metaphysical and explanatory grounds of such mechanisms. My third remark is methodological. In this paper I will privilege arguments based on inference to the best explanation. I will assess different accounts of reason accrual based on how well they fare in explaining different sorts of cases. Moreover, I shall assess such accounts based on (i) how well they preserve intuitive differences between the considered cases, and (ii) whether they are well suited to provide a complete explanation of all cases. As we will see, various available accounts offer good explanations of some such cases but not of others. My contention is that no available account can satisfactorily explain all the different ways in which reasons accrue or fail to accrue. This will urge us to rethink the phenomenon of reason accrual as based on a plurality of independent mechanisms, and eventually to endorse a pluralist account of such a phenomenon.

1 The 'Weighing' Account

According to the 'weighing' account, each reason possesses a specific strength. This can be understood as a sort of weight or alternatively as a force, by analogy with physical forces.[5] When reasons interact, their weights are compared and combined, and the outcome of a combination is a further weight that is an increasing function of the weights of its components.[6] An agent should (overall) perform the action (hold the attitude) that receives the greatest weight.

The weighing conception is well illustrated by an example suggested by John Horty (2012, p.59):

> **(Two Aunts)** I have been invited to the wedding of a distant relative at a difficult time of year. I am not particularly close to this relative, and, since the wedding falls at such an inconvenient time, I would rather not go. But suppose I learn that the guests will include

[3] The relations between derivative and source reasons are commonly expressed by weight transmission principles. An example is the Sufficient Means Transmission Principle according to which if the weight of A's reason to *x* is *n* and *y*-ing is a sufficient means to *x*-ing, then there is a reason to *y* with weight *m*, where *m* is not greater than *n* (Lord and Maguire 2016, p.13-14). For overviews of reason transmission mechanisms see (Kolodny and Brunero 2018, 2) and (Lord and Maguire 2016, 2). It is worth mentioning that some philosophers do not consider 'instrumental reasons' to be reasons at all. See in particular Horty's austere view of reasons (e.g., (Horty 2012, p.44)).

[4] See (Horty 2012, p.18-19 and ch.5).

[5] An early presentation of this model can be found in a letter of Benjamin Franklin (1772, p.348-349). For a recent endorsement of the weighing conception see (Broome 2004). For the 'force' interpretation see (Ross 2002) (though Ross' focus is on prima facie duties). For a discussion of the 'weight' and 'force' interpretations see (Horty 2012, p.3-5).

[6] One may conceive this increasing function as completely additive: where r^i and r^{ii} are reasons, their combined weight is W(r^{ii}+r^{ii}). The example in the text seems to suggest such kind of fully additive accrual. However the increasing function must not be necessarily understood as completely additive. I've here in mind, in particular, a model of accrual for probabilistic weights, particularly apt to formalize the interaction of epistemic reasons for credences (partial beliefs). Probabilities are only partially additive: the combination of two probabilities is not identical to their sum when the probabilities are not independent. Another example in which the resulting weight may be an increasing function of the initial weights without being their additive outcome is the non-linear variation of utility due to the effect of risk aversion – at least if one interprets risk-aversion as due to intrinsic properties of utilities; of course, this is not the only available interpretation of this phenomenon. See (Buchack 2013) for an alternative view.

> my two old aunts, Olive and Petunia, whom I enjoy and who I know would like to see me. Here it is perfectly sensible to imagine that, even though I would still choose not to attend the wedding if only one of the two aunts were going, the chance to see both Aunt Olive and Aunt Petunia in the same trip offers enough value to compensate for the inconvenience of the trip itself.

We can easily imagine a representation of this situation within a weighing conception: the inconvenience of the trip has a certain numerical weight that outweighs the benefits of seeing each aunt alone. However, when the weight of the value of seeing Aunt Olive is added to the weight of seeing Aunt Petunia, the additive weight of the two reasons together outweighs the disvalue of the inconvenient trip.[7]

Despite its intuitive appeal, this account doesn't seem sufficient to explain all cases of reason accrual. Here it is an example from Horty (2012, p.61):

> **(Heat+Rain)** Suppose I am deliberating about an afternoon run, and that both heat and rain, taken individually, function as negative reasons, arguing against a run; still, the combination of heat and rain together might function as a positive reason, favoring the run as refreshing.

In this example, heat and rain, taken individually, provide negative weights against a run. If the weighing account were right, we should then expect that their combined weight is a higher disvalue of a run. But this is not the case. In fact the two reasons combined together provide positive weight favoring running. Consider a similar but funnier example, inspired by Dancy (2004, p.16): the fact that food is terrible may be a reason against going in a restaurant, the fact that portions are too small is another reasons to avoid that restaurant. However if food is terrible, small portions seem better than big ones. Differently from (Heat+Rain), here the reasons' accrual does not change the polarity of the weight (from negative to positive). Nonetheless, the weight resulting from the combination of the two reasons is less than that of each reason taken individually: other things being equal, I would prefer to go in a restaurant where food is terrible and portions are small (of course, assuming that for reasons of etiquette I should finish the meal). In this example, the accrual works as an attenuator of the respective reasons.[8] This is incompatible with a simple weighing conception of the accrual of reasons.

2 The 'Independent Reason' Account

According to the 'independent reason' account, accrual of two or more reasons is not a function of these reasons. Rather, the facts that individually constitute basic reasons when conjoined together constitute a more complex independent reason. This complex reason may well recommend what the more basic reasons recommend and with a higher degree of strength. According to this account, this is precisely what happens in (Two Aunts) where, while the individual facts that Aunt Olive will

[7] Another interesting example is the following (from (Nair 2016, p.56)): suppose that in order to get to that side of town I must cross a bridge that has a $25 toll. The toll is a reason not to cross the bridge. The movie is a reason to cross the bridge and the restaurant is also a reason to cross the bridge. It may be that if there were just the movie to see, it wouldn't be worth it to pay the toll and if there were just the restaurant, it wouldn't be worth it to pay the toll. But given that there is both the movie and the restaurant, it is worth it to pay the toll.

[8] For further examples and discussion see (Lord and Maguire 2016, 2.2; Nair 2016, 1). See also (Dancy 2004, ch.1, 2) for other interesting examples illustrating the inadequacy of a simple weighing model.

be at the wedding and that Aunt Petunia will be at the wedding are not strong enough reasons to outweigh the inconvenience of the trip, the further fact that the two Aunts will be present successfully outweighs the inconvenience. In other cases the complex reason can bear a weaker strength than the more basic reasons, and sometimes even possess an inverted polarity with respect to them. This is what happens in (Heat+Rain) where, while heat supports not to run and rain supports not to run, the joint presence of heat and rain supports to run.

Observe that, according to this account, in *all* cases of accrual the complex reason is not a factor of the weights of the more basic reasons, but a completely new, independent reason. Strictly speaking, according to this account it isn't quite right to talk of 'accrual' of reasons. As a matter of fact, there is no accrual of basic reasons into a complex one. Rather, complex reasons are further independent normative considerations which add up to the basic ones and in normal circumstances fully outweigh them.

The 'independent reason' account has enjoyed a certain degree of popularity,[9] and it is quite resistant to counterexamples. Unfortunately, also this account leads to a number of implausible consequences.[10] In most cases involving reason accrual it doesn't seem right to think of complex reasons as completely independent from basic ones. Consider again (Two Aunts). Does it make sense to say that the reason constituted by the presence of the two aunts has absolutely nothing to do with the separate more basic reasons constituted by the presence of Aunt Olive and the presence of Aunt Petunia? This sounds strikingly counterintuitive. It seems pretty obvious that the normative force of the complex reason strictly depends on the normative import of the basic ones.

Another related problem is that this account unduly multiplies the number of reasons. It seems that in (Two Aunts) it is inappropriate to describe the case as one involving four reasons bearing on the choice whether to attend the wedding: the inconvenience of the trip, the presence of Aunt Olive, the presence of Aunt Petunia, and the presence of Aunt Olive and Aunt Petunia. This seems a vicious double counting of the relevant normative considerations. Similar considerations are valid for (Heat+Rain). It doesn't seem right to describe the case as one in which there are three reasons: that it's hot, that it rains, and that it's hot and it rains why then not to add further even more complex reasons to the list, such as the complex fact that [(it's hot and rains) and that it rains]?

Another problem for this account is that if the complex reason is completely different from the basic ones, we should expect the latter reasons to survive in the background of our normative considerations and preserve their force and polarity, even though merely as outweighed, defeated reasons.[11] However often this is clearly not the case. Consider again (Heat+Rain). I wouldn't describe this situation as one in which, despite the facts that it is hot and it rains individually clearly speak against running, the fact that it's hot and rainy is a stronger reason to run and outweighs the other two considerations for not running. This description of the case sounds pretty odd, at least to my hears. Rather, I would say that in a context in which heat and rain are both present and speak in favor of running, each of the components alone, the heat and the rain, both contribute positively to the normative status of running. In other words, the co-presence of rain and heat doesn't merely defeat the basic reasons against running, but completely changes the polarity of these individual

[9] See, for example, Pollock (1995, p.101102), who treats complex reasons as independent of basic ones. (Schroeder 2007, Ch. 7) can be interpreted as endorsing a similar approach.

[10] Some such consequences have been put forward by Horty and Nair. See (Horty 2012, p.60; Horty and Nair 2018, 5).

[11] As Dancy has observed, "a contributory reason on one side is not necessarily destroyed by the presence of a reason on the other side. This does happen sometimes, I agree, but it is far from the standard case" (Dancy 2004, p.15).

considerations from negative to positive: we can say that in this scenario these considerations are reasons for running (though not sufficient if taken individually), not reasons against. To reinforce this point, think of the absurdity of the following considerations, which would be appropriate if the independent reasons account were right: "There are good reasons not to run: that it is hot and that it's raining. However fortunately there is a stronger reason to run: that it's hot and rainy!".

Indeed it is important to stress that a few cases could be properly described as involving complex reasons independent of their constituent basic ones and weighed against them. Here is an example:

> **(Two promises)** Tom promised to Mary that he would go to her birthday party. He also promised Matt that he would go to Mary's party. However a terrorist threatens to kill Tom's family if Tom makes two promises and keeps one of them.

In this case it seems that the fact that Tom made the two promises is a decisive reason not to go to the party, even though each premise individually remains a (defeated) reason to go. Importantly, in this case the promises do not change their polarity as in (Heat+Rain). Rather, they provide *pro tanto* reasons to go to the party, even though their force is completely defeated by the terrorist's threat. This is confirmed by the fact that, even if Tom should not go to the party, he should nonetheless call Mary and Matt and tell them that he will not be able to keep his promises and apologize for this. Such reparative duties are clues that the reasons constituted by each promise preserve their force and polarity and clash with the further separate reason-fact that Tom made the two promises.

While the 'independent reason' account fits well with this sort of examples, the important differences between this case and others such as (Two Aunts) and (Heat+Rain) should make us reconsider once more the account as inadequate to explain all cases of reason accrual. This reinforces my hypothesis that this model of reason accrual, as others discussed in the literature, is not completely out of track, but is incomplete insofar it only accounts for a specific aspect of the more general phenomenon.

3 The Mixed Account

So far we considered two accounts of reason accrual: the 'weighing' account and the 'independent reason' account. We found problematic both accounts: while they provide correct diagnoses of some cases, they are unable to account for others. However one may suggest a mixed account according to which in some cases reasons accrue according to what the weighing account predicts, while in others an independent reason determines the outcome by outweighing other basic reasons.

How to decide which accrual model we should apply in a given circumstance? When exactly would an independent complex reason enter in the weight? Answer: when there is such a reason. In (Two Promises) there is such an independent reason (namely, that Tom made two promises) which outweighs the strength of basic reasons. However in other cases such as (Two Aunts) there is no complex independent reason and the basic reasons accrue by combining their respective weights. The general idea behind the mixed account is that, *by default*, reasons accrue by combining their weight according to the weighing model. This default rule is defeated when the combination of facts constituting basic reasons is itself a further independent reason relevant for the same decision.[12]

[12] To my knowledge nobody ever suggested such an account. (Prakken 2005) could be interpreted as defending a version of it, though some passages in that article seem to suggest otherwise. While I shall argue that the present account doesn't seem to work in special cases, I must admit my sympathy for it. Compared to other non-pluralist accounts, this is the one I find most promising.

This account can avoid the problem put forward by Horty and Nair. In (Two Aunts) the reasons to attend the wedding are just the two basic ones: that Aunt Olive will be at the wedding and that Aunt Petunia will be at the wedding. As we saw above, the account has also no difficulties in explaining cases like (Two Promises). Such explanation mirrors the one of the independent reason account. Moreover, the account doesn't unduly multiply the number of reasons. In (Two Aunts), the account predicts that there are exactly three reasons: the inconvenience of the trip, the presence of Aunt Olive, and the presence of Aunt Petunia. The joint presence of the two aunts doesn't count as further reason. The account also predicts that in (Two Promises) there are three reasons: the promise that Tom made to Mary, the one he made to Matt, and the two premises together. Here it seems appropriate to assume the existence of a further independent complex reason in addition to the two basic ones, also given its different normative source the former being the result of acts of promising, while the latter being the proper response to a life threat.

However the account seems to founder on the further problems advanced against the 'independent reason' account. The mixed account has no easy explanation of cases such as (Heat+Rain). We already saw that in such cases the accrual cannot be fully explained by a mere weighing conception. Therefore the mixed account must explain the case with a further independent reason. However this would amount to interpret (Heat+Rain) as a case in which each of heat and rain speak against running, but the fact that it's hot and rainy is a stronger reason to run and outweighs the other two considerations for not running. As we said above, this description of the case sounds deeply wrong. Heat and rain both seem to contribute positively to the normative status of running. The two considerations change their normative polarity when they are jointly present. This cannot be easily explained by the mixed account.

4 The Holistic Account

There is an account of reason accrual that can explain (Heat+Rain). According to the holistic account, all things considered obligations are never factors of the conglomeration of a specific subset of reasons; rather, they are always determined by the totality of facts. According to a specfic 'deflationary' version of this view, talk about reasons is a mere pragmatic device. What we call 'reasons' are just specific facts included in a more general net of normative explanations that we pick up and separate from other conditions only because of their salience in a conversational context. They are not ontologically substantive entities carving normative nature at its joints. Alternatively, we may think of reasons as the set of facts standing in an holistic net of normative relations, ontologically substantive but overdeterminated by the context as a whole. In both versions of the view, reason accrual is not an aggregative phenomenon, but a way of describing a certain local configuration of facts determined by the totality of normative relations. According to this account, reasons do not literally accrue; rather, they stand in specific relations as part of a more general normative explanation.[13]

As I just said, the holistic account can easily explain cases such as (Heat+Rain). In circum-

[13] Holistic accounts of reasons and reason accrual seem to be presupposed, at least implicitly, by ought-first views (e.g., (Broome 2013)), and by specific normative frameworks such as Bayesianism in formal epistemology. (Schroeder 2011) could be interpreted as defending a holistic account of reason accrual, even though for him the context dependent holistic features are the weights of reasons, not reasons themselves. Dancy (2004) famously used reasons holism to defend particularism. However, it is important not to conflate holism about the grounds of reasons and holistic accounts of reason accrual. While the two things may go together, they are quite independent. In particular, one can be a reason holist without endorsing an holistic account of reason accrual.

stances in which there is heat but not rain, or rain without heat, the obtaining fact counts against running given the totality of facts (at least other things being equal). However, in circumstances in which heat and rain are both present, other things being equal, the heat and the rain will be both reasons to run – individually insufficient but jointly sufficient. Holism can explain well why and how reasons sometimes completely change their force and polarity when embedded in different contexts. This account assumes that what an agent ought to do is not a factor of reasons; rather, reasons are 'read off' from the totality of normative facts and relations. They can be used as devices for explaining the structure of overall obligations but cannot determine and ground the latter. In this perspective, a change in what one ought to do may have effects on the overall network of normative relations, possibly changing polarity and force of virtually any reason.[14]

One worry with such account is that it doesn't respect our intuitions about how normal people reason about normative matters. It's clear that our deliberative practice rarely follow holistic patterns. In our deliberation about what to do, we almost never proceed via a general consideration of the totality of facts or all the facts relevant to a certain decision. Rather, we proceed from defeasible premises and conclude to what we should do. The overall normative conclusion is almost always constructed from more fundamental contributory elements, by combining together specific considerations with mere *pro tanto* weight. We normally start assuming by default that some fact counts as a reason for some act or attitude, and then we proceed bottom-up modifying our conclusion on the basis of further information. If, as it seems plausible, our deliberative practice is supposed to reflect the structure of normative reasons, we must conclude that the holistic account doesn't provide a realistic account of the accrual of reasons in general.

A related problem is that according to this account, whether something is a reason would depend on a fully qualified, complete list of facts. This would commit us to consider in our reasoning every possible relevant fact in the circumstance before concluding that something is a reason for something else. But this is an unrealistically complex, never-ending work for a human being.[15] Consider a specific example of defeasible reasoning: if Tweety is a bird, then Tweety can fly. If we were to reason holistically about whether the fact that Tweety is a bird counts as a reason to believe that Tweety can fly, we should consider all possible defeasible circumstances such as whether Tweety is not a penguin, that Tweety is not a duck, that is not unable to fly due to particular circumstances such as that it is a baby bird, a sick bird, his wings are clipped, its fits are stuck in cement, and so on. Only after considering all facts we will be in a position to conclude whether the fact that Tweety is a bird really counts as a reason in favor of believing that Tweety can fly. The list of relevant facts is open-ended, just impossible to consider for any human being. The conclusion would be that human beings are almost never in a position to know whether some fact is a reason for something, or whether the fact counting as reason promotes that thing or adverse it.

Finally, notice that a holistic explanation of other cases considered above would be far less intuitive than that provided by other accounts. As way of example consider (Two Aunts). Here it seems odd to say that the decision to attend the wedding is not motivated by the consideration of the facts that my aunts will be at the wedding, but by an infinity of other considerations which I am not attending now, such as whether my aunts have not been substituted by robot replicants or

[14] The holistic account seems to be particularly well suited to describe specific normative properties such as aesthetic and prudential values. See (Moore 1993, p.70-80)'s principle of organic unities and his application to the aesthetic domain. Other normative domains, such as the moral and legal ones, seem less susceptible to a holistic interpretation.

[15] This worry is inspired by the discussions in (Horty 2012, p.55 and ff). The worry is related to the "potato in the tailpipe problem" much discussed in the literature on artificial intelligence.

whether there will be a terrorist attack during the wedding. The case seems to be better described as one in which what the agent ought to do is the product of the interaction and weight of just three reasons.

The above considerations indicate that the holistic account doesn't provide a realistic account of the accrual of reasons in general. This doesn't mean that we can't apply holistic criteria to small sets of facts in order to determine the set of reasons operative in a given context. (Heat+Rain) seems precisely to be a case in which we reason in this holistic way, by considering the meteorological situation as a whole, not as the combination of specific meteorological aspects such as heat and rain. Such kind of deliberation, moving from the consideration of whole sets of information about specific matters, would help explaining a range of cases, while leaving open the possibility of accounting for other cases by means of different mechanisms.[16]

5 Pluralism about Reason Accrual

Let's take stock. So far we have considered four accounts of reason accrual. I argued that none of these accounts is satisfactory. Each account fails to explain the phenomenon in all its complexity. We saw that some of these accounts have seriously implausible consequences. It also emerged that some accounts can explain certain cases better than others, and some accounts cannot explain certain cases at all. This doesn't mean that these accounts are completely wrong. On the contrary, each of them seems particularly well fit to explain a restricted range of cases.

Someone may suggest that these failures are sufficient reasons to give up any attempt to determine how reasons accrue. Maybe there is something deeply wrong with the ideology of reasons and we should abandon the idea of construing rational and normative frameworks by appealing to this notion. Or maybe there is nothing wrong with reasons in themselves. Rather the problem is with our epistemic power to understand the complexity of certain aspects of reasons' logics. However I do not think it is already time to surrender to pessimism. This because I think that there is room for a correct account of reason accrual. My tentative suggestion is that we should endorse a pluralist model of reason accrual. If we want an account of this phenomenon, I think that we should start recognizing that there is not a single mechanism regulating the ways in which reasons combine to determine overall obligations. Rather, there is a plurality of mechanisms, each one appropriate in different contexts. Sometimes the mechanism could consist in the simple weight of the reasons' strength; other times it would involve the presence of independent complex reasons; still others it would be explainable by an holistic approach to a specific range of facts.

How would the different mechanisms interact in the various contexts? Which conditions would trigger certain mechanisms rather than others? Unfortunately I've not an answer to these difficult questions. However, my guess is that the right model starts from something like the mixed account described in 3 and implements further holistic mechanisms. In cases in which there are no independent complex reasons as in (Two Aunts) specific reasons would accrue following a weighing model. This would be the default mechanism of accrual. However, when an independent complex

[16] In this survey and critical discussion of models of reasons' accrual I didn't consider the account provided by (Nair 2016), according to which accrual mechanisms would be influenced by whether reasons are derivative or not. The reason of this omission is that I consider Nair's discussion and proposal orthogonal to the present discussion. In particular, Nair doesn't advance a specific account of the mechanisms of accrual for derivative reasons (which, from how Nair characterizes the distinction, constitute the very wide majority of reasons). I also think that the derivative/non-derivative reason distinction is orthogonal to the distinction between different mechanisms of reason accrual. Unfortunately I must postpone a critical discussion of Nair's proposal to a further occasion.

reason enters in the game, this should also be weighed with the others and could determine a different outcome, defeating more basic reasons as in (Two Promises) or weakening their strength as in Dancy's restaurant case. Moreover, in some circumstances the structure of a specific set of reasons could be re-shaped by holistic considerations about a relevant range of facts. For example, in (Heat+Rain), where facts about weather are considered holistically, the reasons' strength and polarity are not a factor of specific weather conditions, but of the whole meteorological situation.

This pluralistic approach to reason accrual is supported, on the one hand, by the failure of other accounts, and, on the other hand, by the observation that each of these accounts is successful in specific contexts but not others. The new approach is here advanced as a mere hypothesis. It would be interesting to test it on a wider range of cases across different normative domains. Unfortunately I must postpone a comprehensive discussion of the approach to a future occasion. For now, the more modest conclusion is that specific accounts of reason accrual, though ingenious and capable of explaining a wide range of cases, have been so far unsuccessful, and that a possible solution may lie at their intersection.

References

Alvarez, M. (2010). *Kinds of Reasons: An Essay in the Philosophy of Action.* Oxford University Press.

Broome, J. (2004). Reasons. In R. J. Wallace (Ed.), *Reason and Value: Themes From the Moral Philosophy of Joseph Raz*, (pp. 28–55). Oxford University Press.

Broome, J. (2013). *Rationality Through Reasoning.* Wiley-Blackwell.

Buchak, L. (2013). *Risk and Rationality.* Oxford University Press.

Conee, E., and Feldman, R. (2004). *Evidentialism.* Oxford University Press.

Dancy, J. (2000). *Practical Reality.* Oxford University Press.

Dancy, J. (2004). *Ethics Without Principles.* Oxford University Press.

Franklin, B. (1772). Letter to Joseph Priestly. In F. Mott and J. Chester (Eds.), *Benjamin Franklin: Representative Selections* (pp. 348349). American Book Company.

Horty, J. F. (2012). *Reasons as Defaults*. Oxford: Oxford University Press.

Horty, J. F., and Nair, S. (2018). The Logic of Reasons. In D. Star (Ed.), *The Oxford Handbook of Reasons and Normativity.* Oxford University Press.

Kearns, S., and Star, D. (2008). Reasons: Explanations or Evidence? *Ethics*, 119(1), 3156.

Kolodny, N. (2005). Why be rational? *Mind*, 114(455), 509563.

Kolodny, N. and Brunero, J. (2018). Instrumental Rationality. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.).

Lord, E., and Maguire, B. (2016). An Opinionated Guide to the Weight of Reasons. In Lord, E., and Maguire, B. (Eds.), *Weighing Reasons*. Oxford University Press.

Mitova, V. (2015). Truthy psychologism about evidence. *Philosophical Studies*, 172(4), 11051126.

Moore, G. E. (1993). *Principia Ethica*, Cambridge University Press.

Parfit, D. (2011). *On What Matters: Volume One*. Oxford University Press.

Pollock, J. (1974). *Knowledge and Justification*. Princeton University Press.

Pollock, J. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press.

Prakken, H. (2005). A study of accrual of arguments, with applications to evidential reasoning. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law* (ICAIL-05) (pp. 8594). The ACM Press.

Raz, J. (1999). *Engaging Reason: On the Theory of Value and Action*. Oxford University Press.

Ross, W. D. (2002). *The Right and the Good*. Clarendon Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Scanlon, T. (2014). *Being Realistic About Reasons*. Oxford University Press.

Schroeder, M. (2007). *Slaves of the Passions*. Oxford University Press.

Schroeder, M. (2011). Holism, Weight, and Undercutting. *Noûs* 45(2): 328–344.

Schroeder, M. (2015). Knowledge is belief for sufficient (objective and subjective) reason. In J. Hawthorne and T. S. Gendler (Eds.), *Oxford Studies in Epistemology* (Vol. 5, pp. 226252). Oxford University Press.

Skorupski, J. (2010). *The Domain of Reasons*. Oxford University Press.

Sylvan, K. (2014). Reasons in epistemology. In D. Pritchard (Ed.), *Oxford Bibliographies Online.*

Sylvan, K. (2016). Epistemic Reasons I: Normativity. *Philosophy Compass*, 11(7), 364376.

Turri, J. (2009). The ontology of epistemic reasons. *Noûs*, 43(3), 490512.

Reason-based Probabilistic Preference Lifting

Xiaoxuan Fu

Tsinghua University

1 Motivation: Lifting by Sampling

A preference order $\leq$ on a set W of worlds is a reflexive, antisymmetric and transitive relation on W. As has been discussed in the literature (e.g., (Liu 2011b; van Benthem 2011)), there are four kinds of rules for lifting preference orders from worlds to set of worlds, which are usually defined as follows:

a. $\forall\exists$-rule: a set Y is preferred to a set X if $\forall x \in X\, \exists y \in Y : x \leq y$

b. $\forall\forall$-rule: a set Y is preferred to a set X if $\forall x \in X\, \forall y \in Y : x \leq y$

c. $\exists\forall$-rule: a set Y is preferred to a set X if $\exists x \in X\, \forall y \in Y : x \leq y$

d. $\exists\exists$-rule: a set Y is preferred to a set X if $\exists x \in X\, \exists y \in Y : x \leq y$

Generally, the first two rules are often used in practice. Halpern (1997) provided an axiomatization for the lifted preference by the first rule. Van Benthem et al. (2011) argued that the notion of preference defined with $\forall\forall$-rule was what Von Wright had in mind when he studied preference in this book (von Wright 1963). However, in many real life scenarios, an agent α often forms his preference order without exhausting the worlds that are in the scope of $\forall$-quantifier. In such case, α only selects a certain number of worlds as samples from X and Y respectively, and then lifts his preference order after a stepwise comparison between those paring samples. Accordingly, instead of comparing all the worlds in X or Y, α's job changes to compare paring samples from these two sets step by step. This stepwise comparing procedure can be illustrated as follows:

> Without loss of generality, suppose that there are two sets $A = \{a_0, a_1, ...\}$ and $B = \{b_0, b_1, ...\}$. Now an agent α has to choose a more preferred set between them, and his strategy is using the following repeated random sampling method:

step 1: α randomly chooses a_i and b_j from A and B separately. Then he finds out a more preferred one between a_i and b_j.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

step 2: repeat step 1.

... ...

step n: α finds out which set is more preferred and terminate the comparing procedure.

where $n \in \mathbb{N}^+$.

For the purpose of making this method concrete, we provide a simple example here:

Example 1. *There are two baskets, $A = \{a_0, a_1, ...a_m\}$ and $B = \{b_0, b_1, ...b_m\}$, of apples* [1]*. An agent α chooses a more preferred basket between them by the use of the following method:*

step 1. *α chooses an apple from each basket separately, and then decides which one is more preferred, and finally, put them back;*

step 2. *repeat step 1;*

... ...

step n: *α make his decision and terminate the comparing procedure.*
($n \in \mathbb{N}^+$)

During the comparing procedure, agent α has his own criteria of what kinds of properties are better. For example, α's preference among apple's colors, size and so on. Without loss of generality, suppose that α's criteria on apple's properties is as follow:

$$\textbf{\textit{red}} \gg \textbf{\textit{big}} \gg ... \gg \textbf{\textit{round}}$$

It's obvious that α will choose basket A if the following conditions meet:

1. *with finitely random sampling, α gets more red apples in A; Or, α gets the same number of red samples in both A and B, then*
2. *α gets more big apples in A; Or, α gets the same number of big apples in both A and B, then*

... ...

n. *α gets more round apples in A.*

This comparing procedure ends whenever α finds out a better basket on the basis of his criterion.

Since A and B have the same number of apples, the result in Example 1 says that:

A is more preferred than B if

1. the probability of getting a red apple in A is higher than B; Or, the probabilities are the same, then
2. the probability of getting a big apple in A is higher than B; Or, the probabilities are the same, then

[1] The two basket of apples are supposed to have the same size, otherwise α is taking the risk of choosing one basket that he does not really prefer.

For example, $A = \{a_1, a_2\}$ and $B = \{b_1, b_2, \ldots, b_{10}\}$, where

- a_1 has the most preferred property, yet a_2 is extremely bad.
- $b_1, b_2, \ldots, b_{10}$ have the secondary preferred property.

In this case, α may still choose B under mostly practical circumstance.

... ...

n. the probability of getting a round apple in A is higher than B.

In light of the above description, in what follows, we will provide a formal semantics and a logic system to capture the ideas behind it.

2 Language and Semantics

Definition 1 (Language). *Given a nonempty set* $\mathbf{P}$ *of propositional letters and a nonempty finite set* G *of agents, the language* $\mathscr{L}_{\text{pref}}$ *is defined as follows:*

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi \wedge \phi \mid \phi \prec_\alpha \psi$$

where $p \in \mathbf{P}$ and $\alpha \in G$, and moreover, $\phi \& \psi$ are propositional letters or only constructed by boolean operations respectively.

Semantically, every propositional letter p denotes a property P of an object, and $\phi \prec_\alpha \psi$ means that agent α prefers the property ψ to ϕ.

Definition 2 (Criterion Order). *Given a nonempty set* $\mathbf{P}$ *of propositional letters and an agent* α*, a criterion order for* α *is a tuple* $\mathfrak{C}_\alpha =< \Pi, \gg >$*, where*

- $\Phi \subset \mathbf{P}$ *and* $|\Pi| \in \omega$*,*
- $\gg$ *is a strict linear order on* Π.

Definition 3 (Preference Probabilistic Models with $\mathfrak{C}_\alpha$). *A preference probabilistic model for an agent* α *with respect to* $\mathfrak{C}_\alpha$ *is a tuple* $\mathfrak{M} = \langle W, V, P \rangle$ *where*

- W *is a non-empty countable set of objects.*
- V *assigns to every* $w \in W$ *a subset of* $\mathbf{P}$.
- P *is a probability distribution from* W *to* $[0.1]$.

There are some thoughts to be noted here:

- propositional letters are taken as properties here, and all of them can be resulted in a preference order. Recall our choosing-apples example, and suppose that a and b are two apples from separate basket. Clearly, when we compare them, their properties (such as redness) are the criteria of comparison. The idea is similar to (Liu 2011a).
- according to Definition 3, an apple a, which is composed of lots of properties, can be taken as an object in W.
- for any w in W, $P(w)$ is the probability of choosing w in W. Moreover, the probability $P(A)$ of choosing a subset A of W is $\sum_{w \in A} P(w)$. While taking an arbitrary property into consideration, we can use a conditional probability to represent for the occurrence of that property in A: $P(\pi \mid A) = \dfrac{P(\pi(A))}{P(A)}$, where $\pi \in \Pi$ and $\pi(A) = \{w \in A : w \, \text{has the property} \, \pi\}$. For instance, assume that π is the property of red and A is a basket of apples. Then $P(\pi \mid A)$ stands for the probability of red apples in A.

Definition 4. *Give a nonempty finite set G of agents and $\alpha \in G$, and a preference probabilistic models with $\mathfrak{C}_\alpha$ for the language $\mathscr{L}_{\text{pref}}$. The satisfaction relation $\vDash$ between pointed models and formulas is defined as follows:*

$[\![\phi]\!] := \{w \in W : \mathfrak{M}, w \models \phi\}$
$|\,[\![\phi]\!]\,|$ *is the cardinality of* $[\![\phi]\!]$
$\pi([\![\phi]\!]) := \{w \in [\![\phi]\!] : \mathfrak{M}, w \models \pi\}$, *where* $\pi \in \Pi$

$$\begin{array}{lll}
\mathfrak{M}, w \models \top & \textit{iff} & \textit{always} \\
\mathfrak{M}, w \models p & \textit{iff} & w \in V(p) \\
\mathfrak{M}, w \models \neg\phi & \textit{iff} & \mathfrak{M}, w \not\models \phi \\
\mathfrak{M}, w \models \phi \wedge \psi & \textit{iff} & \mathfrak{M}, w \models \phi \textit{ and } \mathfrak{M}, w \models \psi \\
\mathfrak{M}, w \models \phi \prec_\alpha \psi & \textit{iff} & \mathfrak{M}, w \models \phi \prec^n_\alpha \psi, \textit{ where } \phi \prec^n_\alpha \psi \textit{ are defined as:} \\
 & & \bullet \phi \prec^1_\alpha \psi := P(\pi_1 \mid [\![\phi]\!]) \times |\,[\![\phi]\!]\,| < P(\pi_1 \mid [\![\psi]\!]) \times |\,[\![\psi]\!]\,| \\
 & & \bullet \phi \prec^{k+1}_\alpha \psi := \phi \prec^k_\alpha \psi \textit{ or,} \\
 & & \qquad \phi \simeq^k_\alpha \psi \,\&\, P(\pi_{k+1} \mid [\![\phi]\!]) \times |\,[\![\phi]\!]\,| < P(\pi_{k+1} \mid [\![\psi]\!]) \times |\,[\![\psi]\!]\,| \\
 & & (k \in n,\ \pi_i \gg \pi_{i+1} \textit{ in } \mathfrak{C}_\alpha, \textit{ and } \phi \simeq^k_\alpha \psi \textit{ whenever neither } \phi \prec^k_\alpha \psi \textit{ nor } \psi \prec^k_\alpha \phi)
\end{array}$$

The last clause satisfied at w if and only if it holds in the whole model. It implies that agent α prefers the ψ-set to the ϕ-set. In the ψ-set, he can get a greater number of preferred objects by the use of his criterion order. Since the cardinality of ϕ-set may be different from ψ-set, both cardinalities of sets and probabilities of properties should be taken into account. However, in our sampling example, we always compare two sets with the same size. Based on this presupposition, we provide a new rule of lifting preference order here.

Definition 5 (Sampling Rule)**.** *Suppose that $\pi_1 \gg \ldots \gg \pi_n$ ($n \in \omega$) is a criterion order. Then a set Y is preferred to a set X ($X \lhd Y$) if the following recursive conditions meet:*

1. $X \lhd_1 Y := P(\pi_1 \mid X) < P(\pi_1 \mid Y)$
2. $X \lhd_{k+1} Y := X \lhd_k Y \vee ((X =_k Y) \wedge (P(\pi_{k+1} \mid X) < P(\pi_{k+1} \mid Y))$
3. $X \lhd Y := X \lhd_n Y$
   where $k \in n$, $X =_k Y$ whenever neither $X \lhd_k Y$ nor $Y \lhd_k X$.

3 Axiomatization

Let L_{pref} be the logic which is complete for the above semantics. In what follows, we provide some axioms in L_{pref}:

A1 $(\top \prec_\alpha \phi) \to \phi$
A2 $(\phi \prec_\alpha \psi) \to \top \prec_\alpha (\phi \prec_\alpha \psi)$
A3 $(\bot \prec_\alpha \phi) \vee (\bot \simeq_\alpha \phi)$
A4 $(\phi \prec_\alpha \top) \vee (\phi \simeq_\alpha \top)$
A5 $(\phi \prec_\alpha \psi) \wedge (\psi \prec_\alpha \chi) \longrightarrow \phi \prec_\alpha \chi$
A6 $\neg(\phi \prec_\alpha \psi) \wedge \neg(\psi \prec_\alpha \phi) \to (\phi \simeq_\alpha \psi)$
A7 $(\phi \wedge \top \prec_\alpha \psi \wedge \top) \longrightarrow \phi \prec_\alpha \psi$
A8 $(\phi_1 \prec_\alpha \psi_1) \wedge (\phi_2 \prec_\alpha \psi_2) \longrightarrow (\phi_1 \wedge \phi_2 \prec_\alpha \psi_1 \wedge \psi_2)$
A9 $\neg(\phi \prec_\alpha \psi) \longrightarrow (\top \prec_\alpha \neg(\phi \prec_\alpha \psi))$
As rules of inference we take Modus Ponens (MP).

The definitions of theorems and consistency are as usual.

4 Further Issues

The further use of sampling rule in sequential random sampling cases. Actually, there is nothing really matter if we use repeated random sampling rather than sequential random sampling here. Because both of them will approximates the overall preference order which is obtained by exhausting all the possibilities. It can be proved that we would gain the same expected result in both ways with adequate times of sampling. The main purpose of this part is to show that our logic can be used as a simply way to deal with the sequential sampling cases.

References

Halpern, J. Y. (1997). Defining Relative Likelihood in Partially-Ordered Preferential Structure. *Journal of Artificial Intelligence Research*, 7:1–24.

Liu, F. (2011a). A Two-level Perspective on Preference. *Journal of Philosophical Logic*, 15 40(3):421–439.

Liu, F. (2011b). *Reasoning about Preference Dynamics*. Springer Science Publishers.

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

van Benthem, J., Girard, P., and Roy, O. (2009). Everything Else Being Equal: A Modal Logic for Ceteris Paribus Preferences. *Journal of Philosophical Logic*, 38: 83–125.

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

Evidence-based Expectation and Doubt

Haibin Gui

Tsinghua University

1 Motivation

Doubt is an attitude that has been highly appreciated in both philosophy and sciences. Our history abounds with great cases in which doubt leads to amazing new ideas and discoveries. However, there is rarely any logical study about doubt, as an attitude, itself. We aim to make a first attempt in this paper. Let's first consider the following example:

Example 1. *An ornithologist stayed in the southern hemisphere and observed that a group of penguins do not fly. Nevertheless, they do have feathers, airbags, and are warm-blooded, hence are considered to be birds. In other words, he had expected that those penguins can fly too. Given the new observation, he started to doubt that all birds can fly.*

We can see that the ornithologist gathered new evidence that is contrary to his "expectation". Without those evidence, he would have still believed that all birds can fly (p). But now he is in a mental state of doubt, where he believes neither p nor $\neg p$.

Clearly, the notion of doubt is closely related to evidence, belief and expectations. We will adopt the evidence-based belief framework proposed in [Ben and Pac11] and extend it to study the new notions.

2 Preliminary Definitions

Definition 1 (evidence language). *Let At be a set of atomic propositions, $p \in At$. $\mathcal{L}_0$ is the smallest set of formulas generated by the grammar below.*

$$\mathcal{L}_0 := p \mid \neg\phi \mid \phi \wedge \psi \mid B\phi \mid \Box\phi \mid A\phi$$

"$B\phi$" is read as "an agent believes that ϕ", "$\Box\phi$" is read as "an agent has evidence for ϕ, A is a universal modality.

Definition 2 (evidence model). *An evidence model is a tuple $\mathcal{M}$ = (W, E, V), where W is an non-empty set, $E \subseteq W \times \mathcal{P}$(W), V is a propositional valuation: At $\longrightarrow \mathcal{P}$(W).*

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

$E(w)$ is the set $\{X\ |wEX\}$ with $\emptyset \notin E(w)$ and $W \in E(w)$. A family $\mathcal{X}$ of subsets of W has the finite intersection property (f.i.p) if $\bigcap \mathcal{X} \neq \emptyset$. We say $\mathcal{X}$ has the maximal f.i.p if $\mathcal{X}$ has the f.i.p but no proper extension of $\mathcal{X}$ does.

Definition 3 (truth conditions)**.** *Given an evidence model, the truth condition of a formula is defined as follows.*

- *$\mathcal{M}$, w $\models \Box\phi$ iff there is an X with wEX and for all v $\in X$, $\mathcal{M}$, v $\models \phi$*
- *$\mathcal{M}$, w $\models B\phi$ iff for each max f.i.p $\mathcal{X} \subseteq$ E(w) and for all v $\in \bigcap \mathcal{X}$, $\mathcal{M}$, v $\models \phi$*
- *$\mathcal{M}$, w $\models A\phi$ iff for all v $\in W$, $\mathcal{M}$, v $\models \phi$.*

According to the above definitions, the fact concerning the relationship between evidence and belief holds:

Fact 1. *Let $\mathcal{M}$ = (W, E, V) be an evidence model, if X $\in$ E(w) and X is some evidence for ϕ, then $\mathcal{M}, w \models \neg B\neg\phi$.*

Proof. Let $\mathcal{X}$ be an arbitrary max f.i.p family of $\mathcal{M}$:

Case 1: If X is an evidence set for ϕ and it does not intersect with other sets in the model, then $\{$ X $\}$ is a max f.i.p family itself. Then it must be that case that for all $v \in X$, $\mathcal{M}, v \models \phi$. Then it does not hold that $\mathcal{M}, w \models B\neg\phi$, that is, $\mathcal{M}, w \models \neg B\neg\phi$.

Case 2: If X is an evidence set for ϕ and it intersects with other sets in the model, then there is a max f.i.p family $\mathcal{X}$ s.t. for all u $\in \bigcap \mathcal{X}$, $\mathcal{M}$, u $\models \phi$. Similarly, we have $\mathcal{M}$, w $\models$ ¬B¬ϕ.

□

It essentially says that if there is some evidence for ϕ in a model, $\neg\phi$ cannot be believed.

3 From Expectation to Doubt

To formalize the ideas of doubt in Section 1, we need to extend the language $\mathcal{L}_0$.

Definition 4 (extended language)**.** *We extend language from $\mathcal{L}_0$ to $\mathcal{L}_1$:*

$$\mathcal{L}_1 := p \mid \neg\phi \mid \phi \wedge \psi \mid B\phi \mid \Box\phi \mid A\phi \mid \mathbb{E}\phi$$

"$\mathbb{E}\phi$" expresses that ϕ is expected. Now in this language we will define "ϕ is doubtful" ($D\phi$) as $\mathbb{E}\phi \wedge \Box\neg\phi$.

Definition 5 (truth conditions)**.** *Let $\mathcal{M} = (W, E, V)$ be an evidence model, for each $w \in W$ such that $E(w)$ is defined as before. Let $E^{\neg\phi-}$(w) = E(w) – $\{$ X $|$ X is an evidence for $\neg\phi$ $\}$ and we assume that $E^{\neg\phi-}$(w) $\neq \emptyset$.*[1]

[1] Suppose $E^{\neg\phi-}$(w) is empty, all evidence in E(w) is $\neg\phi$-evidence. Then the agent would believe that $\neg\phi$ previously, which is not the case we are interested.

- *$\mathcal{M}$, w $\models \mathbb{E}\phi$ iff for each maximal f.i.p. family $\mathcal{X} \subseteq E^{\neg\phi-}$(w) and for all v $\in \bigcap \mathcal{X}$, $\mathcal{M}$,v $\models \phi$.*
- *$\mathcal{M}$, w $\models D\phi$ iff $\mathcal{M}$, w $\models \mathbb{E}\phi \wedge \Box\neg\phi$*

For us, expectation is not a belief but can be thought to be a virtual state where one might believe ϕ if some evidence regarding $\neg\phi$ is filtered out in a given evidence model. "ϕ being doubtful" is more an attitude about situation with uncertainties in which one expects ϕ but having evidence supporting $\neg\phi$.

Return to Example 1, disregarding the new observations, the ornithologist would have believed that all birds can fly, because that is what the old evidence had suggested. But after one month's observation, he can no longer say that he believes that all birds can fly, instead he might say "I had expectated that all birds can fly, but...".

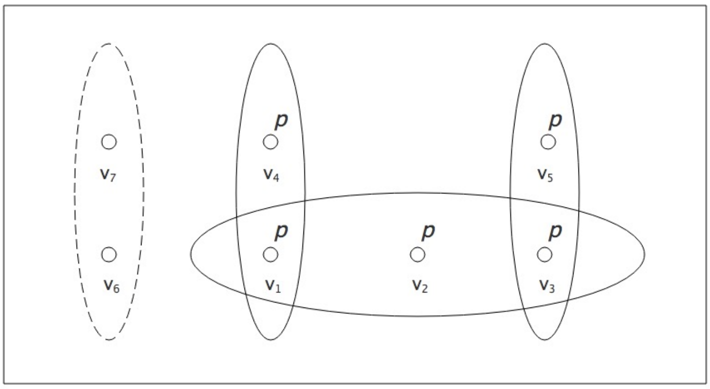


Figure 1: Example 1

As shown in Figure 1, barring evidence against p, an agent would believe that p, but the thing is that he does have evidence against p.

Remark. [Pro and Ols13] mentioned a possible definition of doubt when they discuss group beliefs, namely, thinking of "ϕ is doubtful" as $\neg B\phi \wedge \neg B\neg\phi$. Interestingly, for them, "$\neg\phi$ is doubtful" would be $\neg B\phi \wedge \neg B\neg\phi$, too. This would not be the case in our evidence model. Our language can distinguish doubt about ϕ and about $\neg\phi$, as they are expressed differently in the language.

4 Valid Principles

Fact 2. *The following principles are valid.*

(1) $\mathbb{E}\phi \to \neg A\neg\phi$

(2) $D\phi \to \mathbb{E}\phi$

(3) $B\phi \to \mathbb{E}\phi$

(4) $B\phi \to \neg D\phi$

(1) reflects our assumption that $E^{\neg\phi-}(w)$ is not empty. (2), (3) and (4) are describing the relation between belief, expectation and doubt, and the other direction of these implications do not hold. In particular, we have the following:

Fact 3. *$\mathbb{E}(\phi \to \psi) \to (\mathbb{E}\phi \to \mathbb{E}\psi)$ is not valid.*

A counter-example is as follows, one can see easily that the implication does not hold:

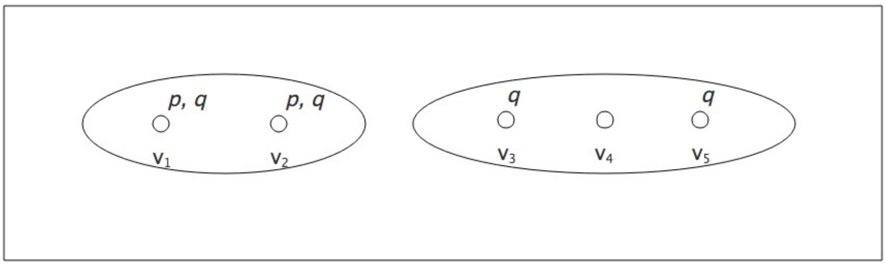


Figure 2: A counter-example for $\mathbb{E}(\phi \to \psi) \to (\mathbb{E}\phi \to \mathbb{E}\psi)$

This means, we may need non-normal logics to characterize the operator of expectation.

5 Further Issues

We have made a first step, but there are immediate issues to be studied,here we list a few:

- The current text has treated doubt as a static operator. It seems natural to think of doubt as a dynamic process with changing in evidence. It is intriguing to study how one's belief or expectation changes before and after the evidence-based doubt takes place. We have started to look at it that way.
- A new notion of monotonic bisimulation was introduced in [Han, Kup and Pac09] and then adopted in [Ben and Pac11] w.r.t the evidence operator $\Box$. We now have an extended language, and we would like to study its expressive power.
- We have looked at a few valid and invalid principles of the logic in this paper. Some of them are very interesting, expressing the relationship between belief, doubt and expectations. Their further connection will be studied in a more systematic manner in terms of possible axiomatization.
- Our discussion has been restricted to a single case, it makes a lot of sense to consider multi-agent scenarios where we can say that agent a doubts that agent b doubts that p, or agent a doubts that agent b believes that p. Moreover, we do want to look at some particular doubt that is shared in a group or a community.

We leave all these issues for future.

References

van Benthem, J. and Pacuit, E. (2011). Dynamic Logics of Evidence-Based Beliefs. *Studia Logica*, 1(3):61–92.

Proietti, C. and Olsson, E. J. (2013). A DDL Approach to Pluralistic Ignorance and Collective Belief. *Journal of Philosophical logic*, 2(3):499–515.

Hansen, H. H., Kupke, C., and Pacuit, E. (2009). Neighborhood Structures: Bisimularity and Basic Model Theory. *Logic Methods in Computer Science*, 5(2):1–38.

Game Equivalence and Expressivity of Game Description Language

Guifei Jiang [*] Laurent Perrussel [†] Dongmo Zhang [‡] Heng Zhang [§]

[*] Nankai University [†] IRIT, Université Toulouse Capitole [§] Tianjin University

[‡] Western Sydney University

Abstract

This paper investigates the equivalence between games represented by state transition models and its applications. We first define a concept of bisimulation equivalence between state transition models and prove that it can be logically characterized by Game Description Language (GDL). Then we demonstrate with real games that bisimulation equivalence can be generalized to characterize more general game equivalent. We also introduce a concept of quotient state transition model. As the minimum equivalent of the original model, it allows us to improve the efficiency of model checking for GDL. Finally we establish a characterization for the definability of GDL and show that exactly the properties of state transition models closed under k-bisimulation are definable in GDL. This provides a powerful tool to identify the expressive power of GDL.

1 Introduction

General Game Playing (GGP) is concerned with creating intelligent agents that understand the rules of previously unknown games and learn to play these games without human intervention (Genesereth et al. 2005). To represent the rules of arbitrary games, a formal game description language (GDL) was introduced as an official language for GGP in 2005. GDL is originally a machine-processable, logic programming language (Love et al. 2006). Most recently, it has been adapted as a logical language for game specification and strategic reasoning (Zhang and Thielscher

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

2015b). The epistemic and dynamic extensions have been also developed (Zhang and Thielscher 2015a; Jiang et al. 2016).

Although GDL is a logical language for representing game rules and specifying game properties, its logical properties, especially its expressive power have not been fully investigated yet. For instance, which game properties are definable or non-definable in GDL? How to show a game property is not definable in GDL? When two game descriptions are equivalent? In this paper, we will address these questions through a *bisimulation* approach.

The notion of *bisimulation* plays a pivotal role to identify the expressive power of a logic. It was independently defined and developed in the areas of theoretical computer science (Park 1981; Henness and Milner 1985) and the model theory of modal logic (van Benthem 1977, 1984). Since bisimulation-equivalent structures can simulate each other in a stepwise manner, they cannot be distinguished by the concerned logic. An appropriate notion of bisimulation for a logic allows us to study the expressive power of that logic in terms of structural invariance and language indistinguishability (Grädel and Otto 2014).

Besides identifying the expressivity of a logic, bisimulation equivalence also allows us to obtain the minimum equivalent of the original model, called *the quotient model*, which can be used to improve the efficiency of model checking (Baier et al. 2008). Moreover, in terms of GDL, bisimulation equivalence tells us when two game structures are essentially the *same*, and thus gives us a natural criterion on the equivalence between games. Exploiting game equivalence may provide a bridge for knowledge transfer between a new game and a well-studied game in GGP (Zhang et al. 2017).

Based on the above consideration, we will use in this paper a concept of bisimulation as a tool to investigate the expressive power of GDL. We first define a concept of bisimulation equivalence between state transition models and prove that it coincides with the invariance of GDL-formulas on state transition models. This justifies that the notion of bisimulation equivalence is appropriate for GDL. Then we demonstrate with real games that bisimulation equivalence can be generalized to capture a wider range of game equivalence. We also provide a characterization for the definability of GDL, and show that a class of state transition models is definable in GDL iff they are closed under k-bisimulation. This allows us to establish non-definability of a property in GDL. Finally we introduce a concept of quotient state transition model and show that it is bisimulation-equivalent to its original model.

The rest of this paper is structured as follows: Section 2 introduces the framework for game description. Section 3 defines the concept of bisimulation equivalence and introduces the notion of quotient model. Section 4 generalizes bisimulation equivalence to characterize more general game equivalence. Section 5 identifies the expressive power of GDL. Finally, we conclude with related work and future work.

2 The Framework

All games are assumed to be played in multi-agent environments. Each game is associated with a *game signature*. A *game signature* $\mathcal{S}$ is a triple $(N, \mathcal{A}, \Phi)$, where

- $N = \{1, 2, \cdots, m\}$ is a non-empty finite set of agents,
- $\mathcal{A}$ is a non-empty finite set of *actions* such that it contains $noop$, an action without any effect , and

- $\Phi = \{p, q, \cdots\}$ is a finite set of propositional atoms for specifying individual features of a game state.

Through the rest of the paper, we will consider a fixed game signature $\mathcal{S}$, and all concepts are based on the game signature unless otherwise specified.

2.1 State Transition Models

In this paper, we focus on synchronous games where all players move simultaneously. These games can be specified by *state transition models* defined as follows:

Definition 1. *A* state transition (ST) model *M is a tuple $(W, w_0, T, L, U, g, \pi)$, where*

- *W is a non-empty finite set of* possible states.
- *$w_0 \in W$, representing the unique* initial *state.*
- *$T \subseteq W$, representing a set of* terminal *states.*
- *$L \subseteq W \times N \times \mathcal{A}$ is a* legality *relation, specifying legal actions for each agent at game states. Let $L_r(w) = \{a \in \mathcal{A} : (w, r, a) \in L\}$ be the set of all legal actions for agent r at state w. To make a game playable, we assume that (i) each agent has at least one available action at each state, i.e., $L_r(w) \neq \emptyset$ for any $r \in N$ and $w \in W$, and (ii) each agent can only do action $noop$ at terminal states, i.e., $L_r(w) = \{noop\}$ for any $r \in N$ and $w \in T$.*
- *$U : W \times \mathcal{A}^{|N|} \rightarrow W \backslash \{w_0\}$ is an* update *function, specifying the state transition for each state and legal joint action, such that $U(w, \langle noop^r \rangle_{r \in N}) = w$ for any $w \in W$.*
- *$g : N \rightarrow 2^W$ is a* goal *function, specifying the winning states of each agent.*
- *$\pi : W \rightarrow 2^\Phi$ is a standard valuation function.*

Note that the turn-based game structure involved in (Zhang and Thielscher 2015b) is a special case by allowing a player only to do "noop" when it is not her turn. For convenience, let D denote the set of all joint actions $\mathcal{A}^{|N|}$. Given $d \in D$, we use $d(r)$ to specify the action taken by agent r.

The following notion specifies all possible ways in which a game can develop.

Definition 2. *Let $M = (W, w_0, T, L, U, g, \pi)$ be an ST-model. A path δ is an infinite sequence of states and joint actions $w_0 \xrightarrow{d_1} w_1 \xrightarrow{d_2} \cdots \xrightarrow{d_j} \cdots$ such that for any $j \geq 1$ and $r \in N$,*

1. *$w_j \neq w_0$ (that is, only the first state is initial.)*
2. *$d_j(r) \in L_r(w_{j-1})$ (that is, any action that is taken by each agent must be legal.)*
3. *$w_j = U(w_{j-1}, d_j)$ (state update)*
4. *if $w_{j-1} \in T$, then $w_{j-1} = w_j$ (self-loop after reaching a terminal state.)*

Let $\mathcal{P}(M)$ denote the set of all paths in M. For $\delta \in \mathcal{P}(M)$ and a stage $j \geq 0$, we use $\delta[j]$ to denote the j-th state of δ and $\theta_r(\delta, j)$ to denote the action taken by agent r at stage j of δ.

2.2 The Language

Let us now introduce a language based on GDL from (Zhang and Thielscher 2015b) for game specification.

Definition 3. *The language $\mathcal{L}$ for game description is generated by the following BNF:*

$$\varphi ::= p \mid initial \mid terminal \mid legal(r,a) \mid wins(r) \mid does(r,a) \mid \neg\varphi \mid \varphi \wedge \psi \mid \bigcirc\varphi$$

where $p \in \Phi$, $r \in N$ and $a \in \mathcal{A}$.

Other connectives $\vee, \rightarrow, \leftrightarrow, \top, \bot$ are defined by $\neg$ and $\wedge$ in the standard way. Intuitively, $initial$ and $terminal$ specify the initial state and the terminal states of a game, respectively; $does(r,a)$ asserts that agent r takes action a at the current state; $legal(r,a)$ asserts that agent r is allowed to take action a at the current state, and $wins(r)$ asserts that agent r wins at the current state. Finally, the formula $\bigcirc\varphi$ means that φ holds in the next state.

We use the following abbreviations in the rest of paper. For $d = \langle a_r\rangle_{r\in N}$, $does(d) =_{def} \bigwedge_{r\in N} does(r,a_r)$, and $\bigcirc^k\varphi =_{def} \underbrace{\bigcirc\cdots\bigcirc}_{k}\varphi$. To help the reader capture the intuition of the language, let us consider the following example.

Example 1 (Number Scrabble)**.** *Two players take turns to select numbers from 1 to 9 without repeating any numbers previously used. The first player who selects three numbers that add up to 15 wins.*

The game signature $\mathcal{S}_{NS}$ is given as follows: $N_{NS} = \{\mathsf{b}, \mathsf{w}\}$ denoting two game players; $\mathcal{A}_{NS} = \{\alpha(n) \mid 1 \leq n \leq 9\} \cup \{noop\}$, where $\alpha(n)$ denotes selecting number n, and $\Phi_{NS} = \{s(r,n), turn(r) \mid r \in \{\mathsf{b}, \mathsf{w}\}$ and $1 \leq n \leq 9\}$, where $s(r,n)$ represents the fact that number n is selected by player r, and $turn(r)$ says that player r has the turn now. The rules of Number Scrabble can be naturally formulated by GDL-formulas as shown in Figure 1 (where $r \in \{\mathsf{b}, \mathsf{w}\}$ and $-r$ represents r's opponent).

1. $initial \leftrightarrow turn(\mathsf{b}) \wedge \neg turn(\mathsf{w}) \wedge \bigwedge_{i=1}^{9} \neg(s(\mathsf{b},i) \vee s(\mathsf{w},i))$

2. $wins(r) \leftrightarrow \big(\bigvee_{i=2}^{3}(s(r,i) \wedge s(r,4) \wedge s(r,11-i)) \vee \bigvee_{i=1}^{2}(s(r,i) \wedge s(r,6) \wedge s(r,9-i)) \vee$
$\bigvee_{l=1}^{4}(s(r,5-l) \wedge s(r,5) \wedge s(r,5+l))\big)$

3. $teminal \leftrightarrow wins(\mathsf{b}) \vee wins(\mathsf{w}) \vee \bigwedge_{i=1}^{9}(s(\mathsf{b},i) \vee s(\mathsf{w},i))$

4. $legal(r,\alpha(n)) \leftrightarrow \neg(s(\mathsf{b},n) \vee s(\mathsf{w},n)) \wedge turn(r) \wedge \neg terminal$

5. $legal(r,noop) \leftrightarrow turn(-r) \vee terminal$

6. $\bigcirc s(r,n) \leftrightarrow s(r,n) \vee (\neg(s(\mathsf{b},n) \vee s(\mathsf{w},n)) \wedge does(r,\alpha(n)))$

7. $turn(r) \wedge \neg terminal \rightarrow \bigcirc\neg turn(r) \wedge \bigcirc turn(-r)$

Figure 1: A GDL description of Number Scrabble.

Formula 1 says at the initial state, player b *has the first turn and all numbers are not selected.*

The next two formulas specify winning states of each player and the terminal states, respectively. The player who succeeds in selecting three numbers that add up to 15 *wins the game (Formula 2). The game ends if one of the players wins or all numbers are selected (Formula 3).*

The preconditions of each action (legality) are specified by Formula 4 and Formula 5. The player who has the turn at a non-terminal state can select any number that is not selected before. A player can do action $noop$ *if it is not her turn or the game terminates.*

Formula 6 is the combination of the frame axioms and the effect axioms (Reiter 1991). It states that a number is selected by a player in the next state if the player selects that number at the current state or the number has been selected by her before. The last formula specifies the turn-taking.

2.3 The Semantics

The semantics of this language is based on ST-models with respect to a path and a stage of the path.

Definition 4. *Let* $M = (W, w_0, T, L, U, g, \pi)$ *be an ST-model. Given a path* δ *of* M*, a stage* $j \geq 0$ *and a formula* $\varphi \in \mathcal{L}$*, we say* φ *is true (or satisfied) at* j *of* δ *under* M*, denoted* $M, \delta, j \models \varphi$*, according to the following definition*:

$$\begin{array}{lll}
M, \delta, j \models p & \textit{iff} & p \in \pi(\delta[j]) \\
M, \delta, j \models \neg\varphi & \textit{iff} & M, \delta, j \not\models \varphi \\
M, \delta, j \models \varphi_1 \wedge \varphi_2 & \textit{iff} & M, \delta, j \models \varphi_1 \textit{ and } M, \delta, j \models \varphi_2 \\
M, \delta, j \models initial & \textit{iff} & \delta[j] = w_0 \\
M, \delta, j \models terminal & \textit{iff} & \delta[j] \in T \\
M, \delta, j \models wins(r) & \textit{iff} & \delta[j] \in g(r) \\
M, \delta, j \models legal(r, a) & \textit{iff} & a \in L_r(\delta[j]) \\
M, \delta, j \models does(r, a) & \textit{iff} & \theta_r(\delta, j) = a \\
M, \delta, j \models \bigcirc\varphi & \textit{iff} & M, \delta, j+1 \models \varphi
\end{array}$$

A formula φ is *valid* in an ST-model M, written $M \models \varphi$, if $M, \delta, j \models \varphi$ for any $\delta \in \mathcal{P}(M)$ and $j \geq 0$. A formula φ is called *satisfied at a state* w in M, written $M, w \models \varphi$, if it is true for all paths going through w, i.e., $M, \delta, j \models \varphi$ for any $\delta \in \mathcal{P}(M)$ and any $j \geq 0$ with $\delta[j] = w$. It follows that $M, w_0 \models \varphi$ iff $M, \delta, 0 \models \varphi$ for all $\delta \in \mathcal{P}(M)$.

3 Bisimulation Equivalence

In this section, we define the concept of bisimulation equivalence over state transition models and show it coincides with the invariance of GDL-formulas. We also introduce the quotient state transition model in terms of such relation.

3.1 Bisimulation

Inspired by the notion of bisimulation in (Blackburn et al. 2002, 2006), we define the concept of state-based bisimulation equivalence between ST-models as follows:

Definition 5. *Let* $M = (W, w_0, T, L, U, g, \pi)$ *and* $M' = (W', w'_0, T', L', U', g', \pi')$ *be two ST-models. We say* M *and* M' *are bisimulation-equivalent, (bisimilar, for short), written* $M \approx M'$*, if there is a binary relation* $Z \subseteq W \times W'$ *such that* $w_0 Z w'_0$*, and for all states* $w \in W$ *and* $w' \in W'$ *with* wZw'*, the following conditions hold:*

1. $\pi(w) = \pi'(w')$*;*

2. $w = w_0$ *iff* $w' = w'_0$*;*

3. $w \in T$ *iff* $w' \in T'$*;*

4. $a \in L_r(w)$ *iff* $a \in L'_r(w')$ *for any* $r \in N$ *and* $a \in \mathcal{A}$*;*

5. $w \in g(r)$ *iff* $w' \in g'(r)$ *for any* $r \in N$*;*

6. If $U(w, d) = u$*, then there is* $u' \in W'$ *s.t.* $U'(w', d) = u'$ *and* uZu'*;*

7. If $U'(w', d) = u'$*, then there is* $u \in W$ *s.t.* $U(w, d) = u$ *and* uZu'*.*

Note that $\approx$ is an equivalence relation over ST-models. When Z is a bisimulation linking two states w in M and w' in M', we say that w and w' are *bisimilar*, written $M, w \leftrightarroweq M', w'$. In particular, if $M \approx M'$, then their initial states are bisimilar, i.e., $M, w_0 \leftrightarroweq M', w'_0$.

Another way to understand bisimulation equivalence is to observe that M *is bisimilar to* M' *iff each path that can be developed in one model can also be induced in the other.* To formalize this idea, we need generalize the notion of bisimilar over states to paths as follows:

Definition 6. *Consider two ST-models* M *and* M'*. Given two paths* $\delta := w_0 \overset{d_1}{\rightarrow} w_1 \overset{d_2}{\rightarrow} \cdots$ *in* M *and* $\delta' := w'_0 \overset{d'_1}{\rightarrow} w'_1 \overset{d'_2}{\rightarrow} \cdots$ *in* M'*, we say* δ *and* δ' *are bisimilar, written* $M, \delta \leftrightarroweq M', \delta'$*, iff for every* $j \geq 0$ *and* $r \in N$*,* $M, \delta[j] \leftrightarroweq M', \delta'[j]$ *and* $\theta_r(\delta, j) = \theta_r(\delta', j)$*.*

That is, two paths are bisimilar if (i) all the corresponding states are bisimilar, and (ii) each agent takes the same action at every stage. With this, the above idea is restated as follows:

Lemma 1. *Given two ST-models* M *and* M'*,* $M \approx M'$ *iff (1) for every* $\delta \in \mathcal{P}(M)$*, there is* $\delta' \in \mathcal{P}(M')$ *such that* $M, \delta \simeq M', \delta'$*, and (2) for every* $\delta' \in \mathcal{P}(M')$*, there is* $\delta \in \mathcal{P}(M)$ *such that* $M, \delta \simeq M', \delta'$*.*

Proof. $(\Rightarrow)$ This direction holds directly by Condition 6 & 7 of Def 5.

$(\Leftarrow)$ Let $Z = \{(w, w') \mid$ there are $\delta \in \mathcal{P}(M), \delta' \in \mathcal{P}(M')$ and $j \geq 0$ such that $\delta[j] = w, \delta'[j] = w'$, the local properties Condition 1-5 in Def 5 hold for $\delta[j]$ and $\delta'[j]$, and $\theta_r(\delta, j) = \theta_r(\delta', j)\}$. Such a relation Z exists due to the assumption. It is easy to show that Z is a bisimulation between M and M'. □

3.2 Invariance

Let us now turn to the logical characterization of bisimulation equivalence. We begin with the invariance of GDL-formulas under path-bisimulation.

Proposition 1. *Let* M*,* M' *be two ST-models. For every* $\delta \in \mathcal{P}(M)$ *and* $\delta' \in \mathcal{P}(M')$*, if* $M, \delta \leftrightarroweq M', \delta'$*, then* $(M, \delta, j \models \varphi$ *iff* $M', \delta', j \models \varphi)$ *for any* $j \geq 0$ *and* $\varphi \in \mathcal{L}$*.*

It is routine to prove this by induction on φ. That is, two bisimilar paths preserve GDL-formulas at each stage. Note that the other direction does not hold. Here is a simple counter-example. Let M and M' be two ST-model depicted in Figure 2, where $N = \{r\}$ and $\Phi = \emptyset$. Now consider two paths $\delta = w_0 \overset{a}{\rightarrow} w_1 \overset{b}{\rightarrow} \cdots$ in M and $\delta' = w'_0 \overset{a}{\rightarrow} w'_1 \overset{b}{\rightarrow} \cdots$ in M'. As $w_3 \notin T$ and $w'_3 \in T'$, then

$M, w_3 \not\simeq M', w'_3$, i.e., the successors of w_1 and w'_1 are not bisimilar, so $M, w_1 \not\simeq M', w'_1$. Thus, δ and δ' are not bisimilar, i.e., $M, \delta \not\simeq M', \delta'$. But it is easy to check that at each stage, δ and δ' satisfy the same GDL-formulas.

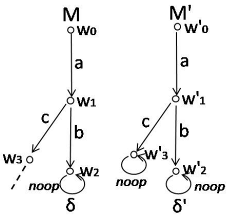


Figure 2: δ and δ' are not bisimilar.

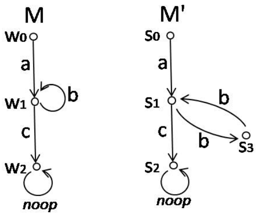


Figure 3: M and M' are not bisimulation-equivalent.

Next we provide the logical characterization of bisimulation equivalence as follows:

Theorem 1. *Let M and M' be any two ST-models. Then $M \approx M'$ iff they satisfy the same GDL-formulas.*

Proof. Assume $M \approx M'$. For symmetry, it suffices to prove one case. For every $\varphi \in \mathcal{L}$, assume φ is satisfied in M. then there is $\delta \in \mathcal{P}(M)$ and stage $j \geq 0$ such that $M, \delta, j \models \varphi$. By the assumption and Lemma 1, there is $\delta' \in \mathcal{P}(M')$ such that $M, \delta \underline{\leftrightarrow} M', \delta'$. And by Proposition 1, we have $M', \delta', j \models \varphi$. Thus, φ is satisfied in M'.

To prove the other direction, we need one additional notion. For each path $\delta := w_0 \xrightarrow{d_1} w_1 \xrightarrow{d_2} \cdots \xrightarrow{d_j} \cdots$ in M, we induce a *trace* $V(\delta) = V(w_0) \cdot does(d_1) \cdot V(w_1) \cdots does(d_{j-1}) \cdot V(w_j) \cdots$. Let $trace(M)$ denote the set of all traces in M, i.e., $trace(M) = \{V(\delta) \mid \delta \in \mathcal{P}(M)\}$. Then it holds that *for two ST-models M and M', $M \approx M'$ iff $trace(M) = trace(M')$*. Now assume $M \not\approx M'$, then by above fact there is $\delta \in \mathcal{P}(M)$ for all $\delta' \in \mathcal{P}(M')$ $V(\delta) \neq V(\delta')$. It follows that for each $\delta' \in \mathcal{P}(M)$ there is $k \geq 0$ such that either $does(d_{k+1}) \neq does(d'_{k+1})$ or $V(\delta[k]) \neq V(\delta'[k])$. From the former, we obtain a formula $does(r, a_{k+1})$ (for $r \in N$ and $a_{k+1} \in \mathcal{A}$) such that $M, \delta, k \models does(r, a_{k+1})$ and $M', \delta', k \not\models does(r, a_{k+1})$. From the latter, we obtain a formula $\chi \in Atm$ such that either (i) $M, \delta, k \models \chi$ and $M', \delta', k \not\models \chi$, or (ii) $M, \delta, k \models \neg\chi$ and $M', \delta', k \not\models \neg\chi$. Let $\varphi_{\delta'}$ be the formula of the form $\bigcirc^k does(r, a_{k+1})$, $\bigcirc^k \chi$ or $\bigcirc^k \neg\chi$ to distinguish δ from δ'. It follows by the construction that $M, \delta, 0 \models \varphi_{\delta'}$ and $M', \delta', 0 \not\models \varphi_{\delta'}$. Let Δ be the conjunctions of all such obtained formulas for all paths in M', i.e., $\Delta := \bigwedge_{\delta' \in \mathcal{P}(M')} \varphi_{\delta'}$. Note that Δ is well-formed due to the fact that M' is finite-branching. Let us now consider formula $initial \wedge \Delta$. Then it is satisfied in M, i.e., $M, \delta, 0 \models initial \wedge \Delta$. But it is unsatisfied in M'. Otherwise, there are some $\delta' \in \mathcal{P}(M')$ and $j \geq 0$ such that $M', \delta', j \models initial \wedge \Delta$, then $M', \delta', 0 \models \Delta$, so $M', \delta', 0 \models \varphi_{\delta'}$, contradicting with $M', \delta', 0 \not\models \varphi_{\delta'}$. Thus, M and M' fail to satisfy the same set of GDL-formulas. □

This theorem asserts that bisimulation equivalence and the invariance of GDL-formulas match on ST-models. On the one hand, this result justifies that the notion of bisimulation equivalence is natural and appropriate for GDL; On the other hand, it allows us to show the failure of bisimulation-equivalence easily. *Two ST-models are not bisimulation-equivalent if there is a GDL-formula that holds in one model and fails in the other.* For instance, let us consider two ST-models depicted in Figure 3, where $N = \{r\}$ and $\Phi = \emptyset$. One can find formula $initial \wedge \bigcirc^2 does(r, c)$ that holds in M, but fails in M'. This leads to $M \not\approx M'$.

Let $\mathcal{T}(M) = \{\varphi \in \mathcal{L} \mid M \models \varphi\}$ be the set of all valid GDL-formulas in M, called the *theory* of M. It follows that

Corollary 1. *Let M and M' be two ST-models. Then $M \approx M'$ iff $\mathcal{T}(M) = \mathcal{T}(M')$.*

Thus, two ST-models are bisimulation-equivalent if and only if they enjoy exactly the same properties. Alternatively, two ST-models are not bisimulation-equivalent if one has a property that the other does not have.

3.3 Bisimulation Quotient

In this subsection, we provide an alternative perspective to consider bisimulation as a relation between states within a single ST-model. Then we introduce the quotient ST-model under such relation.

Definition 7. *Let $M = (W, w_0, T, L, U, g, \pi)$ be an ST-models. A bisimulation is a binary relation $Z \subseteq W \times W$ s.t. for all states $w_1, w_2 \in W$ with $w_1 Z w_2$,*

1. *$\pi(w_1) = \pi(w_2)$;*
2. *$w_1 = w_0$ iff $w_2 = w_0$;*
3. *$w_1 \in T$ iff $w_2 \in T$;*
4. *$a \in L_r(w_1)$ iff $a \in L_r(w_2)$ for any $r \in N$ and $a \in \mathcal{A}$;*
5. *$w_1 \in g(r)$ iff $w_2 \in g(r)$ for any $r \in N$;*
6. *If $U(w_1, d) = u_1$, then there is $u_2 \in W$ s.t. $U(w_2, d) = u_2$ and $u_1 Z u_2$;*
7. *If $U(w_2, d) = u_2$, then there is $u_1 \in W$ s.t. $U(w_1, d) = u_1$ and $u_1 Z u_2$.*

States w_1 and w_2 are bisimulation-equivalent, denoted by $w_1 \sim_M w_2$, if there is a bisimulation Z for M with $w_1 Z w_2$.

It follows that a bisimulation over states for ST-model M is a bisimulation over ST-models for the pair (M, M). Clearly, $\sim_M$ is an equivalence relation on W. For $w \in W$, let $[w]_{\sim_M}$ be the equivalence class of state w under $\sim_M$, i.e., $[w]_{\sim_M} = \{w' \in W \mid w \sim_M w'\}$. We next define the quotient ST-model under such bisimulation equivalence.

Definition 8. *For an ST-model $M = (W, w_0, T, L, U, g, \pi)$ and a bisimulation equivalence $\sim_M$, the quotient ST-model $M/\sim_M = (W', w'_0, T', L', U', g', \pi')$ is defined as follows:*

- *$W' = \{[w]_{\sim_M} \mid w \in W\}$ is the set of all $\sim_M$-equivalence classes;*
- *$w'_0 = [w_0]_{\sim_M}$;*
- *$T' = \{[w]_{\sim_M} \mid w \in T\}$;*
- *$a \in L'_r([w]_{\sim_M})$ iff $a \in L_r(w)$ for any $r \in N$ and $a \in \mathcal{A}$;*
- *$U'([w]_{\sim_M}, d) = [u]_{\sim_M}$ iff $U(w, d) = u'$ for some $u' \in [u]_{\sim_M}$;*
- *$[w]_{\sim_M} \in g'(r)$ iff $w \in g(r)$ for any $r \in N$;*
- *$p \in \pi'([w]_{\sim_M})$ iff $p \in \pi(w)$ for any $p \in \Phi$.*

Note that the defined quotient ST-model is indeed a state transition model, and it is minimum as $\sim_M$ is the coarsest bisimulation for M. Moreover, an ST-model and its quotient ST-model are bisimulation-equivalent.

Proposition 2. *For any ST-model M, $M \approx M/{\sim_M}$.*

Proof. It follows from the fact that $Z = \{(w, [w]_{\sim_M}) \mid w \in W\}$ is a bisimulation between M and $M/{\sim_M}$. □

Combining this result and Theorem 1 allows us to perform model checking on the bisimulation-equivalent quotient ST-model. *A GDL-formula holds for the quotient if and only if it also holds for the original ST-model.* This provides a way to improve the efficiency of model checking for GDL in (Ruan et al. 2009; Jiang et al. 2016). Note that an adaption of bisimulation-quotienting algorithms for a finite transition system in (Baier et al. 2008) can be used to compute the quotient ST-model.

4 Bisimulation and Game Equivalence

State transition models may be viewed as representations of games, and bisimulation equivalence tells us when two state transition models are essentially the same. Thus, bisimulation equivalence provides a criterion on the equivalence between games, i.e., two games are equivalent if their state transition models are bisimulation-equivalent. In this section, we generalize this concept to capture more general game equivalence.

4.1 Structure Bisimulation

Let us consider two games: Number Scrabble in Example 1 and Tic-Tac-Toe specified as follows:

Example 2 (Tic-Tac-Toe)**.** *Two players take turns in marking either a cross 'x'or a nought 'o' on an 3×3 board. The player who first gets three consecutive marks of her own symbol in a row (horizontally, vertically, or diagonally), wins this game.*

The game signature for Tic-Tac-Toe, written $\mathcal{S}_{TT}$, is given as follows: $N_{TT} = \{\mathsf{x}, \mathsf{o}\}$ denoting the two game players; $\mathcal{A}_{TT} = \{a_{i,j} \mid 1 \leq i, j \leq 3\} \cup \{noop\}$, where $a_{i,j}$ denotes filling cell (i, j), and $\Phi_{TT} = \{p^r_{i,j}, turn(r) \mid r \in \{\mathsf{x}, \mathsf{o}\}$ and $1 \leq i, j \leq 3\}$, where $p^r_{i,j}$ represents the fact that cell (i, j) is filled by player r. The rules of this game is given in Figure 6.

The initial state, each player's winning states, the terminal states and the turn-taking are given by formulas 1-3 and 7, respectively. Two players, x and o take turns marking the spaces in a 3×3 board (Formula 7). All cells are empty in the initial state, and player x has the first turn (Formula 1). The player who succeeds in placing three respective marks in a horizontal, vertical, or diagonal row wins the game (Formula 2). The game ends if one of the players wins or all cells are filled (Formula 3).

The preconditions of each action (legality) are specified by 4 and 5. The player who has the turn at non-terminal states can fill any cell that is empty. The player can only do action $noop$ if she does not have the turn or the game terminates.

Formula 6 is the combination of the frame axioms and the effect axioms (Reiter 1991). It states that a cell is marked with a player's symbol in the next state if the player takes the corresponding action at the current state or the cell has been filled by her symbol before.

1. $initial \leftrightarrow turn(\mathsf{x}) \wedge \neg turn(\mathsf{o}) \wedge \bigwedge_{i,j=1}^{3} \neg(p^{\mathsf{x}}_{i,j} \vee p^{\mathsf{o}}_{i,j})$
2. $wins(r) \leftrightarrow \bigvee_{i=1}^{3} \bigwedge_{l=0}^{2} p^r_{i,1+l} \vee \bigvee_{j=1}^{3} \bigwedge_{l=0}^{2} p^r_{1+l,j} \vee \bigwedge_{l=0}^{2} p^r_{1+l,1+l} \vee \bigwedge_{l=0}^{2} p^r_{1+l,3-l}$
3. $teminal \leftrightarrow wins(\mathsf{x}) \vee wins(\mathsf{o}) \vee \bigwedge_{i,j=1}^{3} (p^{\mathsf{x}}_{i,j} \vee p^{\mathsf{o}}_{i,j})$
4. $legal(r, a_{i,j}) \leftrightarrow \neg(p^{\mathsf{x}}_{i,j} \vee p^{\mathsf{o}}_{i,j}) \wedge turn(r) \wedge \neg terminal$
5. $legal(r, noop) \leftrightarrow turn(-r) \vee terminal$
6. $\bigcirc p^r_{i,j} \leftrightarrow p^r_{i,j} \vee (does(r, a_{i,j}) \wedge \neg(p^{\mathsf{x}}_{i,j} \vee p^{\mathsf{o}}_{i,j}))$
7. $turn(r) \wedge \neg terminal \rightarrow \bigcirc \neg turn(r) \wedge \bigcirc turn(-r)$

Figure 6: A GDL description of Tic-Tac-Toe.

Although the two games appear different in their game descriptions, they are actually equivalent (isomorphic) (Michon 1967; Pell 1993). Unfortunately, bisimulation equivalence is not able to capture such game equivalence as they are based on different game signatures. Considering this, we generalize the notion of bisimulation equivalence as follows:

Definition 9. *Consider two ST-models $M_{\mathcal{S}} = (W, w_0, T, L, U, g, \pi)$ with $\mathcal{S} = (N, \mathcal{A}, \Phi)$ and $M'_{\mathcal{S}'} = (W', w'_0, T', L', U', g', \pi')$ with $\mathcal{S}' = (N', \mathcal{A}', \Phi')$. $M_{\mathcal{S}}$ and $M'_{\mathcal{S}'}$ are structure-equivalent, written $M_{\mathcal{S}} \sim M'_{\mathcal{S}'}$, if there are bijections $f_1 : N \mapsto N'$, $f_2 : \mathcal{A} \mapsto \mathcal{A}'$, $f_3 : \Phi \mapsto \Phi'$, and a relation $Z \subseteq W \times W'$ such that $w_0 Z w'_0$ and for all states $w \in W$ and $w' \in W'$ with wZw', the following conditions hold:*

1. *$p \in \pi(w)$ iff $f_3(p) \in \pi'(w')$;*
2. *$w = w_0$ iff $w' = w'_0$;*
3. *$w \in T$ iff $w' \in T'$;*
4. *$a \in L_r(w)$ iff $f_2(a) \in L'_{f_1(r)}(w')$ for any $r \in N$ and $a \in \mathcal{A}$;*
5. *$w \in g(r)$ iff $w' \in g'(f_1(r))$ for any $r \in N$;*
6. *If $U(w, d) = u$, then there is $u' \in W'$ s.t. $U'(w', \langle f_2(d(r))\rangle_{r \in N}) = u'$ and uZu';*
7. *If $U'(w', d') = u'$, then there is $u \in W$ s.t. $U(w, \langle f_2^{-1}(d'(r'))\rangle_{r' \in N'}) = u$ and uZu'.*

Note that $\sim$ is an equivalence relation over ST-models (with different game signatures). Clearly, $\approx$ is a special case of $\sim$ when $\mathcal{S} = \mathcal{S}'$. We say that two games are *equivalent* if their ST-models are structure-equivalent.

Let us back to the examples. As we expected, the equivalence between Number Scrabble and Tic-Tac-Toe can be captured by structure equivalence. The mapping between their state transition models is demonstrated in Table 1. The basic idea is that *filling a cell corresponds to selecting* *the number in the cell, and the fact that a cell is filled amounts to the fact that the corresponding number is selected.* For instance, filling the left-bottom cell corresponds to selecting number 4, i.e., $f_2(a_{1,1}) = \alpha(4)$, and the fact that the center is filled by player x maps the fact that number 5 is selected by player b, i.e., $f_3(p^{\mathsf{x}}_{2,2}) = s(\mathsf{b}, 5)$. And the structure-bisimulation relation starts from their initial states and can be constructed step by step according to the mapping. For example, the states depicted in Table 2 and Figure 7 are structure-bisimilar.

2	7	6
9	5	1
4	3	8

Table 1: The Mapping.

	X	

Table 2: Filling the Center.

$\{1,2,3,4,\underline{5},6,7,8,9\}$
b

Figure 7: Selecting Number 5.

Similarly, we say that w and w' are structure-bisimilar, written $M_{\mathcal{S}}, w \underline{\leftrightarrow}_s M'_{\mathcal{S}'}, w'$, if Z links two states w in $M_{\mathcal{S}}$ and w' in $M'_{\mathcal{S}'}$. In particular, for two paths $\delta := w_0 \xrightarrow{d_1} w_1 \xrightarrow{d_2} \cdots \xrightarrow{d_j} \cdots$ in $M_{\mathcal{S}}$ and $\delta' := w'_0 \xrightarrow{d'_1} w'_1 \xrightarrow{d'_2} \cdots \xrightarrow{d'_j} \cdots$ in $M'_{\mathcal{S}'}$, we say that δ and δ' are *structure-bisimilar*, written $M_{\mathcal{S}}, \delta \underline{\leftrightarrow}_s M'_{\mathcal{S}'}, \delta'$, iff for every $j \geq 0$ and $r \in N$, $M_{\mathcal{S}}, \delta[j] \underline{\leftrightarrow}_s M'_{\mathcal{S}'}, \delta'[j]$ and $f_2(\theta_r(\delta, j)) = \theta_{f_1(r)}(\delta', j)$. Similar to Bisimulation Equivalence, the following result displays that two ST-models are structure-equivalent if and only if each path that can be developed in one model can be also *simulated* in the other.

Lemma 2. *Given two ST-models $M_{\mathcal{S}}$ and $M'_{\mathcal{S}'}$, $M_{\mathcal{S}} \sim M'_{\mathcal{S}'}$, iff for every $\delta \in \mathcal{P}(M_{\mathcal{S}})$, there is $\delta' \in \mathcal{P}(M'_{\mathcal{S}'})$ such that $M_{\mathcal{S}}, \delta \underline{\leftrightarrow}_s M'_{\mathcal{S}'}, \delta'$, and vice versa.*

4.2 Logical Characterization of Structure Equivalence

Let us turn to the logical characterization of structure equivalence. We begin with the transformation of GDL-formulas. The translation between languages is defined as follows:

Definition 10. *Consider two game signatures $\mathcal{S} = (N, \mathcal{A}, \Phi)$ and $\mathcal{S}' = (N', \mathcal{A}', \Phi')$ with the same bijections $f_1 : N \mapsto N'$, $f_2 : \mathcal{A} \mapsto \mathcal{A}'$ and $f_3 : \Phi \mapsto \Phi'$ of Definition 9. A translation* tr *is a bijective mapping from $\mathcal{L}_{\mathcal{S}}$ onto $\mathcal{L}_{\mathcal{S}'}$ such that for $p \in \Phi$, $r \in N$ and $a \in \mathcal{A}$,*

$$\begin{array}{ll}
\mathrm{tr}(p) = f_3(p) & \mathrm{tr}(initial) = initial \\
\mathrm{tr}(terminal) = terminal & \mathrm{tr}(wins(r)) = wins(f_1(r)) \\
\mathrm{tr}(legal(r, a)) = legal(f_1(r), f_2(a)) & \mathrm{tr}(does(r, a)) = does(f_1(r), f_2(a)) \\
\mathrm{tr}(\neg\varphi) = \neg\mathrm{tr}(\varphi) & \mathrm{tr}(\varphi \wedge \psi) = \mathrm{tr}(\varphi) \wedge \mathrm{tr}(\psi) \\
\mathrm{tr}(\bigcirc\varphi) = \bigcirc\mathrm{tr}(\varphi) &
\end{array}$$

Note that such a translation exists as there is a bijective mapping between the game signatures. The following result holds that if two paths are structure-bisimilar, then they preserve the corresponding GDL-formulas at each stage.

Lemma 3. *Let $M_{\mathcal{S}}$, $M'_{\mathcal{S}'}$ be two ST-models. For every $\delta \in \mathcal{P}(M_{\mathcal{S}})$ and $\delta' \in \mathcal{P}(M'_{\mathcal{S}'})$, if $M_{\mathcal{S}}, \delta \underline{\leftrightarrow}_s M'_{\mathcal{S}'}, \delta'$, then $M_{\mathcal{S}}, \delta, j \models \varphi$ iff $M'_{\mathcal{S}'}, \delta', j \models \mathrm{tr}(\varphi)$ for any $\varphi \in \mathcal{L}_{\mathcal{S}}$ and $j \geq 0$.*

Note that the converse to this proposition does not hold. Please refer to Figure 2 for a counterexample. We now provide the following logical characterization result that structure equivalence and the invariance of the corresponding GDL-formulas coincide on ST-models.

Proposition 3. *Let $M_{\mathcal{S}}$, $M'_{\mathcal{S}'}$ be two ST-models. The following are equivalent.*

1. *$M_{\mathcal{S}} \sim M'_{\mathcal{S}'}$*
2. *for every $\varphi \in \mathcal{L}_{\mathcal{S}}$, φ is satisfied in $M_{\mathcal{S}}$ iff $\mathrm{tr}(\varphi)$ is satisfied in $M'_{\mathcal{S}'}$.*

Proof. The direction from Clause 1 to Clause 2 follows from Lemma 2 and Lemma 3. To prove the other direction, we need the following notion.

Consider two ST-models $M_{\mathcal{S}} = (W, w_0, T, L, U, g, \pi)$ with $\mathcal{S} = (N, \mathcal{A}, \Phi)$ and $M'_{\mathcal{S}'} = (W', w'_0, T', L', U', g', \pi')$ with $\mathcal{S}' = (N', \mathcal{A}', \Phi')$. Given bijections $f_1 : N \mapsto N'$, $f_2 : \mathcal{A} \mapsto \mathcal{A}'$ and $f_3 : \Phi \mapsto \Phi'$, let tr be a translation defined in Definition 10. A translation Tr is a bijection from $trace(M)$ to $trace(M')$. For every $V(\delta) \in trace(M)$, $\mathrm{Tr}(V(\delta)) = \mathrm{tr}(V(w_0)) \cdot \mathrm{tr}(does(d_1)) \cdot \mathrm{tr}(V(w_1)) \cdots \mathrm{tr}(does(d_{e-1})) \cdot \mathrm{tr}(V(w_e))$ where $\mathrm{tr}(V(w_j)) = \{\mathrm{tr}(\varphi) \in \mathcal{L}_{\mathcal{S}'} : \varphi \in V(w_j)\}$ for any $0 \leq j \leq e$, and $\mathrm{tr}(does(d_j)) = \bigwedge_{r \in N} \mathrm{tr}(does(r, d_j(r)))$ for any $1 \leq j \leq e$. Let $\mathrm{Tr}(trace(M)) = \{\mathrm{Tr}(V(\delta)) \mid V(\delta) \in trace(M)\}$. Then the fact holds that $M_{\mathcal{S}} \sim M'_{\mathcal{S}'}$ iff $\mathrm{Tr}(trace(M)) = trace(M')$. With this, let us now prove the direction from Clause 2 to Clause 1.

Assume $M_{\mathcal{S}} \not\sim M'_{\mathcal{S}'}$, then by the fact $\mathrm{Tr}(trace(M)) \neq trace(M')$, so there is a path $\delta \in \mathcal{P}(M)$ such that for any path $\delta' \in \mathcal{P}(M')$, $\mathrm{Tr}(V(\delta)) \neq V(\delta')$. It follows that for each $\delta' \in \mathcal{P}(M)$ there is $k \geq 0$ such that either $\mathrm{tr}(does(d_{k+1})) \neq does(d'_{k+1})$ or $\mathrm{tr}(V(\delta[k])) \neq V(\delta'[k])$. From the former, we obtain a formula $does(r, a_{k+1})$ (for $r \in N$ and $a_{k+1} \in \mathcal{A}$) such that $M, \delta, k \models does(r, a_{k+1})$ and $M', \delta', k \not\models \mathrm{tr}(does(r, a_{k+1}))$. From the latter, we obtain a formula $\chi \in Atm$ such that either (i) $M, \delta, k \models \chi$ and $M', \delta', k \not\models \mathrm{tr}(\chi)$, or (ii) $M, \delta, k \models \neg\chi$ and $M', \delta', k \not\models \neg\mathrm{tr}(\chi)$. Let $\varphi_{\delta'}$ be the formula of the form $\bigcirc^k does(r, a_{k+1})$, $\bigcirc^k \chi$ or $\bigcirc^k \neg\chi$ to distinguish δ from δ'. It follows by the construction that $M, \delta, 0 \models \varphi_{\delta'}$ and $M', \delta', 0 \not\models \mathrm{tr}(\varphi_{\delta'})$. Let Δ be the conjunctions of all such obtained formulas for all paths in M', i.e., $\Delta := \bigwedge_{\delta' \in \mathcal{P}(M')} \varphi_{\delta'}$. Note that Δ is well-formed due to the fact that M' is finite-branching. Let us now consider formula $initial \wedge \Delta$. Then it is satisfied in M, i.e., $M, \delta, 0 \models initial \wedge \Delta$. But $\mathrm{tr}(initial \wedge \Delta)$ is unsatisfied in M'. Otherwise, there are some $\delta' \in \mathcal{P}(M')$ and $j \geq 0$ such that $M', \delta', j \models \mathrm{tr}(initial \wedge \Delta)$, then $M', \delta', 0 \models \mathrm{tr}(\Delta)$, so $M', \delta', 0 \models \mathrm{tr}(\varphi_{\delta'})$, contradicting with $M', \delta', 0 \not\models \mathrm{tr}(\varphi_{\delta'})$. Thus, there is $\varphi \in \mathcal{L}_{\mathcal{S}}$, φ is satisfied in $M_{\mathcal{S}}$ but $\mathrm{tr}(\varphi)$ is unsatisfied in $M'_{\mathcal{S}'}$. □

We end this section with the interesting observation that the GDL-descriptions of Tic-Tac-Toe and Number Scrabble are logically equivalent in terms of the translation.

Observation 1. *Let Σ_{TT} and Σ_{NS} denote the GDL-descriptions of Tic-Tac-Toe (Figure 6) and Number Scrabble (Figure 1), respectively. Then $\models \bigwedge \mathrm{tr}(\Sigma_{TT}) \leftrightarrow \bigwedge \Sigma_{NS}$, where $\mathrm{tr}(\Sigma_{TT}) = \{\mathrm{tr}(\varphi) \in \mathcal{L}_{NS} \mid \varphi \in \Sigma_{TT}\}$.*

Proof. We first give the detailed mapping between their game signatures. The bijection $f_1 : N_{TT} \mapsto N_{NS}$ is that $f_1(\mathsf{x}) = \mathsf{b}$ and $f_1(\mathsf{o}) = \mathsf{w}$. The bijections $f_2 : \mathcal{A}_{TT} \mapsto \mathcal{A}_{NS}$ and $f_3 : \Phi_{TT} \mapsto \Phi_{NS}$ is defined according to the mapping in Table 1. In particular, $f_2(noop) = noop$, and $f_3(turn(r)) = turn(f_1(r))$ for $r \in N_{TT}$.

With this, we translate each formula in Σ_{TT} according to Definition 10. Let us take the first formula in Figure 6 as an example.

$\mathrm{tr}(initial \leftrightarrow turn(\mathsf{x}) \wedge \neg turn(\mathsf{o}) \wedge \bigwedge_{i,j=1}^{3} \neg(p^{\mathsf{x}}_{i,j} \vee p^{\mathsf{o}}_{i,j}))$

= $\mathrm{tr}(initial) \leftrightarrow \mathrm{tr}(turn(\mathsf{x})) \wedge \mathrm{tr}(\neg turn(\mathsf{o})) \wedge \mathrm{tr}(\bigwedge_{i,j=1}^{3} \neg(p^{\mathsf{x}}_{i,j} \vee p^{\mathsf{o}}_{i,j})))$

= $initial \leftrightarrow turn(\mathsf{b}) \wedge \neg turn(\mathsf{w}) \wedge \bigwedge_{i=1}^{9} \neg(s(\mathsf{b}, i) \vee s(\mathsf{w}, i))$

This exactly corresponds to the first formula in Figure 1. With all the translated formulas, the result holds immediately. □

5 Bisimulation and Expressivity

To show that a property of ST-models is definable in GDL, it suffices to find a defining formula. However, showing that a property is not definable in GDL is not so straightforward. It is well known that the expressive power of basic modal logic with respect to Kripke semantics can be completely characterized in terms of k-bisimulation (Blackburn et al. 2006). In this section, we provides an analogous characterization for the expressive power of GDL.

5.1 k-Bisimulation

Here we consider the definability of properties that are satisfied at the initial state of an ST-model. For $\varphi \in \mathcal{L}$, let $\|\varphi\|$ be the set of all ST-models that satisfy φ at the initial state. i.e., $\|\varphi\| := \{M \mid M, w_0 \models \varphi\}$. The concept of the *definability* is specified as follows:

Definition 11. *A class $\mathcal{M}$ of ST-models is definable in $\mathcal{L}$, if there is a formula $\varphi \in \mathcal{L}$ such that $\mathcal{M} = \|\varphi\|$.*

The concept of k-*bisimulation* is defined as follows:

Definition 12. *Let $M = (W, w_0, T, L, U, g, \pi)$ and $M' = (W', w'_0, T', L', U', g', \pi')$ be two ST-models. We say M and M' are k-bisimilar, written $M \approx_k M'$, if there exists a sequence of binary relations $Z_k \subseteq Z_{k-1} \cdots \subseteq Z_0$ such that the following hold: for $i + 1 \leq k$, any $w \in W$ and $w' \in W'$,*

1. $w_0 Z_k w'_0$
2. *If wZ_0w', then*
   (a) $\pi(w) = \pi'(w')$;
   (b) $w = w_0$ *iff* $w' = w'_0$;
   (c) $w \in T$ *iff* $w' \in T'$;
   (d) $L_r(w) = L'_r(w')$ *for any* $r \in N$;
   (e) $w \in g(r)$ *iff* $w' \in g'(r)$ *for any* $r \in N$.
3. *If $wZ_{i+1}w'$ and $U(w, d) = u$, then there is $u' \in W'$ s.t. $U'(w', d) = u'$ and uZ_iu';*
4. *If $wZ_{i+1}w'$ and $U'(w', d) = u'$, then there is $u \in W$ s.t. $U(w, d) = u$ and uZ_iu'.*

The intuition is that if $M \approx_k M'$, then their initial states w_0 and w'_0 bisimulate up to depth k. Clearly, if $M \approx M'$, then $M \approx_k M'$ for all $k \in \mathbb{N}$. We say a class $\mathcal{M}$ of ST-models is *closed under k-bisimulation* if $M \in \mathcal{M}$ and $M \approx_k M'$ implies $M' \in \mathcal{M}$.

Similarly, when Z is a k-bisimulation linking two states w in M and w' in M', we say that w and w' are k-*bisimilar*, written $M, w \,\underline{\leftrightarrow}_k\, M', w'$. In particular, if $M \approx_k M'$, then their initial states are k-bisimilar, i.e., $M, w_0 \,\underline{\leftrightarrow}_k\, M', w'_0$. Given two paths $\delta := w_0 \overset{d_1}{\to} w_1 \overset{d_2}{\to} \cdots$ in M and $\delta' := w'_0 \overset{d'_1}{\to} w'_1 \overset{d'_2}{\to} \cdots$ in M', we say δ and δ' are k-*bisimilar*, written $M, \delta \,\underline{\leftrightarrow}_k\, M', \delta'$, iff for every $j \leq k$ and $r \in N$, $M, \delta[j] \,\underline{\leftrightarrow}_{k-j}\, M', \delta'[j]$ and $\theta_r(\delta, j) = \theta_r(\delta', j)$. Then it is routine to prove the following result.

Lemma 4. *Given two ST-models M and M', if $M \approx_k M'$, then for every $\delta \in \mathcal{P}(M)$, there is $\delta' \in \mathcal{P}(M')$ such that $M, \delta \,\underline{\leftrightarrow}_k\, M', \delta'$, and vice versa.*

5.2 Characterization of the Definability of GDL

Before providing the characterization result, we need some basic notions and results. The degree of a formula $\varphi \in \mathcal{L}$, written $deg(\varphi)$, is inductively defined as follows: $deg(\widehat{p}) = 0$ for $\widehat{p} \in \Phi \cup \{initial, terminal, wins(r), legal(r, a)\}$, $deg(does(r, a)) = 1$, $deg(\neg\varphi) = deg(\varphi)$, $deg(\varphi \wedge \psi) = Max\{deg(\varphi), deg(\psi)\}$, and $deg(\bigcirc\varphi) = deg(\varphi) + 1$.

We have the following invariance result in terms of k-bisimilar paths.

Lemma 5. *Let M, M' be two ST-models. For any $\delta \in \mathcal{P}(M)$, $\delta' \in \mathcal{P}(M')$ and $k \in \mathbb{N}$, if $M, \delta \leftrightarroweq_k M', \delta'$, then $(M, \delta, 0 \models \varphi$ iff $M', \delta', 0 \models \varphi)$ for any $\varphi \in \mathcal{L}$ with $deg(\varphi) \leq k$.*

Proof. This is proved by induction on k. For $k = 0$, it is straightforward by definition; For $k = l+1$ where $l \leq k - 1$, it suffices to consider formulas $does(r, a)$ and $\bigcirc\varphi$.

When $\varphi := does(r, a)$, assume $M, \delta, 0 \models does(r, a)$ iff $\theta_r(\delta, 0) = a$ (by the truth condition) iff $\theta_r(\delta', 0) = a$ (by definition) iff $M, \delta, 0 \models does(r, a)$ (by the truth condition).

When $\varphi := \bigcirc\psi$ with $deg(\psi) < l$, assume $M, \delta, 0 \models \bigcirc\psi$, then $M, \delta, 1 \models \psi$. By assumption $M, \delta \leftrightarroweq_{l+1} M', \delta'$, we have $M, \delta[1, \infty] \leftrightarroweq_l M', \delta'[1, \infty]$. And by Induction Hypothesis, we have $M', \delta', 1 \models \psi$, so $M', \delta', 0 \models \bigcirc\psi$. The other direction is proved in a similar way. □

Definition 13. *Let M, M' be two ST-models and $k \in \mathbb{N}$. We say M and M' are k-equivalent, written $M \equiv_k M'$, if at the initial states, they satisfies the same GDL-formulas of degree at most k, i.e., $\{\varphi \in \mathcal{L} \mid deg(\varphi) \leq k$ and $M, w_0 \models \varphi\} = \{\psi \in \mathcal{L} \mid deg(\psi) \leq k$ and $M', w'_0 \models \psi\}$.*

We also use the fact that *for every ST-model M and every $k \in \mathbb{N}$ there is a formula that completely characterizes M up to k-equivalence.* To show this, we take the following steps.

1. Redefine the set of *atomic propositions*, written Atm, as follows: Atm $=\Phi \cup \{initial, terminal\} \cup \{wins(r), legal(r, a) \mid r \in N, a \in \mathcal{A}\}$.

2. Encode the atomic propositions through a valuation V rather than through separate relations or functions. For every $w \in W$, let $V(w) = \{p \in \Phi \mid p \in \pi(w)\} \cup \{initial \mid w = w_0\} \cup \{terminal \mid w \in T\} \cup \{wins(r) \mid w \in g(r)\} \cup \{legal(r, a) \mid a \in L_r(w)\}$. Note $V(w)$ is finite since N, $\mathcal{A}$ and Φ are all finite.

3. For each path $\delta := w_0 \overset{d_1}{\rightarrow} w_1 \overset{d_2}{\rightarrow} \cdots \overset{d_j}{\rightarrow} \cdots$ in M, induce a *trace* $V(\delta) = V(w_0) \cdot does(d_1) \cdot V(w_1) \cdots does(d_j) \cdot V(w_j) \cdots$. Let φ^k_δ be the syntactical representation of δ up to depth k, i.e., $\varphi^k_\delta := (\bigwedge V(\delta[0]) \wedge does(d_1)) \wedge \bigcirc(\bigwedge V(\delta[1]) \wedge does(d_2)) \wedge \cdots \wedge \bigcirc^k(\bigwedge V(\delta[k]) \wedge does(d_{k+1}))$.

4. Define the *k-th characteristic formula* Γ^k_M of M as the disjunctions of all the syntactical representations of paths in M up to depth k, i.e.,

$$\Gamma^k_M := \bigvee_{\delta \in \mathcal{P}(M)} \varphi^k_\delta.$$

Note that Γ^k_M is well-formed as M is finite-branching and all paths are bounded to depth k. It is easy to check that $deg(\Gamma^k_M) = k$ and $M, w_0 \models \Gamma^k_M$.

To illustrate this idea, let us consider the ST-model M depicted in Figure 4, where $N = \{r\}$, $\Phi = \emptyset$, $T = \{w_{22}, w_{23}\}$ and $g(r) = \{w_{23}\}$. Then the 2-th characteristic formula of M is $\Gamma^2_M = \varphi^2_{\delta_1} \vee \varphi^2_{\delta_2} \vee \varphi^2_{\delta_3}$, where $\varphi^2_{\delta_1} = initial \wedge \bigwedge^2_{i=1} legal(r, a_i) \wedge does(r, a_1) \wedge \bigcirc(\bigwedge^2_{i=1} legal(r, b_i) \wedge does(r, b_1)) \wedge$

$\bigcirc^2(legal(r,c) \wedge does(r,c))$, $\varphi^2_{\delta_2} = initial \wedge \bigwedge_{i=1}^2 legal(r,a_i) \wedge does(r,a_1) \wedge \bigcirc(\bigwedge_{i=1}^2 legal(r,b_i) \wedge does(r,b_2)) \wedge \bigcirc^2(terminal \wedge legal(r,noop) \wedge does(r,noop))$, and $\varphi^2_{\delta_3} = initial \wedge \bigwedge_{i=1}^2 legal(r,a_i) \wedge does(r,a_2) \wedge \bigcirc(legal(r,b_3) \wedge does(r,b_3)) \wedge \bigcirc^2(wins(r) \wedge terminal \wedge legal(r,noop) \wedge does(r,noop))$.

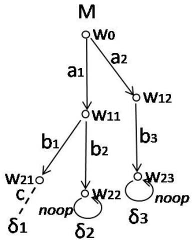


Figure 4: Characteristic Formula Γ^2_M.

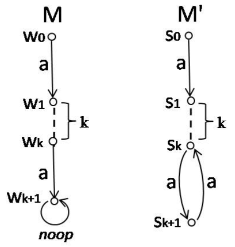


Figure 5: A non-definable property.

The following result shows that the characteristic formula Γ^k_{M} captures the essence of k-bisimulation.

Proposition 4. *Let M, M' be two ST-models, and $k \in \mathbb{N}$. The following are equivalent.*

1. $M \approx_k M'$
2. $M \equiv_k M'$
3. $M', w'_0 \models \Gamma^k_{M}$

Proof. We first show Clause 1⇒ Clause 2. Assume $M \approx_k M'$ and $M, w_0 \not\models \varphi$ for $deg(\varphi) \leq k$, then there is $\delta \in \mathcal{P}(M)$ such that $M, \delta, 0 \not\models \varphi$. By Lemma 4, we have there is $\delta' \in \mathcal{P}(M')$ such that $M, \delta \leftrightarroweq_k M', \delta'$, then by Lemma 5, we get $M', \delta', 0 \not\models \varphi$, so $M', w'_0 \not\models \varphi$. The other direction is proved in a similar way.

Clause 3 follows from Clause 2 and the fact that $M, w_0 \models \Gamma^k_{M}$ and $deg(\Gamma^k_{M}) \leq k$.

Finally, we show Clause 3⇒ Clause 1. Assume $M', w'_0 \models \Gamma^k_{M}$. Let $trace(M \upharpoonright_k) = \{V(\delta[0,k]) \mid \delta \in \mathcal{P}(M)\}$ be the set of all the prefix traces up to depth k in M. Then by the assumption, we have $trace(M \upharpoonright_k) = trace(M' \upharpoonright_k)$. We define a sequence of relations as follows: for any $0 \leq l \leq k$, let $Z_l = \{(w, w') \in W \times W' \mid$ there are $\delta \in \mathcal{P}(M), \delta' \in \mathcal{P}(M')$ and $0 \leq j \leq k-l$ such that $\delta[j] = w, \delta'[j] = w'$ and $V(\delta[0,k]) = V(\delta'[0,k])\}$. It suffices to show the sequence satisfies the conditions of being a k-bisimulation in Definition 12. It holds by definition that $Z_k \subseteq Z_{k-1} \cdots \subseteq Z_0$, and $w_0 Z_k w'_0$ (Condition 1). Let us verify Condition 2. Assume $w Z_0 w'$, then by definition there are $\delta \in \mathcal{P}(M), \delta' \in \mathcal{P}(M')$ and $0 \leq j \leq k-l$ such that $\delta[j] = w, \delta'[j] = w'$ and $V(\delta[0,k]) = V(\delta'[0,k])$, so $V(w) = V(w')$. Thus, conditions (a)-(e) hold by the construction of V. Let us next check Condition 3. Assume $w Z_{i+1} w'$ (for $i+1 \leq k$) and $U(w,d) = u$, then by definition there are $\delta \in \mathcal{P}(M), \delta' \in \mathcal{P}(M')$ and $0 \leq j \leq k-i-1$ such that $\delta[j] = w, \delta'[j] = w'$ and $V(\delta[0,k]) = V(\delta'[0,k])\}$. We extend the segment $\delta[0,j] \overset{d}{\to} u$ to a complete path λ in M. Such a complete path always exists, since each agent has at least one legal action at each state. Then $V(\lambda[0,k]) \in trace(M \upharpoonright_k)$. And by the assumption, we have $V(\lambda[0,k]) \in trace(M' \upharpoonright_k)$, so by the assumption there is $\lambda' \in \mathcal{P}(M')$ such that $V(\lambda[0,k]) = V(\lambda'[0,k])$. Since the update is deterministic and the initial state is unique, so we have $\delta'[0,j]$ is the initial segment of λ'. It follows that $\lambda'[j] = \delta'[j] = w'$ and $w' \overset{d}{\to} \lambda'[j+1]$. Let $\lambda'[j+1]$ be u'. Then by definition we have $(u, u') \in Z_i$. Condition 4 is verified in a similar way of Condition 3. Thus, $M \approx_k M'$. This completes the proof. □

This result asserts that (i) k-bisimulation coincides with k-equivalence on ST-models, and (ii) two ST-models are k-bisimilar if and only if for any path developed in one model, its k-th prefix can also be developed in the other.

We are now in the position to provide a characterization for the definability of GDL with respect to k-bisimulation.

Theorem 2. *A class $\mathcal{M}$ of ST-models is definable in GDL iff there is $k \in \mathbb{N}$ such that $\mathcal{M}$ is closed under k-bisimulation.*

Proof. Assume a class $\mathcal{M}$ of ST-models is definable in GDL, then there is $\varphi \in \mathcal{L}$ such that $\mathcal{M} = \|\varphi\|$. Let $k = deg(\varphi)$. Further assume $M \in \mathcal{M}$ and $M \approx_k M'$, then $M \in \|\varphi\|$, so $M, w_0 \models \varphi$. And by Proposition 4 and the assumption, we have $M', w'_0 \models \varphi$, then $M' \in \|\varphi\|$, so $M' \in \mathcal{M}$. Conversely, assume there is $k \in \mathbb{N}$ such that $\mathcal{M}$ is closed under k-bisimulation. Let $\mathcal{N} = \bigcup_{M \in \mathcal{M}} \{M' \mid M \approx_k M'\}$ be the set of all ST-models that are k-bisimilar to a member of $\mathcal{M}$. Note $\mathcal{N}$ is finite. Let $\Lambda = \bigvee_{M' \in \mathcal{N}} \Gamma^k_{M'}$ be the disjunctions of the k-th characteristic formulas of all ST-models in $\mathcal{N}$. We next show that $\mathcal{M} = \|\Lambda\|$. First assume $M \in \mathcal{M}$, then $M \in \mathcal{N}$. And $M, w_0 \models \Gamma^k_M$, so $M, w_0 \models \Lambda$. Thus, $M \in \|\Lambda\|$. Conversely, assume $M \in \|\Lambda\|$, then $M, w_0 \models \Lambda$, so $M, w_0 \models \Gamma^k_{M'}$ for some M' in $\mathcal{N}$. It follows from Proposition 4 that $M \approx_k M'$. And by the construction $M' \approx_k M''$ for some $M'' \in \mathcal{M}$, so $M \approx_k M''$. Thus, $M \in \mathcal{M}$. □

This theorem indicates that *exactly the properties of ST-models that are closed under k-bisimulation for some $k \in \mathbb{N}$ are definable in GDL.* This provides a feasible approach to test the non-definability of GDL. We can show, for instance, that GDL can express that a player r will win in i steps, i.e., $\bigcirc^i wins(r)$, but it cannot express that *a player has a winning strategy* in general. Indeed for an arbitrary $k \in \mathbb{N}$, we can always construct two ST-models depicted in Figure 5, where $N = \{r\}$, $\Phi = \emptyset$, $w_{k+1} \in g(r)$ and $s_i \notin g'(r)$ for all $i \in \{0, \cdots, k+1\}$. It is easy to check that $M \simeq_k M'$, but player r has a winning strategy in M while she does not have in M'. By a slight change of M and M' with $w_{k+1} \in T$ and $s_{k+1} \notin T'$, we obtain another GDL-undefinable property that *a game will always reach a terminal state.* It is worth noting that this makes GDL different from other strategic logics such as ATL (Alur et al. 2002) and Strategy Logic (Chatterjee et al. 2010), where those properties are expressible.

6 Conclusion

We have investigated the expressive power of GDL in terms of bisimulation relations over state transition models and obtained two characterization results. The first result shows the coincidence between invariance of GDL-formulas and bisimulation equivalence, and the second characterizes the definability of GDL-formulas in terms of k-bisimulation. We have also introduced the quotient model to improve the efficiency of model checking for GDL. Moreover, we have generalized the notion of bisimulation equivalence to capture more general game equivalence.

The existing literature focuses on the relationships of GDL with other strategic logics. For instance, Ruan *et al.* studied the relationship between GDL and ATL by transferring a GDL game specification into an ATL specification (Ruan et al. 2009). Lorini and Schwarzentruber investigated the relation between GDL and Seeing-to-it-that Logics (STITs) by providing a polynomial embedding of GDL into STIT (Lorini and Schwarzentruber 2017). Differently, we use a bisimulation approach to study the expressive power of Game Description Languages. Our characterization results show that compared to other strategic logics, such as ATL, Strategy Logic, STITs, GDL is light-weight yet sufficient to define game rules and describe bounded game properties. Zhang *et al.* investigated game equivalence for knowledge transfer in GGP. They consider two games are equivalent if their state transition models are isomorphic (Zhang et al. 2017). Our notion of game equivalence is more general as it is based on the bisimulation relation.

Directions of future research are manifold. We intend to explore the van Benthem Characterization Theorem for GDL (van Benthem 1984). More recently, GDL has been extended to GDL-II and epistemic GDL for representing and reasoning about imperfect information games (Thielscher 2010; Jiang et al. 2016). We plan to study the expressiveness of these extended languages. Besides structure equivalence, it would be also interesting to investigate different types of game equivalence in GGP (Zhang et al. 2017; Goranko 2003; Pauly 2001; van Benthem et al. 2017).

References

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

Blackburn, P., De Rijke, M., and Venema, Y. (2002). *Modal Logic*. Cambridge University Press.

Baier, C., Katoen, J.-P. and Larsen, K. G. (2008). *Principles of model checking*. MIT press.

Blackburn, P., van Benthem, J., and Wolter, F. (2006). *Handbook of modal logic*, volume 3. Elsevier.

Chatterjee, K., Henzinger, T. A., and Piterman, N. (2010). Strategy logic. *Information and Computation*, 208(6):677–693.

Genesereth, M., Love, N., and Pell, B. (2005) General game playing: Overview of the AAAI competition. *AI magazine*, 26(2):62–72.

Grädel, E. and Otto, M. (2014). The freedoms of (guarded) bisimulation. In *Johan van Benthem on Logic and Information Dynamics*, pages 3–31. Springer.

Goranko, V. (2003) The basic algebra of game equivalences. *Studia Logica*, 75(2):221–238.

Hennessy, M. and Milner, R. (1985). Algebraic laws for nondeterminism and concurrency. *Journal of the ACM (JACM)*, 32(1):137–161.

Jiang, G., Zhang, D., Perrussel, L., and Zhang, H. (2016). Epistemic GDL: A logic for representing and reasoning about imperfect information games. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*, pages 1138–1144.

Love, N., Hinrichs, T., Haley, D., Schkufza, E., and Genesereth, M. (2006) General game playing: Game description language specification. Stanford Logic Group Computer Science Department Stanford University. `http://logic.stanford.edu/reports/LG-2006-01.pdf`.

Lorini, E. and Schwarzentruber, F. (2017). A path in the jungle of logics for multi-agent system: On the relation between general game-playing logics and seeing-to-it-that logics. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS'17)*, pages 687–695.

Michon, J. A. (1967). The game of jam: An isomorph of tic-tac-toe. *The American Journal of Psychology*, 80(1):137–140.

Park, D. (1981) Concurrency and automata on infinite sequences. In *Theoretical computer science*, pages 167–183. Springer.

Pauly, M. (2001). *Logic for Social Software*. PhD thesis, University of Amsterdam.

Pell, B. (1993) *Strategy generation and evaluation for meta-game playing*. PhD thesis, University of Cambridage.

Reiter, R. (1991). The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, 27:359–380.

Ruan, J., van Der Hoek, W., and Wooldridge, M. (2009) Verification of games in the game description language. *Journal of Logic and Computation*, 19(6):1127–1156.

Thielscher, M. (2010). A general game description language for incomplete information games. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, pages 994–999.

van Benthem, J. (1977). *Modal Correspondence Theory*. PhD thesis, University of Amsterdam.

van Benthem, J. (1984). Correspondence theory. In *Handbook of philosophical logic*, pages 167–247. Springer.

van Benthem, J., Bezhanishvili, N., and Enqvist, S. (2017). A new game equivalence and its modal logic. In *Proceedings Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK'17)*, pages 57–74.

Zhang, H., Liu, D., and Li, W. (2015). Space-consistent game equivalence detection in general game playing. In *Workshop on Computer Games*, pages 165–177. Springer.

Zhang, D. and Thielscher, M. (2015a). A logic for reasoning about game strategies. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, pages 1671–1677.

Zhang, D. and Thielscher, M. (2015b). Representing and reasoning about game strategies. *Journal of Philosophical Logic*, 44(2):203–236.

Modelling Legal Systems with Conflicts

Fengkui Ju * Karl Nygren † Tianwen Xu ‡

* Beijing Normal University † Stockholm University
‡ China University of Political Science and Law

A legal conflict arises for a case if more than one norms are applicable to this case but they can not all be applied. There are two types of legal conflicts: the conflicts between a prohibition to do something and a permission to do it and the conflicts between a prohibition to do something and a prohibition not to do it. Legal conflicts are very common in legal practice and many of them are practically reasonable.

Here are some examples of legal conflicts.

Example 1. *The Fourteenth Amendment of the U.S. Constitution states: "No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law . . . ". In a series of cases, the U.S. Supreme Court established that the right to privacy is one of those liberties protected by this amendment.*

Texas Penal Code generally forbids abortion. Article 1191 of it states: "If any person shall designedly administer to a pregnant woman or knowingly procure to be administered with her consent any drug or medicine . . . and thereby procure an abortion, shall be confined in the penitentiary not less than two nor more than five years".

Jane Roe, a single woman who lived in Dallas County, Texas and was willing to terminate her pregnancy but failed because of the prohibition, filed a suit against the county's District Attorney, Henry Wade, with a claim that Texas' prohibition on abortion is unconstitutional. This case, known as Roe v. Wade, was finally heard by the U.S. Supreme Court, and was decided by the Court that Texas' prohibition on abortion indeed violated the Fourteenth Amendment, "which protects against state action the right to privacy, including a woman's qualified right to terminate her pregnancy".

This example is from Case 410 U.S. 113 in U.S. Supreme Court Cases and (Davis 2008).

Example 2. *The 2nd section of the Intellectual Property Law of Norway states: "Intellectual property gives exclusive rights to produce copies, temporary or permanent". Later the 11st section specifies an exception to the previous statute: "If a temporary representation of a work is essential to a process whose sole purpose is to facilitate the legitimate use of the work then the 2nd section is suspended". Then an exception to the previous exception follows: "This provision does not apply to computer programs and databases".*

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

This example is from (Stolpe 2010).

Example 3. *Tort Law of China came into force in 2010, which was meant to revise the part about tort of General Principles of Civil Law of China that came into force in 1987. However, the part about tort of the latter can not be clearly isolated, so it is still in force completely.*

The 133nd article of General Principles of Civil Law of China states: "If a person who has property but is without or with limited capacity for civil conduct causes damages to others, the expenses of compensation shall be paid from his property. Shortfalls in such expenses shall be appropriately compensated for by the guardian unless the guardian is a unit[1] ". By this article, the guardian does not have to pay all the shortfalls in some cases, and if the guardian is a unit, it does not have to pay the shortfalls at all.

The 32nd article of Tort Law of China says: "If a person who has property but is without or with limited capacity for civil conduct causes damages to others, the expenses of compensation shall be paid from his property. Shortfalls in such expenses shall be compensated for by the guardian". This article just simply requires the guardian to pay all the shortfalls.

As far as we observe, the general way that legal conflicts are resolved in legal practice is as follows. There are a few principles. Some of them are prior to some others. Each principle induces a priority relation among norms. A chain of principles is a sequence of principles such that its first element is prior to its second element, its second element is prior to its third element, and so on. A maximal chain of principles is a chain of principles which can not be extended. Assume there is a case where more than one norms are applicable but they can not all be applied. The conflict is resolved as follows. Firstly parallelly apply all the maximal chains of principles. Then observe the results. If no maximal chain can resolve the conflict, then the conflict is unsolvable. Suppose that some maximal chain can resolve the conflict. Then if all the results of applying the maximal chains of principles that resolve the conflict are coincident, then the conflict is resolved, otherwise the conflict is unsolvable.

In this work we try to formalize the way legal conflicts are resolved. In this formalization we assume that there is only one agent, we just focus on the norms concerning actions and we do not consider obligations.

By this formalization we can present an answer to the question asked by (Lewis 1979): *what exactly is the effect of permitting something that was forbidden previously*? By this formalization we can also make a distinction between statical equivalence and dynamical equivalence of legal systems, which is proposed in (Hansson 2013).

References

Davis, S. (2008). *Corwin and Peltason's Understanding the Constitution*. Belmont: Thomson Wadsworth.

Hansson, S. (2013). The varieties of permission. In Gabbay, D., Horty, J., Parent, X., van der Meyden, R., and van der Torre, L., editors, *Handbook of Deontic Logic and Normative Systems*, pages 195–240. College Publications.

Lewis, D. (1979). A problem about permission. In Saarinen, S., Hilpinen, R., Niiniluoto, I., and Hintikka, M., editors, *Essays in Honour of Jaakko Hintikka*, pages 163–175. Springer Netherlands.

[1] Here "shortfalls shall be appropriately compensated" means "shortfalls shall be payed in a reasonable amount". "Unit" means "organization".

Stolpe, A. (2010). A theory of permission based on the notion of derogation. *Journal of Applied Logic*, 8(1):97–113.

Social Interaction and Pointwise Intersection in Neighborhood Modal Logic

Dominik Klein

University of Bayreuth / University of Bamberg

Abstract

Modal logic has proven a powerful tool for studying social interaction. It has been applied to a wide range of contexts, ranging from strategic powers of coalitions to belief formation in networks. In incorporating influence from several sources, the intersection of several modalities becomes crucial. We study pointwise intersection as a way of forming distributed belief, calculating coalitional abilities or tracking the combination of evidence. Towards a general understanding of pooling modalities, we provide a class of soundness and completeness characterization results for intersection modalities in neighborhood modal logic. We present a general technique for completeness proofs that applies to a variety of neighborhood logics with intersection operations.

1 Introduction

The[1] study of social interaction often leads to neighborhood semantics, a well-established tool for studying generalizations and variants of Kripke-semantics for modal logic. Neighborhood logics arise naturally when studying strategic powers in games (van Benthem et al. 2018), more generally for analyzing the logic of ability (Brown 1988; Pauly 2002), in tracking agents' evidence, social

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

[1] This abstract is based on joint work with Frederik Van De Putte, cf. (Van De Putte and Klein 2018). While all results presented here are contained in this paper, the proof of Theorem 1 is slightly different. The present proof keeps the canonical model small (only adding countably many copies of each maximally consistent subset), but cannot be straightforwardly generalized to cases where some infinite groups G of agents are allowed. The proof in (Van De Putte and Klein 2018) allows for such a generalization, at the expense of adding uncountably many copies of each maximally consistent subset. Moreover, the proofs follow slightly different intuitions.

or not, (van Benthem and Pacuit 2011), but also in deontic logic and the analysis of obligations, especially when these come from various sources (Goble 2005; Klein and Marra 2019).

Formally, a neighborhood function $\mathcal{N}$ attaches a set of accessible sets $X_1, X_2, \ldots$ of worlds to any world w in a possible worlds model. $\Box\varphi$ is then true iff there is some such X in the neighborhood set $\mathcal{N}(w)$, that coincides with the truth set of φ. The move from Kripke semantics to neighborhood semantics allows us to invalidate certain schemata that are problematic for certain interpretations of the modal operator $\Box$, but also to include other schemata that would trivialize any normal modal logic. Apart from that, neighborhood models can also be used as a purely technical tool in order to prove completeness or incompleteness w.r.t. other possible worlds semantics.[2]

In studying social interaction, we are often interested in the interaction of several modalities. Agents learning on a network may combine different pieces of evidence they acquire from different sources, thus accessing the *distributed belief* of their surroundings. Similarly, when accepting new trends or social norms, agents will combine a variety of observations from different situations. Also within game theory and, more generally, strategic interaction, players may team up and hence be interested in their joint strategic abilities, rather than their individual powers alone. All these applications, hence, are interested in logical combinations of various neighborhood sets. The general logic of neighborhood models where certain neighborhood functions are obtained by operations on (one or several) neighborhood functions is still largely unknown.

2 The Base Logic

The current paper is a first step towards axiomatizing general intersection modalities. For simplicity of notation, we focus on a case where we combine information from different sources, each denoted by their own modal operator $\Box_1, \Box_2, \ldots$. There, we study logics that are interpreted in terms of the *pointwise intersection* of the different neighborhoods. The result will be new neighborhoods $\Box_G$ for each group G of agents, denoting the shared information (or powers or norms...) of G. Formally, This concept is defined as follows, for a fixed (finite or infinite) index set $I = \{1, 2, \ldots\}$.

Definition 1. *$i)$ A* **model** *$\mathfrak{M}$ is a triple $\langle W, \langle \mathcal{N}_i \rangle_{i \in I}, V \rangle$, where $W \neq \emptyset$ is the* domain *of $\mathfrak{M}$, for every $i \in I$, $\mathcal{N}_i : W \to \wp(\wp(W))$ is a* neighborhood function for *i, and $V : W \to \wp(\mathfrak{P})$ is a* valuation function.

$ii)$ Where $\mathfrak{M} = \langle W, \langle \mathcal{N}_i \rangle_{i \in I}, V \rangle$ is a model and $G = \{i_1, \ldots, i_n\} \subseteq I$, the **neighborhood function for** *G is given by*

$$\mathcal{N}_G(w) = \{X_{i_1} \cap \ldots \cap X_{i_n} \mid \textit{each } X_{i_j} \in \mathcal{N}_{i_j}(w)\}$$

So, in the context of neighborhood semantics, pointwise intersection takes as input any intersection of neighborhoods, one for each agent $i \in G$, to form the new neighborhood set for G. This new neighborhood set is then used to interpret expressions of the type $\Box_G\varphi$, using the standard semantic clause, but plugging in the neighborhood function $\mathcal{N}_G$.

The main goal of this contribution is to provide a sound and complete axiomatization for the logics of intersecting neighborhoods. More specifically, we present a variety of soundness and

[2] One prototypical example of a completeness proof via neighborhood semantics is (Lewis 1973). In (Governatori and Rotolo 2005), neighborhood semantics are used to prove the incompleteness of Elgesem's modal logic of agency (Elgesem 1997). We refer to (Pacuit 2017) for a critical introduction to the many forms, uses and advantages of neighborhood semantics.

completeness results, depending on the exact properties of the underlying neighborhood modality. For the sake of brevity, let us stick with a minimal setting here. Let $\mathfrak{L}$ be the language obtained by closing a countable set of propositional variables $\mathfrak{P} = \{p, q, \ldots\}$ under the classical connectives and all unary modal operators of the type $\Box_G$, where G is a finite subset of I, written $G \subseteq_f I$. To interpret $\mathfrak{L}$, we use the following (standard) semantic clauses:

Definition 2. *Let $\mathfrak{M} = \langle W, \langle \mathcal{N}_i \rangle_{i \in I}, V \rangle$ be a model, $w \in W$, $\varphi, \psi \in \mathfrak{L}$, and $G \subseteq_f I$. Then*

1. *$\mathfrak{M}, w \models p$ iff $w \in V(p)$ for all $p \in \mathfrak{P}$*
2. *$\mathfrak{M}, w \models \neg\varphi$ iff $\mathfrak{M}, w \not\models \varphi$*
3. *$\mathfrak{M}, w \models \Box_G\varphi$ iff $\|\varphi\|^{\mathfrak{M}} \in \mathcal{N}_G(w)$*
4. *$\mathfrak{M}, w \models \varphi \vee \psi$ iff $\mathfrak{M}, w \models \varphi$ or $\mathfrak{M}, w \models \psi$*

where $\|\varphi\|^{\mathfrak{M}} = \{w \in W \mid \mathfrak{M}, w \models \varphi\}$.

For this exposition, we restrict ourselves to the *base logic*, i.e. the logic on intersection modalities for the class of *all* neighborhod models. To characterize this logic syntactically, we will need the following axioms:

$(\Box_G\varphi \wedge \Box_H\psi) \rightarrow \Box_{G\cup H}(\varphi \wedge \psi)$ if $G \cap H = \emptyset$ (B1)

$(\Box_G\varphi \wedge \Box_{G\cup H\cup J}\varphi) \rightarrow \Box_{G\cup H}\varphi$ (B3)

$\Box_{G\cup H}\top \rightarrow \Box_G\top$ (B2)

$(\Box_G\varphi \wedge \Box_H(\varphi \vee \psi)) \rightarrow \Box_{G\cup H}\varphi$ (B4)

and, as usual, replacement of equivalents (RE) and modus ponens (MP):

if $\varphi \vdash \psi$ and $\psi \vdash \varphi$, then $\Box_G\varphi \vdash \Box_G\psi$

if $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$, then $\vdash \psi$

Our first result and starting point of the talk is:

Theorem 1. *[Strong Completeness for the Base Logic] A sound and strongly complete axiomatization of the base logic is obtained by adding (B1), (B2), (B3), and (B4) to any sound and complete axiomatization of* **CL***, and closing the result under (RE) and (MP).*

Proof. We only provide a restricted proof for the case where we cannot allow for infinite groups G. For a full proof, allowing for countably many infinite groups, slightly different techniques than those presented here are necessary, cf. (Van De Putte and Klein 2018)

Soundness is trivial. For completeness, we provide an extended canonical model construction. Contrary to classic approaches, this canonic model will have infinitely many points $\Gamma_1, \Gamma_2 \ldots$ for every consistent subset Γ of the logic that is obtained by adding (B1), (B2), (B3), and (B4) to any sound and complete axiomatization of **CL**, and closing the result under (RE) and (MP). The main task is then to equip agents with sufficiently large neighborhoods to $i)$ create all intersection sets desired while $ii)$ not creating *too many* intersection sets in order not to validate more formulas of the shape $\Box_G\varphi$ than desired.

We beginn by constructing the canonical model $\langle W, n_G, V \rangle$. We first define W. For this, let ω^x denote the positive integers, i.e. $\omega \setminus \{0\}$.

$$W := \{(\Gamma, i) \mid i \in \omega^x \text{ and } \Gamma \subseteq \mathcal{L} \text{ is max. const. w.r.t the base logic}\}.$$

Next, the valuation function $V : W \rightarrow \mathcal{P}(\mathfrak{P})$ is defined as $p \in V(\Gamma, i) \Leftrightarrow p \in \Gamma$. Finally, we have to construct the neighborhood functions $n_i : W \rightarrow \mathcal{P}(\mathcal{P}(W))$ for $i \in I$. Neighborhood functions

n_G for $G \subseteq_f I$ will then be constructed through the intersection operation defined above. Before we define these functions, note that by our assumption that the set of agents is at most countable, also $\{G \mid G \subseteq_f I\}$ and $\mathcal{L}$ are. Hence, we can fix a bijection

$$\mu : \{G \mid G \subseteq_f I\} \times \mathcal{L} \to \{p \mid p \text{ is prime}\}$$

Without loss of generality, we can assume that $\mu(G, \varphi) < |G|$ for all $(G, \varphi) \in \{G \mid G \subseteq_f I\} \times \mathcal{L}$. Moreover, by the axiom of choice, we can fix a well-order $<_w$ of I. With these tools, we can finally define the neighborhood sets n_i. To do so, we first define auxiliary sets $X_i^{G,\varphi}$ for $(G, \varphi) \in \{H \mid H \subseteq_f I\} \times \mathcal{L}$ and $i \in G$. For this, note that the ordering $<_w$ induces a well-ordering on G. Let $k(i)$ the position at which i occurs in this well-ordering. Moreover, for each (G, φ) pick a partition $M_1^{(G,\varphi)}, \ldots, M_{|G|}^{(G,\varphi)}$ of $\{1, \ldots, \mu(G, \varphi)\}$ into $|G|$ non-empty sets. Then define

$$X_i^{G,\varphi} := \{(\Lambda, i) \mid \varphi \in \Lambda\} \cup \left\{(\Lambda, i) \mid i \mod \mu(G, \varphi) \notin M_{k(i)}^{(G,\varphi)}\right\}.$$

By the assumption that $\mu(G, \varphi) < |G|$, it follows that $X_i^{G,\varphi} \neq X_j^{H,\rho}$ if $i \neq j$ or $G \neq H$ or $\varphi \neq \rho$ unless φ and ρ are both tautologies. We set

$$n_i(w) := \{X_i^{G,\varphi} \mid \Box_G \varphi \in w, i \in G\}.$$

To finish the proof, we have to show that the hence constructed model is a canonical model, i.e. that $W, (\Gamma, i) \vDash \varphi$ iff $\varphi \in \Gamma$. We do so by induction over the complexity of φ. For the induction base, φ atomic, this is trivially true by construction. Moreover, if φ is of the form $\psi \vee \rho$ or $\neg\psi$, the proof follows by a standard argument. We only have to show the case where φ is of the form $\Box_G \rho$ and the statement has already been shown for ρ.

For the left-to right direction (i.e., $\varphi \in \Gamma \Rightarrow W, (\Gamma, i) \vDash \varphi$) fix some (Γ', j) with $\varphi = \Box_G \rho \in \Gamma'$. By construction, this implies that $X_i^{G,\rho} \in n_i(w)$ for all $i \in G$. By definition, it holds that $\bigcap_{i \in G} X_i^{G,\rho} \in n_G(w)$. Moreover,

$$\begin{aligned}\bigcap_{i\in G} X_i^{G,\rho} =& \bigcap_{i\in G}\{(\Lambda, i) \mid \rho \in \Lambda\} \cup \bigcap_{i\in G}\{(\Lambda, i) \mid i \mod \mu(G, \rho) \in \{k(j) \mid j \in G, j \neq i\}\} \\ =&\{(\Lambda, i) \mid \rho \in \Lambda\}\end{aligned}$$

By induction hypothesis, $\{(\Lambda, i) \mid \rho \in \Lambda\} = \{w \mid W, w \vDash \rho\}$. Hence $W, (\Gamma', j) \vDash \Box_G \rho$ as desired.

For the right-to-left direction, we have to show that $W, (\Gamma, i) \nvDash \varphi$, i.e. $\{(\Gamma', j) \mid W, (\Gamma', j) \vDash \rho\} \notin n_G(\Gamma, i)$ whenever $\Box_G \rho \notin \Gamma$. Assume the latter. Of course, we have that $X_i^{G,\rho} \notin n_i(\Gamma, i)$. However, this does not yet imply that $\bigcap_{i\in G} X_i^{G,\rho} \notin n_i(\Gamma, i)$, as $\{(\Gamma', j) \mid W, (\Gamma', j) \vDash \rho\}$ could have entered $n_G(\Gamma, i)$ as some other $\bigcap_{k\in G} Y_k$ with $Y_k \in n_k(\Gamma', j)$. We have to show that this is not the case.

Towards a contradiction, assume that such $\{Y_k \mid k \in G\}$ with all $, Y_k \in n_k(\Gamma', j)$ exist. We distinguish two cases. In the first case, ρ is a tautology. Using the induction assumption and that $\bigcap_{k\in G} Y_k = \{(\Gamma, k) \mid W, (\Gamma, k) \vDash \rho\}$, this implies that all $Y_k = W$ for all W. By construction, this implies that for each $k \in G$ there is some G^k with $k \in G^k$ and $\Box_{G^k} \varphi^k \in \Gamma$ where φ^k is a tautology. By (B2), this implies that $\Box_k \varphi^k \in \Gamma$ for all $k \in G$. (B1) then implies that $\Box_G \bigwedge_{k\in G} \varphi^k \in \Gamma$. Since both ρ and $\bigwedge_{k\in G} \varphi^k$ are tautologies we hence have $\Box_G \rho \in \Gamma'$, contradicting our assumption.

We move to the second case where ρ is not a tautology. We split $\{Y_k \mid k \in G\}$ in three disjoint classes C_1, C_2, C_3. The first, C_1 contains those all those $Y_k \in \{Y_l \mid l \in G\}$ with $Y_k = W$. By construction, for all $X \in \{Y_k \mid k \in G\} \setminus C_1$ there is are unique i, ρ_i, G_i with $i \in G, i \in G_i, \rho_i \in \mathcal{L}$ such that $Y_i = X_i^{G_i,\rho_i}$ and $\Box_{G_i}\rho_i \in \Gamma'$. Define

$$C_2 := \{Y_k \mid \text{ for all } l \in G_k : X_l^{G_k,\rho_k} \in \{Y_k \mid k \in G\}\}$$
$$C_3 := \{Y_k \mid \text{ for some } l \in G_k : X_l^{G_k,\rho_k} \notin \{Y_k \mid k \in G\}\}$$

Obviously, $\{C_1, C_2, C_3\}$ is a partition of $\{Y_k \mid k \in G\}$. Now consider C_2. Let $R := \{\rho_k \mid X_k^{G_k,\rho_k} \in C_2 \text{ for some } k, G_k\}$ and $R_3 := \{\rho_k \mid X_k^{G_k,\rho_k} \in C_3 \text{ for some } k, G_k\}$ and let

$$\overline{\rho} = \bigwedge_{\rho_k \in R} \rho_k$$

Note that $\rho \to \overline{\rho}$ as by assumption $\{(\Gamma, k) \mid W, \rho \in \Gamma\} \subseteq \{(\Gamma, k) \mid \rho_k \in \Gamma\}$ for all $\rho_k \in R$. Now, we have to make one last case distinction about whether or not $\vdash \overline{\rho} \leftrightarrow \rho$ holds.

In the first case, $\vdash \overline{\rho} \leftrightarrow \rho$. By construction, we have that $\Box_{G_k}\rho_k \in \Gamma'$ for all $\rho_k \in R$. Moreover, it holds by construction that $G_k \cap G_l = \emptyset$ whenever $\rho_k \neq \rho_l$. By (B1), this implies that $\Box_{\overline{G}}\overline{\rho} \in \Gamma'$ where $\overline{G} = \bigcup_{\rho_i \in R} G_i$. Note that $\overline{G} \subseteq G$ by our construction of C_2 and the $X_i^{G,\varphi}$ Moreover, the fact that $Y_k \supset \{(\Gamma, l) \mid W, (\Gamma, l) \vDash \rho\}$ for all $k \in G$ implies that $\vdash \rho \to \rho_k$ for all $k \in G$. By an iterated application of (B4), we get that $\Box_{\overline{\overline{G}}}\overline{\rho} \in \Gamma'$, where $\overline{\overline{G}} = \bigcup_{\rho_i \in R_2 \cup R_3} G_i$. Moreover, (B2) implies that $\Box_k \top \in \Gamma'$ whenever Y_k in C_1. Hence, by (B4) again, we have $\Box_{\overline{\overline{G}} \cup G}\overline{\rho} \in \Gamma'$. (B3) then implies $\Box_G\overline{\rho} \in \Gamma'$ and hence, by (RE), $\Box_G\overline{\rho} \in \Gamma'$, contradicting our assumption.

The second case is that $\nvdash \overline{\rho} \to \rho$. Hence, $\overline{\rho} \wedge \neg\rho$ is consistent and there is some maximally consistest $\Gamma^{\overline{\rho} \wedge \neg\rho}$ with $\overline{\rho} \wedge \neg\rho \in \Gamma^{\overline{\rho} \wedge \neg\rho}$. Moreover, for all $Y_k \in C_3$, there is some $r(k) \in G_K$ such that $X_{r(k)}^{G_k,\rho_k} \notin \{Y_i \mid i \in G\}$. By the Chinese remainder theorem, there is some natural number q such that $q \mod \mu(G_k, \rho_k) \in M_{r(k)}^{(G_k,\rho_k)}$ for all $Y_k \in C_3$. We then get that $(\Gamma^{\overline{\rho} \wedge \neg\rho}, q) \in \bigcap_{j \in G} Y_j$ contradicting our assumption that $\bigcap_{j \in G} Y_j = \{(\Gamma, i) \mid W, (\Gamma, i) \vDash \rho\}$.

□

3 Generalizations

Naturally, we may want to expand this base logic in various directions. Seen from a semantic perspective, we might impose various frame conditions on the individual neighborhood sets. For instance, we may assume these neighborhoods to be closed under supersets, making the corresponding modalities closed under logical weakening. A related condition may demand the intersection of any two neighborhood sets to be non-empty, thus expressing a certain minimal coherence among neighborhoods. This condition is for instance imposed by (Stalnaker 2006), see also (Klein et al. 2017). Lastly, we could demand neighborhoods to contain or not contain the trivial (i.e., full) and or the empty set. Table 1 lists a number of possible extensions.

As it turns out, the present completeness approach is modular with respects to the generalizations in Table 1. Formally, let K be any subset of the frame conditions in (the first column of) Table 1. Let $\mathsf{Ax}(K)$ the corresponding axioms from the second column of Table 1 and call the logic of intersection modalities for the class of K-frames Λ_K. We then get:

Theorem 2 (Strong Completeness for Extensions of the Base Logic)**.** *A sound and strongly complete axiomatization of the logic Λ_K is obtained by adding (B1), (B2), (B3), (B4) and $\mathsf{Ax}(K)$ to any sound and complete axiomatization of* CL*, and closing the result under (RE) and (MP).*

semantic	syntactic	
$W \in \mathcal{N}_i(w)$	$\vdash \Box_i \top$	NEC
$W \notin \mathcal{N}_i(w)$	$\vdash \neg\Box_i \top$	CO-NEC
$\emptyset \notin \mathcal{N}_i(w)$	$\vdash \neg\Box_i \bot$	P
$\emptyset \in \mathcal{N}_i(w)$	$\vdash \Box_i \bot$	CO-P
indiv. Factivity	$\vdash \Box_i\varphi \to \varphi$	T
indiv. Upward closure	$\vdash \Box_i\varphi \to \Box_i(\varphi \vee \psi)$	RM
indiv. Binary cons.	$\vdash \Box_i\varphi \to \neg\Box_i\neg\varphi$	D

Table 1: Possible extensions for individual modalities

See (Van De Putte and Klein 2018) for details. Finally, we should note that some but not all of the axioms in Table 1 will generalize to the level of intersection modalities.

Remark: Assume all individual modalities $\Box_i$ satisfy NEC, CO-NEC, CO-P, T or RM. Then also all group modalities satisfy the corresponding schemes.

Notably, this does not hold for the remaining two modalities T and P. There are models where all individual neighborhoods satisfy $\emptyset \notin \mathcal{N}_i(w)$ or $\vdash \Box_i\varphi \to \neg\Box_i\neg\varphi$ respectively, but not all (or even none) of the group modalities $\Box_G$ do.

4 Outlook

For this contribution, we have concentrated on intersection modalities $\Box_G$ combining evidence of different sources $i_1, \ldots, i_n$ Notably, within neighborhood logic also a second question could be asked. Given that individual neighborhoods are usually not closed under intersection, we could ask about intersections of pooling operations among evidence from *the same* source. Schematically, we would hence also be interested in pooling modalities of the type $\Box_{\{i,i\}}$ (standing for binary intersection of i-neighborhoods), rather than only $\Box_{\{i,j\}}$. On the formal side, this change requires the transition from sets to *multisets* of indeces. The present logic generalizes straightforwardly to the multi-set case, with the sole exception that an additional axiom $\Box_M\varphi \to \Box_{M^\infty}$ comes into play for infinite multisets, where M^∞ denotes the infinite intersection of members from the M-neighborhood. Our main reason for restricting the present exposition to set-indeces only is simplicity, as multi-sets bring in some additional complexities in notation.

Secondly, there are various possible directions of generalization. In one way, the analysis can be expanded to other boolean set-operations than just intersection. We may, for instance, analyse pointwise *union* of neighborhoods (standing e.g., for uncertain success) or even set-complements of neighborhood sets. In a different direction of generalization, we may analyze special types of intersection-like modalities. In analyzing evidence based belief, for instance, (van Benthem and Pacuit 2011) focus on *non-empty* intersection of neighborhoods. Other applications, partially related to the epistemic concepts of true belief or knowledge, may be interested in intersecting only those neighborhoods that contain the present world (Özgün 2017). Time permitting, we may hint at either of these two directions of generalization in the presentation.

References

van Benthem, J., Bezhanishvili, N., and Enqvist, S. (2018). A new game equivalence, its logic and algebra. *Journal of Philosophical Logic*.

van Benthem, J. and Pacuit, E. (2011). Dynamic logics of evidence-based beliefs. *Studia Logica*, 99(1):61–92.

Brown, M. (1988). On the logic of ability. *Journal of Philosophical Logic*, 17(1):1–26.

Elgesem, D. (1997). The modal logic of agency. *Nordic J. Philos. Logic*, 2(2):146.

Goble, L. (2005). A logic for deontic dilemmas. *Journal of Applied Logic*, 3:461–483.

Governatori, G. and Rotolo, A. (2005). On the axiomatisation of elgesem's logic of agency and ability. *Journal of Philosophical Logic*, 34(4):403–431.

Klein, D. and Marra, A. (2019). From oughts to goals. a logic for enkrasia. *Studia Logica*.

Klein, D., Roy, O., and Gratzl, N. (2017). Knowledge, belief, normality, and introspection. *Synthese*, pages 1–30.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge, Mass.

Özgün, A. (2017). *Evidence in epistemic logic: a topological perspective*. PhD thesis, Université de Lorraine.

Pacuit, E. (2017). *Neighbourhood Semantics for Modal Logic*. Springer.

Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of Logic and Computation*, 1(1):149–166.

Van De Putte, F. and Klein, D. (2018). Pointwise Intersection in Neighborhood Modal Logic *Advances in Modal Logic*, 12 :591–610.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199.

Virtue Ethics for Autonomous Cars (Short Version)

Piotr Kulicki Michael P. Musielewicz Robert Trypuz

The John Paul II Catholic University of Lublin

Abstract

In this paper, we propose virtue ethics as a metaethical theory for autonomous cars, as an alternative to the utilitarian and deontological paradigms. We believe that this theory is more suitable for situations when the mechanisms behind the steering of the car are based on machine learning techniques rather than rule based algorithms. We present the main idea of this solution and discuss some virtues that can be applied to cars, namely: justice, benevolence and courage. We also investigate the way ethics can be incorporated into the learning procedures for cars and how deontic logic can contribute to this endeavour.

1 Introduction

Autonomous cars are one of the emerging technologies that will have a significant impact on society in the upcoming years. Self-driving vehicles are already present in traffic. Most of the tests that have been done so far are successful but also some problems including serious accidents in which people—a passenger of a car or a pedestrian—were killed occurred.

Although predictions estimate that traffic safety will be significantly improved (different estimations varies in this respect, but reduction of serious accidents on the level of 90-95% is expected), many people are still afraid and prefer a human driver's control over vehicles, or at least a human driver's possibility to take control over the car. One of the reasons behind this is that many people want to be sure that, in case of hazardous situations or accidents, a self-driving car will behave in a *proper way*. What does, however, a *proper way* mean? There are several levels that can be considered, but at the end there is the level of values, especially moral values. We believe that

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

the best way to define the acceptable behaviour of a self-driving car in hazardous situation is to introduce a logical formalisation taking into account all aspects of the device's behaviour.

We take advantage of some preliminary results concerning formal ethics for autonomous vehicles presented in (Kulicki et al. 2018), where we justified the approach to self-driving cars as normative agents. In this work, we will focus particularly on two issues: first choosing a metaethical theory that is most appropriate for our considerations, and second finding the best way to introduce ethics in the actual technical processes of designing autonomous cars. The latter issue may lead to determining the role of logic, especially deontic logic, in these processes and setting constraints that the prospective deontic logic for autonomous cars should fulfil. In the following two sections, we shortly discuss the two aforementioned issues.

2 Virtue Ethics as the Metaethics for Autonomous Cars

2.1 Choosing Metaethics for Self-Driving Cars

There are two major metaethical approaches in contemporary ethics: utilitarian (consequentialist) and deontological (e.g. Kantian ethics). In the first metaethical approach, we look at the various states which are consequences of our choices and try to choose the state (and the actions) that maximises utility or minimises harm. In the other one, we confront our choices with the rules of a chosen normative system. These approaches appear usually also in the context of ethical issues concerning self-driving cars (see e.g. (Jean-François Bonnefon 2016; Bringsjord and Sen 2016)). Both of these approaches encounter several difficulties when applied to artificial agents, especially robots and autonomous vehicles (see (Musielewicz 2019) for an extensive discussion of those problems). In this paper we introduce an alternative approach, namely virtue ethics.

Virtue ethics has its roots in Aristotle (Aristotle 2004) and in recent years we can observe some kind of its revival, see e.g. (Swanton and Press 2003; van Hooft 2014). Within the field of robot ethics, this conception has existed on the periphery of discussions for quite some time, but it has recently been picked up in greater length by Nicolas Berberich and Klaus Diepold in their article *The Virtuous Machine - Old Ethics for New Technology?* (Berberich and Diepold 2018), where the conception of a virtue ethics for robots has recently been expanded upon. In this text, Berberich and Diepold set out to answer the question *How can we build a machine that, owing to its constitution, acts appropriately in arbitrary situations* (Berberich and Diepold 2018, p.4)? The answer to which they find in virtue ethics.

We are in agreement with these authors that this is indeed the best place to start looking if we want to build a virtuous machine, as this approach is compatible with the sort of agents that these devices are. Artificial agents, such as autonomous cars, learn specific tasks through trial and error and the building up of vast data sets, with the application of some variety of machine learning tools. As a result of this, we end up with an agent who learns to do something and builds up habits on how to react when it perceives a certain sort of situation. At the beginning of its training, driverless cars often fail, yet after millions of miles of practice, they are able to perform their task with results that are often far better than the typical human driver.

Berbrich and Diepold also pick up on this feature of what they call autonomous moral agents (AMA), of which a driverless car is one example. Here they also root the sense of using virtue ethics in AI for runner discipline cybernetics, which returned a teleological understanding of things from its exile in the last century and in particular of a teleological understanding of living being and machines (Berberich and Diepold 2018, p.5). This reliance on teleology introduces their reliance upon Aristotle for the formation of an ethics for AMA.

Musielewicz in (Musielewicz 2019) agrees with Berbrich and Diepold in that virtue ethics is broadly speaking the best suited to AMA, but differs in the sort of virtue ethics to be employed. He relies upon a target centred virtue ethics, as developed by Christine Swanton in (Swanton and Press 2003), where we can say that an agent is virtuous when it hits the target of virtue. A virtuous action is a successful response in a particular situation. Still it is based on previous experience and a kind of *moral training*.

2.2 A Virtuous Car

Since autonomous cars are clearly different agents than humans, we need to adjust a human oriented ”original” virtue ethics to find out which virtues are appropriate for such agents. Different virtues of a cognitive machine are analysed in (Berberich and Diepold 2018). We are here interested directly in the virtues applicable to autonomous vehicles. The ones that most clearly can be applied here are, in our opinion, justice, benevolence and courage.

In regards to driverless cars, we need to address the field, mode and target of justice. The field of this virtue primarily (though not necessarily) relates to following the universal law and the various norms that relate to the car itself as a normative agent. The targets of the car’s duties of justice include itself, its passengers, other road occupants, and the state (or some other normative authority) within which it is driving. The modes of response to this universal justice depend upon the particular patient in question and so will depend upon the particular context of some virtuous act. The hitting of the targets of this virtue also is contextually dependent upon the laws related to the situation at hand. At any given time the driverless car, being the bearer of the various norms related to driving, owes something to the state: such as the obligation to obey the speed limit or to stop for pedestrians and other laws inscribed in the criminal code and statutes— or to private individuals—in matters related to civil law.

Benevolence aims at the promotion of what is good. What counts as a benevolent act, much like what is a just act, depends upon the context within which the act is undertaken. Considering a driverless car, the targets of benevolence would be related to its sphere of activity, i.e. driving, and the patients of these acts can include its occupants, other road users, pedestrians, etc. where the autonomous vehicle has the appropriate duty of care relevant to its relationship with the patient in question.

Crucial to the understanding of benevolence towards others is the Kantian notion of not treating others as mere means to one’s own ends, conjoined with a recognition of the moral value of other humans as entities that are ends in themselves (Swanton and Press 2003, p.107). To this universal understanding of benevolence, particular considerations can be added due to particular circumstances. So a nurse has a special duty of care for their patient, a guardian to their ward, a parent to their child. In the case of the autonomous car, the special position is given to its passengers and owner. However this kind of *preference* cannot cross the line of treating other humans as means to this end and, for example, by depriving other traffic participants of their rights (right-of-way) when a passenger is in a hurry.

Courage is the willingness to confront danger. In the case of self-driving cars, it may seem not to be the right guide of behaviour since one of the most desirable features here is safety. However, it is not possible to participate in traffic without any danger. Thus, courage is here about the appropriate assessment of a risk occurring while driving. Already Aristotle defined it as a balance between cowardice and foolhardiness. Manifestations of courage may be for example driving with a speed appropriate to the traffic conditions – not too slow (which would be cowardice) and not too fast (foolhardiness). Moreover, especially in situations that are not typical, looking for creative solutions of a problem rather then following safe rules also may be understood as courage.

2.3 Practising Virtues

Driverless cars being presently trained are done so by means of machine learning techniques, where wrong acts are punished and right (or correct) acts are rewarded, allowing for the machine to learn through habituation, and through trial and error. Both computer simulators and real road training can be used in the process of training. Moral virtues are likewise gained through trial and error and habituation. The more a normative agent acts and successfully hits the targets of virtues, the easier it is for them to successfully hit similar targets of that virtue in the future.

3 Ethical Norms and lLgic in the Process of Designing of Autonomous Cars

3.1 Ethical Control in the Operating System

In (Bringsjord and Sen 2016) we read that the self-driving car should have implemented an *operating-system-rooted ethical control* by which the authors mean *logics that are connected to the operating-system level of [...] cars, and that ensure these cars meet all of their moral and legal obligations, never do what is morally or legally forbidden, invariably steer clear of the invidious, and, when appropriate, perform what is supererogatory*.

In this approach, decision making, including decisions based on ethical principles, is *hard-coded* into a software operating a self-driving car. Similar ideas are used in (Zhao et al. 2015, 2017) where Advanced Driving Assistant System Ontologies are proposed and combined with logical rules of reasoning that are expressed in the Semantic Web Rule Language (SWRL). An attempt to enrich such a system with ethical principles formalised in the form of a deontic logic is presented in (Kulicki and Trypuz 2019).

3.2 Machine Learning as Practising Virtues

However, most, if not all, actual autonomous car operating systems are based on different variants of machine learning (applying such tools as deep learning and reinforcement learning in neural networks). That way of programming the control over vehicles has shown to be far more successful than rule based expert systems.

How can we introduce various ethical principles into systems that are trained in this way? Virtue ethics seems to provide a fruitful point of view to the problem. Ethics formulated as a set of virtues can be used in the process of learning as a pattern of the positive signals in the process of machine learning.

First of all, we can interpret example based (machine) learning of basic driving skills by an operating system controlling a self-driving car, like keeping the track, accelerating and decelerating according to traffic conditions, avoiding crashes, as practising some kind of virtues. Similarly, we can treat introducing traffic regulations into the operating system through learning as a part of gaining the virtue of justice or promoting certain modes of actions as a part of gaining the virtue of benevolence. As learning is based on the mechanism of punishment and reward, while designing the learning procedure, we need to decide which particular behaviours should be punished and which should be rewarded.

3.3 Transparency of Ethical Decisions and Efficiency of Learning

In most cases, it is not difficult to judge which ones of a car's behaviour are good (virtuous) or bad. Still there are some hard cases quite well recognised when ethical issues concerning self driving cars are considered. They include finding the right balance between different values relevant to driving such as mobility, safety and legality and the question of crash optimisation, which focuses the public interest. Yet, another question to answer that is especially interesting from the perspective of learning and virtue ethics, is whether a self-driving car should, in its decision process, imitate the behaviour of (good) human drivers or follow some ideal choices proposed by moral authorities or democratic procedures (based e.g. on empirical research like the one presented in (Jean-François Bonnefon 2016)).

We do not pretend to solve these problems here. We just want to point out two issues for which we believe that using a logical description of an intended behaviour is useful, no matter what the actual choices are: the need for transparency and efficiency in learning hard cases.

The former of them is present in most documents concerning policies for autonomous cars. United States Department of Transportation states that the implicit ethical values must be made clear so that all stakeholders can ensure that these *ethical judgements and decisions are made consciously and intentionally* (U.S. Department of Transportation, National Highway Traffic Safety Administration 2016, p.26). This claim for transparency is mirrored in the report made by the ethics commission of the *Bundesministerium fur Verkehr und digitale Infrastruktur* (hereinafter BMVI) made in June of 2017. Here the BMVI underscores the importance of maintaining the autonomy of people in making ethical decisions (Federal Ministry of Transport and Digital Infrastructure, Ethics Commission 2017, p.16).

The later of the two issues is recognised by researchers working in machine learning of robots in (Li et al. 2017). The authors notice that rules and experience/knowledge play a very important role in the process of human learning, making it much more efficient than using a pure trial and error method. They claim that robot learning can be analogous and consequently *[b]eing able to formally express these rules as reward functions and incorporate them in the learning process is helpful and often necessary for a robot to operate in the world* (Li et al. 2017).

3.4 Specification of the Desired Formalism

In both cases, some way of specifying the desired behaviour of a self-driving car, in a way that is readable both for humans and computer programs, is required. Some kind of deontic logic seems to be the best option when looking for a way to express explicit ethical rules and respective reward functions.

Which logical formalism should be chosen? How should it be used to gain maximal benefits? These questions remain open and are the challenges for our future work within the project: *Deontic logic for autonomous cars*.

At this moment we can just point out some postulates concerning a satisfactory formalism. An obvious thing is that, within such a formalism, we should be able to represent the possible actions of a car, their desirable and undesirable results, other agents and objects that are involved in traffic and some of their properties. This would allow us to express norms, values and finally, on their basis, also talk about virtues.

Moreover, there is a need for considering alternative solutions to the problems occurring on the road. This issue is mentioned in (Kulicki and Trypuz 2019) where the lack of such possibility is pointed out as a drawback of the ontological approach from (Zhao et al. 2015, 2017). Without such a feature we can just formulate if-then rules.

Another desirable feature of the prospective formalism is the possibility to express probability, or likelihood, of the undesired events to judge risks. Such an approach is present in (Contissa et al. 2017). This would be a step to defining properly the virtue of courage.

Yet another useful aspect of the formalism would be the temporal aspect of events. Let us just notice here that temporal logic is a starting point of the approach to the norm-based reinforcement learning in (Li et al. 2017).

4 Conclusion

We want to know that an autonomous car will behave like we want it to. What we want can be defined in many different ways. We may want it to follow rules or perhaps we may prefer that it decides upon an action by calculating the possible utility of different choices. Here, we advocate for another point of view: we may want to know that a car is well trained and is doing its best to be virtuous.

Still in this approach there is a place for logic. Logic is not, however, a part of a car's operating system, but a tool that is used to describe expected behaviour in the process of machine learning. In this place, logic can provide transparency of the values introduced through this process.

Acknowledgements

The research is supported by the National Science Centre, Poland (HARMONIA, *Deontic logic for autonomous vehicles*: UMO-2017/26/M/HS1/01092)

References

Aristotle (2004). *Nicomachean Ethics*. Cambridge University Press.

Berberich, N. and Diepold, K. (2018). The virtuous machine - old ethics for new technology?

Bringsjord, S. and Sen, A. (2016). On creative self-driving cars: Hire the computational logicians, fast. *Applied Artificial Intelligence*, 30(8):758–786. https://doi.org/10.1080/08839514.2016.1229906.

Contissa, G., Lagioia, F., and Sartor, G. (2017). The ethical knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, 25(3):365–378.

Federal Ministry of Transport and Digital Infrastructure, Ethics Commission (2017). Automated and Connected Driving.

Jean-François Bonnefon, Azim Shariff, I. R. (2016). The social silemma of autonomous vehicles. *Science*, 352(6293):1573–1576.

Kulicki, P. and Trypuz, R. (2019). Judging actions on the basis of prima facie duties. the case of self-driving cars. *Logic and Logical Philosophy*.

Kulicki, P., Trypuz, R., and Musielewicz, M. (2018). Towards a formal ethics for autonomous cars. In Broersen, J., Condoravdi, C., Nair, S., and Pigozzi, G., editors, *Deontic Logic and Normative Systems, 14th International Conference, DEON 2018*, pages 193–209.

Li, X., Vasile, C. I., and Belta, C. (2017). Reinforcement learning with temporal logic rewards. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 3834–3839.

Musielewicz, M. (2019). *On the Application of Norms within Driverless Cars*. PhD thesis, The John Paul II Catholic University of Lublin.

Swanton, C. and Press, O. U. (2003). *Virtue Ethics: A Pluralistic View*. Oxford University Press.

U.S. Department of Transportation, National Highway Traffic Safety Administration (2016). *Federal Automated Vehicle Policy Accelerating the Next Revolution in Road Safety*. US Federal policy concerning AV.

van Hooft, S. (2014). *The Handbook of Virtue Ethics*. Acumen Publishing.

Zhao, L., Ichise, R., Liu, Z., Mita, S., and Sasaki, Y. (2017). Ontology-based driving decision making: A feasibility study at uncontrolled intersections. *IEICE Transactions*, 100-D(7):1425–1439. http://search.ieice.org/bin/summary.php?id=e100-d_7_1425.

Zhao, L., Ichise, R., Mita, S., and Sasaki, Y. (2015). Core ontologies for safe autonomous driving. In Villata, S., Pan, J. Z., and Dragoni, M., editors, *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015.*, volume 1486 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-1486/paper_9.pdf.

Incomplete Preference and Indeterminate Qualitative Probability (Extended Abstract)

Yang Liu

University of Cambridge

Abstract

The notion of qualitative probability defined in Bayesian subjective theory stems from an intuitive idea that, for a given pair of events, one event may be considered "more probable than" the other. Yet it is conceivable that there are cases where it is indeterminate as to which event is more probable, due to, e.g., lack of robust statistical information. This paper provides an axiomatisation of a generalised notion of qualitative probability within the analytic framework of L. J. Savage which includes probabilistic indeterminate cases.

1 Introduction

Modern Bayesian decision theory seeks to ground statistical inference in a logical process of rational decision-making. Central to this goal is the task of specifying how rational agents organise, in a coherent manner, their probabilistic and value judgments in face of uncertainties. As exemplified in classical works of Ramsey (1926), de Finetti (1937), and Savage (1954), the upshot of this approach is a representation theorem, where the decision maker's beliefs and values are characterised, respectively, by a single subjective probability measure and a subjective utility function, provided that various postulates governing rational decision-making are granted.

Common to many classical Bayesian models of rational decision-making, the most demanding assumption is arguably the so-called "completeness axiom." This axiom mandates that an agent's preferences among possible courses of actions in any given decision situation be representable by a *complete* ordering. That is to say, in classical Bayesian decision theory it is assumed that

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

the decision makers are maximally opinionated in their choices of actions in that they are always prepared to compare and rank any two given options in any decision situations.

The completeness assumption is often questioned on the ground that the agent may, for various reasons, lack rational basis for always being able to choose between a given pair of options the preferred one. For instance, the agent may lack robust statistical information in assessing the probabilistic nature of the events under which the acts are to be performed. In this case the decision makers face *probabilistic indeterminacy*. As a consequence, they are rationally justified in suspending judgments on their actions: these acts are *incomparable*.

To be sure, the consideration of incomparable acts in a decision model is a natural one. In fact, Savage himself was tempted by the idea of "analyzing preference among acts as a partial ordering, that is, . . . admitting that some pairs of acts are incomparable" and this, he says, "would seem to give expression to introspective sensations of indecision or vacillation, which we may be reluctant to identify with indifference" (Savage 1972, p. 21). However, it is conceivable that the employment of incomparable acts (represented by an *incomplete* ordering) into a decision model will result in a different preferential structure from what was adopted in Savage's original framework. Savage didn't think that much can be advanced in pursuing this direction, "a blind alley" rather, he said, he nonetheless added that "only an enthusiastic exploration could shed real light on the question."

In the past two decades or so we have seen a number of such 'enthusiastic explorations:' (Seidenfeld et al. 1995; Ok 2002; Galaabaatar & Karni 2013; Ok et al. 2012) to name just a few. These efforts share a common goal of attempting to make classical decision theory more tractable and more realistic by relaxing the completeness axiom in their respective models. It should be emphasised that this direction of research for decision theory, i.e., modelling rational decision making without the completeness assumption, has far-reaching implications than mere theoretic interests. Sen (2004, 2018) recently reiterated the importance of incompleteness in social justice and global politics. And in the field of artificial intelligence (AI), Zaffalon & Miranda (2017, 2019) also pointed out the crucial role incompleteness plays in building robust AIs. It is the goal of this paper to make another 'enthusiastic exploration' in this direction.

Now, the literature on decision theory with incomplete preferences classifies agent's inability to compare certain pair of options in decision situations as coming from two main sources: the uncertainty regarding the likelihood of the event in question (i.e., *probabilistic indeterminacy*) and the uncertainty about the values of the consequences of the acts available to the decision maker (i.e., *value indeterminacy*). The former is sometimes referred to in the economic literature as the decision maker's *indecisiveness in belief*, the latter *indecisiveness in tastes*.[1] We will follow this dichotomy in this paper. However, to simplify things, in what follows we only consider probabilistic indeterminacy as the sole source of incompleteness.

Most recent theoretic work on incompleteness cited above are based on the analytic framework of Anscombe & Aumann (1963) (cf. Remark 1 below). In contrast, the analysis here is set within the framework of Savage (1972), the latter is widely seen as *the* paradigmatic system of subjective decision making, on which a classical theory of personal probability is based. Our eventual goal is to gain a representation of partially ordered preferences among acts in terms of indeterminate probabilities and a utility function, this paper contains the first step towards this goal regarding an

[1] See (Dubra et al. 2004) for a discussion. The notion of decisiveness in beliefs corresponds to the notion of probabilistic sophistication in (Machina & Schmeidler 1992), by which the authors mean that the agent is capable of assigning precise subjective probabilities to events. As pointed out by Levi (1986), the well-known paradoxes of Allais and Ellsberg are, respectively, examples of decision making with indeterminacy in values and indeterminacy in beliefs.

axiomatisation of indeterminate qualitative probability, which we shall explain in more detail next.

1.1 Savage's approach

Savage's theory centres on a binary relation which models a decision maker's preferences over possible courses of actions. A set of axioms is postulated on this preference relation. The culmination of theory is a representation theorem with which an agent's preferences can be represented by expected utilities under the proposed postulates.

More precisely, from the first five of Savage's seven postulates a *comparative* notion of subjective probability is derived which reflects the agent's qualitative probabilistic judgments over possible circumstances under which these actions are performed. With the sixth postulate, the derived qualitative probability (to be defined precisely below) is further precisified with a numerical probability measure and a personal utility function for simple acts (i.e., acts that may potentially lead to finitely many different consequences under different states). The last postulate plays the sole role of extending the utility function for simple acts to all acts.[2]

Table 1: Inferential order in Savage's system.

P1-P5		**+ P6**		**+ P7**
Qualitative probability	⇒	Quantitative probability Utility for simple acts	⇒	Utility for all acts

Remark 1. Savage's method differs from the approaches adopted by Ramsey (1926) and Anscombe & Aumann (1963) in that the agents' subjective probabilities in these latter cases are derived from their personal utilities, which in turn are constructed based on some presupposed chance mechanisms (or, in the case of Ramsey, the notion of ethically neutral propositions, which can be employed to play the same role as an unbiased coin receiving objective probability 1/2). This inferential order is reversed in Savage's theory of subjective utility where the decision makers' preferences over acts is taken as the only primitive notion, from which their personal probabilities and utilities are subsequently revealed. As a result of this methodological reversal, Savage's approach may appear to have some computational disadvantages in the sense that the mathematical representation theorem given in Savage's theory is considerably more involved than many of its alternatives (including Ramsey's and Anscombe and Aumann's systems), yet the theory is conceptually significant in that the system is seen as a purely subjective framework with no reference to objective probabilities.

In this paper, we generalise Savage's notion of qualitative probability to cover probabilistically indeterminate events. That is, we aim to generalise Savage's system and arrive at a representation of the cases where, for two given events E and F, neither E is considered more probable than F nor that F is more probable than E. This will be the main result of this paper. Our generalisation parallels the first part (P1-P5) of Savage's construction illustrated in Table 1, we leave further generalisations of indeterminate quantitative probabilities for future work. In the next section we provide some analysis of basic elements of Savage's system as well as a more precise formulation of the goal of this paper.

[2] An outline of Savage's proofs can be found in Gaifman & Liu (2018), a full exposition in Fishburn (1970); Liu (2016).

1.2 Indeterminate qualitative probability

Recall that a *Savage decision model* is a structure of the form $(S, \mathcal{B}, X, \mathcal{A}, \succcurlyeq)$ where S is an (infinite) set of *states* of the world; $\mathcal{B}$ is a Boolean algebra equipped on S, each element of which is referred to as an *event* in a given decision situation; X is a set of consequences; and a (Savage) *act* is a function f mapping from S to X, the intended interpretation is that $f(s)$ is the consequence of the agent's action f performed when the state of the world is in s. As a primitive notion of the model, $\succcurlyeq$ is a binary relation on the set of all acts, the latter denoted by $\mathcal{A}$. For any $f, g \in \mathcal{A}$, $f \succcurlyeq g$ says that f is *weakly preferred to* g. Say that f is *strictly preferred to* g, written $f \succ g$, if f is weakly preferred to g but not vice versa, and that f is *indifferent to* g, denoted by $f \sim g$, if f is weakly preferred to g and vice versa.

Definition 1 (fused acts). *For any $f, g \in \mathcal{A}$, define the fusion of f and g with respect to an event E (a set of states), written $f|E + g|\overline{E}$, to be such that:*

$$(f|E + g|\overline{E})(s) =_{\text{Df}} \begin{cases} f(s) & \textit{if } s \in E \\ g(s) & \textit{if } s \in \overline{E}, \end{cases} \tag{1}$$

where $\overline{E} = S - E$ is the compliment of E.

In other words, $f|E + g|\overline{E}$ is the act which agrees with f on event E, with g on $\overline{E}$, and it is easily seen that $f|E + g|\overline{E} \in \mathcal{A}$. This notion of fused acts can easily be generalised for a series of acts $\{f_1, \ldots, f_n\}$ and a partition $\{E_1, \ldots, E_n\}$ of the state space such that the following is also a Savage act: $f_1|E_1 + f_2|E_2 + \cdots + f_n|E_n$.

Definition 2 (constant acts). *For any $a \in X$, an act is said to be constant with respect to consequence a, written $\mathfrak{c}_a$, if*

$$\mathfrak{c}_a(s) = a \quad \textit{for all } s \in S. \tag{2}$$

Remark 2. By definition, act $\mathfrak{c}_a$ 'constantly' outputs consequence a no matter which state $s \in S$ transpires. Constant acts play an import role in Savage's proofs. One motivation for having this type of acts is that it can be used to induce a preference ranking $\geq$ among consequences in terms of preferences $\succcurlyeq$ among acts, that is, for any $a, b \in X$, $a \geq b =_{\text{Df}} \mathfrak{c}_a \succcurlyeq \mathfrak{c}_b$.[3] But in order to get such an induced ordering it is necessary to assume that for *any* consequence $a \in X$ there exists a constance act $\mathfrak{c}_a$. The latter is in fact employed in Savage theory as an implicit assumption.

However, unlike vNM's notion of degenerate lotteries, the constant acts assumption is highly problematic. Take, for instance, Savage's own omelette example, it is difficult to imagine what act can constantly result in a six-good-egg-omelette (the consequence) even when the sixth is bad (the state). Elsewhere (Gaifman & Liu 2018), we addressed, among other things, this issue and developed new techniques to show that Savage's theory can be simplified with a weakened assumption of the existence of *two* constant acts (but without mandating that there is a constant act for each consequence). This paper presupposes the assumptions and proof techniques we made there.[4]

[3] This is a similar technical construct inherited from von Neumann-Morgenstern's (vNM) notion of *degenerate lotteries* in their utility theory that a degenerate lottery always yields the same (monetary) reward regardless which state obtains.

[4] Another technical assumption made by Savage is that the background algebra $\mathcal{B}$ is a σ-algebra – in fact he demands $\mathcal{B}$ to be 2^S. The techniques we developed in (Gaifman & Liu 2018) enables us to drop also this implicit assumption. Note that, although we presuppose the proof techniques we developed previously, the proofs used in this paper do not require them and can be read independently.

With notions of fused acts and constant acts in hand, Savage then defines a concept of what it means for an event to be said to be *more probable* than another in terms of preferences among acts.

Definition 3. *For any events $E, F \in \mathcal{F}$, say that E is weakly more probable than F, written $E \succeq F$ (or $F \preceq E$), if, for any constant acts $\mathfrak{c}_a, \mathfrak{c}_b$ with $\mathfrak{c}_a \succcurlyeq \mathfrak{c}_b$, we have*

$$\mathfrak{c}_a|E + \mathfrak{c}_b|\overline{E} \succcurlyeq \mathfrak{c}_a|F + \mathfrak{c}_b|\overline{F}. \tag{3}$$

E and F are said to be equally probable*, written $E \equiv F$, if both $E \succeq F$ and $F \succeq E$ hold.*

An intuitive explanation of (3) is that the act $\mathfrak{c}_a|E + \mathfrak{c}_b|\overline{E}$ is weakly preferred to $\mathfrak{c}_a|F + \mathfrak{c}_b|\overline{F}$ because it is weakly more probably for the former to result in the more preferable consequence of a than the latter. Now, note that Savage assume that $\succcurlyeq$ is a *complete order* (the first axiom of Savage's system), that is, it is assumed that any two given acts are comparable. As a consequence of this strong assumption (and via Definition 3 and the weakened constant act assumption), any two given events E, F are taken to be probabilistically comparable, in other words, any events E and F are assumed to be in one of the following relations: $E \succeq F$, $E \equiv F$, or $F \succeq E$.

In this paper, we will take the step of relaxing the completeness requirement in Savage's system, and consider the possibilities that two acts f, g are *incomparable* under the preference relation $\succcurlyeq$, in symbols, $f \bowtie g$. As mentioned above, we will attribute these incomparabilities solely to the probabilistic indeterminacy involved, where we take that, for a given pair of events E, F it is possible that it is *indeterminate* that one event is more probable than the other, written

$$E \bowtie F. \tag{4}$$

The goal of this project is to provide a Savage-style axiomatisation of a generalised notion of qualitative probability that covers indeterminate cases. This work thus can be seen as providing a decision-theoretic foundation for *indeterminate qualitative probability*.

References

Anscombe, F. & Aumann, R. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, 199–205.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In H. E. Kyburg & H. E. Smokler (Eds.), *Studies in Subjective Probability* (2d ed.). (pp. 53–118). Huntington, N.Y.: Robert E. Krieger Publishing Co., Inc., 1980.

Dubra, J., Maccheroni, F., & Ok, E. A. (2004). Expected utility theory without the completeness axiom. *Journal of Economic Theory*, *115*(1), 118–133.

Fishburn, P. C. (1970). *Utility Theory for Decision Making*. New York: Wiley.

Gaifman, H. & Liu, Y. (2018). A simpler and more realistic subjective decision theory. *Synthese*, *195*(10), 4205–4241.

Galaabaatar, T. & Karni, E. (2013). Subjective expected utility with incomplete preferences. *Econometrica*, *81*(1), 255–284.

Levi, I. (1986). The paradoxes of Allais and Ellsberg. *Economics and Philosophy*, *2*(01), 23–53.

Liu, Y. (2016). *Elements of Bayesian Decision Theory*. Manuscript (available at http://yliu.net/s/SDT.pdf).

Machina, M. J. & Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica: Journal of the Econometric Society*, 745–780.

Ok, E. A. (2002). Utility representation of an incomplete preference relation. *Journal of Economic Theory*, *104*(2), 429–449.

Ok, E. A., Ortoleva, P., & Riella, G. (2012). Incomplete preferences under uncertainty: Indecisiveness in beliefs versus tastes. *Econometrica*, *80*(4), 1791–1808.

Ramsey, F. P. (1926). Truth and probability. In H. E. Kyburg & H. E. Smokler (Eds.), *Studies in Subjective Probability* (pp. 23–52). Robert E. Krieger Publishing Co., Inc. 1980.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons, Inc.

Savage, L. J. (1972). *The Foundations of Statistics* (Second Revised ed.). New York: Dover Publications, Inc.

Seidenfeld, T., Schervish, M., & Kadane, J. (1995). A representation of partially ordered preferences. *The Annals of Statistics*, *23*(6), 2168–2217.

Sen, A. (2004). Incompleteness and reasoned choice. *Synthese*, *140*(1), 43–59.

Sen, A. (2018). The importance of incompleteness. *International Journal of Economic Theory*, *14*(1), 9–20.

Zaffalon, M. & Miranda, E. (2017). Axiomatising incomplete preferences through sets of desirable gambles. *Journal of Artificial Intelligence Research*, *60*, 1057–1126.

Zaffalon, M. & Miranda, E. (2019). Desirability foundations of robust rational decision making. *Synthese*.

Reasoning about Social Choice in Modal Logic Quickly Becomes Undecidable

Erik Parmann [†] Thomas Ågotnes [*]

[†] Equinor ASA [*] University of Bergen / Southwest University

Abstract

In this paper we show that modal logics for reasoning about social choice quickly become undecidable. In particular, we study modal logics that can be used to reason about situations involving both *actual* and *claimed* preferences in the context of a social choice function, and argue that reasoning on this level often occurs in social choice. We formally define a particular logic, interpreted in such situations, that can express the properties involved in the Gibbard-Satterthwaite theorem. We, then, however, demonstrate that any modal logic interpreted in such situations having a certain natural expressive power, in particular a modality quantifying over all possible claimed preferences, becomes undecidable when there are enough agents in the system.

1 Introduction

Modal logic has been proposed as a natural framework for formal reasoning about social choice (Ågotnes et al. 2006; Pauly 2008; Troquard et al. 2009; Ågotnes et al. 2011; Troquard et al. 2011). However, the computational properties of suitable modal logics have not received much attention so far.

How hard is it to reason about social choice? That depends on the expressive power of the logical language: the trade-off between expressive power and computational complexity is one of the main concerns in applied logic. In this paper we are interested in studying the computational properties of a minimal modal language with "adequate" expressive power. What is adequate expressive power? It has been suggested (Ågotnes et al. 2006) that one measure of expressive power

[*] Corresponding author.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

is the ability to express the different properties of *social welfare functions* or *social choice functions* involved in key results such as Arrow's theorem (Arrow 1951) or the Gibbard-Satterthwaite theorem (Gibbard 1973; Satterthwaite 1975), viz. *non-dictatorship*, *Pareto efficiency*, *independence of irrelevant alternatives* and *strategy-proofness*, and thereby being able to formally prove such results using a (complete) reasoning system. However, that test alone can be satisfied trivially, by simply introducing a symbol for each of the considered properties, e.g. d for "the social choice function is dictatorial" (the mentioned results can now be included directly as axioms in the logic and thus trivially proven). However, such a language would only allow us to reason on a very high abstraction level. Interesting formal reasoning about social choice, such as the reasoning demonstrated in proofs of Arrow's theorem or the Gibbard-Satherthwaite theorem in the literature, is on a different level of abstraction.

Social choice theory is concerned with how to select outcomes when two or more agents have (possibly different) preferences. In this setting, an agent is given a set of alternatives, and is asked to order them according to her preferences. A social choice function is a function which, given the orderings of the different agents, selects a winner (the outcome). Note that an agent can "lie", meaning that what he votes does not necessarily reflect what he prefers. A key type of properties found in proofs of, e.g., the Gibbard-Satterthwaite theorem are properties that say something about the relationship between the *actual preferences* and the *claimed preferences* of agents; one particular example is the property called strategy-proofness: is it always best to vote for your favorite? It is on this level of actual and claimed preferences interesting reasoning about social choice is found. Being able to discern between different situations involving different actual and claimed preferences is fundamental in reasoning about social choice.

Now consider a logic that can express the following facts about such situations:

(1) pa_i: "agent i weakly prefers the outcome of the social choice function on the *actual preferences* over the outcome of the social choice function on the *claimed preferences*"

and in addition has a modal operator $\Box_i$ that quantifies over the claimed preferences of agent i, with the following meaning, where ϕ is some formula representing a statement about such situations:

(2) $\Box_i\phi$: "for all possible claimed preferences of agent i, ϕ is true".

If we in addition add a modal operator $\Box^U$ to quantify over all *honest situations*, i.e., situations where the actual preferences and the claimed preferences coincide for each agent, we can express the strategy-proofness property by the formula

$$\Box^U \left(\bigwedge_{1 \le i \le n} \Box_i pa_i \right).$$

The formula says that in all situations where agent i is the only one changing her vote, her actual preference is always at least as good as any claimed preference. However, we show in this paper that *any modal logic for reasoning about arbitrary sets of alternatives that can express (1) and that has the modality in (2), becomes undecidable if there are enough agents.*

Of course, this does not mean that "reasoning about social choice is undecidable". In particular, the undecidability result holds for modal logics where arbitrary nestings of modalities are allowed, and as we illustrated above only limited nestings are needed, e.g., to express strategy-proofness. But we argue that it has some fundamental implications for *modal logic* for social choice. If we want a modal logic that can reason on the level of actual and claimed preferences, and we want it to be able to express, e.g., strategy-proofness it has to be able to express (1) and be able to quantify over all possible claimed preferences (2). Quantification in standard modal logic is obtained by using modalities, which must be allowed to be arbitrarily nested if we want a standard modal language. We therefore argue that the undecidability result is relatively fundamental for the relationship between modal logic and social choice, in the sense that it follows from relatively weak assumptions about the ability to discern between different situations.

The rest of the paper is organized as follows. In the next section we give a brief review of necessary concepts from social choice theory and modal logic. In Section 3 we introduce a formal modal logic interpreted in the types of situations discussed above, and show that it can express interesting properties of such situations. We then, in Section 4, show that the fragment of that logic with operators expressing exactly the minimal properties discussed above, and thus *any* modal logic with such operators, is undecidable given that there are enough agents. We discuss related and future work and conclude in Section 5.

2 Background

We briefly review the main concepts from social choice theory and modal logic that we will use. We refer to (Arrow et al. 2002; Blackburn et al. 2001) for a less concise presentation.

2.1 Social Choice

The key concept we are concerned with in this paper is the *social choice function*, intuitively mapping a profile representing the preferences of each of the agents over a set of *alternatives* to a single alternative.

Formally, given a number n representing the (finite) number of *agents* and a nonempty countable set of *alternatives* A, $L(A)$ denotes the set of all linear orders[1] over A, the possible *preference relations* over A, and $L(A)^n$ denotes the set of all n-tuples of linear orders. An element of $L(A)^n$ will be called a *profile*. For $D \in L(A)^n$ and $1 \leq i \leq n$, we will use D_i to denote the linear order of agent i in D. $(x, y) \in D_i$, henceforth often written xD_iy, is intended to mean that agent i *weakly prefers* y *over* x (x is ordered lower than or equal to y in the linear order representing i's preferences). We will use $<^R$ to denote the *strict* version of a preference relation $R \in L(A)$, i.e., $x <^R y$ holds iff yRx does not hold.

Given n and A, we define a binary relation $\sim_i$ for $1 \leq i \leq n$ on $L(A)^n$ in the following way.

Definition 2.1. $\sim_i$ is the binary relation on $L(A)^n$, such that for all $D, P \in L(A)^n$ we have $D \sim_i P$ if and only if $D_j = P_j$ for all $j \neq i$, $1 \leq j \leq n$. $\dashv$

In other words, $D \sim_i P$ whenever the two profiles are the same except possibly for agent i.[2] We can now proceed to define social choice functions.

Definition 2.2. A *social choice function* (SCF) F over a non-empty set of alternatives A and an integer $n \in \mathbb{N}$ is a function of the form $F\colon L(A)^n \to A$. For a profile D, the outcome $F(D)$ is called the *winning* alternative.

[1] A *linear order* is an antisymmetric, transitive and total binary relation.

[2] This use of the symbol $\sim_i$ must not be confused by the way it is used in epistemic logic (Fagin et al. 1995), in particular in interpreted systems semantics where the relation holds between two tuples if and only if the i-component is *the same*.

SCF(A, n) is the set of all social choice functions $L(A)^n \to A$ over alternatives A and integer n. **SCF**(n) is the union of **SCF**(A, n) for all countable sets A, and **SCF** is the union of **SCF**(n) for all $n \in \mathbb{N}$. $\dashv$

We now recall three properties of social choice functions that we will be interested in and that form the basis of the Gibbard-Satterthwaite theorem. Given an SCF F over A and n, we say that F is:

- *k-winning* when the image of F has at least k elements;
- *strategy-proof (SP)* if for all agents i and profiles $P = (P_1, \ldots, P_n)$, it holds that $F(P')P_iF(P)$ for all $P \sim_i P'$. In other words, a social choice function is strategy-proof if an agent's actual preference is always at least as good as any claimed preference; and
- *i-dictatorial* if for all $D, P \in L(A)^n$, we have that $F(P)D_iF(D)$. In other words, i is a dictator for F if, for every profile D, there is no other possibly winning alternative, i.e., no other alternative $F(P)$ for some profile P, such that i prefers $F(P)$ to $F(D)$ in D.[3] F is *dictatorial* when it is i-dictatorial for some agent $1 \leq i \leq n$.

The following is the most well known result about SCFs.

Theorem 2.3 (Gibbard-Satterthwaite (1973; 1975))**.** *For any SCF F, if F is 3-winning and strategy-proof, then it is dictatorial.*

2.2 The Modal Logic $\mathbf{S5}^m$

In the following we will make use of a particular, well-known, modal logic: the *product logic* $\mathbf{S5}^m$. We now briefly review the concepts and results we need. The reader might want to skip this presentation on a first reading and refer back to it when needed.

Given a natural number m, a *frame* $\mathcal{F}$ is a tuple $(W, R_1, \ldots, R_m)$ where W is a set of states and R_i, $1 \leq i \leq m$, is a binary *accessibility relation* on W.

$\mathcal{L}_m$ is the modal language having m boxes and propositional letters from a countable infinite set P. Formally, it is defined by the following grammar, where $p \in \mathsf{P}$ and $1 \leq i \leq m$:

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid \Box_i\phi$$

Definition 2.4 (Product frames)**.** The product $\mathcal{F}_1 \times \cdots \times \mathcal{F}_m$ of m single-agent frames $\mathcal{F}_1 = (W_1, R_1), \ldots, \mathcal{F}_m = (W_m, R_m)$, is the frame

$$\mathcal{F}_1 \times \cdots \times \mathcal{F}_m \overset{def}{=} (W_1 \times \cdots \times W_m, \overline{R}_1, \ldots, \overline{R}_m)$$

where for each $i \in \{1, \ldots, m\}$, $\overline{R}_i$ is the binary relation on $W_1 \times \cdots \times W_m$ such that

$$(u_1, \ldots, u_m)\overline{R}_i(v_1, \ldots, v_m) \text{ if and only if } u_iR_iv_i \text{ and } u_k = v_k, \text{ for } k \neq i.$$

$\dashv$

[3] Several definition of dictatorship can be found in the literature. The basic idea is, of course, that there is some agent i who can choose the winning outcome. However, several issues come up when this idea is formalised, concerning how the dictator chooses the winning element and how much power the dictator should have. The definition we have used here is a standard definition (Barbera 2001) which avoids many problems with other definitions.

We will by U^m denote the class of all products of frames which are the product of m single-agent frames, each with a relation R_i that is the universal relation (see below) on its domain W_i.

A *model* $\mathcal{M}$ consists of a frame with states W and a valuation function $V\colon \mathsf{P} \to \wp(W)$. A *pointed model* $(\mathcal{M}, w)$ consists of a model and a state w in $\mathcal{M}$. We will use $\mathbf{Mod}^p(U^m)$ to denote all pointed models built on frames in U^m.

Satisfaction of a formula $\phi \in \mathcal{L}_m$ in a pointed model $(\mathcal{M}, w) \in \mathbf{Mod}^p(U^m)$, $(\mathcal{M}, w) \models \phi$, is recursively defined as follows (standard modal logic).

- $\mathcal{M}, w \models p$ if and only if $w \in V(p)$;
- $\mathcal{M}, w \models \Box_i\phi$ if and only if $\mathcal{M}, v \models \phi$ for all $v \in W$ such that $w\overline{R_i}v$;
- $\mathcal{M}, w \models \phi \wedge \psi$ if and only if $\mathcal{M}, w \models \phi$ and $\mathcal{M}, w \models \psi$; and
- $\mathcal{M}, w \models \neg\phi$ if and only if $\mathcal{M}, w \not\models \phi$.

We use $U^m \models \phi$ to denote the fact that ϕ is *valid*, i.e., that $(\mathcal{M}, w) \models \phi$ for all $(\mathcal{M}, w) \in \mathbf{Mod}^p(U^m)$. ϕ is *satisfiable* if there exists some $(\mathcal{M}, w) \in \mathbf{Mod}^p(U^m)$ such that $(\mathcal{M}, w) \models \phi$.

Definition 2.5 ($\mathbf{S5}^m$)**.** $\mathbf{S5}^m \overset{def}{=} \{\phi \in \mathcal{L}_m \mid U^m \models \phi\}$. That is, $\mathbf{S5}^m$ is the set of all $\mathcal{L}_m$-formulae valid in all m-product frames $\mathcal{F}_1 \times \cdots \times \mathcal{F}_m$ where for each $\mathcal{F}_i$, the accessibility relation R_i is the universal relation ($R_i = W_i \times W_i$). ⊣

Note that $\mathbf{S5}^m$ is usually taken to be the logic of products of *arbitrary* equivalence relations, not universal relations. However, it turns out (Gabbay et al. 2003, Proposition 3.12) that the set of validities ($\mathbf{S5}^m$) is in fact the the same for products of frames with arbitrary equivalence relations and with universal relations.

The *satisfiability problem* for a modal logic is the problem of deciding whether or not a given formual is satisfiable, or, equivalently, whether a given formula is valid. A logic is said to be *decidable* if the satisfiability problem is decidable.

The following theorem will be of importance later.

Theorem 2.6. $\mathbf{S5}^m$ *is undecidable for* $m \geq 3$.

This theorem was originally proved in an algebraic setting (Maddux 1980). For a more direct proof, see (Gabbay et al. 2003, p 381).

3 A Logic of Social Choice Functions

In this section we introduce a modal logic for reasoning about social choice functions. In particular, the logical language is interpreted directly in the context of the combination of an SCF and two preference profiles, intuitively representing *actual* and the *claimed* preferences, respectively, which, as we argued in the introduction, is the level of which much reasoning about social choice functions takes place (e.g., in proofs of the Gibbard-Satterthwaite theorem). We show that the logic can express all the properties involved in the Gibbard-Satterthwaite theorem.

Assume a number of agents n. The language $\mathcal{SCL}_n$ is defined by the following grammar, where $1 \leq i \leq n$.

$$\phi ::= pa_i \mid pc_i \mid \neg\phi \mid \phi \wedge \phi \mid \Box_i\phi \mid \Box^U\phi$$

We use the normal abbreviations $\vee$, $\rightarrow$, $\leftrightarrow$, as well as $\Diamond_i$ and $\Diamond^U$ for $\neg\Box_i\neg$ and $\neg\Box^U\neg$ respectively. As mentioned above, formulae of this language are interpreted as statements about the combination of an SCF F, an actual preference profile D and a claimed preference profile P. Informally, the atom[4] pa_i stands for "i prefers the actual profile", or in more detail: agent i actually (i.e., according to her actual preferences) weakly prefers the outcome of the social choice function on the actual preference profile over the outcome on the claimed preference profile. Similarly, pc_i stands for "i prefers the claimed profile", i.e., that agent i actually weakly prefers the outcome of the social choice function on the claimed preference profile over the outcome on the actual preference profile (mnemonic: read pa as "prefers-actual" and pc as "prefers-claimed"). A formula of the form $\Box_i\phi$ is intended to mean that ϕ holds no matter what agent i claims that her preferences are, while $\Box^U\phi$ holds whenever all agents are honest, i.e., claimed preferences are the same as the actual preferences. Examples of formulae follow, but first we formalise the semantics that we just described.

The language is parameterised by the number of agents n, as is the class of social choice functions we will use to interpret it. We will interpret formulae φ of the language in the context of an SCF $F \in \mathbf{SCF}(n)$ and two preference profiles D and P over the same set of alternatives as F. The fact that φ is true in the context of F, D and P is denoted $F,(D,P) \models \varphi$.

Definition 3.1. Let n be a given number of agents. Let $F \in \mathbf{SCF}(A,n)$ for some countable set A, and $D,P \in L(A)^n$. Whether $F,(D,P) \models \varphi$ holds, where $\varphi \in \mathcal{SCL}_n$, is defined inductively as follows.

- $F,(D,P) \models pa_i$ if and only if $F(P)D_iF(D)$;
- $F,(D,P) \models pc_i$ if and only if $F(D)D_iF(P)$;
- $F,(D,P) \models \Box_i\phi$ if and only if $F,(D,P') \models \phi$ for all $P' \in L(A)^n$ such that $P' \sim_i P$;
- $F,(D,P) \models \Box^U\phi$ if and only if $F,(D',D') \models \phi$ for all $D' \in L(A)^n$;
- $F,(D,P) \models \phi \wedge \psi$ if and only if $F,(D,P) \models \phi$ and $F,(D,P) \models \psi$; and
- $F,(D,P) \models \neg\phi$ if and only if $F,(D,P) \not\models \phi$.

$\dashv$

The formula $\Box_i\phi$ is read as saying that for every way agent i can change his part of the claimed preferences (P'), the formula ϕ will still hold. It is also worth noticing that the modality $\Box^U$ only ranges over pairs of profiles (D',D') where both components are equal. That is, it ranges over every possibility where everyone is honest about their preferences.

We lift this definition in the usual way: $F \models \phi$ if and only if $F,(D,P) \models \phi$ for all $(D,P) \in L(A)^n \times L(A)^n$, and $\models \phi$ if and only if $F \models \phi$ for all $F \in \mathbf{SCF}(n)$. Finally we define the "voting logic" (for n agents) as the validities on $\mathbf{SCF}(n)$:

$$\mathbf{VL}_n \overset{def}{=} \{\phi \in \mathcal{SCL}_n \mid \models \phi\}.$$

Here are some examples of formulae and their meanings.

$$\neg pc_1 \tag{1}$$

[4] Or nullary modality.

Formula (1) says that it is not the case that agent 1 actually weakly prefers the outcome of the claimed preferences over the outcome of the actual preferences, i.e., that he actually *strictly* prefers the outcome of the actual preferences over the outcome of the claimed preferences.

$$\Box_1 pa_1 \tag{2}$$

Formula (2) says that agent 1 prefers the outcome of the actual preferences more than or equal to all alternatives that can win if he lies.

$$\Box_1 pa_1 \wedge \Diamond_1 \Diamond_2 (pc_1 \wedge \neg pa_1) \tag{3}$$

Formula (3) says that agent 1 cannot lie about his preferences to get some alternative he prefers more, but if agent 1 *and* agent 2 lie, they can force a winning alternative such that agent 1 strictly prefers it to his honest winning alternative.

$$\Diamond^U \Diamond_1 \neg pa_1 \tag{4}$$

Formula (4) says that there exists a preference profile such that agent 1 is better off by (possibly) lying as long as everyone else is honest.

3.1 The Properties

We will now see that the language $\mathcal{SCL}_n$ can express certain interesting properties of social choice functions; properties involved in the Gibbard-Satterthwaite theorem.

Let:

$$\textit{i-dict} \overset{def}{=} \Box^U \Box_1 \cdots \Box_n pa_i.$$

In the formula *i-dict*, the modality $\Box^U$ universally quantifies over all possible preference profiles D, while the effect of the combination $\Box_1 \cdots \Box_n$ is to universally quantify over all preference relations P_i for each agent, i.e., all preference profiles. Thus, the expression can be read "for every preference profile D, there is no other possibly winning alternative $F(P)$ for such that i prefers $F(P)$ to $F(D)$ in D", which means exactly that F is i-dictatorial as defined in Section 2.

Let:

$$\textit{SP} \overset{def}{=} \Box^U (\bigwedge_{i \leq n} \Box_i pa_i).$$

SP says that, for all possible preference profiles, for any agent, for any other claim that agent can make about his own preferences, the agent weakly prefers the outcome of the original preference profile. Said in another way, for all possible preference profiles, no single agent is better off by claiming he prefers something he does not.

Lemma 3.2. *For any A, $F \in$ **SCF**(A, n) and $D, P \in L(A)^n$:*

1. *$F, (D, P) \models$ i-dict iff F is i-dictatorial;*
2. *$F, (D, P) \models$ SP iff F is strategy-proof.*

We now have two of the three properties involved in the Gibbard-Satterthwaite theorem. Let us move on to the third, namely 3-winning. Consider the following formula:

$$\textit{2p3a} \overset{def}{=} \Diamond^U ((\Diamond_1 \cdots \Diamond_n \neg pa_1 \wedge \Diamond_1 \cdots \Diamond_n \neg pc_1) \vee (\Diamond_1 \cdots \Diamond_n \neg pa_2 \wedge \Diamond_1 \cdots \Diamond_n \neg pc_2))$$

Note that in writing down this formula we assumed that there are at least two agents ($n \geq 2$). This is an implicit assumption anywhere it is used below.

Lemma 3.3. *For any A, $F \in$ **SCF**(A, n) and $D, P \in L(A)^n$, if $F, (D, P) \models$ 2p3a then F is 3-winning.*

Proof. It is easy to see that $F, (D, P) \models$ *2p3a* iff there exist D', P', P'' such that $F(P'') <^{D'_i} F(D')$ and $F(D') <^{D'_i} F(P')$ with $i = 1$ or $i = 2$. That means that $F(D')$, $F(P')$ and $F(P'')$ must all be different. □

While this formula always implies at least three possibly-winning alternatives, it is in fact not valid on all 3-winning SCFs.[5] However, as we show next, it is valid on all 3-winning SCFs that *also are strategy-proof.* It follows that the antecedent of the implication that forms the Gibbard-Satterthwaite theorem, namely "3-winning and strategy-proof", is expressed by the formula *SP* $\wedge$ *2p3a*.

Lemma 3.4. *For any A and $F \in$ **SCF**(A, n), if F is 3-winning and strategy-proof then $F \models$ 2p3a.*

Proof. Let F be 3-winning and strategy-proof. We show that $F \models$ *2p3a*, which (see the proof of Lemma 3.3) holds iff there exist D, P, P' such that

$$F(P') <^{D_i} F(D) \text{ and } F(D) <^{D_i} F(P) \tag{5}$$

with $i = 1$ or $i = 2$.

We say that an alternative $x \in A$ is *possibly-winning* if there is a D such that $F(D) = x$. Let a, b be two different possibly-winning alternatives, and let D^a, D^b be such that $F(D^a) = a$ and $F(D^b) = b$. Let D' be a preference profile such that agent 1's most preferred possibly-winning alternative is a (i.e., she prefers a over all other possibly-winning alternatives), and agent 1's second most preferred possibly-winning alternative is b, and agent 2's most preferred and second most preferred possibly-winning alternatives are b and a, respectively. Let $x = F(D')$. We reason by two cases.

First, consider the case that either agent 1 or agent 2 ranks some other possibly-winning alternative Y *strictly lower* than X in D'; wlog. assume that this holds for agent 1. Consider the following two sub-cases. First, consider the case that X is either a or b. If $x = a$, then (5) holds with $D = D'$, $i = 2$, $P = D^b$, and $F(P')$ is a third possibly-winning alternative (which must exist since F is 3-winning). Note that $F(P')$ must be ranked below a by agent 2, by construction of D'. If $x = b$, then (5) holds for agent 1 by a similar argument. Second, consider the case that X is different from a and b. Then (5) holds for $i = 1$, $D = D'$, $P = D^a$ and P' is such that $F(P') = Y$.

Second, consider the case that X is the lowest-ranked possibly-winning alternative for both 1 and 2 in D'. Let the preference profile D'' be exactly like D', except that b and X change places in the linear order for agent 1. It must be the case that $F(D'') = X$: otherwise, if D'_i represents agent 1's actual preferences he was able to improve his outcome from $F(D')$ by claiming that his preferences was actually D''_i, which contradicts the fact that F is strategy-proof (this can be seen more directly by taking $P' = D''$ and $P = D'$ in the definition of strategy-proofness given in Section 2). Thus, $F(D'') = X$. But then (5) holds with $i = 1$, $D = D''$, $P = D^a$ and $P' = D^b$. □

[5] We leave a counterexample as an exercise to the reader!

Corollary 3.5. *For any A, $F \in$ **SCF**(A, n) and $D, P \in L(A)^n$, $F, (D, P) \models$ SP $\wedge$ 2p3a iff F is strategy-proof and* 3*-winning.*

Thus, we can express the condition "strategy-proof and 3-winning" in the statement of the Gibbard-Satterthwaite theorem.

Combined, the formulae above give us the ability to express the Gibbard-Satterthwaite theorem itself, for $n \geq 2$. Let:

$$GS \overset{def}{=} \textit{SP} \wedge \textit{2p3a} \rightarrow \bigvee_{i \leq n} \textit{i-dict}$$

The Gibbard-Satterthwaite theorem is exactly that GS is valid.

Note that these formulae are independent of the set of alternatives A; they express the corresponding properties no matter what A is.

4 Undecidability

We now show that modal logics expressing properties of actual and claimed preferences in the context of a social choice function very quickly become undecidable. In particular, we show that any modal logic that can express that an agent actually prefers the outcome of the actual preferences over the outcome of the claimed preferences and that has a modality quantifying over possible claimed preferences (properties (1) and (2) from the introduction, and used in the formalisation in the previous section), is undecidable. As discussed in the introduction, this is a level of reasoning that is common in social choice. Technically, we do this by identifying an undecidable fragment of the logical language discussed in the previous section.

The language is, informally, $\mathcal{SCL}_n$, without the $\Box^U$ modality and the propositional letters pc_i. Formally, the language $\mathcal{SCL}_n^-$ is defined as follows, where $1 \leq i \leq n$.

$$\phi ::= pa_i \mid \neg\phi \mid \phi \wedge \phi \mid \Box_i \phi$$

A formula in $\mathcal{SCL}_n^-$ is evaluated in a situation SCF $F, (D, P)$ in the same way as in Definition 3.1. We will by $\mathbf{VL}_n^-$ denote all formulae of $\mathcal{SCL}_n^-$ true on every $F \in \mathbf{SCF}(n)$, that is, $\mathbf{VL}_n^- \overset{def}{=} \{\phi \in \mathcal{SCL}_n^- \mid \models \phi\}$.

We now proceed to show that $\mathbf{VL}_n^-$ is undecidable, by translating $\mathbf{S5}^m$ to $\mathbf{VL}_n^-$. First, we give a formula translation t_Φ (relative to some finite set of propositional letters Φ) from $\mathcal{L}_m$ to $\mathcal{SCL}_n^-$, and then a model translation which preserves satisfiability through the formula translation, enabling us to show that for any $\phi \in \mathcal{L}_m$, we have $\phi \in \mathbf{S5}^m$ if and only if $t_{\Phi(\phi)}\phi \in \mathbf{VL}_n^-$ (where $\Phi(\phi)$ are the propositional letters occurring in ϕ).

A major difference between the two languages is that $\mathcal{L}_m$ is associated with a countably infinite set of propositional letters, while in $\mathcal{SCL}_n^-$ there are only n "propositional letters" available. Still, every $\mathcal{L}_m$-formula contains only a finite number of propositional letters, and thus there is some $n \in \mathbb{N}$ such that $\mathcal{SCL}_n^-$ has enough propositional letters available. This is the reason for the dimensional change from m to n in the translation below.

4.1 Formula translation

Given an $\mathcal{L}_m$-formula ϕ, for some m, let $\Phi(\phi)$ denote the (finite) set of propositional letters occurring in ϕ. As is clear from the discussion above, we will use at least as many agents as $|\Phi(\phi)|$; more precisely we will translate ϕ to a formula $t_{\Phi(\phi)}(\phi) \in \mathcal{SCL}_n^-$ with $n = 1 + \max(m, |\Phi(\phi)|)$.

We assume that for any finite set $\Phi \subseteq \mathsf{P}$ there is an *enumeration* $h_\Phi \colon \Phi \to \mathbb{N}$ of the propositional letters in Φ.

Definition 4.1. Let $\phi \in \mathcal{L}_m$ and let $\Phi \subseteq \mathsf{P}$ be a finite set of propositional letters such that $\Phi(\phi) \subseteq \Phi$. The translation $t_\Phi(\phi) \in \mathcal{SCL}_n^-$ where $n = 1 + \max(m, |\Phi|)$ is defined recursively over subformulae of ϕ as follows:

$$t_\Phi(p) = pa_{h_\Phi(p)} \qquad t_\Phi(\Box_i \phi) = \Box_i t_\Phi(\phi)$$
$$t_\Phi(\phi \wedge \psi) = t_\Phi(\phi) \wedge t_\Phi(\psi) \qquad t_\Phi(\neg\phi) = \neg t_\Phi(\phi)$$

$\dashv$

Essentially, the translation maps propositional letters p to the "prefers-actual" symbol for agent $h_{\Phi(\phi)}(p)$, and leaves the rest unchanged. Note that a given ϕ might be translated into $\mathcal{SCL}_n^-$ with $n > m$, but the resulting formula will only contain *modalities* for the first m agents.

4.2 Model translation

The next step is to translate models of $\mathbf{S5}^m$ into social choice functions with actual and claimed preferences such that satisfiability of formulae over the $t_{\Phi(\phi)}$-translation is preserved.

Reflecting our remarks on the restricted set of propositional letters above, we will translate an $\mathbf{S5}^m$-model modulo some finite set of propositional letters. Our goal is that all formulae using only those propositional letters will have satisfiability preserved.

For the translation we need to choose an appropriate set A of alternatives, find some point (D, P), and finally construct an SCF F.

The essential insight is that we are free in choosing A, and we can choose A to have as many alternatives as there are states in the model satisfying ϕ. We will then make a profile D which will be the first coordinate of the point (D, P). In D we will place the alternatives depending on the associated state in the satisfying $\mathbf{S5}^m$-model. Letting $\wp_{fin}(\mathsf{P})$ denote the finite subsets of the set P, we define the model translation as follows.

Definition 4.2. Given a pointed $S5^m$-model $(\mathcal{M}, w)$ with frame $\mathcal{F} = (W, R_1, \ldots, R_m)$ and valuation function $V \colon \mathsf{P} \to \wp(W)$, and some *finite* set Φ of propositional letters, we let $n = 1 + \max\{m, |\Phi|\}$ and $A = W \cup \{l\}$ where l is a new element not in W. We now define the tuple

$$\theta((\mathcal{M}, w), \Phi) = (F, (D, P))$$

consisting of an SCF F and profiles D and P over A and n, induced by $(\mathcal{M}, w)$ and Φ.

Recall that we assume that $\mathcal{F}$ is a universal product frame, so there are $W_1, \ldots, W_m$ such that each point $w \in W$ is of the form $(w_1, \ldots, w_m)$, where $w_i \in W_i$, and the relations R_i are universal on W_i.

We now define the profile D. We will then define F such that $F(D) = l$, but first we place the other alternatives in A relative to l in D_i according to the valuation of the propositional letters in $\mathcal{M}$. Recall that in the formula translation we associated each propositional letter with an agent. For each alternative $a \in A \setminus \{l\}$ and for each $i \leq |\Phi|$ we place a below l in D_i if and only if $a \in V(h_\Phi^{-1}(i))$, and above otherwise:

for all $a \in A \setminus \{l\}$ let aD_il if and only if $a \in V(h_\Phi^{-1}(i))$.

Note that $h_\Phi \colon \Phi \to \mathbb{N}$ is the assumed enumeration of the propositional letters Φ, and that $h_\Phi^{-1}(i) \in \Phi$. As $A \setminus \{l\}$ is exactly the domain W, the test $a \in V(h_\Phi^{-1}(i))$ is well defined.

This defines two "bags" for every D_i, one for alternatives above l, and one for those below, for every $i \leq |\Phi|$. This allows for many different profiles $D \in L(A)^n$, we choose one and fix it as D for the remainder of the construction.

Fix a family $(s_i)_{i \leq m}$ of surjective mappings $s_i \colon L(A) \to W_i$. These mappings exist, as $L(A)$ is exponentially larger than W, which itself is larger than each W_i. Given $(s_i)_{i \leq m}$, define $s \colon L(A)^n \to W$ in the following way:

$$s(P) = (s_1(P_1), \dots, s_m(P_m)).$$

Note that the domain of s is $L(A)^n$, but its definition uses only the first m linear orders. This is natural, since W contains m-tuples. As $n > m$ there will be several distinct $P, P' \in L(A)^n$ such that $P \neq P'$, but $s(P) = s(P')$.

Pick a $P \in L(A)^n$ such that $D_k \neq P_k$ for some $k > m$, while satisfying $s(P) = w$, where w is the distinguished point of the pointed $\mathbf{S5}^m$-model $\mathcal{M}$. This is possible, since each s_i is onto W_i, and as s is indifferent about any differences in the coordinates above m.

Note that since D and P differ in some coordinates, the SCF F does not need to assign to them the same winning alternative, and this will hold for all profiles P' accessible from P by changing only the first m coordinates.

Define F by letting $F(D) = l$, and for all P' such that $P_j = P'_j$ for $j > m$ letting $F(P') = s(P')$. By construction of P we know that for all of these P' we have $P' \neq D$. Also note that as $W \subset A$, $s(P')$ is an alternative. Observe that s is surjective, and that for every element $a \in W \setminus \{l\}$ there is a P' such that $P_j = P'_j$ for $j > m$ and $s(P') = a$, resulting in F being onto A. F can assign arbitrary alternatives to other profiles, as they are not reachable by any translated $\mathbf{S5}^m$-formula.

This concludes the model translation. ⊣

To summarize, we used the fact that we are free to choose the set A of alternatives, and then we associated every point in the $\mathbf{S5}^m$-model with an alternative. We then mapped profiles to points, such that the profiles had the associated alternative as its winning alternative.

The following lemma describes the interaction between the formula translation t_Φ and the model translation θ.

Lemma 4.3. *For all* $(M, w) \in \mathbf{Mod}^p(U^m)$ *and* $\phi \in \mathcal{L}_m$:

$$M, w \models \phi \text{ if and only if } \theta((M, w), \Phi(\phi)) \models t_{\Phi(\phi)}(\phi).$$

Proof. Let $(M, w) \in \mathbf{Mod}^p(U^m)$ and $\phi \in \mathcal{L}_m$ for some m, and let n, A, and s be as in Definition 4.2. From the function s we define its preimage $s^\vee \colon W \to \wp(L(A)^n)$,

$$s^\vee(q) = \{P \in L(A)^n \mid s(P) = q\},$$

as all the profiles mapping to some state q.

Let

$$(F, (D, P')) = \theta((M, w), \Phi(\phi)).$$

From P' we construct the set Ξ of profiles agreeing on P' on all "innaccessible" dimensions j:

$$\Xi = \{P \in L(A)^n \mid P_j = P'_j \text{ for all } j > m\}.$$

Henceforth let $\Phi = \Phi(\phi)$; the set of propositional letters in the input formula ϕ.

To get the lemma we prove a stronger statement, namely that for all $q \in W$ and $P \in s^{\vee}(q) \cap \Xi$ we have $\mathcal{M}, q \models \phi$ if and only if $(F, (D, P)) \models t_{\Phi}(\phi)$. As $P' \in s^{\vee}(w)$ by construction of θ, and $P' \in \Xi$ by definition of Ξ, this is sufficient.

We prove this by induction on the complexity of the formula ϕ (we leave out the trivial cases of negation and conjunction here). A subtlety worth menioning is that for subformulae ψ of ϕ we are not proving the result for $t_{\Phi(\psi)}(\psi)$, but rather for $t_{\Phi}(\psi)$. This is natural as the translated model is constructed relative to $\Phi(\phi)$.

Base case, $\phi = p$. Since $t_{\Phi}(p) = pa_{h_{\Phi}(p)}$, we need to show that for arbitrary $q \in W$ and $P \in s^{\vee}(q) \cap \Xi$, we have $(\mathcal{M}, q) \models p$ if and only if $(F, (D, P)) \models pa_{h_{\Phi}(p)}$. Let $i = h_{\Phi}(p)$. Observe that as $P \in s^{\vee}(q) \cap \Xi$, we have that $F(P) = q$.

We have $(\mathcal{M}, q) \models p$ if and only if $q \in V(p)$ if and only if $q \in V(h_{\Phi}^{-1}(i))$ if and only if qD_il, the latter coming from the construction of θ. This is equivalent to $F(P)D_iF(D)$, which holds if and only if $(F, (D, P)) \models pa_i$.

Inductive case, $\phi = \Box_i\psi$. We are assuming the induction hypothesis that for any $q' \in W$ and $P' \in s^{\vee}(q') \cap \Xi$ we have $\mathcal{M}, q' \models \psi$ if and only if $F, (D, P') \models t_{\Phi}(\psi)$. Note that ψ is translated using t_{Φ}, not $t_{\Phi(\psi)}$. Let $q \in W$ and $P \in s^{\vee}(q) \cap \Xi$.

For the left to right direction we assume $\mathcal{M}, q \models \Box_i\psi$, meaning that we have $\mathcal{M}, q' \models \psi$ for all qR_iq'. Since R_i is universal, we know that it holds for *all* q' such that $q \sim_i q'$, that is, all q' possibly differing from q only in their i'th coordinate.

Now assume an arbitrary P' such that $P \sim_i P'$, possibly differing from P in only the i'th coordinate. Note that since $i \leq m$ we have $P' \in \Xi$. By the construction of s, any such P' has the property that $s(P') \sim_i q$.

By assumption we then have that $\mathcal{M}, s(P') \models \psi$ and hence, by the induction hypothesis, we get $(F, (D, P')) \models \psi^{t_{\Phi}}$. Since P' was an arbitrary profile such that $P \sim_i P'$, we obtain $(F, (D, P)) \models \Box_i\psi^{t_{\Phi}}$, which is the same as $(F, (D, P)) \models (\Box_i\psi)^{t_{\Phi}}$.

For the other direction, note that for any q' with $q \sim_i q'$, there exists some P' with $P \sim_i P'$ such that $s(P') = q'$. This follows from the surjectivity of each s_i, and by the coordinate-wise construction of s. By the induction hypothesis, we are done. □

To summarize, we have that if some $\phi \in \mathcal{L}_m$ is satisfiable then $t_{\Phi(\phi)}(\phi) \in \mathcal{SCL}_n^-$ is satisfiable. Contrapositively, we get that if $t_{\Phi(\phi)}\phi \in \mathbf{VL}_n^-$ then $\phi \in \mathbf{S5}^m$. We proceed to show that the converse holds as well. We assume that we have a model satisfying the translated formula, and we will show that we can construct a model satisfying the original formula.

Lemma 4.4. *For any $\phi \in \mathcal{L}_m$, if there is an $(F, (D, P)) \in$ **SCF**$(n) \times L(A)^n \times L(A)^n$ with $n = 1 + \max\{m, |\Phi(\phi)|\}$ such that $(F, (D, P)) \models t_{\Phi(\phi)}(\phi)$ then there is an $(M, w) \in \mathbf{Mod}^p(U^m)$ such that $(M, w) \models \phi$.*

Proof. We start with the SCF F such that $(F, (D, P)) \models t_{\Phi(\phi)}(\phi)$. From this we will extract a model in $\mathbf{Mod}^p(U^m)$.

Given the SCF $F\colon L(A)^n \to A$ and a pair of profiles (D, P) we extract a model in $\mathbf{Mod}(U^m)$ by letting

$$\begin{aligned} W =&\{(D', P') \in L(A)^n \times L(A)^n \mid D = D' \text{ and } P'_j = P_j \text{ for } j > m\} \text{ and} \\ R_i =&\{((D, P'), (D', P'')) \mid D = D' \text{ and } P'' \sim_i P'\} \text{ for } i \leq m. \end{aligned}$$

The valuation function is generated from F by defining

$$(D, P) \in V(p) \text{ if and only if } F(P) D_{h_{\Phi(\phi)}(p)} F(D).$$

The model generated above will be denoted as $(\mathcal{M}, (D, P))$, and is clearly an $R_1 \dots R_m$ model in $\mathbf{Mod}^p(U^m)$.

By structural induction on the complexity of the formula we can show that for all $(D, P) \in W$, $(F, (D, P)) \models t_{\Phi(\phi)}(\phi)$ if and only if $(\mathcal{M}, (D, P)) \models \phi$. Details can be found in the appendix. □

Lemma 4.4 gives us that for $\phi \in \mathcal{L}_m$ we have that if $\phi \in \mathbf{S5}^m$ then $t_{\Phi(\phi)}(\phi) \in \mathbf{VL}_n^-$. Combining this with Lemma 4.3 we get that for all $\phi \in \mathcal{L}_m$ there is an $(\mathcal{M}, w) \in \mathbf{Mod}^p(U^m)$ such that $(\mathcal{M}, w) \models \phi$ if and only if there is an $(F, (D, P)) \in \mathbf{SCF}(n) \times L(A)^n \times L(A)^n$ such that $(F, (D, P)) \models t_{\Phi(\phi)}(\phi)$, where $n = 1 + \max\{m, |\Phi(\phi)|\}$.

Theorem 4.5 (Equisatisfiability)**.** *Let $\phi \in \mathcal{L}_m$ for some m and let $n = 1 + \max\{m, |\Phi(\phi)|\}$. We have that $\phi \in \mathbf{S5}^m$ if and only if $t_{\Phi(\phi)}(\phi) \in \mathbf{VL}_n^-$.*

We immediately obtain the following result.

Theorem 4.6 (Undecidability)**.** *For some n, $\mathbf{VL}_n^-$ is undecidable.*

Suppose otherwise, i.e., that $\mathbf{VL}_n^-$ us decidable for all n, and let $\phi \in \mathcal{L}_3$. From Theorem 4.5 we have that $\phi \in \mathbf{S5}^3$ if and only if $t_{\Phi(\phi)}(\phi) \in \mathbf{VL}_n^-$ with $n = 1 + \max\{m, |\Phi(\phi)|\}$. In order to answer the question "is $\phi \in \mathbf{S5}^3$?", we can use the decision procedure for $\mathbf{VL}_n^-$ to check whether $t_{\Phi(\phi)}(\phi) \in \mathbf{VL}_n^-$. But that contradicts the fact that $\mathbf{S5}^3$ is undecidable (Theorem 2.6).

Unfortunately we can not say at which number of agents the logic $\mathbf{VL}_n^-$ becomes undecidable. This depends on the smallest number of propositional letters one can restrict $\mathbf{S5}^3$ to use for it to still be undecidable.

5 Discussion

Reasoning about social choice, for example in the proofs of key results found in the literature, typically occurs on the level of reasoning about a given actual preference profile compared to a given claimed preference profile, in the context of a social choice function. In this paper we discussed modal logics with formulae ϕ expressing properties of such situations: $F, (D, P) \models \phi$. We presented one such logic, and showed how it can be used to characterize the Gibbard-Satterthwaite theorem.

The trade-off between expressivity and computability is of interest, and the key point we are making in this paper is that very little can be said about such situations before the logic becomes undecidable. In particular, we showed that a minimal logic with a modality ranging over possible claimed preference relations of a single agent, as well as an atom expressing that the agent prefers the outcome on the claimed preferences over the outcome on the actual preferences, becomes undecidable if there are enough agents. It follows that any language with such a modality, which can express the property expressed by the mentioned atom, is undecidable for some number of agents.

Of course, this does not mean that "reasoning about social choice is undecidable"; it is possible that the logic can be made decidable for instance by restricting the nesting of modalities. However, then the language would, strictly speaking, no longer be a standard modal language, and we argue that the undecidability result is quite fundamental for modal logics of social choice. It follows from relatively weak assumptions about the ability to discern between different situations.

We conjecture that the logic presented in Section 3 is decidable in 2-agent case, and that it is finitely axiomatizable. We also conjecture that it is not finitely axiomatizable for $n > 2$, similarly to the case for $\mathbf{S5}^n$. We leave these conjectures for future work.

When it comes to related work, (Troquard et al. 2011) also proposes a logic interpreted in social choice functions. The logic can express properties of social choice functions such as strategy-proofness, and it is decidable. There are two fundamental differences between the modalities and logics studied in (Troquard et al. 2011) and in the current paper, however. In the former, first, both the class of models and the language is parameterized by the set of alternatives A, and, second, the set of alternatives is required to be finite. In particular, the expression for strategy-proofness is *different* for different sets of alternatives, and the property cannot be expressed if the set of alternatives is infinite. In contrast, in the current paper we have focused on combinations of modalities that allows us to express properties of social choice functions in the strong sense that the same logical formula expresses the same property no matter what the set of alternatives is. Compare to informal statements of these properties found in the literature: they are not stated by explicitly referring to each of the alternatives and iterating over them (as in in (Troquard et al. 2011)); they are general expressions that describe the properties no matter what the set of alternatieves are (as we do). The fact that the set of models under consideration is parameterized by a fixed finite set of alternatives in (Troquard et al. 2011) makes the logic trivially decidable of course. The notion of validity, or satisfiability, used in the current paper is stronger; it considers the class of all social choice functions rather than only those over a fixed finite set of alternatives. A similar logic for social *welfare* functions with similar differences to the logic in the current paper has already been proposed (Ågotnes et al. 2011). A first-order logic version of expressing properties of social welfare functions has also been discussed (Grandi and Endriss 2013).

References

Ågotnes, T., van der Hoek, W., and Wooldridge, M. (2006). Towards a logic of social welfare. In Bonanno, G., van der Hoek, W., and Wooldridge, M., editors, *Proceedings of The 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT)*, pages 1–10.

Ågotnes, T., van der Hoek, W., and Wooldridge, M. (2011). On the logic of preference and judgment aggregation. *Autonomous Agents and Multi-Agent Systems*, 22:4–30.

Arrow, K. J. (1951). *Social Choice and Individual Values*. Wiley.

Arrow, K. J., Sen, A. K., and Suzumura, K., editors (2002). *Handbook of Social Choice and Welfare Volume 1*. Elsevier Science Publishers B.V.: Amsterdam, The Netherlands.

Barbera, S. (2001). An introduction to strategy-proof social choice functions. *Social Choice and Welfare*, 18(4):619–653.

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1995). *Reasoning About Knowledge*. The MIT Press, Cambridge, Massachusetts.

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

Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601.

Grandi, U. and Endriss, U. (2013). First-order logic formalisation of impossibility theorems in preference aggregation. *Journal of Philosophical Logic*, 42(4):595–618.

Maddux, R. (1980). The equational theory of CA_3 is undecidable. *The Journal of Symbolic Logic*, 45(2):311–316.

Pauly, M. (2008). On the role of language in social choice theory. *Synthese*, 163:227–243.

Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217.

Troquard, N., van der Hoek, W., and Wooldridge, M. (2009). A logic of propositional control for truthful implementations. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK '09, pages 237–246, New York, NY, USA. ACM.

Troquard, N., van der Hoek, W., and Wooldridge, M. (2011). Reasoning about social choice functions. *Journal of Philosophical Logic*, 40:473–498.

Justified True Belief Revisited in Topological Argumentation Model

Chenwei Shi

Tsinghua University

Since Gettier's paper (Gettier 1963) challenges the definition of knowledge as justified true belief (JTB), different remedies of the JTB theory of knowledge have been proposed in epistemology. On the other hand, the challenge by skepticism also prompts epistemologists to reconsider what constitutes knowledge. In this talk I will make use of the topological argumentation model (Shi et al. 2017) to formalise three different interpretations of the notion of justified true belief. Moreover, we show that the last one of the three interpretations can serve as a definition of knowledge by demonstrating that it can resolve the problem of skepticism.

References

Gettier, E. L. (1963). Is Justified True Belief Knowledge. *Analysis*, 23(6):121–123.

Shi, C., Smets, S., and Velázquez-Quesada, F. R. (2017). Argument-based Belief in Topological Structures. *In Proceedings of TARK 2017*, EPTCS 251: 489-503.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

The Logic of Collective Acceptance

Frederik Van De Putte

University of Bayreuth / Ghent University

Abstract

We present a logic that features three types of modalities: classical modal operators $\Box_i$ that express the norms endorsed by a given agent i; normal modalities $\Box^s_G$ that express what is required for the group G in view of the norms that are commonly endorsed by all its members; and normal modalities $\Box^u_G$, that express what is required for an option to be universally acceptable within G. Although our semantics is relatively simple, it poses some challenges where axiomatization is concerned. We argue that this logic is useful as a stepping stone towards richer semantics that can express various types of collective acceptance and social norms.

1 Introduction

In criticizing Philip Pettit's republican philosophy of popular control (Pettit 2012), Sean Ingham (2015) introduces a distinction between two senses in which a given policy can be accepted by a group. According to Ingham's argument, Pettit confuses these two senses, using the stronger one in formulating general desiderata for popular control, but the other, weaker one in evaluating and defending a specific model of democratic decision-making. Our aim in this presentation will not be to question Ingham's attack itself, but rather to take a closer look at his distinction from a formal logic point of view, connecting it to known work on the logic of shared and distributed (doxastic or epistemic) propositional attitudes.

2 Ingham's Distinction

Ingham explains his distinction using a simple set-theoretic model. In line with social choice theory, he starts from a set X of potential policies, from which the government is supposed to

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

select one policy within the boundaries set by popular control. Such popular control should in turn be a function of the norms endorsed by the citizens. Ingham models those norms in terms of the alternatives they rule out. That an agent i endorses a basic norm $Y \subseteq X$ means, hence, that i does not accept any alternative outside Y. An alternative is then acceptable relative to $\mathscr{N}_i \subseteq \wp(X)$, the set of norms endorsed by i, if and only if the alternative is permitted by every member of $\mathscr{N}_i$.[1]

Even with this relatively simple representation of norms and their relation to acceptance held fixed, there are at least two non-equivalent ways to specify the notion of group acceptance. On the first, an alternative is acceptable to a group iff it is acceptable in view of the set of all shared norms, where shared norms are norms endorsed by each agent in the group. Call this *shared norm acceptability*. On the second specification, which Ingham refers to as *universal acceptability*, an alternative is acceptable for the group if and only if it is accepted by each member of the group, in view of their respective individual norms.

Let us illustrate the difference between these two specifications by means of two simple examples. Suppose that there are three alternatives x, y, z to choose from, that Antje's norms are $\{x, y\}$ and $\{x, z\}$, and that Billie has a single norm, viz. $\{x\}$. In this example, there are no shared norms among Antje and Billie; consequently, all alternatives are shared norm-acceptable within the group consisting of Antje and Bob. However, only x is universally acceptable. That is, given Antje's norms, she excludes all alternatives but x.

Consider, as a second example, a case where Antje still endorses the norms $\{x, y\}$ and $\{x, z\}$, but Billie now endorses the norms $\{x, y\}, \{y, z\}$. In this case, the norm $\{x, y\}$ is shared by both, and hence both x and y are shared-norm acceptable. There is no alternative that is universally acceptable: Antje only accepts x, while Billie only accepts y.

As these examples illustrate, both notions are relevant when thinking about the concept of jointly accepted norms and popular control. Sometimes shared norm acceptance is the only feasible requirement – this is so in particular when aggregating all the norms of the group is impossible, as in our second example. In our first example, one may argue – as Ingham does – that shared norm acceptance is too weak, since it leaves open the possibility that the government chooses a policy that no individual citizen accepts (viz. y or z), even though there is an alternative that is acceptable to each.

It may be tempting to link this distinction to other notions in the theory of democratic decision-making, such as Rawls' *overlapping consensus* (Rawls 1993), Sunstein's *incompletely theorized agreements* (Sunstein 1995), and the notion of *meta-agreement* (Dryzek and Niemeyer 2006; Ottonelli and Porello 2015). One should however be careful not read too much into the formalism just presented. After all, a given citizen may reject a certain subset of the alternatives for various reasons of a very different nature, and one cannot simply reduce such reasons to subsets of alternatives. In particular, the mentioned concepts often crucially refer to the general principles and values that are used to argue for or against accepting a certain alternative, which are not represented in Ingham's formalism.

So rather than an adequate representation of these intricate and rich concepts, we will argue that Ingham's distinction and model provide a useful abstraction to think about group acceptance in exact terms, and may serve as a first step towards more fine-grained accounts. As will become, already this simple model poses interesting difficulties and gives rise to various logical questions.

[1] Cf. (Ingham 2015, p. 106): "[i]f an option x is permitted by every policy-making norm that she [= the agent in question] accepts, then I will say that she finds it acceptable; if it violates a policy-making norm she accepts, I will say that she finds it unacceptable."

What exactly is implied by shared norm acceptability, what does universal acceptance entail? How do these various notions behave and interact, when relativized to different coalitions? Under what conditions do they coincide? How much does Ingham's distinction, and the logic of these notions more generally, depend on the logical properties of norms and the relation between norms and acceptability?

3 The Logic

3.1 Formal Language

Let $\mathfrak{P} = \{p_1, p_2, \ldots\}$ a set of propositional variables. The formal language $\mathfrak{L}$ is given by the following BNF:

$$\varphi := p \mid \bot \mid \neg\varphi \mid \varphi \vee \varphi \mid [\forall]\varphi \mid \Box_i\varphi \mid \Box^u_G\varphi \mid \Box^s_G\varphi$$

See Table 1 for the intended reading of the primitive operators.

$[\forall]\varphi$	$\approx$	every alternative has property φ
$\Box_i\varphi$	$\approx$	i endorses the norm φ
$\Box^u_i\varphi$ / $\Box^s_i\varphi$	$\approx$	φ is required for i.
$\Box^u_G\varphi$	$\approx$	φ is universally required for G
$\Box^s_G\varphi$	$\approx$	φ is a shared-norm required for G

Table 1: Intended reading of the modal operators.

3.2 Semantics

Definition 1 *A* model *is a tuple $M = \langle W, \langle \mathcal{N}_i \rangle_{i \in N}, V \rangle$ where $W \neq \emptyset$ is the* domain *of M, for each $i \in N$, $\mathcal{N}_i \subseteq \wp(\wp(W))$ is such that $W \in \mathcal{N}_i$*[2]*, and $V : W \to \wp(\mathfrak{P})$ is a valuation function.*

Definition 2 *Where $M = \langle W, \langle \mathcal{N}_i \rangle_{i \in N}, V \rangle$ and $w \in W$,*

$M, w \models [\forall]\varphi$ iff for all $w' \in W$, $M, w' \models \varphi$

$M, w \models \Box_i\varphi$ iff $||\varphi||^M \in \mathcal{N}_i$

$M, w \models \Box^u_G\varphi$ iff $\bigcap_{i \in G} \bigcap \mathcal{N}_i \subseteq ||\varphi||^M$

$M, w \models \Box^s_G\varphi$ iff $\bigcap \bigcap_{i \in G} \mathcal{N}_i \subseteq ||\varphi||^M$

Suppose that we introduce a propositional variable p_x to denote the specific alternative $x \in W$. Then $\Diamond^u_i p_x$ coincides with Ingham's reading of "x is acceptable for i", on the above semantics. Likewise, $\Diamond^s_G p_x$ expresses that x is shared-norm acceptable for G, and $\Diamond^u_G p_x$ that x is universally acceptable for G.

Theorem 1 *Suppose that for all $i \in G$, $\mathcal{N}_i$ is closed under supersets. Then $M, w \models \Box^s_G\varphi$ iff $M, w \models \bigwedge_{i \in G} \Box^u_i\varphi$.*

[2] This condition simplifies the meta-theory and semantic clauses of our logic significantly. Alternatively, one could leave it out and rewrite the semantic clauses for $\Box^s_G$ and $\Box^u_G$, by intersecting sets $\mathcal{N}_i(w) \cup \{W\}$ rather than sets $\mathcal{N}_i(w)$.

For the class of monotonic models, $\Box^s_G$ and $\Box^u_G$ can be seen as (deontic counterparts of) shared belief, resp. distributed belief, where $\Box^u_i$ is the underlying operator for individual belief. It is well-known that shared belief is a much stronger notion than distributed belief. The main technical novelty of the current formalism, from the viewpoint of existing multi-agent doxastic logics, is the addition of the basic operators $\Box_i$ that allow us to speak about the source (or evidential base) of each agent's beliefs.

CL	all classical tautologies	
K	**K** for $\Box^u_G$ and $\Box^s_G$	
S5	**S5** for $[\forall]$	
(G)	$[\forall](\varphi \leftrightarrow \psi) \to (\Box_i\varphi \leftrightarrow \Box_i\psi)$	
(MP)	if $\vdash \varphi$ and $\vdash \varphi \to \psi$, then $\vdash \psi$	
(NEC)	if $\vdash \varphi$, then $\vdash [\forall]\varphi$	
(I0)	$[\forall]\varphi \to \Box_i\varphi$	$W \in \mathscr{N}_i(w)$
(I1)	$\bigwedge_{i\in G}\Box_i\varphi \to \Box^s_G\varphi$	$R^s_G(w) \subseteq \bigcap\bigcap_{i\in G}\mathscr{N}_i(w)$
(I2)	if $G \subseteq H$, then $\Box^u_G\varphi \to \Box^u_H\varphi$	if $G \subseteq H$, then $R^u_H(w) \subseteq R^u_G(w)$
(I3)	if $G \subseteq H$, then $\Box^s_H\varphi \to \Box^s_G\varphi$	if $G \subseteq H$, then $R^s_G(w) \subseteq R^s_H(w)$
(I4)	$\Box^s_i\varphi \leftrightarrow \Box^u_i\varphi$	$R^u_i(w) = R^s_i(w)$
(U1)	$\Box_i\varphi \to [\forall]\Box_i\varphi$	for all $w, w' \in W$: $\mathscr{N}_i(w) = \mathscr{N}_i(w')$
(U2)	$\Box^u_G\varphi \to [\forall]\Box^u_G\varphi$	for all $w, w' \in W$: $R^u_G(w) = R^u_G(w')$
(U3)	$\Box^s_G\varphi \to [\forall]\Box^s_G\varphi$	for all $w, w' \in W$: $R^s_G(w) = R^s_G(w')$

Table 2: Axioms and rules for a simple logic of collective acceptance, and corresponding frame conditions on quasi-models. Here, i ranges over N and G,H over $\wp(N) \setminus \emptyset$.

3.3 Axiomatization

A sound and complete axiomatization for the class of all models is given in Table 2. Our completeness proof consists of three steps. We first show that our semantics can be rephrased in terms of quasi-models of the type $M = \langle W, \langle \mathscr{N}_i\rangle_{i\in N}, \langle R^u_G\rangle_{G\subseteq_{\emptyset}N}, \langle R^s_G\rangle_{G\subseteq_{\emptyset}N}, V\rangle$, where the $\mathscr{N}_i$ are neighbourhood functions that are used to interpret $\Box_i$, and $\Box^u_G$ and $\Box^s_G$ are interpreted as normal modalities, using the accessibility relations R^u_G, resp. R^s_G. Next, we prove that the axioms from Table 2 characterize specific (well-known) frame conditions, using standard techniques from modal logic. These frame conditions are shown in the second column of Table 2. Finally, we show that every quasi-model M satisfying such frame conditions can be transformed into an equivalent qausi-model $M' = \langle W', \langle \mathscr{N}_i{}'\rangle_{i\in N}, \langle T^u_G\rangle_{G\subseteq_{\emptyset}N}, \langle T^s_G\rangle_{G\subseteq_{\emptyset}N}, V'\rangle$ such that, for all non-empty $G \subseteq N$ and all worlds $w \in W'$:

$$T^u_G(w) = \bigcap_{i\in G}\bigcap\mathscr{N}_i(w)$$

$$T^s_G(w) = \bigcap\bigcap_{i\in G}\mathscr{N}_i(w)$$

This third (and most difficult step) in the proof then allows us to conclude that the transformed model is equivalent to a model in our original semantics. After sketching this third step, we will consider various frame conditions and corresponding axioms, highlighting where our completeness proof generalizes to such conditions. We will pay specific attention to frame conditions that express certain types of agreement among agents, and the impact they have on the two notions of group acceptance and their interrelation.

4 Variations and Extensions

In the final part of the talk, we will consider several enrichments and variations of the simple framework presented. First, one may relativize the space of alternatives X to the world of evaluation w, thus allowing for a more natural interpretation of the semantics in terms of forward-looking acceptance. Technically, this means that $[\forall]$ is weakened to a normal modal operator of type $\mathbf{K}$, quantifying over the set $R(w)$ of alternatives that are (commonly) deemed "feasible" from the viewpoint of world w. One may then either stipulate that norms are subsets of $R(w)$ – hence, they are assumed to be "directly applicable" at w – or use a more primitive notion of "general" norms, and derive "applied" norms from them, as the (non-empty) intersections of primitive norms with $R(w)$. The former option is technically straightforward, whereas the latter is philosophically more plausible.

Second, one may further relativize the space of alternatives to the agents, thus encoding epistemic/doxastic aspects of individual and collective acceptance. This gives rise to additional philosophical questions: should collective acceptance be defined with respect to the set of alternatives that are commonly taken to be feasible, or rather to the set of all alternatives that are feasible to at least one group member? Whereas the first option would call for interaction between collective acceptance and distributed knowledge/belief, the latter would imply interaction between collective acceptance and common knowledge/belief.

Third, one may represent *reasons* for norms explicitly in the object language and semantics, in order to give a more adequate model of the notion of meta-agreement (cf. Section 2). Thus, where e.g. r is a general moral principle, $\mathcal{N}_r(w)$ may represent the application of that principle to the set of feasible alternatives (whether universal or world-relative, cf. supra). An agent i's norms may then naturally be interpreted as obtained by applying all the principles endorsed by i. This induces a third level of (dis)agreement among agents: although they may agree w.r.t. to the applied norms, they may disagree w.r.t. the principles that motivate those norms. An interesting question is how such a richer semantics would relate to the notion of general norm mentioned above, and whether it gives rise to an altogether different logic.

Fourth and last, one may replace Ingham's notion of (individual and collective) acceptance with one that is conflict-tolerant. Conflicts among norms become particularly pressing in view of the other enrichments, since there is no guarantee that applying general norms or even principles to a specific case will result in applied norms that are all jointly satisfiable. Here, one may draw on ideas from Evidence Logic (van Benthem and Pacuit 2011) to characterize more subtle forms of interaction between (possibly conflicting) endorsed norms and collective acceptance.

Whereas each of these refinements pose additional technical challenges, they also promise to lead to a much richer framework within which one can express and investigate the logic of various types of collective acceptance.

Acknowledgements

Frederik Van De Putte is a Marie Skłodowska-Curie Fellow at the Chair of Philosophy I of the University of Bayreuth and a Postdoctoral Fellow of the Research Foundation of Flanders (FWO-Vlaanderen), at the Department of Philosophy and Moral Sciences of Ghent University. Thanks to Olivier Roy and Dominik Klein for discussions on this topic.

References

Chellas, B. (1980). *Modal Logic: an Introduction*. Cambridge university press, Cambridge.

Dryzek, J. and Niemeyer, S. (2006). Reconciling pluralism and consensus as political ideals. *American Journal of Political Science*, 50(3):634–649.

Ingham, S. (2015). Theorems and models in political theory: An application to Pettit on popular control. *The Good Society*, 24(1):98–117.

Ottonelli, V. and Porello, D. (2013). On the elusive notion of meta-agreement. *Politics, Philosophy & Economics*, 12(1):68–92.

Pettit, P. (2012). *On the People's Terms. A Republican Theory and Model of Democracy*. Cambridge University Press, Cambridge.

Rawls, J. (1993). *Political Liberalism*. Columbia University Press, New York.

Sunstein, C. R. (1995). Incompletely theorized agreements. *Harvard Law Review*, 108(7):1733–1772.

van Benthem, J. and Pacuit, E. (2011). Dynamic logics of evidence-based beliefs. *Studia Logica*, 99(1):61.

Van De Putte, F. and Klein, D. (2018). Pointwise Intersection in Neighborhood Modal Logic *Advances in Modal Logic*, 12 :591–610.

A New Modification of the Halpern-Pearl Definition of Causality

Xiaoan Wu

Tsinghua University

This paper is trying to the search for a satisfying solution to the problem of actual causality, which is just a specialized topic within the study of causation, peripheral to the issues of legal and moral responsibility. I will focus on the causal modelling approach developed by (Pearl 2000, 2009) whose seminal contributions are widely known. This approach arises at the beginning of the new century, the last twenty years have seen an explosion of research directed at it, a lot of considerable progress have been made, but also been criticized. Because of the different understanding of causal modelling, there are three closely-related approaches for using structural equations to model actual causation: Hitchcock-approach (Hitchcock 2001, 2007), Halpern-Pearl-Hitchcock-approach (Halpern 2016) and Ned Hall-approach (Hall 2007), in this paper I will show the whole development of HP-definition, (Halpern and Pearl 2001) introduced the first version of the definition using structural equations, but there is an obvious problem with this approach, it was updated in the journal version of the paper (Halpern and Pearl 2005). Further counterexamples were given to the updated definition, to deal with these examples, Halpern taken into account considerations of normality and defaults. But these approaches do not always seem so satisfactory, so Halpern further modify the HP-definition (Halpern 2015). New definition means new frame, also means that we need to "test" the definition by the counterexamples, I will do that as a brave man. Also I will show that even (Halpern 2015) is not very satisfactory, inspired by Ned Hall's theory of causation (Hall 2007), I get a improved version of (Halpern 2015), it works well , can dealt with the symmetric overdetermination problems that (Halpern 2015) cannot do, and can dealt with other problems that it can do, I think I've done it sucessfully.

References

Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, 132(1):109–136.

Halpern, J. Y. (2015). A modification of the halpern-pearl definition of causality. In *IJCAI*, pages 3022–3033.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Halpern, J. Y. and Pearl, J. (2001). Causes and explanations: A structural-model approach. In *In Proceedings IJCAI-01*. Citeseer.

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 56(4):843–887.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6):273–299.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, 116(4):495–532.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press, first edition edition.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. New York: Cambridge university press, second edition edition.

Formalization of Argumentation Schemes for Slippery Slope Argument

Zhe Yu

Sun Yat-Sen University

Abstract

Formal argumentation is seen as a promising tool to bridge the gap between human reasoning and machine reasoning. As an attempt to portray natural language arguments based on formal argumentation systems, this paper aims to give a formalism for a commonly-used defeasible argument - slippery slope argument (SSA). To evaluate SSA, this paper defines the critical questions for SSA, and discussed why value judgement important in evaluating a SSA by taking the gene-edited babies case in China as an example.

1 Introduction

Argumentation is a cross-disciplinary topic involving multiple subjects such as philosophy, cognitive science, logic, linguistics and computer science. There are several research directions in the field of artificial intelligence, such as natural language processing and argumentation mining, can be combined with argumentation and benefits from it (Cabrio et al. 2016). As an approach for non-monotonic reasoning, formal argumentation is promising to bridge the gap between human reasoning and machine reasoning. To achieve this goal, a key problem is how to model natural language arguments by formal argumentation.

Argumentation schemes can be seen as a "semi-formal" generalization of arguments (Verheij). Many researchers have shown their interests in the formalization of argumentation schemes, for example, the schemes for argument from expert opinion (Prakken et al. 2015; Gabbay and Thiruvasagam 2017; Atkinson and Bench-Capon).

In (Walton 1992), Walton mentioned that slippery slope argument, as a subclass of argument from negative consequences, is commonly used in the context of deliberation, to persuade people

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

not to take an action under consideration. Typically, SSA can be found in the discussions about legal, biomedical, or ethical issues, such as topics like abortion, gay marriage, euthanasia, human gene therapy, etc.

This paper aims to fomalise and evaluate slippery slope argument based on formal argumentation theory. Firstly, we give a formal model of slippery slope argument based on the structured argumentation framework $ASPIC^{+}$ (Besnard et al. 2014; Modgil and Prakken 2013)(calls SSAT), by consulting the argumentation schemes for slippery slope argument presented by Walton (Walton 2015, 2017). Then we discuss how to evaluate a slippery slope argument by asking critical questions and value judgement in formal systems. For illustration, this paper take the gene-edited babies case in China[1] as an example.

The rest of this paper is structured as follows. Section 2 will firstly summarize the basic features of SSA according to Walton's basic argumentation scheme for SSA, then construct a slippery slope argumentation theory based on formal argumentation system. Then, in section 3 we will model the critical questions for SSA and discuss the value judgement in the evaluation of SSA. In section 4, we conclude this paper.

2 SSAT

In this section, we model slippery slope argument based on the basic scheme for it and the structured argumentation framework $ASPIC^{+}$.

2.1 Basic Scheme for SSA

According to (Walton 2015), we use a_0 to denote an action under consideration, a_n to denote a catastrophic outcome. a_1, a_2, …, a_x, …, a_y denotes a sequence of action or events between a_0 and a_n, each causes the next one, so that the slippery slope argument has the following basic argumentation scheme:

Initial Premise An agent α is considering carrying out an action a_0.

Sequential Premise Carrying out a_0 would lead to a_1, which would in turn lead to carrying out a_2, and so forth, through a sequence $a_2, \ldots, a_x, \ldots, a_y, \ldots, a_n$.

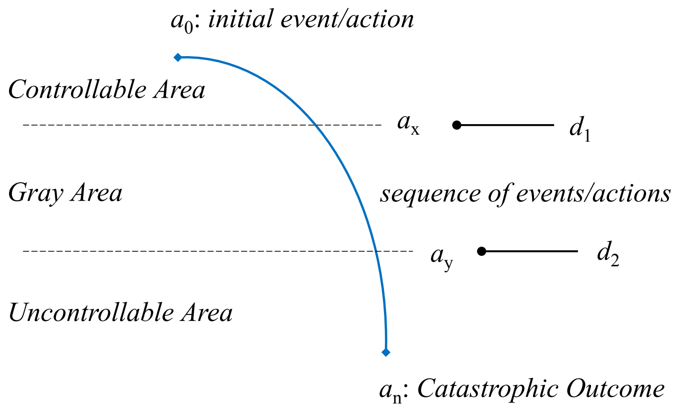
Indeterminacy Premise There is a sequence a_0, a_1, a_2, …, a_x, …, a_y, …, a_n that contains a subsequence a_x,…, a_y called the gray zone where x and y are indeterminate points.

Control Premise α has control over whether to stop carrying out the actions in the sequence until α reaches some indeterminate point in the gray zone a_x, ,…, a_y.

Loss of Control Premise Once α reaches the indeterminate point in the gray zone a_x, ,…, a_y, α will lose control and will be compelled to keep carrying out actions until she reaches a_n.

Catastrophic Outcome Premise a_n is a catastrophic outcome that should be avoided if possible.

[1] see websites
https://www.aljazeera.com/news/2018/11/chinese-scientist-pauses-gene-edited-baby-trial-outcry-181128095039618.html,
https://edition.cnn.com/2018/11/26/health/china-crispr-gene-editing-twin-babies-first-intl/index.html,
https://edition.cnn.com/2019/01/21/health/china-gene-editing-babies-intl/index.html, etc.



a_0: *initial event/action*
Controllable Area
a_x
d_1
Gray Area
sequence of events/actions
a_y
d_2
Uncontrollable Area
a_n: *Catastrophic Outcome*

Figure 1: A basic SSA

Conclusion a_0 should not be brought about.

Walton also mentioned that "there are factors that help to propel the argument and series of consequences along the sequence, making it progressively harder for the agent to resist continuing to move ahead" (Walton 2015). He calls these factors "Drivers". Then, based on Walton's opinion, we conclude that a slippery slope argument has the following 8 basic components

1. An initial event/action a_0

2. A sequence of events/actions: as the sequence proceeds, the consequences tend to become more serious.

3. Drivers: catalyst that helps to propel the argument along the sequence in the argument. Drivers could be factors like precedent, public acceptance, vagueness, climate of social opinion, public acceptance, etc. In this paper drivers are denoted by d_i ($d = 1, 2, 3, \ldots$).

4. Gray area: area start at an undetermined point x (denoted by a_x), end at another undetermined point y (denoted by a_y). In this area a slippery slope argument is turning form controllable to uncontrollable.

5. Controllable area: the area between initial event/action and the gray area.

6. Uncontrollable area: the area between the gray area and the catastrophic outcome.

7. Catastrophic outcome a_n

8. Conclusion: not to take the initial step.

Based on this idea, the developing process of a slippery slope argument can be illustrated by figure 1.

2.2 Argumentation Theory for SSA

The current work is mainly based on the structured argumentation framework $ASPIC^+$, which proposed by Prakken et al. in (Modgil and Prakken 2013). $ASPIC^+$ is not a system but a framework, so that people can specify or extend it as an instantiation, as long as meeting some specific requirements.

Based on the above summarization of the characters of slippery slope arguments, we use a symbol "$\perp$" to denote "bad/unwanted (consequence)", $\mathcal{R}_{sl}$ and $\mathcal{R}_j$ to denote two kinds of rules used in slippery slope arguments respectively, $\mathcal{K}_0 = \{a_0, b_0, c_0, \ldots\}$ to denote a set of initial actions/events, C to denote a set of actions/events, and $D = \{d_1, \ldots, d_n\}$ to denote a set of drivers. Then an argumentation theory for SSA can be defined as following.

Definition 1 (SSAT). *A slippery slope argumentation theory ($SSAT$) is a tuple $SSAT = (\mathcal{L}, ^-, \mathcal{R}, n, \mathcal{K}, C, D)$, where:*

- *$\mathcal{L}$ is a logical language; $\perp \in \mathcal{L}$.*
- *$^-$ is a function from $\mathcal{L}$ to $2^{\mathcal{L}}$, such that*
  1. *φ is a contrary of ψ if $\varphi \in \overline{\psi}$, $\psi \notin \overline{\varphi}$;*
  2. *φ is a contradictory of ψ, [2] if $\varphi \in \overline{\psi}$, $\psi \in \overline{\varphi}$;*
  3. *each $\varphi \in \mathcal{L}$ has at least one contradictory.*
- *n is a partial function such that n: $\mathcal{R}_d \rightarrow \mathcal{L}$.*
- *$\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules of the form $\varphi_1, \ldots, \varphi_n \rightarrow \varphi$ and $\varphi_1, \ldots, \varphi_n \Rightarrow \varphi$ respectively (where φ_i and φ are meta-variables raging over wff in $\mathcal{L}$), and $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$. $\mathcal{R}_{sl} \subseteq \mathcal{R}_d$ is slippery slope rule of the form $\varphi_1, \ldots, \varphi_n \Rightarrow_{sl} \varphi$ (φ_i, φ are elements in $\mathcal{L}$), $\mathcal{R}_{sl} \neq \emptyset$; $\mathcal{R}_j \subseteq \mathcal{R}_s$ is consequence judging rule of the form $\varphi_1, \ldots, \varphi_n \rightarrow_j \varphi$, there is at least one consequence judging rule form as $r_j = \varphi_1, \ldots, \varphi_n \rightarrow_j \perp$.*
- *$\mathcal{K} \subseteq \mathcal{L}$ is a knowledge base in an argumentation system, consisting of three disjoint subsets $\mathcal{K}_n$, $\mathcal{K}_p$ and $\mathcal{K}_0$ (i.e. $\mathcal{K} = \mathcal{K}_n \cup \mathcal{K}_p \cup \mathcal{K}_0$), where:*
  1. *$\mathcal{K}_n$ is a set of the axioms;*
  2. *$\mathcal{K}_p$ is a set of the ordinary premises, s.t. $\neg\perp \in \mathcal{K}_n \cup \mathcal{K}_p$;*
  3. *$\mathcal{K}_0$ is a set of initial steps in a slippery slope argument of the form $\mathcal{K}_0 = \{a_0, b_0, c_0, \ldots\}$, where a_0, b_0, c_0 are initial actions or events.*
- *C is a set of actions or events in a slippery slope argument of the form $C = \{a_0, \ldots, a_n, b_0, \ldots, b_m, c_0, \ldots, c_q, \ldots\} \subseteq \mathcal{L}$, where a_i, b_j, c_k are actions or events; $\mathcal{K}_0 \subseteq C$.*
- *D is a set of drivers, $D = \{d_1, \ldots, d_n\} \subseteq \mathcal{K}_p$, while d_i is a driver.*

Accoring to $ASPIC^+$, an argument in $SSAT$ can be defined as following.

[2] denoted by '$\varphi = -\psi$'

Definition 2 (Arguments). *An **argument** A on the basis of a $SSAT = (\mathcal{L}, ^{-}, \mathcal{R}, n, \mathcal{K}, C, D)$ is defined as:*

1. *φ if $\varphi \in \mathcal{K}$ with: $Prem(A) = \{\varphi\}$, $Conc(A) = \varphi$, $Sub(A) = \{\varphi\}$, $DefRules(A) = \emptyset$, $TopRule(A) = undefined$.*

2. *A_1 , ... ,$A_n \rightarrow \psi$ if A_1, ..., A_n $(n \geq 1)$ are arguments such that there exists a strict rule $Conc(A_1)$, ..., $Conc(A_n) \rightarrow \psi$ in $\mathcal{R}_s$ with: $Prem(A) = Prem(A_1) \cup \ldots \cup Prem(A_n)$; $Conc(A) = \psi$; $Sub(A) = Sub(A_1) \cup \ldots \cup Sub(A_n) \cup \{A\}$; $DefRules(A) = DefRules(A_1) \cup \ldots \cup DefRules(A_n)$; $TopRule(A) = Conc(A_1) \ldots Conc(A_n) \rightarrow \psi$.*

3. *A_1 , ... ,$A_n \Rightarrow \psi$ if A_1, ..., A_n $(n \geq 1)$ are arguments such that there exists a defeasible rule $Conc(A_1)$, ..., $Conc(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$ with: $Prem(A) = Prem(A_1) \cup \ldots \cup Prem(A_n)$; $Conc(A) = \psi$; $Sub(A) = Sub(A_1) \cup \ldots \cup Sub(A_n) \cup \{A\}$; $DefRules(A) = DefRules(A_1) \cup \ldots \cup DefRules(A_n) \cup \{Conc(A_1), \ldots, Conc(A_n) \Rightarrow \psi\}$; $TopRule(A) = Conc(A_1) \ldots Conc(A_n) \Rightarrow \psi$.*

Then a slippery slope argument in an argumentation system can be defined as following.

Definition 3 (SSA). *If an argument A in $SSAT = (\mathcal{L}, ^{-}, \mathcal{R}, n, \mathcal{K}, C, D)$ such that:*
$Prem(A) \cap \mathcal{K}_0 \neq \emptyset$, $Prem(A) \cap D \neq \emptyset$, $SlRule(A) \neq \emptyset$, $JRule(A) \neq \emptyset$, $Conc(A) = \perp$, for every $A' \in Sub(A)$ and $A' \neq A$, $Conc(A') \in C \cup D$,
then A is a SSA.

By claiming that the bad outcome is unacceptable, slippery slope arguments always attempt to draw the conclusion that the initial step should not be taken. To capture this feature, in addition to transposition under strict rules required by $ASPIC^{+}$, we define a weak transpositon for the slippery slope rules used in SSA.

Definition 4 (Transposition & Weak Transpositon). *Let $SSAT = (\mathcal{L}, ^{-}, \mathcal{R}, n, \mathcal{K}, C, D)$ be a $SSAT$, $SSAT$ is closed under transposition and weak transposition:*

1. *if $\varphi_1, \ldots, \varphi_n \rightarrow \psi \in \mathcal{R}_s$, then for each $i = 1 \ldots n$, there is*
   $\varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots \varphi_n \rightarrow -\varphi_i \in \mathcal{R}_s$;

2. *if $\varphi_1, \ldots, \varphi_n \Rightarrow_{sl} \psi \in \mathcal{R}_{sl}$, then for eachi $= 1 \ldots n$, iff $\varphi_i \in C$, there is*
   $\varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots \varphi_n \Rightarrow_{slt} -\varphi_i \in \mathcal{R}_d$; the set of transpositon rule denotes as $\mathcal{R}_{slt} \subseteq \mathcal{R}_d$; transpositon rule of a slippery slope rule $r_i \in \mathcal{R}_{sl}$ $(i = 1 \ldots n)$ denotes as $r_{it} \in \mathcal{R}_{slt}$.

In $ASPIC^{+}$, arguments could be attacked in three ways: 1) undermining attack on the ordinary premises; 2) rebutting attack on the conclusions (only when the last rule is defeasible); 3) undercutting attack on the defeasible rules. Because in this paper we add a special set of premises $\mathcal{K}_0$, whose elements are actually presumptions that be supposed to take place, so that we define undermining attack slightly different from in $ASPIC^{+}$. Besides, we defined the weak transposition, so that the undercutting attack will also become different. Then the attack relation for SSAT is defined as following.

Definition 5 (Attack). *Let A, B and X be arguments in $SSAT = (\mathcal{L}, ^{-}, \mathcal{R}, n, \mathcal{K}, C, D)$, $\varphi, \psi \in \mathcal{L}$. A attacks B, iff A undercuts, rebuts or undermines B, where:*

- *A undercuts B on B' iff:*

    1. *$Conc(A) = \overline{n(r)}$* [3] *for some $B' \in Sub(B)$ such that $TopRule(B') = r$ where $r \in \mathcal{R}_d$;*
    2. *$\exists X$, $X' \in Sub(X)$, such that $TopRule(X') = r_i (i = 1, \ldots, n)$, $r_i \in \mathcal{R}_{sl}$ and $r_{it} \in \mathcal{R}_{slt}$, $Conc(A) = \overline{n(r_i)}$, i.e. A undercuts X on X', where $B' \in Sub(B)$, such that $TopRule(B') = r_{it}$.*

- *A rebuts B on B' if and only if $Conc(A) = \overline{Conc(B')}$ for some $B' \in Sub(B)$ of the form $B''_1, \ldots, B''_n \Rightarrow \varphi$; A contrary-rebuts B iff $Conc(A)$ is a contrary of φ.*

- *A undermines B on B', iff:*

    1. *$B' = \varphi$ and $\varphi \in Prem(B) \cap \mathcal{K}_p$, such that $Conc(A) \in \overline{\varphi}$ and if $A = \psi$, then $\psi \notin \mathcal{K}_0$;*
    2. *$B' = \varphi$ and $\varphi \in Prem(B) \cap \mathcal{K}_0$, such that $Conc(A) \in \overline{\varphi}$.*

    A contrary-undermine B iff $Conc(A)$ is a contrary of φ.

In $ASPIC^+$, whether an attack from A to B (on its sub-argument B') succeeds as a defeat may depend on the relative strength of A and B'. In (Modgil and Prakken 2013), this is determined by a binary ordering $\preceq$ on the set of all arguments. With arguments and the defeat relation, we can evaluate the arguments by Dung style abstract argumentation frameworks and the semantics (Dung 1995) and decide the set of arguments that acceptable together (called an extension) under specific semantics. We omit the definition of defeat relation in $ASPIC^+$ and the abstract argumentation framework here, the reader is refer to paper (Dung 1995) and (Modgil and Prakken 2013) to find more details.

3 Evaluating SSA

In argumentation schemes, each scheme is matched with a different sequence of critical questions. Basically, there are two way to evaluate a given argument: 1) use schemes to check the form of the argument; 2) ask the corresponding critical questions, to see if the questions can be answered satisfactorily.

In this section we try to give some way to evaluate a slippery slope argument based on formal argumentation. The main idea is to formalize the critical questions of argumentation scheme for slippery slope argument, thus we can involve the critical questions into an argumentation framework and evaluate all the arguments together.

3.1 Critical Questions

We define the critical questions for slippery slope argument as following.

Definition 6 (Critical Question)**.** *Let argument A, B be arguments in $SSAT = (\mathcal{L}, ^-, \mathcal{R}, n, \mathcal{K}, C, D)$, $\varphi, \psi \in \mathcal{L}$. Let A be a SSA, such that $d_i \in Prem(A)$, $r_{sli} \in SlRule(A)$, $Conc(A) = \bot$. B is an argument of critical question for A (denotes by **CQA**) iff $TopRule(B) = \varphi_1, \ldots, \varphi_n \rightarrow / \Rightarrow \psi$, while $\psi = \overline{n(r_{sli})}$, $\psi = \overline{d_i}$ or $\psi = \overline{\neg\bot}$.*

In (Walton 2017), the author gave the following 5 critical questions for the basic scheme of slippery slope argument.

[3] '$n(r)$' means that rule r is applicable.

CQ1 What intervening links in the sequence of events a_1, a_2, $\ldots$, a_i needed to drive the slope forward from a_0 to a_n are explicitly stated?

CQ2 What missing steps are required as links to fill in the sequence of events from a_0 to a_n, to make the transition forward from a_0 to a_n plausible?

CQ3 What are the weakest links in the sequence, where additional evidence needs to be given on whether one event will really lead to another?

CQ4 Is the sequence of argumentation meant to be deductive, so that if the first step is taken, it is claimed that the final outcome a_n must necessarily come about?

CQ5 Is the final outcome a_n shown to be catastrophic by the value-based reasoning needed to support this claim?

Suppose that a proposed slippery slope argument cannot answer CQ1, CQ2 or CQ4, it means that (at least one of) the link between the initial step a_0 to the bad outcome a_n is too weak, in other words, the defeasible rule between premises to the conclusion is too weak to apply ($\overline{n(r_d)}$). And if a proposed slippery slope argument cannot answer CQ3, it perhaps that there lack a driver to back up the slope, or the given driver is not good enough. For the first situation, it also can be seen as that the related link is too weak; for the second situation, it means at least one given driver has been attacked ($\overline{d_i}$). At last, if a proposed slippery slope argument cannot answer CQ5, it means that the final outcome of this argument is not really unacceptable as it has been claimed to ($\overline{\neg\bot}$).

3.2 Value Judgement: a Case Study

Slippery slope argument is considered to be a subspecies of argument from negative consequences, which reject a proposal for action by point out that the action will lead to a negative consequence. While argument from negative consequences is considered to be a subspecies of argument from values (Walton 2017; Walton et al. 2008). In this paper, we model the negative value by introduce a symbol "$\bot$" into the language $\mathcal{L}$ of an argumentation system, correspondingly, introduce a symbol "$\neg\bot$" in the knowledge base $\mathcal{K}$ to represent the intrinsic unacceptability of something bad. Thus a slippery slope argument can be attacked by a statement as "the final outcome is not as bad as it has been claimed", i.e., based on definition 3.1, an argument of critical question (CQA) with the conclusion $\overline{\neg\bot}$.

As a SSA and the CQA that questioning the value judgement the SSA are obviously conflict with each other, in a scenario of deliberation, conflicts can be resolved by comparing these arguments according to priorities. A slippery slope argument work by claim that take the first step will lead to a highly undesirable consequence, which means that the consequence strongly contravenes values held by the audience (Walton 2017). Based on this idea, when we consider the priorities in a SSAT, value judgment deserved to be taken into account.

We can see this point of view more clearly through an example from real life: the gene-edited babies experiment in China.

On November 26, 2018, Chinese researcher He Jiankui claims that his lab had been editing embryos' genetic codes for seven couples undergoing in-vitro fertilization, and twin girls had been born with DNA altered to make them resistant to HIV. He used a tool known as CRISPR-cas9, which can insert or deactivate certain genes. As claimed, the twin babies are immune to HIV.

Editing the genes of embryos intended for pregnancy is banned in many countries, in the other countries, editing of embryos may be permitted for research purposes with strict regulatory approval. He's experiment is the world's first case of germline gene therapy performed on humans, which is likely to spark significant ethical questions around gene editing and so-called designer babies. Mang scientists are outraged. A professor in genetics and human embryology at University College London described this research as "premature, dangerous and irresponsible" [4].

A key point in evaluating this case is whether the advantages it brings outweighs the disadvantages. In this regard, the general tendency of scientists is that the benefits are extremely small compared to the risks that this move will bring.

First of all, due to the uncontrollability of the gene editing process, this experiment may bring unpredictable risk of genetic disease. Meanwhile, there are many effective ways to prevent HIV in healthy individuals.

More seriously, people are worried that this move may open the "Pandora's Box". In other words, this move may become the first step of a slippery sloping, leads to the acceptence of genetic enhancement technologies, and finally come to a catastrophic outcome: "eugenic" - a fearful term that reminds people Nazi Germany and have long been concerned in the ethical discussion of genetic technology.

Despite of the form of this slippery slope argument, people hold negative opinion in this case mainly because that by value judgement, the possible final consequence is too scary and disastrous that should be avoided at all cost.

There are already some related formal systems that we can use to handle the value judgment of agents. For instance, in (Liao et al. 2016), the authors introduced a hierarchical abstract theory of normative system (called HANS) to resolve conflicts amongst norms. In simple terms, this system associated numbers that indicating priorities of norms to an abstract theory of normative system defined by Tmsatto et al. (Tosatto et al. 2012). When conflicts arise between norms, HANS resolve it by the priorities assigned to them, and derive extensions according to different detachment procedure. Based on formal argumentation, the HANS can be extended to a value based system by introduce value assignments from different agent(Liao et al. 2016; Beishui Liao 2019). Which can be defined as following.

Definition 7 (VHANS). *An $VHANS$ is a tuple $\mathcal{H}_v = \langle L, N, C, V, Ag, \rho \rangle$, where*

- $L = E \cup \{\neg e | e \in E\} \cup \{\top\}$[5] *is the universe, a set of literals based on some finite set E of atomic elements;*
- $N \subseteq L \times L$ *is a finite set of regulative norms;*
- $C \subseteq L$ *is a subset of the universe, called a context, such that $\top \in C$ and for all e in E, $\{e, \neg e\} \nsubseteq C$;*
- V *is a set of values;*
- Ag *is a set of agents representing different participants;*
- $\rho : N \rightarrow V \times Ag \times IN$ *is a function from norms to a triple of values, agents and natural numbers.*

[4] https://edition.cnn.com/2018/11/26/health/china-crispr-gene-editing-twin-babies-first-intl/index.html

[5] $\top$ used to represent the body of a body- free norm.

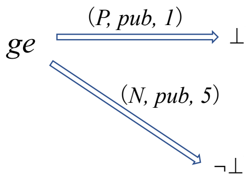


Figure 2: A VHANS

Based on definition 3.2, suppose there are two conflicting norms in the discussion about gene-edited babies case:

1. Immune to HIV is a positive consequence, gene-edited babies experiment should be accepted: $(ge, \neg\bot)$ (ge reperents gene-edited babiew experiment)

2. Eugenic is a negative consequence, gene-edited babies experiment should be rejected: $(ge, \bot)$

The public may assign different value to these two norms. We use P to represent positive, N to represent negative, pub to represent the public, assuming that $\rho(ge, \neg\bot) = (P, pub, 1)$, $\rho(ge, \bot) = (N, pub, 5)$, we can get a VHANS as figure 2.

By apply different principles to lift priorities (for example, the last link priciple or the weakest link principle (Modgil and Prakken 2013) in $ASPIC^{+}$), we can get the priorities on arguments based on the value of norms and decide whether an argument can successfully defeat its counter-arguments (Liao et al. 2016).

4 Conclusion

In this paper we introduced an argumentation theory for a basic form of slippery slope argument (SSAT) mainly based on the argumentation scheme for slippery slope argument given by (Walton 2015) and the formal argumentation framework $ASPIC^{+}$ (Modgil and Prakken 2013).

For evaluating a slippery slope argument, we defined the critical questions for slippery slope argument, then people can model and evaluate a slippery slope argument (as well as its sub-arguments) by formal argumentation systems.

Besides, the slippery slope argument could also be seen as an approach to achieve practical reasoning. The powers of persuasion of a slippery slope argument should be resting on a premise that the ultimate consequence is catastrophic, which could be seen as resting on various kinds of value judgments. In this paper, we illustrate this idea by the gene-edited babies case, and point out that it is possible to combine SSAT to some value-based argumentation systems, for example, the argumentation based HANS.

References

Atkinson, K. and Bench-Capon, T. Abstract argumentation scheme frameworks. in Danail Dochev, Marco Pistore and Paolo Traverso eds., Artificial Intelligence: Methodology, Systems, and Applications, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 220–234.

Beishui Liao, Zhe Yu, L. v. d. T. (2019). Practical reasoning about normsvalues and preferences. *Journal of Tsinghua University(Philosophy and Social Sciences)*, 34(2):140–149+201.

Besnard, P., García, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., and Toni, F. (2014). Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4.

Cabrio, E., Hirst, G., Villata, S., and Wyner, A. (2016). Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (Dagstuhl Seminar 16161). Technical report, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321 – 357.

Gabbay, D. M. and Thiruvasagam, P. K. (2017). Reasoning schemes, expert opinion and critical questions. sex offenders case study. *FLAP*, 4.

Liao, B., Oren, N., van der Torre, L., and Villata, S. (2016). Prioritized norms and defaults in formal argumentation. In *in Proceedings of the 13th International Conference on Deontic Logic and Normative Systems (DEON2016)*, pages 139–154.

Modgil, S. and Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397.

Prakken, H., Wyner, A., Bench-Capon, T., and Atkinson, K. (2015). A formalization of argumentation schemes for legal case-based reasoning in aspic+. *Journal of Logic and Computation*, 25(5):1141–1166.

Tosatto, S. C., Boella, G., van der Torre, L., and Villata, S. (2012). Abstract normative systems: Semantics and proof theory. In G. Brewka, T. E. and McIlraith, S. A., editors, *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR 2012*, pages 358–368, Rome, Italy.

Verheij, B. *The Toulmin Argument Model in Artificial Intelligence*. Guillermo Simari and Iyad Rahwan eds., Argumentation in Artificial Intelligence, Chapter 11, Boston, MA: Springer US, 2009: 219-238.

Walton, D. (1992). *Slippery Slope Arguments*. Oxford UniversityPress, Oxford.

Walton, D. (2015). The basic slippery slope argument. *Informal Logic*, 35(3):273–311.

Walton, D. (2017). The slippery slope argument in the ethical debate on genetic engineering of humans. *Science and Engineering Ethics*, 23(6):1507–1528.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press, Cambridge.

Solution Complexity of Local Variants of Sabotage Game

Tianwei Zhang

Tsinghua University

Abstract

The study of graph games serves as a way to analyze an existing logic as well as an inspiration for designing new logics. Given the fact that game-theoretic analysis is regaining attention in AI study, a new stress in the study of graph games can be the performance of standard algorithmic tasks conducted on graphs. In this paper we carry out a case study on the respective graph game for three main local variants of sabotage modal logic, which have a broad range of applications in various other fields. We analyze the solution complexity for each game and show the implications these results have on their respective matching logic. This work is a first attempt to understand why similar-looking variants of a graph game and their matching logics can have drastically different computational complexity, and hopefully brings up a more general topic that requires more studies, namely to identify the parameters of games and logic that crucially affect complexity.

1 Motivation

Sabotage modal logic was proposed in 2003 as a format for analyzing games that modify graphs they are played on. Since its first introduction, it has inspired a line of systems in a dynamic-epistemic spirit, (van Benthem 2011). Along with different variants that stems from the minimal language such as ‘graph modifier logic’ (Aucher et al. 2009), ‘swap logic’ (Areces et al. 2014), ‘arrow update logic’ (Areces et al. 2012; Kooi and Renne 2011) and so on, various graphs games

In: B. Liao, F. Liu, and H. Dong (eds.): Proceedings of the Workshop on Logics for the Formation and Dynamics of Social Norm (LFDSN 2019), Hangzhou, China, 4-5 May 2019, published at http://ceur-ws.org.

of interactive settings are proposed to capture the essences of such logics. Meanwhile examples of how real life scenarios can be depicted by graph games can be found in a broad collection of areas including theoretical computer science (Grüner et al. 2013; Radmacher and Thomas 2008), learning theory (Gierasimczuk et al. 2009), logics for social networks (Liu et al. 2014; Seligman et al. 2013), argumentation (Grossi 2010b,a) and many more.

In light of the the avid interest that has recently arisen in game-theoretic analysis of AI study, a new stress in the study of graph games can be the performance of standard algorithmic tasks conducted on graphs. This paper is a case study. We particularly focus on localized sabotage games, whose global versions can have either link deletion or node deletion. We take three of the exisiting variants of localized sabotage game: local link-cutting sabotage(adjacent sabotage game in (Rohde 2005)), local node-deleting sabotage (van Benthem and Klein 2018) as well as poison game (Blando et al. 2018), and focus on their solution complexity. We find that local link-cutting lowers complexity for game solution to PTIME while the corresponding logic being undecidable. Meanwhile, local node-deletion and poison game do not have this effect: model checking and game solution are in both cases are PSPACE-complete.

2 Local Link-Cutting Variant

In this section, we show that the solution problem of local link-cutting sabotage game is in the complexity class of PTIME. For this aim, we introduce an algorithm that solves the problem in PTIME and prove the correctness of the algorithm.

2.1 A Solution Algorithm

Here we present the polynomial-time algorithm for solving the local link-cutting sabotage in pseudo-code. V is the set of all the vertices in the given graph. F is the set of all the goal points of the graph (*a fortiori* a subset of V). η is a function from $V \times V$ to $\{true, false\}$ used to indicate whether two nodes in the graph are connected. Here we are assuming there are no multiple links between nodes, but with some modifications, this algorithm can also be use to calculate all the winning positions in in a multi-linked graph with a new η from $V \times V$ to the set of natural numbers.

Algorithm 1 Algorithm to solve local link-cutting sabotage

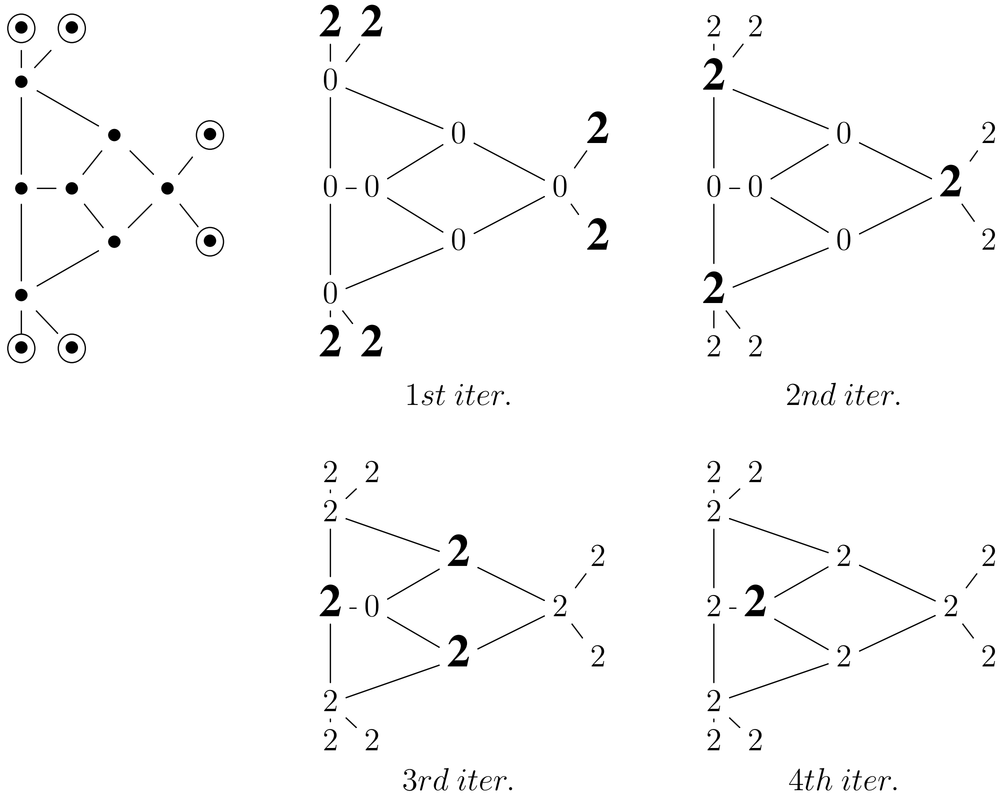
```
procedure LocalLinkCutting($V, F, \eta$)
    input: arena($V, \eta$); $F \subseteq V$ final vertices
    output: $L \subseteq V$ winning positions
    $L = F$
    $count = |F|$
    while $count > 0$ do
        $T =$ the set of $v$ in $V - L$ with two or more neighbors in $L$
        $count = |T|$
        $L = L \cup T$
    end while
end procedure
```

In practice, if provided with $O(|V|^2)$ space to store the pointer list of neighbors for each node and $O(|V|)$ space for a queue for scheduling, the algorithm can solve the problem in $O(|V|^2)$ time.

Intuitively, the algorithm can be understood as the propagation of a signal. At the very beginning, the only sources of the signal are the goal points. They emit the signal to each of their immediate neighbors. The other nodes in the graph turn themselves into a source and in turn emit the signal to their neighbors as soon as they have received two signals themselves. This way the algorithm goes in iterations until it reaches an equilibrium.

Also note that in the multi-linked version of the algorithm, signals transmit through links, so a multiple link should be counted multiple times, that is, if a non-source node has only one neighbor emitting signals but has a multiple link to this neighbor, then this node should also be considered a source and start emitting signals itself. The reason for this is easy to understand, since Demon can only cut one link at each of his turn.

An example of how the signal propagates in each iteration is shown in Figure 1, with the circled nodes being goal points. In each iteration, the number on each node denote how many links it has to an already active neighbor. The goal points are also denoted 2 for the consistency of notation. In the beginning, the only sources are the goal points. In the second iteration, a, b and c also become source since they each have two links that connects to a source. In the third iteration, d, e and f are also converted to sources because of their connection to a, b and c. Eventually, in the fourth iteration, g is converted to a source because of its connection tp d, e and f. At this point, there are no more nodes to be converted, and thus the situation reaches an equilibrium and the algorithm ends. As a result, all the nodes in the example are winning positions for Traveler since they all become source in the end.



1st iter. *2nd iter.*

3rd iter. *4th iter.*

Figure 1: example for the algorithm

2.2 Correctness Proof of the Algorithm

Proposition 2.1. A node $v \in V$ in a local link-cutting sabotage game $((V, \eta), F)$ is a winning position for Traveler if and only if $v \in L = \text{LocalLinkCutting}(V, F, \eta)$.

Proof. $\leftarrow$: Suppose Traveler is currently at $v \notin L$. Then by the description of the algorithm, at most one of v's neighbors is in L. In light of that, Demon can make sure that Traveler next lands on a node $v' \notin L$ by cutting the only (if any) link to L. This way, if Traveler starts from a node in $V - L$, then there's a strategy for Demon that guarantees that Traveler always stays in $V - L$. Eventually. Traveler finds herself at a node with no successor and Demon wins.

$\rightarrow$: We prove the this direction of the proposition by proving a slightly stronger statement:

Definition 2.2. Define the degree of a node v in L: $\xi : L \rightarrow \mathbb{N}, \forall v \in L\ \xi(v) = n$ if and only if v is add to L in the nth iteration. We stipulate that for all $v \in F$, $\xi(v) = 0$.

Proposition 2.3. For all $v \in L$, there is a winning strategy in $((V, \eta), F, v)$ for Traveler where if $v_1, v_2, ...v_n$ are the nodes visited by Traveler in a sequential order, then $\xi(v_1) > \xi(v_2) > ... > \xi(v_n)$. For convenience, we will call such a strategy a 'decreasing strategy' in the following paragraphs.

Proof. We prove the proposition by induction on $\xi(v)$.

If $\xi(v) = 0$, Traveler wins at the very beginning, so the proposition is trivially true.

Suppose $\forall v \in L$ with $0 \leq \xi(v) \leq n$ the statement is true. For any $v \in L, \xi(v) = n + 1$, by the description of the algorithm, two of more of its neighbors has degree less than $n + 1$. Suppose Demon cuts l and changes the game to $((V, \eta'), F, v)$. By the algorithm, v has two or more neighbors with degree less than $n + 1$, so Traveler can then opt to land on a node u of degree less than $n + 1$. By the assumption,Traveler has a decreasing strategy in $((V, \eta), F, u)$. Easy to see that the strategy only involves links between nodes in $\{v \in L \mid \xi(v) \leq n\}$, and therefore does not involve l. This shows that this strategy is also a decreasing strategy for $((V, \eta'), F, u)$.

Now we can create a decreasing strategy for $((V, \eta), F, v)$. In the first round, after Demon cuts a link and changes the game to $((V, \eta'), F, v)$, move to any node u of degree less than $n + 1$. In all the following rounds, move according to the decreasing strategy for $((V, \eta'), F, u)$.

Thus, for all $v \in L$ Traveler has a decreasing strategy in $((V, \eta), F, v)$. □

Since a decreasing strategy is by nature a winning strategy, the $\rightarrow$ direction is thus proven. □

.

3 Local Node-Deleting Variant

In this section, we establish the PSPACE-Completeness of the problem of solving the local node-deleting variant of sabotage game. For this aim, we provide an alternating algorithm for solving the decision problem of the variant that runs in polynomial time, along with a polynomial-time reduction from the problem of Quantified Boolean Formula, which is known to be PSPACE-complete.

3.1 PSPACE Upper Bound by an Alternating Algorithm

In this and the next section, we use the existence of a alternating algorithm for the problems to show the upper bounds of the problems. *Alternating Turing machines*, ATM for short, are a generalization of non-deterministic Turing machines. Their state sets are divided into two components: existential states and universal states. An existential state is accepting if and only if some transition leads to an accepting state; while a universal state is accepting if and only if every transition leads to an accepting state. For more information on alternating turing machines and alternating algorithms, please see (Balcázar et al. 2012). We use APTIME to denote the class of problems that can be solved by Alternating Turing machines. According to (Balcázar et al. 2012), we have the following result:

Theorem 3.1. APTIME = PSPACE

In this and next section, we use pseudocode to specify alternating algorithms. In codes, we use instruction 'guess' to signify non-deterministic choices of successor corresponding the existential states of an ATM, and instruction 'choose all' to signify the branching of universal states of an ATM. Apart from inheriting notations V, F, η, we also introduce new ones: v for denoting the starting point of Traveler and L an initially empty set for keeping track of the visited nodes at current state.

Algorithm 2 Alternating algorithm to solve local node-deleting sabotage

```
procedure LocalNodeDeleting($V, F, \eta, v, L$)
    input: arena $(V, \eta)$; $F \subseteq V$ final vertices; $v \in V$; $L \subseteq V$
    if $v \in F$ then
        accept
    else if $v \in L$ or $\forall v' \in V, \eta(v, v') = false$ then
        reject
    else
        guess $v' \in V$ with $\eta(v, v') = true$
        universally choose $u \in V$ with $\eta(v', u) = true$
        let $V' := V - \{u\}$
        let $F' = F - \{u\}$
        let $L' := L \cup \{v\}$
        let $\eta' = \eta_{|V' \times V'}$
        call LocalNodeDeleting($V', F', \eta', v', L'$)
    end if
end procedure
```

We start the algorithm with the initial call

$$\text{LocalNodeDeleting}(V_{in}, F_{in}, \eta_{in}, v_{in}, \emptyset).$$

Lemma 3.2. Let $\mathcal{G} = ((V_{in}, \eta_{in}), F_{in}, v_{in})$ be a local node-deleting sabotage game. The alternating algorithm LocalNodeDeleting accepts its initial input if and only if Traveler has a winning strategy in $\mathcal{G}$. The running time is polynomial with respect to $|V|$.

Proof. First we prove the correctness of the algorithm.

Let $c_0c_1...$ be some computation branch of LocalNodeDeleting and assume that $(V_i, F_i, \eta_i, v_i, L_i)$ is the parameter of call c_i. Let

$$\pi := (0, V_0, F_0, v_0, \eta_0)(1, V_1, F_1, v_1, \eta_1)(0, V_2, F_2, v_2, \eta_2)(1, V_3, F_3, v_3, \eta_3)...$$

By the update of parameters in Line 8-13, π forms a legal play of the sabotage game. In fact, by the universal choice in Line 9, there is a one-to-one correlation between computation trees and game trees of Traveler. Further, by Line 12, we have $L_i := \{v_j | j < i\}$, that is, L_i gives the set of already visited nodes during the prefix of the play π from v_0 to v_i.

If Traveler has a winning strategy from v_{in} in the game, then she can ensure that she reaches some final vertex starting from v_in. By a lemma yet to be proven, Lemma 3.11, she can do so without visiting any vertex twice. The algorithm can guess the successor v_{i+1} of v_i in Line 8 according to the winning strategy. Since it is winning, there is some natural number n such that $v_n \in F$. Further, no node v_i for $i < n$ is a dead end at the moment when it is Traveler's turn to continue the play from v_i, i.e., at position $(0, v_i, \eta_i)$. Thus, the algorithm does not reject its input in Line 6 during some call c_i with $i < n$, before it finally terminates in an accepting state in Line 4 during call c_n. Therefore, every computation branch terminates in an accepting state and the algorithm accepts its input.

Conversely, if Demon has a winning strategy, then he can delete nodes in such a way that every play leads either to some already visited node, or to some dead end. Thus, a winning strategy of Demon ensures that for each guess of successors in Line 8, the corresponding computation tree contains at least one non-accepting branch. Hence, the algorithm rejects its input.

Because parameter L in each embedded call is increasing, it restricts the depth of recursion up to $|V| + 1$. It is also clear that in each layer of recursion, only polynomial space is required. Finally, checking the conditions and updating the parameters take a time that is polynomial with respect to $|V|$. By the correspondence of alternating polynomial time and deterministic polynomial space, the last part of the proposition is proven. □

The lemma and theorem 3.1 automatically gives us the following theorem:

Theorem 3.3. Solving the local node-deleting sabotage game is at most PSPACE.

3.2 Reduction from QBF

Definition 3.4. (QBF) Instances of the problem Quantified Boolean Formula, or QBF for short, are quantifier-free Boolean formulae in conjunctive normal form over a set of Boolean variables $x_1...x_n$ for some $n \in \mathbb{N}$. Let φ be such a formula. Then φ is a positive instance if

$$\exists x_1 \forall x_2 \exists x_3 ... Q x_n : \varphi$$

where Q is equal to $\exists$ for n odd and to $\forall$ otherwise. Note that our definition of QBF requires the formulae to begin with an existential quantification, but this is no loss of generality.

Stockmeyer and Meyer (1973) determined the complexity of this problem:

Theorem 3.5. The problem QBF is PSPACE-complete.

In the remainder of this section, we present a polynomial time reduction from QBF to the solution problem of a local node-deleting sabotage game. This method of reduction is inspired by the proof of PSPACE-Completeness of the global link-cutting variance of the sabotage game by (Rohde 2005).

Let φ be an instance of QBF in conjunctive normal form with Boolean variables $x_1 ... x_n$. We construct an arena $\mathcal{A}_\varphi$ for a local node-deleting sabotage game $\mathcal{G}_\varphi$ such that

Proposition 3.6. Let φ, $\mathcal{G}_\varphi$ be as in Definition 3.3. Traveler has a winning strategy in $\mathcal{G}_\varphi$ if and only if there exists an assignment for x_1 such that for all assignments for x_2 ... φ is satisfied.

3.2.1 Construction of the Arena

The arena consists of $n + 2$ components, one for each variable plus two verification components. Generally, the traversal of the arena starts with a existential component, which is modified to be slightly different from the standard existential component, so that Traveler can make the first move in the game. Following it is a universal component, and then again an existential component. At the bottom of the arena, we find two verification components. In each verification component, C_is are called clause vertices. Connected to them below are literal vertices L_{ijk}s, each representing one literal in the clause. The curve line stemming from each literal vertex represents a link that connects the literal vertex to one of the variable vertices. Each of these lines leads to variable vertex X_i if x_i is positive in l, and to $\neg X_i$ if x_i is negative in l. The complete lines are omitted for the sake of neatness.

Note that in each existential and universal component, X_i and $\neg X_i$ are 'duos'.

Definition 3.7. (a duo) Let $\mathcal{A} = (V, \eta)$ be the arena of $\mathcal{G}$. A duo $D \subseteq V$ contains two nodes x and y such that $\forall z \in V, \eta(x, z) = \eta(y, z)$ and $\eta(x, y) = false$. Duos are represented by X_i and $\neg X_i$ in figures.

Intuitively, a duo is just two nodes that are not connected and that have the exact same connections to the other nodes. The introduction of duos serves the purpose of not only simplifying the arena in display but also ensuring Proposition 3.6. Note that an arena with a duo is bisimilar to the arena of the same construction except for only a single node in place of the duo. We can establish a bisimulation between the arena and an arena of the similar construction with all the duos replaced by single nodes. In the local node-deleting sabotage game, this bisumulation will be valid as we do not delete nodes belonging to a duo, which is the case most of the time, as we will see later in this subsection. In light of that, for the rest of the section, without further explanation, we will address the duos as if they were single nodes. In 3.2.4, we will encounter situations where the deletion of one of the nodes of a duo should be considered, which in turn will explain why we introduce duos in the first place.

Example 3.8. Let $\varphi = c_1 \wedge c_2 \wedge c_3 \wedge c_4$ be a Boolean formula in conjunctive normal form with Boolean variables x_1, x_2, x_3. We assume that each clause consists of exactly three literals, i.e., for $k \in [1, 4]$, we have $c_k = l_{k1} \vee l_{k2}$ with $i_{kj} \in \{x_1, \neg x_1, x_2, \neg x_2, x_3, \neg x_3\}$. Thus, φ belongs to QBF if and only if we have

$$\exists x_1 \forall x_2 \exists x_3 : \varphi$$

The arena for example 3.8 is depicted in Figure 2.

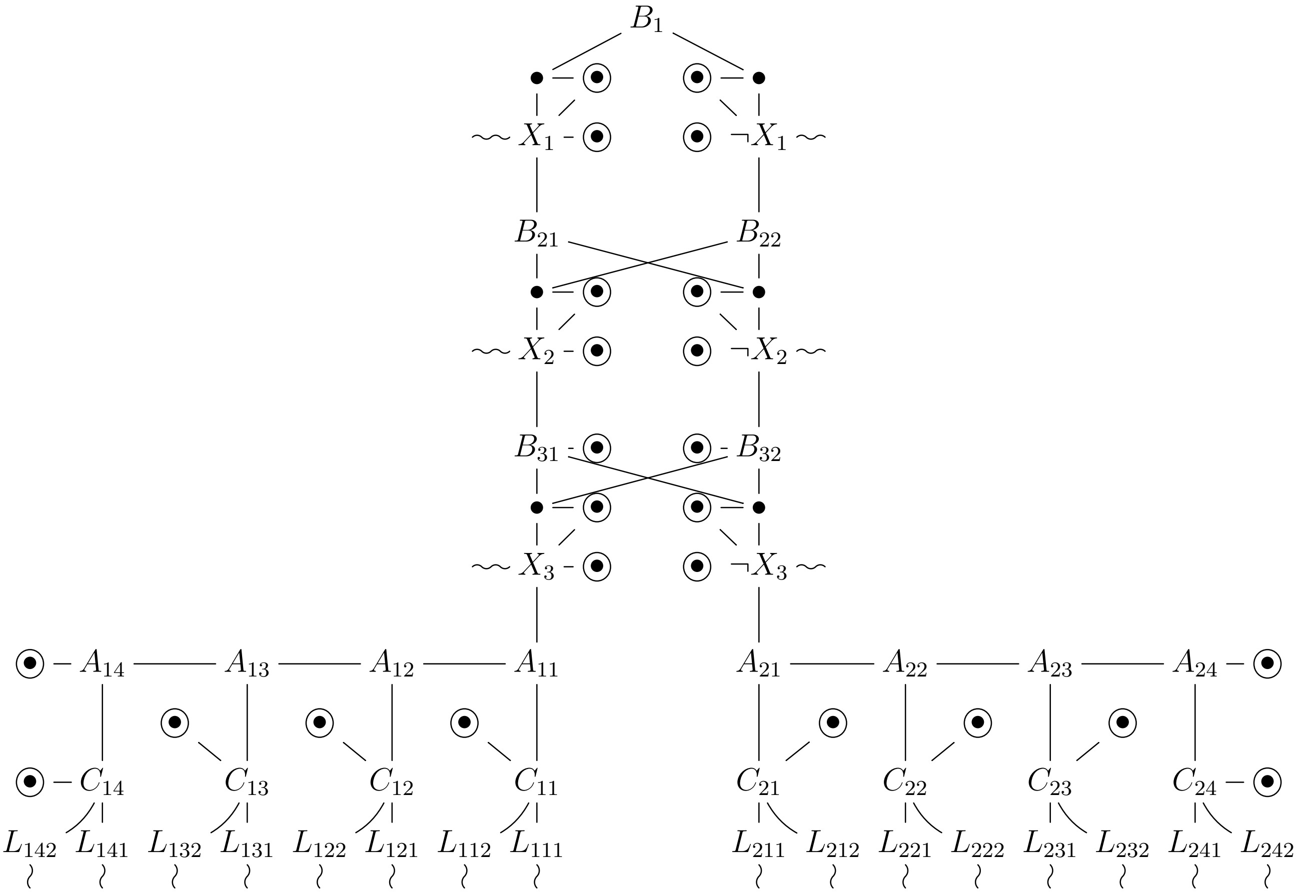


Figure 2: local node-deleting :arena for example 3.8

3.2.2 Properties of the game

Before we delve into the details of each component, we first some useful properties of the game. These properties are also discussed in (van Benthem and Liu 2018). Along with several others, the discovery of these properties might help eventually understand problem of the disparity of the computational complexity in similarly-appearing graph games, which we will discuss in greater details in section 5.

Definition 3.9. Given two local node-deleting sabotage games $\mathcal{G}_1 = (\mathcal{A}_1, F_1, x_1)$ and $\mathcal{G}_2 = (\mathcal{A}_2, F_2, x_2)$ where $\mathcal{A}_1 = (V_1, \eta_1)$, $\mathcal{A}_2 = (V_2, \eta_2)$, then we say $\mathcal{G}_1 \leq \mathcal{G}_2$ if and only if there exists an injective map φ from V_1 to V_2 such that
i)$\varphi(x_1) = x_2$
ii)$\forall v \in V_1, \varphi(v) \in F_2$ if and only if $v \in F_1$
iii)$\forall y_1, y_2 \in V_1, \eta_1(y_1, y_2) \leq \eta_2(\varphi(y_1), \varphi(y_2))$

Corollary 3.10. *(Monotonicity)* Given two local node-deleting sabotage game $\mathcal{G}_1$ and $\mathcal{G}_2$ as in Definition 3.8, if Traveler has a winning strategy in $\mathcal{G}_1$,and $\mathcal{G}_1 \leq \mathcal{G}_2$, then Traveler has a winning strategy in $\mathcal{G}_2$.

Proof. This is obvious since Traveler can simply use the map φ in Definition 3.8 to transform the winning strategy in $\mathcal{G}_1$ to a winning strategy in $\mathcal{G}_2$. During the transformation, we first simply apply φ to the strategy for $\mathcal{G}_1$. This gives us a fragment of the strategy for $\mathcal{G}_2$ in which we only consider the situations where Demon deletes nodes in $\varphi(V_1)$. Then we augment the fragment in the following way: when Traveler is at $u \in \varphi(V_1) - F_2$, if Demon deletes a node $v \notin \varphi(V_1)$, Traveler then (randomly) picks a node $v' \in \varphi(V_1)$ such that $\eta_2(v', u) = true$ (the existence of such v' is guaranteed by the fact that Traveler has a winning strategy in $\mathcal{G}_1$) and follows the strategy as if Demon had deleted v'. Note that in this way Traveler will always stay in$\varphi(V_1)$. It's easy to see that the augmented fragment is a winning strategy for $\mathcal{G}_2$ since it provides Traveler with reaction for all the possible situations and by playing by the strategy Traveler can always win. □

Lemma 3.11. *(Non-Cyclicity)* If Traveler has a winning strategy in $\mathcal{G}_\varphi$, she can win without visiting any node twice.

Proof. If Traveler has a winning strategy that involves visiting a node v twice, suppose when she first visits v, the game is $\mathcal{G}_1 = (\mathcal{A}_1, F_1, v)$, and when she visits v the second time, the game is $\mathcal{G}_2 = (\mathcal{A}_2, F_2, v)$. It is obvious that $\mathcal{G}_1 \leq \mathcal{G}_2$ with φ being the inclusion map. Since Traveler has a winning strategy in $\mathcal{G}_2$, we can use φ to transform this strategy into a winning strategy in $\mathcal{G}_1$. In this transformed strategy, the number of circles is reduced by (at least) 1 by avoiding the journey in the original strategy from $\mathcal{G}_1$ to $\mathcal{G}_2$. Since any winning strategy in a local node-deleting sabotage game is finite in terms of steps, it consequently has a finite number of circles. By removing circles for a finite number of times, we can obtain a winning strategy that does not involve visiting a node twice. □

Given Lemma 3.11, we will assume that Traveler never visits any node twice in the whole game and will not repeat this assumption unless necessary. Note that by our discussion before, the two nodes of a duo should be considered as one single node, and thus visiting one node of a duo and then later visiting the other should count as visiting a node twice as well.

3.2.3 Components and their functions

In the rest of this section, we describe the details of each component (existential, universal and verification component), show a few consequences of the structure and how these consequences lead to Proposition 3.6.

The structures of each component(existential, universal and verification) are shown in Figure 3 on the left, on the right and in Figure 4 respectively.

Corollary 3.12. When traversing an existential component, Traveler has the choice not to visit either X_i or $\neg X_i$.

Proof. We prove the statement by analyzing how the two players interact in the competitive setting. Depending on the choice of Demon in the previous component, Traveler enters an existential component from B_{i1} or B_{i2}. We will see that this does not affect her strategy. Without loss of generality, suppose Traveler starts from B_{i1} and wants to avoid passing through X_i. Due to the way components are assembled, we assume that Demon makes the first move. At B_{i1}, Demon has to delete the immediate goal point or Traveler can move straight to it and win. Then Traveler can move to d_2. At d_2, Demon again has to remove the immediate goal point for the same reason. Then Traveler moves to $\neg X_i$. At $\neg X_i$, similarly, Demon has to delete the immediate goal point. Now Traveler can go down and exit the component, leaving X_i not visited. □

The previous corollary shows that the existential component is a place where Traveler get the chance to choose. Later, as we will see, when Traveler comes back to verify the QBF, the unvisited duo will provide Traveler with access to goal points. This way Traveler chooses the valuation of the variable of the existential quantifier by choosing which duo to not visit.

Corollary 3.13. When traversing a universal component, Demon can force Traveler to visit either X_i or $\neg X_i$ and prevent her from reaching any goal point inside the component.

Proof. Similarly, we prove the statement by analyzing the process of traversal. Depending on the choice of herself in the previous component, Traveler enters a universal component from B_{i1} or B_{i2}. We will see that this does not affect Demon's strategy. Without loss of generality, suppose Traveler starts from B_{i1} and Demon wants to make sure that she passes through X_i. Due to the way components are assembled, we assume that Demon makes the first move. At B_{i1}, Demon deletes d_2. Then Traveler can only move to d_1. At d_1, Demon has to remove the immediate goal point. Then Traveler moves to X_i. At X_i, similarly, Demon has to remove the immediate goal point. Now Traveler can only exit the component. □

Here the idea is similar. Demon get to choose the valuation of the variable of the universal quantifier and he does that by forcing Traveler to visit the opposite duo. This way the duo Traveler doesn't visit remains access to goal points and in turn corresponds to the literal of the chosen valuation being true.

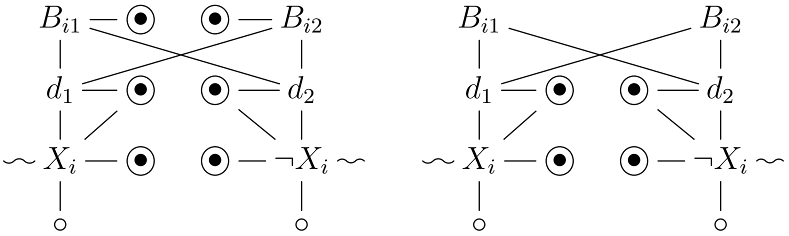


Figure 3: Local Node-Deleting:Existential and Universal Component

Corollary 3.14. When traversing a verification component, Demon can force Traveler to visit one of C_1, C_2,...,C_m with m being the number of clauses in the QBF, and prevent her from reaching any goal point inside the component.

Proof. Again, we prove the statement by showing the reasoning of both players. Due to the way the components are put together, we assume that Demon acts first. Suppose Demon wants Traveler to visit C_i. Traveler starts from A_1. For all $1 \leq j < i$, when Traveler arrives at A_j, Demon deletes C_j. Then Traveler has to move to A_{j+1}. If $i < m$, When Traveler arrives at A_i, Demon deletes A_{i+1}, else Demon deletes the immediate goal point. When Traveler arrives at C_i, Demon deletes the immediate goal point. □

Corollary 3.15. At C_i, Traveler can choose any one of the literal nodes directly linked to C_i to move to.

Proof. This is obvious since when Traveler arrives at C_i, Demon has to delete the immediate goal point and this leave Traveler to choose freely which literal nodes to move to. □

The previous two corollaries shows that we are able to mimic a verification game in a verification component. Demon, the falsifier, chooses the clause in which to falsify the formula by choosing the C_i for Traveler to visit. Traveler, the verifier, in turn chooses the literal that is true (if any) in the clause Demon has just chosen.

Now it is time to give a proof of Proposition 3.6.

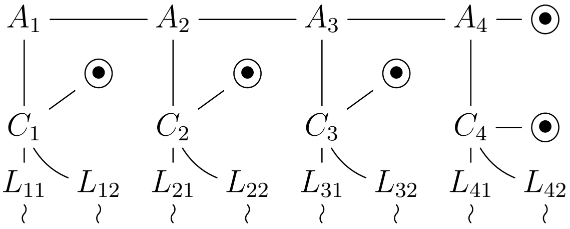


Figure 4: Local Node-Deleting:Verification Component

3.2.4 Correctness of the Reduction

Now with Corollary 3.12 to 3.15, we can prove an equivalent version of Proposition 3.6:

Proposition 3.16. Let φ, $\mathcal{G}_\varphi$ be as in Definition 3.3.

1) If there exists an assignment for x_1 such that for all assignments for x_2 ..., φ is satisfied, then Traveler has a winning strategy in $\mathcal{G}_\varphi$.

2) If for all assignments for x_1 there exists an assignment for x_2 such that..., $\neg\varphi$ is satisfied, then Demon has a winning strategy in $\mathcal{G}_\varphi$.

Proof. 1)If there exists an assignment for x_1 such that for all assignments for x_2 ... φ is satisfied, then a winning strategy for Traveler can be as following: Traverse the existential and universal components by the order that they are put together (if at any X_i or $\neg X_i$ Demon delete the node directly below, move to the immediate goal point and win), and when at each existential component, avoid X_i if x_i is true in the assignment and avoid $\neg X_i$ if false. When at C_i, move to the literal node that corresponds to the literal that makes the clause true according to the assignment. The duo connected to the literal node should have not been visited.

Demon now may delete one node of the duo, but Traveler can move to the remaining node of the duo. Note that this is the only chance for Demon to delete a node that belongs to a duo without allowing Traveler to move to a goal point in the immediate turn. This is also why we need such structures as 'duos' in the first place: if there are only single nodes in place of the duos in the arena, then at this point of the game Demon can simply delete the unvisited X_i or $\neg X_i$ to prevent Traveler from winning.

When Traveler moves to the unvisited X_i or $\neg X_i$ (now may consist of only one node instead of two), there are two immediate goal points. Demon may delete one of then, but Traveler can still move to the remaining one and win.

2)We first show the following fact: If Traveler moves through a curve link from X_i or $\neg X_i$ to a literal node, since Demon can then delete the immediate clause node, Traveler will have no choice but go back through the curve link. Therefore, suppose Traveler does not visit any node twice, she should always exit a universal or an existential component by going down through the straight link, that is, Traveler should traverse the existential and universal components by the order that they are put together.

If for all assignments for x_1 there exists an assignment for x_2 such that... $\neg\varphi$ is satisfied, then a winning strategy for Demon can be as following: When Traveler traverses the existential and universal components, at each universal component Demon forces Traveler to visit the duo corresponding to the overall assignment that satisfies $\neg\varphi$. Then when Traveler arrives at a verification component, there should be at least one clause node that is only connected to literal nodes whose immediate duo have already been visited. Demon then forces Traveler to that clause node, and then Traveler has to visit a duo for the second time, which indicates that she loses. □

Theorem 3.17. The solution problem of local node-deleting sabotage games is PSPACE-complete.

Proof. Since the reduction only involving generating verification components that has certain number of clause nodes and literal nodes and assembling all the components, the time complexity of the reduction is $O(l)$, where l is the length of the QBF. Thus we have a polynomial-tmie reduction from the problem of true QBF to the solution problem of local node-deleting sabotage games. This establish the PSAPCE-hardness of the problem. With Lemma 3.2, we conclude that the problem is PSPACE-complete. □

4 Poison Game

In this section, we establish the PSPACE-completeness of the problem of solving poison game. To this end, we provide an alternating algorithm for solving the problem that runs in polynomial time, along with a polynomial-time reduction from the problem of Quantified Boolean Formula.

4.1 PSPACE Upper Bound by an Alternating Algorithm

Here we present the alternating algorithm for solving poison game. Aside from the familiar notations, we also utilize an integral variable *count* to keep record of for how many steps Demon hasn't poisoned a new node. As we will see later, this extra variable will ensure that this simulation of poison game, which is potentially infinite, will eventually terminate.

Algorithm 3 Alternating algorithm to solve poison game

```
 1: procedure SolvePoison($V, \eta, v, L, count$)
 2:     input: arena $(V, \eta)$; $v \in V$; $L \subseteq V$
 3:     if $count = |L|$ then
 4:         accept
 5:     else
 6:         universally choose $u \in V$ with $\eta(u, v) = true$
 7:         guess $v' \in V$ with $\eta(v, v') = true$
 8:         if $v' \in L$ then
 9:             reject
10:         end if
11:         let $L' := L \cup \{u\}$
12:         if $u \in L$ then
13:             call SolvePoison($V, \eta, v', L', count + 1$)
14:         else
15:             call SolvePoison($V, \eta, v', L', 0$)
16:         end if
17:     end if
18: end procedure
```

We start the algorithm with the initial call

$$\text{SolvePoison}(V, \eta, v_{in}, \emptyset, 0).$$

Lemma 4.1. Let $\mathcal{G} = ((V, \eta_{in}), v_{in})$ be a poison game. The alternating algorithm SolvePoison accepts its initial input if and only if Traveler has a winning strategy in $\mathcal{G}$. The running time is polynomial with respect to $|V|$. In particular, the problem of solving poison games belongs to PSPACE.

Proof. First we prove the correctness of the algorithm.

Note Traveler wins if she can play infinitely, and since the arena itself is finite, in this situation there is bound to be repetition. By the pigeonhole principle, when Demon hasn't poisoned any new nodes in the most recent $|L| + 1$ steps, he must have visited/poisoned a node for at least twice. Since no nodes were added between the repeated visits, we can prune the branches that have such repetition.

Let $c_0c_1...$ be some computation branch of SolvePoison and assume that $(V, \eta, v_i, L_i, count_i)$ is the parameter of call c_i. Let

$$\pi := (0, v_0, L_0, count_0)(1, v_1, L_1, count_1)(0, v_2, L_2, count_2)(1, v_3, L_3, count_3)$$

By the update of parameters in Line 6-16, π forms a legal play of the poison game. In fact, by the universal choice in Line 6, there is a one-to-one correlation between computation tree and game trees of Traveler(pruned in the way mentioned above). Further, by Line 11, L_i gives the set of nodes already poisoned by Demon during the prefix of the play π from v_0 to v_i.

If Traveler has winning strategy from v_in in the game, then she can survive for infinite number of steps. The algorithm can guess the successor v_{i+1} of v_i in Line 6-7 according to the winning strategy. Since it is winning, for any natural number i, $v_{i+1} \notin L_i$, and since after pruning the algorithm always terminates after finite steps of computation, it follows that every computation branch terminates in an accepting state and the algorithm accepts its input.

Conversely, if Demon has a winning strategy then he can poison nodes in such a way that there is some natural number n such that $v_{n+1} \in L_n$. Thus, a winning strategy of Demon ensures that for each guess of successors in Line 7, the corresponding computation tree contains at least one non-accepting branch. Hence, the algorithm rejects its input.

Complexity-wise speaking, since $|L| \leq |V|$ and every time Demon poisons a new node $count$ is reset, the depth of recursion is at most $|V|^2$, which is polynomial with respect to $|V|$. As in the proof for Proposition 3.2, in each layer of recursion also only polynomial space is required, and checking the conditions and updating the parameters also take a time that is polynomial with respect to $|V|$. By the correspondence of alternating polynomial time and deterministic polynomial space, the last part of the proposition is fulfilled. □

4.2 Reduction from QBF

In this part, our aim is to construct an arena $\mathcal{A}_\varphi$ for a poison game $\mathcal{G}_\varphi$ such that Traveler has a winning strategy in $\mathcal{G}_\varphi$ if and only if there is an assignment for x_1 such that for all assignments for x_2 there exists... and φ is satisfied by the overall assignment, in which case φ belongs to QBF.

4.2.1 Construction of the arena

The arena also consists of $n + 1$ components of three kinds (existential, universal and verification component). Starting from the existential component, the traversal of the arena alternates between existential and universal components, until the verification component is reached. For sake of convenience, we continue to use example 3.8 for illustration in this section.

The arena for example 3.8 is depicted in Figure 5. Each of the wavy arrows in the verification component represents multiple outward arrows, the number of which equals the number of literals in each clause. Each of these arrows ends in the variable vertex X_i if x_i is positive in l, and in $\neg X_i$ if x_i is negative in l. The complete arrows are omitted for the sake of neatness.

4.2.2 Components and their function

In this subsection, we go into details and describe the structure and function of each component.

Proposition 4.2. When Demon poisons a node with no successors, Traveler wins.

Proof. It follows directly from the winning conditions for Traveler. □

Existential Component. Shown in Figure 6. To show how this component fits in the whole arena, the hollow bullet at the bottom represents a node in the following component. The same representation is also used in the universal component. Traveler enters the existential component from s_i. Given Prop.4.1, Demon will avoid esc and chooses to poison a. Now, Traveler can choose the assignment of x_i simply by choosing which node between b_1 and b_2 to travel to. She moves towards b_1 if she wants x_i to be false and towards b_2 if she wants x_i to be true. Whichever branch she chooses, the variable vertex on that branch is poisoned by Demon in the next step, which is his only option. Traveler moves onto the hollow bullet and exits the component.

Universal Component. Shown in Figure 7. Similarly, Traveler enters the component from s_i, and for the same reason, Demon does not poison esc but instead chooses the assignment for x_i by

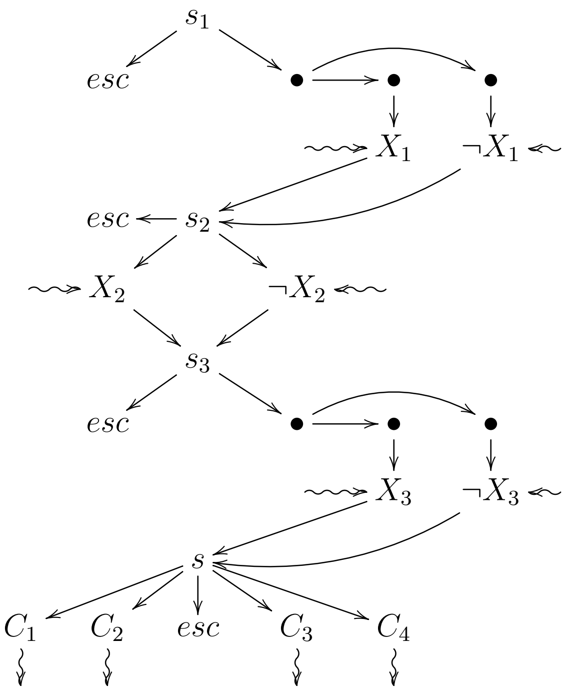


Figure 5: poison game: arena for example 3.8

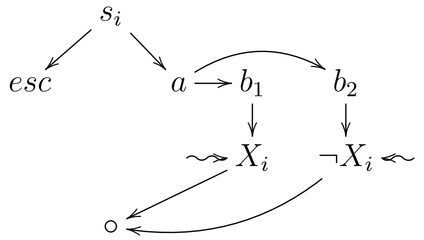


Figure 6: Poison Game:Existential Component

poisoning either of the two variable vertices. He poisons b_1 if he wants x_i to be false and b_2 if he wants x_i to be true. Then Traveler moves onto the hollow bullet and exists the component.

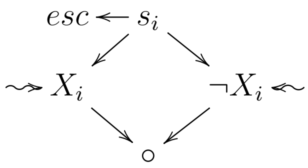


Figure 7: Poison Game:Universal Component

Verification Component. Shown in Figure 8. Traveler enters the verification component from s. Given Prop.4.1, Demon only considers poisoning one of the clause vertices (marked as C_i in the figures). Since Example 3.8 has four clauses in the quantifier-free formula, there are four clause vertices in Figure 8. Generally, the number of clause vertices in the verification component equals the number of clauses in the QBF. Demon chooses in which clause to falsify the QBF by poisoning the corresponding clause vertex. Suppose Demon chooses to poison C_i, given the way clause vertices point to variable vertices, Traveler chooses among the literals in c_i by moving towards the corresponding variable vertex. If the variable vertex of choice has already been poison during the traversal of that component, Traveler loses. Otherwise, the game continues and Demon has no choice but poison the s_i/s of the next component, as there is only one outward arrow at a variable vertex. Then, from that s_i/s, Traveler can move towards esc, meaning to escape, and win the game.

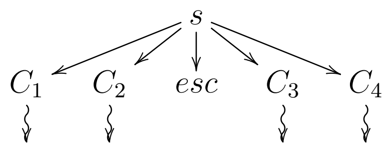


Figure 8: Poison Game:Verification Component

4.2.3 Correctness of the Reduction

Lemma 4.3. For any QBF φ, Traveler has a winning strategy in $\mathcal{G}_\varphi$ if and only if φ is a true QBF.

Proof. If φ is a true QBF, then there is an assignment for x_1 such that for all assignments for x_2 there exists... φ is true, i.e., there is at least one true literal in each clause. If Traveler follows the choice of such assignments in each existential component, then it is guaranteed that each C_i will leads to at least one literal vertex that has not been poisoned. Traveler can move to that literal vertex and then win by moving to esc.

If φ is a false QBF, then for all assignments for x_1 there is an assignment for x_2 such that... φ is false, i.e., there is at least one clause of QBF in which all the literals are false. If Demon follows the choice of such assignments in each universal component, then it is guaranteed that in the verification component, Traveler will have to choose to move from a point whose successors(one or more) have all been poisoned, and thus loses the game. □

Theorem 4.4. The solution problem of poison games is PSPACE-complete.

Proof. Since the reduction only involving generating verification components that has certain number of clause nodes and assembling all the components, the time complexity of the reduction is $O(l)$, where l is the length of the QBF. Thus we have a polynomial-tmie reduction from the problem of true QBF to the solution problem of poison games. This establish the PSAPCE-hardness of the problem. With Lemma 4.1, we conclude that the problem is PSPACE-complete. □

5 Logical Aspect

There's long been an intriguing entanglement between a modal language of games and its corresponding evaluation game, or so call 'logic game' (van Benthem 2014; van Benthem and Klein 2018). All the games mentioned above have their corresponding modal languages. For the modal language for local link-cutting sabotage game, see *loc*SML in (Aucher et al. 2017). For a modification of *loc*SML that considers deleting links of certain property called *definable sabotage logic* and its corresponding logic game, see (Li 2018). For the modal language that captures the idea of local node-deleting sabotage game and its modification considering deleting nodes with certain property please refer to (Chen 2018). For formalization for poison game, see (Blando et al. 2018).

Considering the corresponding modal languages, the proceeding results can have further implications. Note that in all three of the logic, the winning positions of a graph game can be defined using the modal μ-calculus, see (van Benthem and Liu 2018; Areces and van Benthem 2019; Blando et al. 2018).

In finite graph game, the each μ-calculus formulation of the winning positions above is equivalent to a finite formula of the language with μ-calculus and thus is a valid formula of respective language.[1] In case of local node-deleting sabotage and poison game, whose upper bound for model checking can be deducted to be PSPACE-complete from (Areces et al. 2009; Blando et al. 2018) [2] and (Chen 2018) respectively. Since the complexity for model checking should be no less than the complexity for checking any specific formula, we have the follow theorems:

Theorem 5.1. Model checking of PSL, PML and $ML_{\emptyset}$ is PSPACE-complete.

Theorem 5.2. Model checking of (whatever logic for local node deletion is) is PSPACE-complete.

If we take a closer look at the specific μ-calculus formulation for the winning positions for a local link-cutting sabotage game shown as follows,

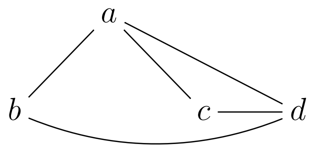
$$\nu p \cdot (\gamma \vee \Box \perp \vee \blacksquare \Diamond p)$$

given the polynomial-time solution algorithm in Section 2, we find that locality induces a significant change in the modal fixed-point definition for Traveler's winning positions in the sabotage game. Instead of using the dynamic link deletion modality, we can make do with the following formula:

$$\nu p \cdot (\gamma \vee \Box \perp \vee \langle \geq 2 \rangle p)$$

Here $\langle \geq 2 \rangle$ is a static graded modality stating the existence of at least two successors of the current point satisfying φ. Proving the equivalence between these two modal fixed-point formulas requires care with comparing winning strategies for Traveler at the same point in different graphs, and the argument typically seems to fail for the global sabotage game, since the solution complexity for the latter is Pspace (Löding and Rohde 2003).

Some might suspect that the algorithm in Section 2 can somehow be expended to produce a polynomial-time algorithm for the purpose of model checking. However, this idea is not viable given the fact that model checking for local link-cutting sabotage game is PSPACE-complete (Rohde 2005).



The failure of a similar recursive algorithm for model checking of the local link-cutting sabotage can be illustrated in the following example: Suppose we want to use a similar recursive algorithm to determine the set of nodes in the arena($V(p) = \{a\}$) above that satisfy

$$\blacksquare \Diamond \Diamond \Diamond p$$

[1] Actually (Blando et al. 2018) provides three modal languages for poison game, for each of which a μ-calculus formulation of the winning positions is given

[2] In (Areces et al. 2009), it is been proven that the model checking for $ML(\textcircled{k}, \textcircled{r}, \textcircled{e})$ is PSPACE-complete and since $ML(\textcircled{k}, \textcircled{r})$(which is in turn the same language as $ML_{\emptyset}$ in (Blando et al. 2018)) is a fragment of $ML(\textcircled{k}, \textcircled{r}, \textcircled{e})$ and thus the upper bound for the model checking for $ML_{\emptyset}$ is PSPACE, and by the inclusion of expressive power of the three modal languages for poison game and the clearly polynomial-time translation between the languages, all three of the languages has PSPACE as the upper bound for model checking.

First, in the spirit of the analysis above, we replace $\blacksquare\Diamond$ with $\langle\geq 2\rangle$ and get

$$\langle\geq 2\rangle\Diamond\Diamond p$$

Then in a recursive manner we can calculate:
$V(p) = \{a\},\ V(\Diamond p) = \{b, c, d\},\ V(\Diamond\Diamond p) = \{a, b, c, d\},\ V(\langle\geq 2\rangle\Diamond\Diamond p) = \{a, b, c, d\}$
At the same time, starting from a, if Demon cuts the link between a and d , a will no longer satisfy $\Diamond\Diamond p$. Hence, $a \notin V(\blacksquare\Diamond\Diamond\Diamond p)$, which in turn means that $V(\blacksquare\Diamond\Diamond\Diamond p) \neq V(\langle\geq 2\rangle\Diamond\Diamond p)$.

This inequality comes from the fact that there's a two-way dependency between a and d in that if we cuts the link between them, at both of them there is a sub-clause of $\blacksquare\Diamond\Diamond\Diamond p$ changes from true to false, with $\Diamond\Diamond\Diamond p$ for a and $\Diamond p$ for d. This indicates that we can no longer establish a tree-like dependence relation among the nodes like we did in section 2 since, for example, the link between a and d bears a two-way dependency. The property of having a tree-like hierarchy of dependence among nodes might be key to the polynomial-time complexity of the solution problem, and after proper abstraction and formalization, can be a promising candidate parameter to account for the discrepancy of complexity between superficially similar games.

6 Related Work

This paper focuses on the complexity analysis of several localized variants of sabotage games and its implication. (Rohde 2005) provides yet another proof of the solution problem of local link-cutting sabotage being PTIME using double tractor. The work also provides insights for the design of arenas in Section 3 and 4.

For a thorough introduction of SML, (Aucher et al. 2017) formulates the latest results in the minimal sabotage modal logic of arbitrary edge deletion. Stemming from this minimal logic, a number of authors have studied other graph games using matching modal logics. The logic of (Li 2018) characterizes the scenario where Demon at each of his move, can delete all the local links that lead to nodes of a certain property. (Chen 2018) provides a logic for analyzing a similar game in which Demon deletes all the local nodes of a certain property instead of links. For Poison game, which we have mentioned, (Blando et al. 2018) provides three languages of different expressive power to characterize the game play. All of the three language take inspiration from memory logics (Mera 2009) of the hybrid tradition. (Thompson 2018) proposes a scenario called Boolean network games, combined with a logic of local fact change that can characterize Nash equilibria. All of these studies provide new ways of looking at the interaction between graph games, network games and logics of control.

For more logical systems inspired by the dynamic-epistemic tradition, see ‘graph modifier logic’ (Aucher et al. 2009), swap logic’ (Areces et al. 2014) and ‘arrow update logic’ (Areces et al. 2012; Kooi and Renne 2011). To see how the study of sabotage modal logic infuses with other areas of study, for theoretical computer science see (Grüner et al. 2013; Radmacher and Thomas 2008), for learning theory see (Gierasimczuk et al. 2009), for logics of social networks see (Liu et al. 2014; Seligman et al. 2013) and for argumentation see (Grossi 2010b,a). For a much broader inspection of analysis and design for graph games in tandem with matching modal logics, see (van Benthem and Liu 2018), which proposes various meaningful new games and identifies general questions behind the match between logic and game.

7 Conclusion

In this paper, we take on a case study of three of the local variants of sabotage game. For the local link-cutting variant, we propose a polynomial-time algorithm for calculating the winning positions for Traveler in the corresponding game. The static nature of the algorithm in turn inspires the substitution of the dynamic operator with a static graded one in the mu-calculus formula in *loc*SML expressing that Traveler has a winning strategy. We provide alternating algorithms simulating the process of the games that run in polynomial time for both local node-deleting sabotage game and poison game to show that the complexity of game solution is within PSPACE. We also describe ways to reduce instances of Quantified Boolean Formula to instances of both of the games. Since it is well established that QBF is a PSPACE-complete problem, it shows that the solution problems of both games are PSPACE-hard. This combined with the upper bound we just found establishes the fact that both local node-deleting sabotage and poison game are PSPACE-complete in terms of game solution. Using existing results that limit the complexity of model checking for the matching logics of both languages up to PSPACE, it is easy to deduce that model checking for these matching logics PSPACE-complete.

In a broader perspective, what these results show more generally is that varying graph games seemingly slightly can have big complexity effects on game solution, and on the matching logics. In light of this fact, a next more general research topic after this paper can be to understand this phenomenon in a more systematic way and to identify which parameters of games and logics crucially affect complexity.

$$\mathcal{F} = (V = \{p, q\},\ \theta = p \to q,\ \Omega_1 = \neg p \wedge \neg p' \wedge (\neg q \leftrightarrow q'),\ \Omega_2 = q \wedge q' \wedge (\neg p \leftrightarrow p'))$$

$$\mathcal{F} = (V = \{p\}),\ \theta = \top,\ \Omega_L = \top,\ \Omega_M = \top)$$

8 Acknowledgement

I thank Johan van Benthem for his invaluable advice throughout the development of this project. I would also like to thank my parents and several teachers and friends, whose company has given me the courage to discover and create, and has made this journey meaningful.

References

Areces, C., Fervari, R., and Hoffmann, G. (2012). Moving arrows and four model checking results. *Logic, Language, Information and Computation*.

Areces, C., Fervari, R., and Hoffmann, G. (2014). Swap logic. *Logic Journal of Igpl*, 22(2):309–332.

Areces, C., Figueira, D., Gorn, D., and Mera, S. (2009). Tableaux and model checking for memory logics. In *International Conference on Automated Reasoning with Analytic Tableaux and Related Methods*.

Areces, C. and van Benthem, J. (2019). The logic of stepwise removal.

Aucher, G., Balbiani, P., Cerro, L. F. D., and Herzig, A. (2009). Global and local graph modifiers. *Electronic Notes in Theoretical Computer Science*, 231(none):293–307.

Aucher, G., Benthem, J. v., and Grossi, D. (2017). Modal logics of sabotage revisited. *Journal of Logic and Computation*, 28(2):269–303.

Balcázar, J. L., Díaz, J., and Gabarró, J. (2012). *Structural complexity II*, volume 22. Springer Science & Business Media.

Blando, M., Areces, Franchesca Zaffora, K., and Carlos (2018). The modal logics of the poison game.

Chen, Y. (2018). Modal logics of definable point deletion.

Gierasimczuk, N., Kurzen, L., and Velzquezquesada, F. R. (2009). Learning and teaching as a game: A sabotage approach. In *International Conference on Logic*.

Grossi, D. (2010a). Argumentation in the view of modal logic. In *International Conference on Argumentation in Multi-agent Systems*.

Grossi, D. (2010b). On the logic of argumentation theory. In *International Conference on Autonomous Agents and Multiagent Systems*.

Grüner, S., Radmacher, F. G., and Thomas, W. (2013). Connectivity games over dynamic networks. *Theoretical Computer Science*, 493:46–65.

Kooi, B. and Renne, B. (2011). Arrow update logic. *Review of Symbolic Logic*, 4(4):536–559.

Li, D. (2018). Losing connection: the modal logic of definable link deletion.

Liu, F., Seligman, J., and Girard, P. (2014). Logical dynamics of belief change in the community. *Synthese*, 191(11):2403–2431.

Löding, C. and Rohde, P. (2003). Model checking and satisfiability for sabotage modal logic. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 302–313. Springer.

Mera, S. (2009). *Modal memory logics*. PhD thesis, Citeseer.

Radmacher, F. G. and Thomas, W. (2008). A game theoretic approach to the analysis of dynamic networks 1 abstract. *Electronic Notes in Theoretical Computer Science*, 200(2):21–37.

Rohde, P. (2005). On games and logics over dynamically changing structures.

Seligman, J., Liu, F., and Girard, P. (2013). Facebook and the epistemic logic of friendship. *Computer Science*.

Stockmeyer, L. J. and Meyer, A. R. (1973). Word problems requiring exponential time (preliminary report). In *Proceedings of the fifth annual ACM symposium on Theory of computing*, pages 1–9. ACM.

Thompson, D. (2018). Local fact change logic. *Manuscript*.

van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge University Press.

van Benthem, J. (2014). *Logic in games*. MIT press.

van Benthem, J. and Klein, D. (2018). Interfaces of logic and games.

van Benthem, J. and Liu, F. (2018). Graph games and logic design.

清華大學 Tsinghua University

浙江大學 ZHEJIANG UNIVERSITY

浙大东方论坛 ORIENT FORUM OF ZHEJIANG UNIVERSITY