From arbitrary sets to social groups:
group agency in multi-agent modal logic

Daniil Khaitovich

From arbitrary sets to social groups:
group agency in multi-agent modal logic

ILLC Dissertation Series DS-2026-11

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Haoran Zhi

Printed and bound by Ipskamp Printing

ISBN: 978-94-6536-223-6

From arbitrary sets to social groups: group agency in multi-agent modal logic

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 29 oktober 2026, te 10.00 uur

door Daniil Khaitovich
geboren te Minsk

Promotiecommissie

Promotor:	prof. dr. S.J.L. Smets	Universiteit van Amsterdam
Copromotor:	dr. A. Özgün	Universiteit van Amsterdam
Overige leden:	prof. dr. A. Betti	Universiteit van Amsterdam
	dr. D. Grossi	Universiteit van Amsterdam
	prof. dr. J.F.A.K. van Benthem	Universiteit van Amsterdam
	prof. dr. O. Roy	University of Bayreuth
	prof. dr. V. Goranko	Stockholm University
	dr. B.R.M. Gattinger	Universiteit van Amsterdam
	dr. Z.L. Christoff	Rijksuniversiteit Groningen

Faculteit der Geesteswetenschappen

Contents

Acknowledgments **1**

1 Introduction **3**

2 Philosophical background **7**
2.1 Group agency: Bratman's account 10
2.2 Team reasoning 21
2.3 Envoi 28

3 Logical background: multi-agent modal logics (MAMLs) **31**
3.1 Multi-agent epistemic logics 35
3.2 Multi-agent modal logics of agency 40
3.2.1 Coalition logic 40
3.2.2 Group STIT logic 46
3.3 Envoi 57

4 Logics of collective intention: Bratmanian version **59**
4.1 Setting the stage: the logic of conditional intention 60
4.1.1 Philosophical theories of conditional intention 61
4.1.2 Logical challenges 64
4.1.3 Formal semantics 69
4.1.4 Logic, soundness, completeness 74
4.1.5 Time, action, and and side-effects 79
4.2 Simpler multi-agent framework 95
4.2.1 Formalizing participatory intentions in $\mathcal{L}^{I}_{CL_\alpha}$ 101

4.2.2 Envoi 110

5 Collective intentions and decision problems **113**
5.1 Problem-sensitive intention: single-agent case 114
5.1.1 Logics of Intension and The Problem of Side-Effects 115
5.1.2 Formalising decision problems as partitions 118
5.1.3 Hyperintensional logic of intention 122
5.2 Problem-sensitive collective intentions 129
5.2.1 Logic 136
5.2.2 Envoi 138
5.2.3 Appendix: soundness and completeness of Log 138
5.2.4 Appendix: soundness and completeness of Log_G 144

6 Fixing MAMLs: the case of STIT logic **151**
6.1 Group STIT logic and its technical problems 152
6.2 Plan-restricted group STIT logic 156
6.3 Metalogical results 162
6.3.1 Relativisations and plans 163
6.3.2 Temporal group STIT logic with disjoint groups 173
6.3.3 Atemporal STIT with disjoint groups 199
6.3.4 Discussion: is our definition of plans too weak? 204
6.4 Solving coordination problems via communication 204
6.4.1 Coordination in deontic games: preliminaries 205
6.4.2 Weakening the announcements 209
6.4.3 Formal language, semantics 210
6.4.4 When semi-private communication is enough: formal proof 216
6.4.5 Envoi 218
6.4.6 Appendix: proofs for Section 6.4 219

7 Conclusion **225**

Bibliography **229**

Abstract **246**

Samenvatting **247**

Acknowledgments

I believe that the greatest way to acknowledge others is to save them some time, so I will do my best to stay laconic. First and foremost, I thank my supervisors, Aybüke Özgun and Sonja Smets. Aybüke taught me a great deal about doing formal philosophy, on the level of both the technicalities and the phronesis of this endeavour. The hours we spent in front of the whiteboard wrestling with messy proofs while discussing logic and epistemology have shaped most of the work done in this thesis. Hundreds of pages of my barely readable and mostly nonsensical drafts would have never turned into any significant piece of work, had they not been read and commented on by Aybüke. Sonja was always there for advice, helping to strategise about the optimal research direction, navigate the bureaucracy, arrange research visits, and choose venues for publications. I am grateful for their wise guidance throughout these four years.

I express my gratitude to the committee members – Arianna Betti, Davide Grossi, Johan van Benthem, Olivier Roy, Valentin Goranko, Malvin Gattinger, and Zoé Christoff – for dedicating their precious time to reading and evaluating this thesis.

Warm thanks go to the faculty, staff, and students at the ILLC for creating such a friendly and productive research environment. Ever since joining the ILLC, whenever I think of Pythagoreans and their communities, I envision them at SP107. I am particularly grateful for the Amsterdam Dynamics Group and the LIRa team: Alexandru Baltag, Sonja Smets, Malvin Gattinger, Aybüke Özgun, and Mina Young Pedersen. Helping with the technical setup and some errands around the LIRa site was the least I could have done for the seminar. Not a single LIRa talk I have attended has been unremarkable, so I am also grateful for all the speakers we have had. Thank you to Alexandra Zieglerová, Ewout

Arends, Peter van Ormondt, and Minghui Wu for helping my colleagues and me with a variety of administrative stuff: this vital work often stays invisible and does not get as much credit as it deserves. Thanks to my fellow promovendi with whom I have crossed paths either as office mates or on other occasions: Tianyi Chu, Haitian Wang, Woxuan Zhou, Fei Xue, Sophie Nagler, Soeren Knudstorp, and Tomasz Klochowicz. I wish them success with the research and graduation process, although I am sure they will have it in virtue of their work independently of my redundant wishes.

Thanks to my coauthors and collaborators: Hein Duijf, Jan Broersen, Lorenz Hornung, and Valentin Goranko. Without them, I would have been engaged in a dubious enterprise: studying collective reasoning in total solitude. I also express my gratitude to Agi Kurucz and Balder ten Cate, who helped me with several questions about cylindric algebras and product modal logics via email correspondence.

I am grateful to my former teachers and colleagues: Pavel Barkouski – for helping students with their first steps in philosophy and actively supporting the student strike in 2020; Anton Prakapchuk – for sharing the path throughout the years, staying faithful to the discipline no matter what; Vitaliy Dolgorukov, Elena Dragalina-Chornaya, Konstantin Shishov, and Elena Popova – for the Promethean effort of maintaining formal philosophy in Russia against growing obscurantism.

Outside of my academic life, I thank Boris, Wolf and Lennart Vaders for their warm friendship and the perfect exemplification of *Heldhaftigheid, Vastberadenheid en Barmhartigheid*. Hartelijk bedankt voor jullie geduld om het Nederlands met mij te praten, ondanks mijn lelijke taalgebruik. Thanks, Paul Baneke, for intellectually stimulating chats, although they appeared annoying on the surface for both of us. Thanks, Aneesa Razak, for teaching me not to be so hateful. Thank you, Leonid Kotelnikov, for all the disastrous adventures and graceful failures. Thanks, Haoran Zhi, for designing the cover for this thesis and tolerating my pointless sombre monologues once in a while. Thanks, Robert Hartley, for all the curiosities of Dutch and Scouse, as well as wild vegan soups.

Thanks to all the bacteria that inhabit my body, corrode my teeth and help me to digest food. Thanks to all the animals whose flesh I have eaten. Thanks to all the living and non-living, natural and artificial agents whose contribution I failed to register and acknowledge. Thank you, Grigoriy Khaitovich and Galina Bolshakova, for the unsolicited, grotesque yet wonderful gift of life.

Daniil Khaitovich
August, 2026.

Chapter 1

Introduction

In early 2021, *GameStop*, a slowly dying gaming retail company, rose sharply in price: its stock that cost $20 skyrocketed up to $400 at its peak. There was no financial stimulus for people to begin investing in GameStop: such behaviour was mostly driven by ironic discussions and jokes on the Reddit community r/WallStreetBets. The question arose: was it a market manipulation or just an instance of simultaneous economically irrational behaviour of individuals? The U.S. Securities and Exchange Commission had a hard time investigating this case and did not find any evidence of an organised illegal collusion [Man21].

In 2023, Russia's Supreme Court recognised the "international LGBT movement" as an "extremist organization". Participating in such extremist organizations, as well as collaborating with them, is severely punishable: according to Russian criminal law, one can get up to 12 years in prison for such offenses. However, the organization that the government outlawed, the "international LGBT Movement", simply has never existed! The Supreme Court decision allowed persecution of arbitrary LGBT people and their advocates [Hum23].

These two cases demonstrate that the difference between random sets of agents and groups is very important, especially when we try to hold someone responsible as a group member. Much ink has been spilled across various disciplines to define what makes sets of agents social groups. Nevertheless, logicians seem to mostly overlook the problem. Several group notions are studied logically: group forms of knowledge and belief [Fag+04], group strategic abilities and actions over time [AHK02], collective preference [ÅHW11], collective obligation and responsibility [Gro+04], and the list goes on. Strikingly, in most of such formalisms, a group is defined as nothing but an arbitrary set of agents. Using the most popular

formalisms, we must treat all Russian LGBT people, all buyers of GameStop stocks, all left-handed people, the king of the Netherlands, together with the reader of this text, and the union of all aforementioned people as groups. We have the formal tools to reason about such groups' common knowledge, their strategic abilities, collective preferences, or even attribute collective responsibility to them. As we will show, not only is it conceptually questionable, but it also creates unnecessary mathematical complexity in the mentioned formal systems and related decision procedures. Briefly, given that you have α many agents and every set of agents is a group, you have 2^α-many groups. As the old Cantorian wisdom teaches us, 2^α is a lot.

These simple concerns create a starting point of the thesis. We face two main tasks: come up with a plausible definition of a group and group agency, and properly introduce it to the formal logics of different group notions. It may sound innocent, but (as we hope the reader will see throughout the text) these tasks are far from being trivial.

What follows is an overview of each chapter of the thesis.

Chapter 2 introduces the philosophical background we rely upon throughout the thesis. Namely, we provide a birds-eye view on philosophical theories of group agency. After making a distinction between methodologically individualistic and holistic theories, we closely examine a representative of each camp: Michael Bratman's theory of shared intention in Section 2.1 and Michael Bacharach's theory of team reasoning in Section 2.2.

Chapter 3 introduces the logical background: multi-agent modal logics. There are two aims of this chapter. Firstly, we introduce the kind of formalisms with which we work throughout the thesis, fixing the notation and terminology. Secondly, we demonstrate how groups are defined in them (spoiler: as any set of agents) and to what kind of both conceptual and formal problems this definition leads.

In the next two chapters, we develop logics of collective intention, a key notion for us to distinguish groups from random sets of agents. We do not take sides in the philosophical debate of methodological individualism versus holism, developing two logics of collective intention, both having different philosophical assumptions behind them. So **Chapter 4** and **Chapter 5** are dedicated to formalisation of one of the philosophical theories of collective intention – individualistic theory by Bratman and holistic theory by Bacharach, respectively. In Section 4.1, we set the stage by developing a logic of conditional intention. In Section 4.2, we extend this logic to the multi-agent case and formalise Bratmanian theory in it. Namely, we formally prove key results about ontological, conceptual and normative conservativity of the theory of shared intention over the theory of individual (conditional) intention. Chapter 5 has a similar structure: we start

by developing a single-agent theory of problem-sensitive intention in Section 5.1, extending it to the multi-agent case with groups in Section 5.2.

In **Chapter 6**, we make use of a more elaborate definition of groups: a set of agents is a group iff their actions are guided by a shared motivational attitude. We show how this definition may be integrated in group STIT logic, and which conceptual and formal advantages it brings. First, we delve into the negative results about group STIT logic's axiomatisability and decidability. Then, in Section 6.2, we develop plan-restricted grup STIT logic, where standard group STIT is enhanced with a refined definition of groups. We show what conceptual advantages it brings via the use-case of collective responsibility attribution. Section 6.3 is devoted to positive metalogical results about (fragments of) plan-restricted group STIT logic. Section 6.4 is dedicated to how agents may coordinate on their agreed-upon joint plans via communication, and what are the minimal requirements on communication channels that will make such coordination possible. Finally, we conclude and contemplate over future work directions in Chapter 7.

Sources used in the thesis

Section 4.1 is based on:

> Daniil Khaitovich. "Logic of conditional intention". To appear in Studia Logica (accepted pending minor revisions).

Section 5.1 is based on:

> Daniil Khaitovich and Aybüke Özgün. "Hyperintensional Intention". Proceedings of TARK 2025. Edited by Adam Bjorndahl. EPTCS. Waterloo, Australia: Open Publishing Association, 2025, pages 44–60.

Daniil Khaitovich developed the idea for the paper and initiated the project. Then, the idea was refined in discussion with Aybüke Özgün. Both authors contributed to writing and revising the paper.

Sections 6.2 and 6.3.3 of Chapter 6 are based on:

> Daniil Khaitovich. Plan-restricted group STIT logic. Under review.

Section 6.4 is based on the extended abstract of:

> Daniil Khaitovich, Hein Duijf, Jan Broersen. Solving coordination games with minimal communication.

which was peer-reviewed and accepted for presentation at the *16th Conference on Logic and the Foundations of Game and Decision Theory*. The project was initiated by Daniil Khaitovich during the research visit at Utrecht University. Hein Duijf and Jan Broersen participated in discussing the conceptual grounds

for the paper. The paper was written and finalised by Daniil Khaitovich, after implementing coauthors' comments.

Authorship orders of the co-authored papers are decided according to the contributions of the authors. Daniil Khaitovich is the first author in all co-authored articles that are presented in this dissertation. The rest of the thesis is based on the unpublished work by Daniil Khaitovich.

Chapter 2

Philosophical background

In this chapter, we map the main philosophical positions with respect to the nature of social groups and their agency. The aim of this chapter is twofold: we want to familiarise the reader with several philosophical notions about group agency that we constantly evoke further in the thesis, and show the variety of plausible ways how social groups can be defined.

The questions about the nature of social phenomena puzzling us are in no way novel: see the discussion on political nature of humans in the first book of Aristotle's *"Politics"* [Ari13] and on the nature of language in Plato's "*Cratylus*" [Ree97]. For a concise historical overview of the philosophical inquiry about the nature of sociality, see the corresponding supplement to [Eps16].

To understand when we can attribute agency to a set of individuals, we need a working definition of agency in the first place. Following [LP11, p.20], we regard any system with the following features as an agent:

1. The system has representational states that depict how things are in the environment;
2. The system has motivational states that specify how it requires things to be in the environment;
3. It has the capacity to process its representational and motivational states, leading it to intervene suitably in the environment whenever that environment fails to match a motivating specification.

For example, humans are agents, since we have mental states: knowledge and belief are representational, whereas intentions and desires are motivational. These

mental states guide our actions (or at least that is what the classical causal theory of agency claims [Dav01]). Consequently, we can attribute agency to some group of people if their behaviour is caused by some collective forms of representational and motivational states. Now our task is clearer: we need to define collective forms of representational and motivational states and explain how they can cause group actions.

Before ascribing any representational and motivational states to groups, we need to understand groups' ontological status. Should we postulate the existence of groups, at least partly autonomous from the existence of individuals that constitute them? E.g., does the kingdom of the Netherlands even exist, or are there only ∼18 million of Dutch citizens? There are two opposing positions regarding that question. The most common position is *ontological individualism*. According to ontological individualists, groups as such do not exist separate from their members. Only individual agents are first-class citizens. The opposite view, ontological *holism*, postulates that a group as a whole may be more than a sum of the individuals as its parts. Therefore, groups do exist "over and above" their members. In the most radical forms, ontological holism may even claim groups as first-class citizens: first, there exist civilizations, classes, objective spirit, and so forth, where the individuals are byproducts of such group entities. In the modern social sciences, ontological individualism is widely accepted.[1]

On top of the ontological question about the status of groups (which is considered mostly settled by modern-day scholars), there is a methodological one. Even if groups lack autonomous existence, should we explain group agency in terms of the actions and minds of their members? E.g., even if the kingdom of the Netherlands is nothing but ∼18 million people with Dutch passports, when we say "*the kingdom of the Netherlands wants to prevent North Sea floods*", can we rephrase this claim in terms of Dutch citizens' properties without significant alteration in meaning? According to *methodological individualism*, every meaningful claim about groups – including ascription of representational or motivational states – is an acronym for a claim about the group's members, so such rephrasings are indeed possible:

> *We do, of course, speak freely of the mental properties and acts of a group in the way we do of individual people. Groups are said to have beliefs, emotions, and attitudes and to take decisions and make promises. But these ways of speaking are plainly metaphorical. To ascribe mental predicates to a group is always an indirect way of ascribing such predicates to its members.* [Qui75, p.17]

[1] For a detailed overview of ontological individualism, see, e.g., [Eps14, Chapter 1]. A good overview of ontological holism can be found in, e.g., [She03]. A detailed, engaging introduction into the field of social ontology and its relation to social science can be found in [Tol02].

Methodological holists, on the other hand, argue that at least some group phenomena cannot be explained in individualistic terms. Proper explanations will require some irreducible collectivity. For example, as it was argued, some ascriptions of collective preferences, beliefs, and obligations cannot be translated into the language of the corresponding attitudes of individuals [LP02; DTP21; Lac21; Arr63]. We will encounter some of these impossibility results in the later sections.

An answer to the ontological question does not determine the answer to the methodological one. While the vast majority of modern scholars are ontological individualists – it is nowadays rare to encounter ideas à la Hegelian universal spirit – some of them are still methodological holists. The gap between what objects constitute our social life and in what terms we explain social phenomena seems counterintuitive. If nothing but individuals exist, why do we need collective terms to explain social phenomena? Methodological holists often compare the relation between individual agents and groups with the one between cognitive processes and human physiology. It is a common ground that humans consist of nothing but biological cells. Yet it is doubtful that, e.g., psychological claims about human behaviour or philological claims about the interpretations of texts written by humans are nothing but metaphors for claims about the cells that constitute involved human bodies.

We list some notable representatives of both individualistic and holistic camps, and we start with the former. Max Weber's theory of social action is widely considered a starting point of methodological individualism in the social sciences [Web13]. Karl Popper [Pop02], together with his student John Watkins [Wat52], also explicitly defended individualism, which allowed them to argue in favour of the unity of method across both natural and social sciences. Individualism is at home in the orthodox rational choice theory, where only individual agents, their preferences, and beliefs are first-class citizens [Opp20]. Among the modern-day philosophers of group agency, Michael Bratman's theory of *modest sociality* [Bra13] is one of the most prominent individualistic accounts.

As for holism, the canonical legacy reference point is Emile Durkheim, who stated his methodological position as follows:

> [...]*determining cause of a social fact should be sought among the social facts preceding it and not among the states of the individual consciousness.* [Dur95, p.110]

Among the modern-day authors, the holistic explanations in social epistemology are defended, among others, by Margaret Gilbert [Gil13], Raimo Tuomela [Tuo13], John Searle [Sea95], and Deborah Tollefsen [Tol02]. In decision theory, a respectable heterodoxy emerged – *team reasoning* – which challenges the mainstream individualistic assumptions of the field [Bac06]. This framework has a strong connection with the aforementioned holistic philosophical approaches,

and, as argued in [GS07], is generally compatible with them.

The body of literature on the nature of sociality is enormous and keeps growing, so we cannot address all the existing theories of group agency. We pick one representative for both individualistic and holistic camps: Michael Bratman's theory of modest sociality [Bra13] and Michael Bacharach's theory of team reasoning [Bac06], respectively. This choice is not completely arbitrary. Both authors share the same first name, but this is not the main grounding of our choice. As we are to show, the theories are mostly compatible with other representatives of their camps.

2.1 Group agency: Bratman's account

Bratman is concerned with the shared activity of small groups without any internal hierarchy and with the stable set of members: think of singing duos, small friend groups, coauthors of research papers, and such. The joint intentional activity of such groups constitutes what Bratman calls *modest sociality*. The theory does not aim to explain more complex instances of group agency, such as actions of hierarchically organised corporations, whose set of members can change over time. For example, a development company working on the construction of a building is not involved in modest sociality, while two friends building a treehouse are.

The theory of modest sociality is conceptually, metaphysically, and normatively conservative over theories of individual agency. Neither should we use anything but concepts of individual agency and practical reasoning, nor should we postulate the existence of new, irreducible entities, nor should we impose additional new norms of "collective" practical reasoning to build the theory of modest sociality. Therefore, Bratman's proposal is methodologically individualistic.

Since Bratman is concerned with joint *intentional* actions of groups, the main task is to explain what a *shared intention* is. In Bratman's theory, shared intentions are groups' motivational states, while common knowledge, standardly understood,[2] fulfills the role of a representational state. Following the conservativity principle, the definition of shared intention is given in terms of individual attitudes of the group's members. Bratman's account of a shared intention can be presented as follows.

2.1.1. Definition (Shared intention). Without loss of generality, we take a group of two agents: A and B. According to Bratman [Bra13, Chapter 2], A and B intend to φ together iff:

[2]It is commonly known that φ iff everybody knows that φ, everybody knows that everybody knows that φ, and so on. This standard definition was independently given by David Lewis [Lew08] and Robert Aumann [Aum76]. We address common knowledge in detail in Section 3.1.

1. A intends *that they* φ and B intends *that they* φ;
2. A and B each intend that they φ by way of the intentions of each that they φ;
3. A and B each intend that they φ by way of meshing sub-plans of each of their intentions in favour of φing ;
4. It is commonly known among A and B that (1-3).

From now on, we will call the individual intentions of A and B that constitute their shared intention *participatory*.

We unpack some vague spots in this definition. A and B's intending *that they* φ is an instance of intention *that*. In other words, the object of such an intention is not an action of the intender, but rather a proposition about the world; it is an intention that the world would be a certain way, namely, such that they will φ.

A and B's intending that they φ *by way of the intentions of the other* means that both A and B regard one another as necessary co-participants in their intended shared activity. It presupposes certain beliefs about the intentions and actions of one another and, as we are to show later, the conditionality of A's intentions on the intentions of B and vice versa. By *subplans* of A and B we mean A and B's individual intentions connected with doing their parts of φing. The subplans mesh iff they are mutually consistent with their φing.

We borrow Bratman's example to clarify this definition [Bra13, p.41]. Assume A and B share the intention to go to New York together. First of all, both A and B intend that they go to New York together, as in (1). It is not sufficient: e.g., two criminals may each intend to kidnap the other and forcefully drag her to New York – so, each of the criminals intends that they will go to New York – which is obviously not an instance of joint intentional activity. A and B need to recognise one another as intentional companions in their trip to New York. Therefore, as in (2), A intends that they go, because B intends likewise, and vice versa. Finally, A and B should have some derivable intentions about doing their parts of the trip: to buy tickets, to pack their belongings, to get to the train, etc. Such intentions cannot be inconsistent with their joint trip. For example, it should not be so that A intends to get on a train, while B intends to take a bus. Therefore, as in (3), each A and B should intend that they go to New York by way of their meshing sub-plans to go to New York. Finally, all of the above-mentioned should be commonly known to A and B, as stated in 4. Assume otherwise. For instance, A has a secret intention to go to New York with B. A intends to accept B's invitation to go to New York in case B will invite her. At the same time, B has the mirroring secret intention to go to New York with A. B intends to invite A, in case she will accept Bs invitation. Unless A and B establish common knowledge about their secret longings, they cannot share the

intention to go to New York.

Bratman's account faces two major criticisms. According to the first one, the definition appears to be circular. We explain what it means for A and B to perform a shared intentional action of φing together by the individual intentions of A and B that they φ. But does not the object of their intentions – them φing – already presuppose that they do so as a shared intentional action? Note that the object of the participatory intentions is nothing but a proposition about the joint action. The proposition "A and B will φ" is only about A and B's actions, not mental states. That proposition by itself does not entail that A and B share the intention to φ together. For example, the building blocks of the shared intention of A and B to go to New York together consist of A and B's individual intentions *that they go to New York together*. Therefore, the object of their intention is the following proposition: *we will go to New York together*. From this proposition itself, it does not follow that they will go to New York together intentionally. There is nothing said about A and B's mental states or attitudes at all; we only talk about A and B's actions ontically. Therefore, there is no circularity in the definition of the shared intention so far.

According to the second line of criticism, Bratman's account breaks the *control constraint*: if one intends that φ, then one should (believe to) have control over whether φ holds. Therefore, A and B cannot simultaneously intend that *they* φ without breaking norms of rationality [Vel97]. For example, assume that A and B intend to dance together. Then, A intends that A and B will dance together. By the control constraint, A should be able to settle whether A and B will dance together or not. This might be true: maybe A has control over B's will. But B intends that A and B dance together as well! Hence, B should have control over whether they will dance or not, which contradicts A's ability to settle whether they will dance or not. This seems to be inconsistent.

Bratman responds to this worry by highlighting *interdependence* of participatory intentions that follow from (2). By interdependence of A and B's intentions, we mean that A intends that they φ as long as B intends that they φ, and vice versa. Norms of rationality will force both A and B to act in accordance with their intentions. Therefore, as long as both of them intend that they φ, each of them is rationally pressured to do her part of φing. Given that these intentions are commonly known among A and B, each of them believes that the other one will do her part of φing; therefore, each believes that it is totally in her control to settle whether they will φ or not by either reciprocating or not:

> [T]*hese intentions of each are interdependent: other things equal, each will continue so to intend if, but only if the other continues so to intend. Suppose, more precisely, that there is this interdependence because each will know whether or not the other continues so to intend, and each will adjust to this knowledge in a way that involves respon-*

> *siveness to norms of individual plan-theoretic rationality. Call this persistence interdependence.* [Bra13, p.66]

To conclude, Bratman responds to the concern around the control constraint as follows. Participatory intentions of A and B that they φ do not contradict the control constraint, since neither A nor B intends that they φ no matter what. A can settle whether they will φ or not, given that it is known that B cooperates, and analogously for B. Therefore, given that their participatory intentions are interdependent and commonly known, they do not contradict the control constraint. The interdependence of participatory intentions appears to be problematic, since we simultaneously conditionalise the intentions of agents on one another. To demonstrate it, consider the next example.

Two generals prepare for a coordinated attack on the common enemy. In case only one army attacks, it is doomed to be defeated: the joint, simultaneous effort is necessary to win the battle. The generals can communicate to establish common knowledge about their intentions. Neither generals ever fall under the weakness of will: given that they have a conditional intention[3] and they know the condition of their intention is satisfied, they act upon this intention. Moreover, they are cautious: unless the condition of their intention is satisfied, they will never act upon it.

They come to commonly know the following:

1. The first general intends that they jointly attack, given that the second intends that they jointly attack.
2. The second general intends that they jointly attack, given that the first general intends that they jointly attack.

For convenience, we will rewrite these statements in a pseudo-language as follows:

$$\alpha := Int_1^{(Int_2 attack)} attack$$
$$\beta := Int_2^{(Int_1 attack)} attack$$

where for every $i \in \{1, 2\}$, $Int_i attack$ reads as "*The ith general intends that they attack*, and $Int_i^{\psi}\varphi$ reads as "ith general intends that φ conditional on ψ".

It seems that $\alpha \wedge \beta$ will not force the generals to attack. The condition of the intention of the first general is not satisfied: by α, it is required that $Int_2 attack$, but the second general does not intend that they attack: he only intends that *conditionally*, as in β. Analogously, the second general will not attack, since the first general has only a conditional intention to attack. Therefore, it seems

[3]For now, we rely on the everyday, pre-theoretical understanding of what conditional intention is. An elaborate theory of conditional intention will be developed in Section 4.1.

that they need to reformulate their intentions and conditionalise them on the conditional intentions of one another. The next day, they establish common knowledge about the following:

$$\alpha' := Int_1^{\beta} attack$$
$$\beta' := Int_2^{\alpha} attack$$

They still will not attack, since they face the same problem. The first general does not attack, since the condition of α' requires the second one to intend to attack conditionally on his intention, i.e. $\alpha' := Int_1^{\beta} attack$, while in reality, the second general intends to attack only conditionally on his *conditional* intention, i.e. $\beta' := Int_2^{\alpha} attack$, and vice versa. Obviously, this could go on forever: neither $\alpha'' \wedge \beta''$ would work, nor $\alpha''' \wedge \beta'''$, etc. It seems that we might solve the problem by adopting asymmetrical conditional intentions. For example, one may consider the following:

$$Int_1 attack \wedge (Int_2^{(Int_1 attack)} attack)$$

Then, the antecedent of the second general's conditional intention is exactly the intention of the first general. But then the intention of the first general is not conditional and genuinely breaks the control constraint. Analogous arguments can be made about $\alpha' \wedge \beta$, $\alpha' \wedge \beta''$, $\alpha'' \wedge \beta'$, etc.: if the depth of conditionalization of one general is one or more levels higher than the other, the intention of one of them will break the control constraint.

So it seems that interdependence presupposes the vicious cycle: I intend that we φ, if you intend that we φ, while you intend that we φ, if I intend that we φ, but neither of us can simply intend that we φ, so the conditions of our interdependent intentions are never satisfied.

Bratman briefly addresses this issue, defending the point that participatory intentions in question do not have to be conditional at all. Namely, both generals may simply intend that they attack: such intentions will not break the control constraint, since both generals have as a background of their reasoning that the other intends that they attack. Then, another question arises: given that each may rationally have such an intention if and only if the other has the mirroring one, how can they simultaneously come up with them?

Bratman argues that there might be public events that provoke such participatory intentions:

> *For example, at the end of a wonderful concert each of us might intend that we applaud together. That this was a wonderful concert is, among us, what Robert Stalnaker calls a "manifest event": it is an event "that, when it occurs, is mutually recognised to have occurred." This supports my confidence, given our common knowledge about the*

> *kind of people who attend such concerts, that you also intend that we applaud together. And similarly for you. Each of us arrives at his intention that we applaud primarily in response to the performance and its manifest quality, together with relevant prior common knowledge about the kind of person who attends such concerts.* [Bra13, p.68]

There are two ways to interpret such manifest events. The stronger interpretation is to view them as public announcements.[4] E.g., when the wonderful concert happens, it is publicly announced that the wonderful concert has happened. Therefore, after the concert, it will be commonly known that the concert is wonderful. It seems natural to interpret the "common knowledge about the kind of people who attend such concerts" as common knowledge that if the concert is wonderful, each attendee will intend to applaud. The weaker interpretation, following Stalnaker's original account of manifest events, will not lead to common knowledge, but rather to common belief that the event has happened.[5]

The stronger interpretation leads to the inconsistency of Bratman's story. If, as a result of the manifest event, it is publicly announced that the concert was wonderful, then, given the common knowledge about the nature of attendees, it is commonly known that everyone intends that they all applaud. But then each agent cannot really arrive at her intention *in response* to such an event, since after the event, it is already commonly known what she intends! If we assume that some agent, in response to the manifest event, can form the intention to refuse to applaud, then such an announcement cannot be truthful. For example, assume that the wonderful concert happens, and it is publicly announced that the concert was wonderful. Hence, everyone gets to know that everyone else intends to applaud. As a response to that event, someone decides not to applaud. Then, the knowledge others have gained after the announcement – everyone, including the attendee who decided not to applaud – intends that they all applaud, which is false! So either the manifest event may not be a truthful public announcement, or the attendees do not have a choice but to arrive at their intentions to applaud as a response to that event, which seems to contradict the very idea of intentional agency.

[4]By public announcement, we understand a true message that is sent to everyone publicly, i.e., everyone gets the message and knows it to be true, knows that everyone got the message and knows it to be true, knows that everyone knows that everyone got the message and knows it to be true, etc. "A public communication can be imagined as a statement made in a conference room in which all agents are present. If at a moment t agent i starts a public communication with a message α, then at moment $t+1$ it becomes common knowledge of all agents that i knew that α was true at t" [Pla07, p.170].

[5]In [Sta02], Stalnaker argues that manifest events result in establishing common belief that the event happened. By the common belief of a and b that φ Stalnaker means that a believes that φ and b believes that φ; a believes that b believes that φ and b believes that a believes that φ, and so on. In the same paper, Stalnaker clarifies that by the belief that φ he means not necessarily true, but consistent, positively and negatively introspective attitude.

Apparently, the problem might not be in the manifest event itself, but in our interpretation of the common knowledge about the kind of people who attend the concert. Namely, we interpret it as the knowledge about what everyone will intend under the condition of the external event. We can relax this interpretation so that agents will not definitely know what others will intend. E.g., it is commonly known that everyone *most likely* will intend to applaud the wonderful concert. But these interpretations will conflict with the condition (4). To share the intention to applaud, attendees need to commonly know that each intends that all applaud, not that everyone is *most likely* so to intend.

If we take the weaker interpretation of manifest events as grounds for only common belief, then Bratman's story is consistent, but the kind of information everyone receives from the manifest event is not enough to form a shared intention. Again, by (4), common *knowledge* about the participatory intentions of everyone is required. But manifest events result only in common belief, which is weaker than common knowledge. For example, assume once again that there are two attendees of the concert: $Ags = \{a, b\}$. For simplicity, assume it is commonly known that everyone will applaud the concert if the concert is wonderful. Then, the wonderful concert happens, which results in the common belief that the concert was wonderful. In other words, both a and b believe that the concert was wonderful. Both a and b consider it highly unlikely but possible that the concert was not that wonderful after all. a believes that b believes that the concert was wonderful; a also considers it highly unlikely but possible that b does not believe that the concert was wonderful. Analogously, b believes that a believes that the concert was wonderful, and so on. Given that a considers it possible that b does not believe that the concert is wonderful, she does not know whether b intends to applaud, but only considers it most plausible that b does. Hence, there is no common knowledge among a and b that everyone intends that they all applaud.

So the stronger interpretation of the manifest is inconsistent with the attendees being intentional agents, while the weaker interpretation is inconsistent with the common knowledge condition. Therefore, we proceed to interpret participatory intentions as conditional. The interpretation is not ad hoc. Even Bratman admits that in some cases, the participatory intentions might be conditional:

> *Some such interactions may also involve a prior stage-setting. You might set the stage by forming and announcing a conditional intention that we J if I nonconditionally intend that we J. And I might recognise this and so form and announce my nonconditional intention that we J, fully confident that this will lead you also to such a nonconditional intention.* [Bra13, p.75]

Moreover, the way Bratman talks about interdependent unconditional intentions basically coincides with how other authors define conditional intentions. For example, one of the forms of interdependence Bratman talks about is *feasibility-*

based interdependence:

> *Consider, for example, two gang members, Alex and Ben. Each thinks that their going together to NYC would be desirable, and each intends that they do indeed go together to NYC. Each recognises that given the complexities, and given the limits on the power of each, this joint activity is not going to happen, other things equal, unless both of them so intend, though it will indeed happen if each does so intend. [...] Each sees the other's intention as only a feasibility condition – and not a desirability condition – for what he intends, namely that the two of them go to NYC together.* [...] *Call this feasibility-based interdependence.* [Bra99, p.71]

The formulation virtually corresponds to the definition of *preconditional* intention in [Fer09]. Oftentimes, Bratman describes the participatory intention as unconditional, given that the knowledge about others intending likewise is in the background of the reasoning of the intender. Such cases are also defined as conditional in their *"deep structure"* by Ferrero:

> *The condition C is not erased from the content of the intention; it is rather pushed in what might be called the 'cognitive background' of the agent's plans and projects. This 'cognitive background' is comprised of the set of all the circumstances relevant to the possibility and reasonability of an agent's intentions that the agent takes to obtain either now or at any time prior to the completion of her plans. Once a condition C is moved into the cognitive background, there is usually no longer a need for the agent to entertain the condition explicitly in his practical reasoning. The agent might thus proceed to reason as if the intention were unconditional. In a similar fashion, there is often no need to make the condition explicit in the expression of one's intention, at least in so far as one's audience is expected to believe that C obtains as well.* [Fer09, p.709]

When taking participatory intentions to be conditional, we are not alone in our interpretation. David Velleman, commenting on the earlier version of Bratman's theory and formulating the worry about the control constraint in [Vel97], also proposes to view participatory intentions as conditional. Velleman reduces conditions of participatory intentions to the form "if the others intend *likewise*":

> *I suggest that the statement 'I will if you will' should be understood in this sense. It means, 'I hereby frame an effective intention that's conditional on your framing an effective intention as well' – that is, 'I hereby will it, conditional on your willing likewise'.* [Vel97, p.45]

The same line of thought is continued in [Rac24], where it is clearly proposed to redefine a shared intention as an aggregation of interdependent conditional

intentions:

> *And it has been suggested, although not widely accepted, that we can construct collective intentions by way of combinations of individual conditional intentions towards an action of the same kind.* [...] *My aim here is to defend the coherence of this important kind of collective intention by developing that construal and response.* [Rac24, p.272]

Both [Vel97] and [Rac24] acknowledge the problem of the vicious circle in the conditions of participatory intentions and respond in a similar way. Namely, Velleman and Rachar formulate the participatory intentions in the indeterminate, gappy form of "I intend that φ, in case you intend *likewise*". This move leads to the indeterminacy of the participatory intentions' content. Namely, suppose that the first general intends to jointly attack, if the second intends likewise; the second general intends to jointly attack, if the first intends likewise. If we look at the intention of the first general, to compositionally determine the meaning of his intention, we need to evaluate the condition, i.e., whether the second general intends likewise, which refers us to the intention of the second general. But when we try to evaluate the meaning of the second general's intention, we will be referred back to the first general's intention by the condition of his intending likewise.

To simplify, such participatory intentions may be formalised in our pseudo-language as follows:

$$\alpha := Int_1^{\beta} attack$$
$$\beta := Int_2^{\alpha} attack$$

To determine what α means, we need to determine the meaning of β, which will refer us to back to α, without any termination. In [Vel97, p.45] and [Rac24], the arguments are presented as to why agents may hold intentions with such indeterminate content, appealing to the similarity between the undetermined content of both intentions or to the theory of *gappy* propositions:

> *Each person is aiming to frame intentions similar to the other person's, and if we take 'likewise' to mean 'determinate and indeterminate in the same way, if any', when they frame participatory intentions as proposed here, they succeed. They succeed since 'it is perfectly determinate whether either person's intention has the same potential indeterminacies as the other's' [Vel97, fn.26]. That is, the intentions in our case do have the same determinacy and indeterminacy. The determinate part of the content is the collective behaviour, which is the same in both, and the indeterminate part is the self-referentiality in the antecedent, which is the same in both.* [Rac24, p.282]

Such a solution, although by itself might be coherent, is inconsistent with the core of Bratman's theory. The intentions should not only be interdependent, but, following (3), they should also be consistent with the connected individual subplans. How can one track the consistency of their individual subplans with the interdependent intentions, if the meaning of the latter is partly indeterminate? On top of that, to support Velleman and Rachar's proposal, we need to develop or onboard a theory of propositions with indeterminate content, which strikes as an *ad hoc* solution.

So far, we have spotted a principal problem in the theory of shared agency that concerns the logical form of participatory intentions. If the participatory intentions are unconditional – I intend that we φ and you intend that we φ – than such intentions go against the control constraint. If the participatory intentions are conditional, as Velleman and Rashar argue, then they do not break the control constraint, but create a vicious circle of interdependence.

We think that the second option, i.e., the conditional account, is still salvageable. The problem of the vicious circle appears, since agents cannot intend that they φ unconditionally, so we conditionalise one's intention that they φ on the intention of other(s) that they φ, which, in its turn, also begs for conditionalization. To break out of this circle, at least one agent should intend something unconditionally, and this intention should not break the control constraint, hence, should be about her own, not their joint action. In what follows, we propose exactly such a solution.

We argue that in a lot of cases, the participatory intentions in question may not be completely symmetric, i.e., they may not simultaneously conditionalise on one another in a similar way. Namely, that happens in cases where at least some members – we will call them *initiators* – intend to do their parts of the joint activity no matter what, even if others will not reciprocate. In such cases, the vicious circle of conditionalization is broken, since it terminates on the unconditional intention of initiators to do their share no matter what. The presence of initiators does not entail hierarchy: one may initiate a joint activity by inviting others to participate (unconditionally, even if no one will reciprocate), which does not automatically give the initiator power over others or lead to any other hierarchical imbalance in the group.

To clarify it, let us look at two real-world examples. In Argentinian tango culture, potential dance partners sit on opposite sides of a dance hall. According to the custom, if someone wants to dance together with the other, one is to make eye contact with this person. If they, in their turn, do not look away and nod or smile, one is to approach and offer a dance. Otherwise, if the potential dance partner looks away or acts as if they have not noticed the gaze, it is a signal of a polite refusal, i.e., the initiator is better not to approach with an offer, since it will not be accepted. The whole point of the eye contact protocol is to avoid the

awkwardness of the explicit public rejection of the dance partners. Therefore, for the outside observer, it will seem like the harmony is pre-established, and every time someone invites the other for a dance, they agree, as if the dancers can read each other's minds or if manifest events happen. In reality, in every pair there should be at least one initiator, who intends to invite their peer by establishing eye contact, regardless of whether their potential partner will reciprocate or not. The fact that one has initiated the dance does not entail that one gets any degree of control over their dance partner.

Another illustrative example is the practice of solo protests. Often activists demonstrate alone in public places. Clearly, a single person is not enough for a protest to be successful. Some passersby may share the activists' views, but intend to join the protest only if a significant number of other citizens will participate as well. Therefore, by doing solo protests, activists express their intention to keep protesting regardless of whether others will join or not. Consequently, if like-minded passersby encounter solo protests often enough, they might get the reason to believe that the movement is popular enough, so it makes sense to join. For some, it may be sufficient to see at least dozens of solo protests or stickers in the city. For others, it might be necessary to be sure that the qualified majority of residents of their city are ready to join. Therefore, expressions of the initiators' intentions to protest no matter what may cause *protest cascades*, which sometimes result in mass protest movements [Pea18].

To generalise, given that there are initiators who, on top of their participatory intentions, intend to do their share no matter what, others can conditionalise their intentions towards the joint activity exactly on those unconditional intentions of the initiators. In such a case, neither will they face the vicious circle of interdependence, nor will anyone break the control constraint, since the initiator intends to do *her* share, and others intend their joint activity only conditionally. We do not claim that the initiators are necessary to share intentions. We only claim that in case there are initiators, a shared intention may be established via interdependent conditional intentions without facing the vicious circles.

So far, our discussion has been informal, while the main problem is the *logical* form of participatory intentions. Later in the thesis (Chapter 4), we are going to develop a logic to prove the following formally:

1. Bratman's unconditional account, taken literally – I intend that we φ and you intend that we φ – inevitably breaks the control constraint;

2. Our account with initiators is consistent with the control constraint and can be specified by means of the compositional semantics, hence, it is not self-referential. Moreover, it has the desired property of conservativity over individual practical reasoning. We will explicitly demonstrate that, given that individuals' participatory intentions obey some norms of individual

practical reasoning, their shared intention satisfies the very same norms.

This is all we have to say about Bratman's theory as a characteristic representative of methodologically individualistic accounts of group agency. Next, we wove into the theory of team reasoning, which represents a methodologically holistic camp.

2.2 Team reasoning

In this section, we address a holistic account of group agency: the theory of *team reasoning*. In this framework, we do not postulate the autonomous existence of groups, but ascribe individuals a specific mode of reasoning, where they think of groups as autonomous agents. Our main source will be a book "*Beyond Individual Choice*" by Michael Bacharach [Bac06]. In this work, the core arguments and concepts of team reasoning are introduced. We will also rely on the articles by Bacharach's fellow thinkers to show how team reasoning is comparable with other, more philosophically inclined holistic accounts.

Throughout the section, we are going to talk about *games*.

2.2.1. Definition (Games). A game is a tuple $G = (N, \{A_i, u_i\}_{i \in N})$, where:

1. $N \neq \emptyset$ is a finite set of agents;
2. For each $i \in N$, $A_i \neq \emptyset$ is a finite set of actions available for i. Possible combinations of agents' actions, $\prod_{i \in N} A_i$, are called *action profiles*. We associate each action profile with the unique outcome it yields. Since there is a one-to-one correspondence between action profiles and outcomes, both terms are to be used interchangeably.
3. For each $i \in N$, $u_i : \prod_{i \in N} A_i \to \mathbb{R}$ is a utility function, ascribing a numerical utility value of every possible outcome for i.

Bacharach starts with presenting some puzzling games to demonstrate the limitations of orthodox game theory as both a normative and descriptive tool. In these games, either agents will be better off by *not* following the rationality norms of the orthodox game theory, or intuitively should distinguish between options that the orthodox game theory regard equivalent. Moreover, as experiments show, agents do tend to play those games differently than the standard game-theoretic solution concepts predict.

By the game theory's orthodoxy, Bacharach means the attempt to explain agents' choices of action by modeling the specific mode of practical reasoning: "[The] *rational approach, which is that of classical game theory, sets out to explain behaviour as the outcome of the agent's reasoning or reasons*" [Bac06, p.7]. According to the orthodox assumption, agents choose their actions to maximise their own payoffs, given what they know and believe about other agents. *"The*

agent is modeled as deliberating over seeking the optimum value of what she herself controls — her own act. So she must take what she does not control, but is controlled by the other, as parametric, and make a hypothesis about the value of this parameter" [Bac06, p.61]. Bacharach calls it *best-reply reasoning*: every agent chooses *her* best action to reply on how she believes other agents are going to act. Typically, best-reply reasoning operate under the assumption of the common knowledge about everyone's rationality: it is commonly known that everyone also chooses their actions to maximise their payoffs, implementing the very same best-reply reasoning.

As a result, the orthodox game theory uses *pure Nash equilibrium* as its solution concept. In other words, we predict that rational agents are going to choose their actions in games in a way that will result in Nash equilibrium.

2.2.2. Definition (Nash Equilibrium). Given a game $G = (N, \{A_i, u_i\}_{i \in N})$, an action profile $\alpha \in \prod_{i \in N} A_i$ is a Nash equilibrium iff for every agent $i \in N$, $u_i(\alpha) \geq u_i(\alpha^*)$ for any $\alpha^* \in \prod_{i \in N} A_i$, such that $\forall j \in N \setminus \{i\} : \alpha^*(j) = \alpha(j)$. Informally, an action profile is a Nash equilibrium iff no agent can improve her payoff by choosing a different action, given that everyone else chooses the same actions.

To show the limitations of orthodox game theory, Bacharach presents games in which either the Nash equilibrium is not what agents intuitively should choose, or there are multiple Nash equilibria, one of which stands out as preferable, although orthodox game theory cannot distinguish them.

One of the puzzles is probably the most well-known game, the *Prisoner's Dilemma*.

2.2.3. Example (Prisoner's Dilemma). Two prisoners are being separately interrogated about the crime they have committed together. They can either accuse their partner ($defect$) or not ($cooperate$). If both accuse one another, they will get jailed. If both stay silent, both will get just a fine. If one accuses the other, while the other stays silent, the first will be let free, while the accused one will get jailed for a longer term. See its classic version depicted on Figure 2.1

	$cooperate$	$defect$
$cooperate$	$2, 2$	$0, 3$
$defect$	$3, 0$	$1, 1$

Figure 2.1: Prisoner's Dilemma: normal-form.

There is a unique Nash equilibrium in this game: $(defect, defect)$. Hence, the orthodox game theory predicts (and, if taken to be normative, advices) both players to defect. When presented with the Prisoner's Dilemma, about a half of real-world players choose $(cooperate, cooperate)$ [Sal95]. Real people often play

this game by *not* following the rationality norms imposed by orthodox game theory, and are better off than if they had followed them. This puzzle questions orthodox game theory as both an explanatory and normative tool.

Another puzzle is the *Hi-Lo game*.

2.2.4. Example (Hi-Lo game). Both players have two actions available: hi and lo. If both choose hi, each gets some reward. If both choose lo, each gets half of the reward they would have gotten for simultaneously choosing hi. If they choose different actions, no one gets anything. See Figure 2.2 for the illustration.

	hi	lo
hi	$2, 2$	$0, 0$
lo	$0, 0$	$1, 1$

Figure 2.2: Hi-Lo game: normal-form.

In Hi-Lo game, there are two pure Nash equilibria: (hi, hi) and (lo, lo). One of these equilibria, (hi, hi), yields higher payoffs for both players. So it seems trivial which of the two equilibria should be chosen. Moreover, experimental data shows that real people choose (hi, hi) regularly, in about 96% of cases, as if the problem is indeed trivial for our pre-theoretic reasoning [Bar+10]. Yet there are no way to explain this in orthodox game theory! Following the best-reply reasoning, it is rational for every player to choose hi only if the opponent is going to choose hi, while for the best-reply reasoning opponent, it is again rational to choose hi, only if her opponent is choosing hi. Yet there are no reasons to believe that any of the players will choose hi unconditionally. In other words, best-reply reasoning under the common knowledge about everyone's rationality traps the players in the familiar vicious circle of conditionality. So how do real-world players get out of this vicious circle?

Bacharach conjectures that they step out of the best-reply reasoning. Instead of driving their choice of action by answering the question "*What should I do?*", they entertain the question "*What should **we** do?*" first:

> *I suggest that (Conjecture 1) a typical person facing any game may have an inclination to reason about what to do in it in a certain way, which for the moment I will label reasoning in mode-P.* [..] *Mode-P reasoning has two features:*
>
> 1. *The player ranks all act-profiles, using a Paretian criterion* (see Definition 2.2.5).
>
> 2. *She takes herself to have a reason to enact her component in the highest-ranked act-profile.* [Bac06, p.59]

2.2.5. Definition (Paretian criterion). Given two action profiles $\alpha, \alpha' \in \prod_{i \in N} A_i$, α Pareto-dominates α' iff for every player $i \in N$, $u_i(\alpha) \geq u_i(\alpha')$, and for at least one player $j \in N$, $u_j(\alpha) > u_j(\alpha')$. An action profile α is a Pareto-optimum iff there is no other action profile α' that Pareto-dominates α.

Consequently, by the Paretian criterion, we understand the ranking of action profiles in accordance with Pareto-dominance: given that α Pareto-dominates α', then α should be ranked higher than α'.

In other words, if the agent reasons in the P-mode (or, as Bacharach baptises it later in the book, *team-reasons*), then she first chooses among the *group*'s joint actions, ranking them in terms of common good, asking herself: which joint action of *ours* is the best? Then, she chooses her part in the highest-ranked joint action she has chosen. Clearly, team reasoning is incompatible with the best-reply reasoning: a team-reasoner chooses her action not as the best-reply to what all others are expected to do, but as a component of the joint action that is the best for everyone.

Going back to the Hi-Lo game, there is a single Pareto-optimal action profile, (hi, hi). If both players "P-reason" about the game, then both should choose their own component of the Pareto-optimum, i.e., hi. Although the conjecture successfully predicts the intuitively appealing outcome of (hi, hi) in the Hi-Lo game, we are yet to explain why agents would team-reason in this situation. It is not clear why they should use the Paretian criterion. And even if we accept the Paretian ranking, given that (hi, hi) is Pareto-optimal, how does it ground each player's choice of hi? Even if (hi, hi) is "the best" option, the choice of hi still does not dominate the choice of lo: in case the other player chooses lo, one would be better off choosing lo instead of hi. On the surface, it seems like P-reasoning just instructs agents to ignore the vicious circle of conditionality that is still there.

Bacharach argues that *group identification* fosters team reasoning. If one perceives herself as a member of a group, she is inclined to team-reason: when someone thinks of herself as a part of the group, it is natural for such an agent to wonder what the group's best joint action is. But why or when do agents group-identify? Bacharach refers to the experimental data from psychology and social sciences that show agents are likely to group-identify in a variety of situations: starting from being labeled as an ad hoc group by outsiders and ending with sharing an experience or common fate [Bac06, Chapter 2]. One specific criterion allows agents to simultaneously group-identify in a strategic interaction, based on nothing but the payoff-structure of the game: having a *common interest* that can only be achieved together.

For example, if agents have their individual goals, and everyone can reach her goal independently of everyone else, there is no sense for the agents to identify as a group. If agents' goals are incompatible – e.g., agents play a zero-sum

game – there seems to be no reason to group-identify as well. On the contrary, if all agents have compatible goals that they can achieve simultaneously, while everyone's success in achieving her goal depends on what everyone else is going to do, group identification seems intuitively natural. Moreover, in such situations, agents can do better by implementing team- rather than best-reply reasoning. Bacharach's formal definition of common interest and interdependence reflects these intuitions.

2.2.6. Definition (Common interest, copowers). Let $G = \{N, \{A_i, u_i\}_{i \in N}\}$ be a game, where $C \subseteq N$ is a set of agents and $S, S' \subseteq \prod_{i \in N} A_i$ are the sets of outcomes. Then, there is a common interest among C in S over S' iff every outcome $s \in S$ Pareto-dominates any outcome $s' \in S'$.

C has a copower to S iff there exists a joint strategy $a_C \in \prod_{i \in C} A_i$, such that, for any joint strategy $a_{\overline{C}} \in \prod_{j \in \overline{C}} A_j$, $(a_C, a_{\overline{C}}) \in S$.

2.2.7. Definition (Interdependence). Let $G = \{N, \{A_i, u_i\}_{i \in N}\}$ be a game, where $C \subseteq N$ is a set of agents. Let $sol(G) \subseteq \prod_{i \in N} A_i$ be a solution set of G in standard game theory, e.g., the set of all pure Nash equilibria in G.

Then, C is interdependent iff for some $S, S' \subseteq \prod_{i \in N} A_i$, C has copower for S and S', and common interest in S over S', while $S \subseteq sol(G)$.

Informally, if members of C are interdependent, then they can do better for everyone by *not* choosing some Nash equilibrium. For example, in the standard prisoners' dilemma game, $sol(G) = \{(defect, defect)\}$. Both players have a common interest in $\{(cooperate, cooperate)\}$ over $sol(G)$. Since it is a two-player game, they trivially have copowers for both sets. Hence, players are interdependent. In the Hi-Lo game, players are interdependent as well, given that they have a common interest in (hi, hi) over (lo, lo), while $(lo, lo) \in sol(G) = \{(hi, hi), (lo, lo)\}$.

Interdependence promotes group identification among agents, since reasoning about themselves as a team and choosing their best joint action might do better for everyone than coming to one of the Nash equilibria as individual best-reply reasoners:

> *The fact that some games have* [interdependent set of players] *is the same as the fact that individual rationality can be collectively inefficient or group irrational.* [Bac06, p.86]

Next, we put team reasoning in the context of our philosophical inquiry. Note that we do not postulate the existence of groups, but we do postulate groups as a concept that is crucial for agents' practical reasoning and, consequently, for our explanation of their behaviour. "*Whether or not groups are a fact of life they are certainly a fact of theory*" [Bac06, p.73]. As we have shown, some choices are "group rational", but "individually irrational" (e.g., to choose *cooperate* in the Prisoner's Dilemma), while other choices are "individually rational" but "group

irrational" (e.g., to choose *defect* in the Prisoner's Dilemma). It is impossible to reduce the rational norms of team reasoning to the aggregation of individual, best-reply rationality norms that are imposed on every member of the team. So the theory of team-reasoning is methodologically holistic, but ontologically individualistic.

Team reasoning can be reframed as a theory of shared, or collective intention:

> *An intention is interposed between reasoning and action, so it is natural to treat the intentions that result from team reasoning as collective intentions* [...] *So, if the distinctive feature of collective intentions is to be found in the reasoning by which they were formed, then an analysis that focuses on the intentions themselves will miss the feature that makes collective intentions collective.* [GS07, p.111]

Then, what distinguishes group members from independent individuals is not the content of their motivational attitudes, but the mode of practical reasoning by which they come to their motivational attitudes. For instance, a team-reasoner unconditionally intends to choose *hi* in the Hi-Lo game, just like some best-reply reasoners might do. The main difference is not in *what* they intend, but *why*: the team-reasoner intends to choose *hi* to enable her group to satisfy their common interest, while the best-reply reasoner intends to choose *hi* to satisfy nothing but her own interest, given what she believes about what others are about to do.

Viewed as a theory of shared intention, Bacharach's account provides a strong argument against Bratman's theory. Namely, Bratman's definition of modest sociality is not strict enough: best-reply reasoners choosing Nash equilibrium fall under Definition 2.1.1. For instance, assume two prisoners in the Prisoner's Dilemma are rational best-reply reasoners who adhere to the norms of orthodox game theory. Then, each prisoner intends that both of them defect: it is rational for both of them to expect the other to defect, and the best reply to such an action is to defect as well. We might say that each intends that they both defect by means of the intention of the other that they defect: it is not solely up to any of them whether they will defect or not, they only have a *copower* to ensure that both defect. Each prisoner has a subplan – to defect – and their subplans obviously mesh. Finally, since it is commonly known that they are individually rational and reason in such ways, it should also be commonly known among them that they so intend. Intuitively, defecting prisoners act in the most individualistic way possible, there is nothing social in their behaviour. Analogously, egoistically motivated bargaining individuals who reach equilibrium on the free market should be perceived as a group, or even state leaders who simultaneously ignore international law, commonly knowing that everyone is going to ignore it. Although Bratman tries to distinguish modest sociality from the choices of Nash equilibrium under the common knowledge of individual rationality, proponents of team reasoning find his explanations vague:

> *However, the core analyses provided by* [among others] *Bratman seem to imply that all Nash equilibrium situations are instances of collective intentions. Cases in which Nash equilibria are not joint actions are excluded only by stipulation or by the addition of further conditions which are just as problematic as the original concept of collective intention.* [...] *Bratman adds conditions which require each agent to be responsive to the behavior of the other as the joint action proceeds and if unexpected problems occur, but these conditions are stated only informally, and rely on a pre-analytic understanding of the nature of cooperative activity.* [GS07, p.110]

The theory of team reasoning has another advantage over Bratman's view. Neither do team-reasoners fall into the vicious circle of conditionalization, nor do they break the control constraint. A team-reasoner first chooses the Pareto-optimal joint action, then *unconditionally* intends to do *her share* of the chosen joint action. Neither does she intend that her team reaches Pareto-optimal outcome, nor does she condition her intention on the intentions of others (at least it does not follow from the theory itself). Yet there is a core problem. According to Bacharach, group identification is psychological rather than instrumentally rational. Whether one group-identifies or not is not up to her rational choice, but rather up to one's idiosyncratic psychological features. As Bacharach's argument goes, aspects such as common faith or recognizing interdependence do not necessarily provide rational agents with the justification to team-reason; agents are only *inclined to* team-reason in such situations. Bacharach writes about salient features like interdependence as "promoting" or "favouring" group-identification [Bac06, p.76]; he does not speak about them as providing justifications or grounds for agents to *choose* to group-dentify. In Sugden's development of Bacharach's theory, one can choose whether to group-identify, and should do so if there is a common reason to believe that everyone group-identifies [Sug00; Sug03].

For example, idealised rational agents should commonly know that everyone will be better off if they team-reason in the Hi-Lo game. Therefore, they have a common reason to believe that everyone indeed group-identifies and team-reasons. The manifest events, weakly interpreted, can also provide a common reason for group identification. But what if there is no common reason to believe in members' group identification? We might end up in the same vicious circle, but not about the choice of action, but rather the choice of mode of reasoning. For instance, assume one finds herself in the Prisoner's Dilemma. The interdependence between the prisoners is commonly known among them, but each prisoner considers it equally possible that the other will take it to be the reason to group-identify. One wonders whether she should group-identify. It makes sense to group-identify iff the other prisoner group-identifies, while for the other prisoner, it also makes sense to group-identify iff the first one group-identifies. Unless it is commonly believed among the prisoners that at least one of them unconditionally group-

identifies, the prisoners are trapped in the vicious circle of conditionalization.

The other core problem of team reasoning is the very concept of common interest. In simple cases like the Hi-Lo game, the common interest is clear. Yet there might be no unique meaningful way to aggregate individual preferences into the collective preference [LP02]. Bacharach acknowledged the problem and provides the only minimal requirement: given agents' individual preferences specified by utility functions $\{u_i\}_{i \in N}$, their group preference U_N should be *Paretian*: given two outcomes α, α', if α strictly Pareto-dominates α', then $U_N(\alpha) > U_N(\alpha')$ [Bac06, p.88]. It remains unclear what a team reasoner should do in case she faces, e.g., the Condorcet paradox.[6]

2.3 Envoi

In this chapter, we have non-exhaustively reviewed the literature on the philosophy of group agency. We derived the following insights: for sets of agents to be regarded as a group, they should share some informational and motivational attitudes that cause their actions. Most of the authors regard common knowledge or belief as groups' informational attitudes, while shared intention – as their motivational attitude. While there exist consensual (although not uncontroversial) definitions of common knowledge and belief, what shared intention is remains debated. We have reviewed two groups of theories: methodologically individualistic and holistic. The former reduces shared intention to an aggregation of individuals' interdependent intentions; the latter argues that whether individuals' intentions constitute a shared intention depends on whether they group-identify and reason accordingly. The group-identification requires agents to think of their group as a single irreducible unit of agency. Individualistic theories do not presuppose the existence of groups as either real entities or irreducible concepts, while the holistic theories postulate that groups lack autonomous existence ontologically speaking, but do exist over and above their members in the agents' reasoning.

Despite significant differences between the theories of group action we have touched upon, one remains clear: not any set of agents is a group, and not any aggregation of individuals' actions is a group action. To understand whether the set of agents is a group, we should check what agents know, believe, and intend. Consequently, to understand whether an action profile is a group action, we need to look into individuals' reasons to act. In the following chapter, we will review some multi-agent modal logics of knowledge and action to see how groups are defined there. Via use-cases, we will show how treating random sets of agents as groups affects

[6] A classical formulation of the Condorcet paradox goes as follows. There are three agents – $\{a, b, c\}$ – each of which has three options: $\{1, 2, 3\}$. Agents' preferences are such that $u_A(1) > u_A(2) > u_A(3)$; $u_B(2) > u_B(3) > u_B(1)$; $u_C(3) > u_C(1) > u_C(2)$. In such a situation, no pair of alternatives satisfies Pareto dominance; hence, any total ordering on $\{1, 2, 3\}$ will be Paretian.

the practical reasoning these logics model.

Chapter 3

Logical background: multi-agent modal logics (MAMLs)

In this chapter, we review a family of different formalisms that we call *multi-agent modal logics* (MAMLs). Typically, MAMLs formalise agents' practical reasoning in interactive scenarios. For example, in multi-agent epistemic and doxastic logics, we deal with agents' (higher-order) knowledge and belief, as well as different notions of group epistemic attitudes, such as common and distributed knowledge/belief [Hin62; Fag+04]; In a variety of multi-agent modal logics of actions, we formalise reasoning about the existence of agents' and groups' (joint) actions and long-term strategies that are effective to achieve their goals [Pau02; AHK02; BPX01]. MAMLs widely differ in their fundamental assumptions about agents and about the environment in which they act. MAMLs also differ in their metalogical properties, such as finite axiomatizability, decidability and complexity of the satisfiability problem, expressive power, etc.[1] This chapter has two main aims: 1) to introduce the logical background upon which we will rely further; 2) to demonstrate how the most well-known MAMLs overlook the difference between arbitrary sets of agents and groups, and how it leads to conceptual problems.

First, we outline some common notions and concepts about MAMLs. All MAMLs have similar syntaxes. Let us fix a finite set of agents $Ags = \{1, \ldots, n\}$ and a countable set of propositional variables $Var = \{p_1, p_2, \ldots\}$. Then, a formal language $\mathcal{L}(Ags, Var)$ is recursively defined as follows:

$$\mathcal{L}(Ags, Var) \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box_i\varphi \mid \Box_G\varphi$$

[1] We take the standard definitions of all the mentioned metalogical properties of modal logics, as defined in, e.g., [BRV01].

where $p \in Var$ is a propositional variable, $i \in Ags$ is an agent and $G \subseteq Ags$ is a set of agents, or a group. We use standard acronyms for Boolean connectives and constants: $\varphi \wedge \psi = \neg(\neg\varphi \vee \neg\psi)$, $\varphi \to \psi = \neg\varphi \vee \psi$, $\varphi \leftrightarrow \psi = (\varphi \to \psi) \wedge (\psi \to \varphi)$; $\top = (p \vee \neg p)$; $\bot = \neg\top$. We also use standard abbreviations for modalities' duals: $\Diamond_G\varphi = \neg\Box_G\neg\varphi$. $\Box_i\varphi$ typically expresses either agent's propositional attitude or some fact about the agent's actions. Depending on the exact MAML we are working with, $\Box_i\varphi$ may be read as, e.g., *i knows/believes/intends/prefers that* φ or *i sees to it/can ensure/has a long-term strategy to ensure that* φ. $\Box_G\varphi$ expresses the group's analogue of the corresponding propositional attitude or fact about the group's joint action. Depending on the MAML, $\Box_G\varphi$ may be read as *G commonly knows/commonly believes/share an intention/commonly prefers that* φ [Fag+04; ÅHW11; DV02] or *G jointly sees to it/has a copower to ensure/has a joint long-term strategy to ensure that* φ [BPX01; Pau02; AHK02], respectively. Given any set $\{1, \ldots, j\} \subseteq Ags$, we use the following acronym: $\Box_{1,\ldots,j}\varphi = \Box_{\{1,\ldots,j\}}\varphi$.

Note that the group notions are defined for every subset of agents whatsoever, including the empty set $\emptyset$ and all the singletons $\{\{i\}\}_{i \in Ags}$. Usually, singleton group notions are identical to the individual notions. Hence, $\Box_i\varphi$ is redundant in the syntax and can be redefined as an acronym $\Box_i\varphi = \Box_{\{i\}}\varphi$. As for the empty group, it is either explicitly excluded from the syntax (as in, e.g., [WÅ13b]) or perceived as nature or the environment (as in ,e.g., [Pau02]). MAMLs typically use standard Kripke semantics:

3.0.1. Definition. A Kripke frame is a tuple $F = (W, \{R_i\}_{i \in Ags})$, where $W \neq \emptyset$ is a set of possible worlds; for each $i \in Ags$, $R_i \subseteq W \times W$ is a binary relation on the set of possible worlds. For every group $G \subseteq Ags$, R_G is defined as $\bigoplus_{i \in G} R_i$, where $\oplus$ is some, typically idempotent, associative and commutative, algebraic operation on binary relations.[2] Any Kripke frame $F = (W, \{R_i\}_{i \in Ags})$ can be extended to a model $M = (F, V)$ by a valuation function $V : Var \to 2^W$ that takes a propositional variable and returns a set of possible worlds, where this proposition is true. Given any Kripke model $M = (W, \{R_i\}_{i \in Ags}, V)$ and a world $w \in W$, we call the pair (M, w) a pointed model.

As we will often use operations on binary relations, we list some of the most common ones to fix the notation.

[2]We do not specify which algebraic operation is used, since it varies across MAMLs. Most importantly, R_G should be functionally dependent on $\{R_i\}_{i \in G}$. For the variety of different algebraic operations on binary relations, see [Tar41].

3.0.2. Definition. For any binary relations $R, R_1, R_2 \subseteq W \times W$:

$$
\begin{array}{lcl}
R_1 \cup R_2 & = & \{(w,v) \mid (w,v) \in R_1 \text{ or } (w,v) \in R_2\} \\
R_1 \cap R_2 & = & \{(w,v) \mid (w,v) \in R_1 \text{ and } (w,v) \in R_2\} \\
R_1 \circ R_2 & = & \{(w,v) \mid \exists u \in W : (w,u) \in R_1, (u,v) \in R_2\} \\
R^{-1} & = & \{(w,v) \mid (v,w) \in R\} \\
R^n & = & \begin{cases} \{(w,w) \mid w \in W\} & \text{if } n = 0 \\ R & \text{if } n = 1 \\ R \circ R^{n-1} & \text{otherwise} \end{cases} \\
R^* & = & \bigcup\limits_{n \in \mathbb{N}^+} R^n, \text{ where } \mathbb{N}^+ := \mathbb{N} \setminus \{0\}
\end{array}
$$

For brevity, we use the following acronym $wRv := (w,v) \in R$. For any world $w \in W$, by $R(w) = \{v \in W \mid wRv\}$ we denote the set of worlds that are accessible from w via R.

3.0.3. Definition (Semantic clauses for $\mathcal{L}(Ags, Var)$). Given any Kriple model $M = (W, \{R_i\}_{i \in Ags}, V)$ and any state $w \in W$, we recursively define $\models$ relation as follows:

$$
\begin{array}{lcl}
M, w \models p & \Leftrightarrow & w \in V(p) \\
M, w \models \neg\varphi & \Leftrightarrow & M, w \not\models \varphi \\
M, w \models \varphi \vee \psi & \Leftrightarrow & M, w \models \varphi \text{ or } M, w \models \psi \\
M, w \models \Box_i\varphi & \Leftrightarrow & \forall v \in W(wR_iv \Rightarrow M, v \models \varphi) \\
M, w \models \Box_G\varphi & \Leftrightarrow & \forall v \in W(wR_Gv \Rightarrow M, v \models \varphi)
\end{array}
$$

We can reformulate this definition by inductively defining the truth-set, or intension, of any formula as follows:

$$
\begin{array}{lcl}
[\![p]\!]_M & = & V(p) \\
[\![\neg\varphi]\!]_M & = & W \setminus [\![\varphi]\!]_M \\
[\![\varphi \vee \psi]\!]_M & = & [\![\varphi]\!]_M \cup [\![\psi]\!]_M \\
[\![\Box_i\varphi]\!]_M & = & \{v \in W \mid R_i(v) \subseteq [\![\varphi]\!]_M\} \\
[\![\Box_G\varphi]\!]_M & = & \{v \in W \mid R_G(v) \subseteq [\![\varphi]\!]_M\}
\end{array}
$$

Consequently, $M, w \models \varphi$ iff $w \in [\![\varphi]\!]_M$.

By $M \models \varphi$ we mean that $M, w \models \varphi$ for every world $w \in W$. By $F, w \models \varphi$ we mean that $(F, V), w \models \varphi$ for every valuation function $V \in 2^{W^{Var}}$. By $F \models \varphi$ we mean that $F, w \models \varphi$ for every $w \in W$. By $\models \varphi$ we mean that $F \models \varphi$ for any

Kripke frame F. Given some class of Kripke frames $\mathbf{F}$, $\mathbf{F} \models \varphi$ means that $F \models \varphi$ for any $F \in \mathbf{F}$. Given a class of frames $\mathbf{F}$, a set of formulae Γ and a formula φ, $\Gamma \models_{\mathbf{F}} \varphi$ means that for any $F \in \mathbf{F}$, if $F \models \psi$ for all $\psi \in \Gamma$, then $F \models \varphi$. Consequently, $\mathbf{F} \models \varphi$ iff $\emptyset \models_{\mathbf{F}} \varphi$. The latter will be abbreviated simply as $\models_{\mathbf{F}} \varphi$.

Depending on the intended interpretation, the relation $\{R_i\}_{i \in Ags}$ may have additional properties. For some MAMLs, instead of binary relations, there might be neighborhood functions $N_i : W \to 2^{2^W}$ (as in, e.g., [Pac17; Che80]) or subset spaces $\tau_i \subseteq 2^{2^W}$ (as in, e.g., [PMS07; WÅ13a; Özg17]). What is common to most MAMLs known to us: any set of agents is a group, and the semantic component for any group $R_G/N_G/\tau_G$ is functionally dependent on the semantic components of the group members $\{R_i/N_i/\tau_i\}_{i \in G}$, respectively. Yet, there are some outliers.

In some MAMLs, the definition of group is syntactically restricted: the collection of subsets $\mathcal{G} \subseteq 2^{Ags}$ is introduced, and $\Box_G \varphi$ is a well-formed formula iff $G \in \mathcal{G}$; oftentimes, $\mathcal{G} = \{Ags\}$, i.e., only the set of all agents is regarded as a group [Sch12; Lor13]. This restriction is introduced for purely technical reasons. With the full language, where $\mathcal{G} = 2^{Ags}$, it is often difficult (or even impossible) to find a finite axiomatization or decision procedure for the satisfiability problem, so one hopes to find as expressive fragments as possible with better metalogical and computational behaviour. Another interesting exception is a logic of *intensional groups* [BS23], where the group is associated with agents' dynamic property, not a fixed set of members. Think of, e.g., a group of British MPs. The set of agents that constitute this group changes once every five years. Therefore, which individuals are members of the group varies across possible worlds. One exception that is conceptually closer to our line of work is the modal logic of non-reductionist group belief [Gau+15]: while we still can reason about group belief of any set of agents, group belief is not functionally dependent on the beliefs of group members.

Throughout the thesis, we use the following notation and terminology, unless stated otherwise. For any MAML, by $\mathcal{L}$ we denote its formal language. By $\mathbf{F}$ we denote the class of frames, over which we interpret $\mathcal{L}$. By $\mathbf{L}$ we denote the logic, understood as the following set: $\mathbf{L} = \{(\Gamma, \varphi) \in 2^{\mathcal{L}} \times \mathcal{L} \mid \Gamma \models_{\mathbf{F}} \varphi\}$. When possible, we will define $\mathbf{L}$ as an axiomatic system: a set of axioms and derivation rules. In such cases, by $\Gamma \vdash_{\mathbf{L}} \varphi$ we will denote the fact that there exists a derivation of φ from Γ in $\mathbf{L}$. The notion of derivation is defined standardly for modal logics [BRV01]. We will say that an axiomatic system $\mathbf{L}$ is strongly sound w.r.t. $\mathbf{F}$ iff $\Gamma \vdash_{\mathbf{L}} \varphi$ entails $\Gamma \models_{\mathbf{F}} \varphi$; and it is strongly complete w.r.t. $\mathbf{F}$ iff $\Gamma \models_{\mathbf{F}} \varphi$ entails $\Gamma \vdash_{\mathbf{L}} \varphi$. $\mathbf{L}$ is weakly sound and complete w.r.t. $\mathbf{F}$ iff $\vdash_{\mathbf{L}} \varphi$ entails $\models_{\mathbf{F}} \varphi$ and $\models_{\mathbf{F}} \varphi$ entails $\vdash_{\mathbf{L}} \varphi$, respectively. To distinguish different MAMLs, we are going to use sub- and supersripts: i.e., $\mathcal{L}_{CL}$, $\mathcal{L}_{EL}$, $\mathcal{L}^t_{STIT}$, etc.

In what follows, we address two kinds of MAMLs. The first is multi-agent *epis-*

temic logics: formalisms to reason about agents' (higher-order) knowledge. The second is the modal logics of agency: formalisms to reason about agents' abilities and actions.

3.1 Multi-agent epistemic logics

In this section, we will interpret $\mathcal{L}(Ags, Var)$ in the following way. $\Box_i\varphi$ reads as *i knows that* φ. Its dual, $\Diamond_i\varphi$, reads as *i considers it possible that* φ. $\Box_G\varphi$ reads as *it is commonly known among G that* φ. The resulting formalism is called *multi-agent epistemic logic*. The epistemic logic was baptised as such and initially studied for a single agent by Jaako Hintikka [Hin62], after which it was further developed to the multi-agent settings and popularised beyond philosophical logicians by, among many others, Daniel Lehman, Ronald Fagin, Joseph Halpern, Yoram Moses, and Moshe Vardi [Leh84; Fag+04].

To distinguish epistemic logic from other MAMLs, we introduce the following notation:

$$\mathcal{L}_{EL} \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid K_i\varphi \mid C_G\varphi$$

where $p \in Var$, $i \in Ags$ and $G \in 2^{Ags} \setminus \{\emptyset\}$. $K_i\varphi$ corresponds to $\Box_i\varphi$, and $C_G\varphi$ – to $\Box_G\varphi$ of $\mathcal{L}(Ags, Var)$. Following [Fag+04], we use the following acronym: for any non-empty $G \subseteq Ags$, $E_G\varphi = \bigwedge_{i\in G} K_i\varphi$. $E_G\varphi$ reads as *everyone in G knows that* φ. For any modality $\Box \in \{K_i, C_G, E_G\}$ and any positive natural number $n \in \mathbb{N}^+$, we inductively define $\Box^n\varphi$ as follows:

$$\Box^1\varphi = \Box\varphi$$
$$\Box^n\varphi = \Box(\Box^{n-1}\varphi)$$

As for semantics, we interpret R_i as an *indistinguishability* relation: wR_iv means that at w, agent i considers it possible that the real world is v. We require that every indistinguishability relation R_i is an *equivalence relation*, i.e, it is reflexive, symmetric, and transitive. We define these properties as follows, together with their intuitive epistemic meanings:

1. *Seriality:* for every $w \in W$, there exists a world $v \in W$, such that wR_iv. Every agent considers at least some state of affairs possible;
2. *Reflexivity:* for every $w \in W$, wR_iw. All agents consider the real world to be possible;
3. *Symmetry:* for every $w, v \in W$, if wR_iv, then vR_iw. Any agent either distinguishes two possible worlds or not, independently of which is the real one;

4. *Transitivity*: for every $w, v, u \in W$, if wR_iv and vR_iu, then wR_iu. If any agent does not distinguish w from v and v from u, then she does not distinguish w and u.

As for common knowledge, it is semantically defined in terms of the transitive closure of the union of relations. Formally, for any non-empty $G \subseteq Ags$:

$$R_G = \{(w, v) \mid wR_{i_1}w_1, \ldots, w_nR_nv \text{ for some } w_1, \ldots, w_n \in W, \{i_1, \ldots, i_n\} \subseteq G\}$$

In other words, wR_Gv iff there exists a finite path from w to v, such that every step of the path is via $\bigcup_{i\in G} R_i$. This semantics corresponds to the classical iterative definition of common knowledge, proposed independently in [Lew08] and [Aum76], as formally stated in Proposition 3.1.1 below. Namely, it is commonly known among G that φ iff everyone knows that φ, everyone knows that everyone knows that φ, and so on.[3]

3.1.1. Proposition. *Given any pointed Kripke model (M, w), the following holds:*

$$M, w \models E^n_G\varphi \text{ for any } n \in \mathbb{N}^+ \text{ iff } M, w \models C_G\varphi$$

Proof:
See [LM94, p.89] □

By dropping reflexivity of accessibility relations (and hence factivity of knowledge), we get the simplified semantics for *doxastic* logic, where $\Box_i\varphi$ is interpreted as "*i believes that* φ", and $\Box_G\varphi$ is interpreted as "it is commonly believed among G that φ" [LM94].

The notion of common knowledge is crucial to us, since it often serves as the group's representational state. Common knowledge allows agents to coordinate their actions and act simultaneously without preceding communication. For instance, assume a non-speaking Dutch logician goes to Bulgaria (I promise, this example will start to make sense). In the Netherlands, like in most European countries, nodding your head as a response to a binary question means "yes", while in Bulgaria, it means "no". The Dutch logician knows that. He goes to a grocery store. At the register, the cashier asks the logician if he wants a receipt. The logician does not need it and, knowing the local custom, nods in reply. The cashier notices that her customer is a foreigner, so she expects him not to know that his nodding means no. Hence, she thinks that he wants the receipt and gives

[3] Alternatively, common knowledge among G that φ may be characterised in two other ways: as the greatest fixpoint of the equation $C_G\varphi = \nu X.(E_GX \wedge E_G\varphi)$ or as "shared epistemic situation" in situation semantics [BK88]. On Kripke models, however, all three characterizations coincide. For a detailed overview of the relationship between different notions of common knowledge, see, e.g., [Alb02, Chapter 8].

it to the logician, and he throws it away immediately. The next day, someone tells the cashier that the Dutchman knows that nodding means no. The logician visits the store again, and the cashier asks if he wants a receipt. This time, he needs the receipt, and he still nods in reply. The Dutchman does not know that the cashier knows that he knows the custom. After yesterday's confusion, he expects the cashier to interpret his nodding as affirmative. The cashier, knowing that the Dutchman knows that the nodding means "no", sees him nodding and does not give him the receipt while he waits for it. If someone tells the Dutch logician that the cashier knows that he knows the custom, the story will repeat itself the next day, since the cashier will not know the Dutchman knows that she knows that he knows the custom. The story can go on forever, and every time there will be confusion, since either the cashier or the Dutch logician will lack some higher-order knowledge about the custom. Yet such confusions never happens between locals, since they *commonly* know that nodding means "no".

The axiomatic system $\mathbf{L}_{EL}$, as defined in Table 3.1, was proven to be weakly sound and complete for the class of Kripke frames with equivalence accessibility relations [HM90]. Axiom (K), together with the rule of necessitation (RN), express that agents are logically omniscient: they know all of the tautologies, and whenever φ logically entails ψ, if they know φ, then know ψ as well. Axiom (T) corresponds to factivity of knowledge: whenever it is (commonly) known that φ, φ is true. Axiom (D) expresses the consistency of knowledge: if it is (commonly) known that φ, then it is not (commonly) known that $\neg\varphi$.[4] Axioms (4) and (5) express positive and negative introspection, respectively: whenever it is (commonly) known that φ, it is (commonly) known that it is (commonly) known; and whenever it is not (commonly) known that φ, it is (commonly) known that it is not (commonly) known. In other words, agents' and groups' knowledge is fully transparent for themselves. The fixpoint axiom (FP) and the induction axiom (IN) capture the meaning of iterative definition of common knowledge.

As similar axioms will reoccur for other logics as well, we will fix some terminology and notation here. A minimal modal logic that satisfies (CPL), (K), (RN) and (MP) is called a *minimal normal modal logic*. We will denote the minimal extension of minimal normal modal logic with some axioms by K followed by the name of the axioms we extend it with. E.g., KD is the minimal extension of a normal modal logic with the axiom D. By S5 we will denote KT45, i.e., the minimal extension of the minimal normal logic that contains axiom (T), (4) and (5).

3.1.2. Example (Common knowledge in the Hi-Lo game). Assume that two per-

[4] Axiom (D) is redundant here, as it can be derived from (T). By (T), $\Box\varphi \to \varphi$; by (CPL), $\varphi \to \neg\neg\varphi$. By contraposition of (T), $\neg\neg\varphi \to \neg\Box\neg\varphi$. Then, by transitivity of $\to$, $\Box\varphi \to \neg\Box\neg\varphi$. We still mention (D) here, as it is going to reoccur throughout the thesis, so it is handy to introduce it.

(CPL)	Tautologies of classical propositional logic
	For any $\Box \in \{K_i\}_{i\in Ags} \cup \{C_G\}_{\emptyset\neq G\subseteq Ags}$:
(K)	$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$
(T)	$\Box\varphi \to \varphi$
(D)	$\Box\varphi \to \neg\Box\neg\varphi$
(4)	$\Box\varphi \to \Box\Box\varphi$
(5)	$\neg\Box\varphi \to \Box\neg\Box\varphi$
(FP)	$C_G\varphi \leftrightarrow E_G(\varphi \wedge C_G\varphi)$
(IN)	If $\vdash \varphi \to E_G(\varphi \wedge \psi)$, then $\vdash \varphi \to C\psi$
(RN)	If $\vdash \varphi$, then $\vdash \Box\varphi$
(MP)	If $\vdash \varphi$ and $\vdash \varphi \to \psi$, then ψ

Table 3.1: Logic $\mathbf{L}_{EL}$

fect logicians, a and b, play the Hi-Lo game. The rules are commonly known. The first logician, a, has decided to play hi. Her opponent, b, has not decided upon her strategy yet. After her decision was made, a has publicly announced that she knows what strategy she is going to play, yet she has not disclosed which one. Afterwards, b has publicly announced that she has not decided yet what to do.

What do players know about the game and each other's knowledge after their announcements? What is commonly known?

We represent agents' knowledge with the following Kripke model, illustrated in Figure 3.1. Let $W = \{w_{(hi,hi)}, \ldots, w_{(lo,lo)}\}$ be the set of possible worlds that represent all the possible outcomes of the Hi-Lo game. Since a knows her startegy and does not know b's, she cannot distinguish between two worlds iff she plays the same strategy in these worlds: $R_a = \{(w_{\sigma_a,\sigma_b}, w_{\sigma'_a,\sigma'_b}) | \sigma_a = \sigma'_a\}$. On the contrary, b knows neither her own nor a's choice of actions, so she cannot distinguish between any worlds: $R_b = W \times W$.

For any $i \in \{a, b\}$, let hi_i and lo_i be propositional variables that read as "*i's strategy is hi*" an "*i's strategy is lo*", respectively. Then, $V(hi_a) = \{w_{(hi,\sigma_b)} \,|\, \sigma_b \in \{hi, lo\}\}$, $V(hi_b) = \{w_{(\sigma_a,hi)} | \sigma_a \in \{hi, lo\}\}$. $V(lo_i) = W \backslash V(hi_i)$ for any $i \in \{a, b\}$.

Assume the real world is $w_{(hi,lo)}$. In this world, a knows that she will choose hi and does not know what b will choose: $M, w_{(hi,lo)} \models K_a hi_a \wedge \neg(K_a hi_b \vee K_a lo_b)$; b knows neither her own nor a's choice: $M, w_{(hi,lo)} \models \neg(K_b hi_a \vee K_b lo_a) \wedge \neg(K_b hi_b \vee K_b lo_b)$. Since a has announced that she has made her decision, b knows that a knows her strategy:

$$M, w_{(hi,lo)} \models K_b(K_a hi_a \vee K_a lo_a)$$

Since b has announced that she is yet to decide upon her strategy, a knows that

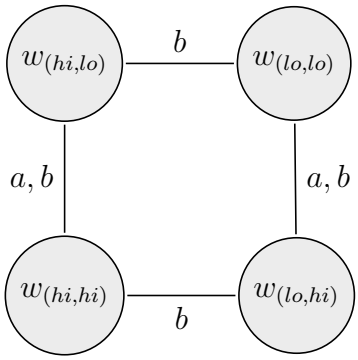

Figure 3.1: Kripke model for Example 3.1.2. Nodes represent possible worlds, and edges – the indistinguishibility relations. The reflexive and transitive arrows are omitted.

b does not know her strategy:

$$M, w_{(hi,lo)} \models K_a(\neg K_b hi_b \wedge \neg K_b lo_b)$$

Moreover, since both announcements were public, it is commonly known that a knows her strategy and b does not:

$$M, w_{(hi,lo)} \models C_{a,b}\big(\neg K_b hi_b \wedge \neg K_b lo_b \wedge (K_a hi_a \vee K_a lo_a)\big)$$

Note that, characteristically for MAMLs, we can reason about common knowledge among any non-empty finite set of agents. Conceptually, there is nothing wrong with that: postulating common knowledge among the arbitrary set of agents G does not entail that G should be a group. For instance, let G be the set of all readers of this text, including the author. By no means is G a group (at least in the sense in which philosophers talk about group agency). Yet we can claim that it is commonly known among members of G that every member has at least some command of English: otherwise, neither the author could write this text, nor the readers could read it.

Although reasoning about common knowledge of any set of agents is semantically coherent, it might be pragmatically redundant. Typically, we are interested in what is commonly known among G when members of G interact in any way, or at least have higher-order beliefs about one another. Standard multi-agent epistemic logic takes agents' interaction for granted. For instance, in standard epistemic logic, we cannot model situations where one agent is unaware of the other's existence. Consider any Kripke model with two agents: the king of the Netherlands (k) and the author of this text (i). While the author knows the king of the Netherlands exists and can reason about what he knows, the king of the Netherlands has no concept of the author whatsoever. Yet in standard

multi-agent epistemic logic, we can say that at least they commonly know tautologies: $\models C_{i,k}\top$. This common knowledge attribution is not false (at least given idealizations of standard epistemic logic about agents' logical omniscience), but pragmatically, it is not that meaningful.[5]

In what follows, we will still utilise the standard notion of common knowledge, as presented here. Since we will model strategically interacting agents, we will not stumble upon pragmatically questionable common knowledge attributions.

3.2 Multi-agent modal logics of agency

In this section, we will look into two interpretations of $\mathcal{L}(Ags, Var)$: a modal logic of coalitional power in games [Pau02] and a modal logic of *seeing to it that* [Hor01]. Both languages are closely related: the latter is more expressive than the former [BHT06]. Both formalisms deal with group actions, where the group is taken to be nothing but a set of agents.

3.2.1 Coalition logic

We start with the logic of coalitional power in games, or, for short, Coalition logic (CL). In CL, $\Box_G\varphi$ is interpreted as *coalition G can ensure that φ holds in the outcome of the game* (following [Pau02], we call groups coalitions in CL). We define the formal language $\mathcal{L}_{CL}$ of CL as follows:

$$\mathcal{L}_{CL} \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid [G]\varphi$$

where $p \in Var$, $G \subseteq Ags$. For every $i \in Ags$, $[\{i\}]\varphi$ corresponds to $\Box_i\varphi$, and for any non-singleton group $G \subseteq Ags$, $[G]\varphi$ corresponds to $\Box_G\varphi$ of $\mathcal{L}(Ags, Var)$. We will use $C, G, C_1, C_2, \ldots, G_1, G_2, \ldots$ as placeholders for coalitions. Note that $[\emptyset]\varphi$ is a well-formed formula of $\mathcal{L}_{CL}$. We interpret the empty coalition as the environment. Namely, $[\emptyset]\varphi$ reads as *φ will be true in any outcome of the game, independently of how agents act.* Semantically, we interpret $\mathcal{L}_{CL}$ over the class of *game models.*

3.2.1. Definition (Game models). Given the finite set of agents $Ags = \{1, \ldots, n\}$ and a countable set of propositional variables $Var = \{p_1, p_2, \ldots\}$, a game model

[5] There are some non-standard ways to avoid the problem. First, we can introduce awareness functions to distinguish between agents' explicit and implicit knowledge, syntactically restricting with formulae agents are aware of [FH87]. Then, we can represent the fact that the king of the Netherlands is not aware of the author by excluding any formulae with $K_i\varphi$ as its subformulae from the king's awareness set. A more recent alternative allows for modeling agents' uncertainty about whether other agents exist (or, using authors' terminology, whether others are dead or alive) in non-Kripkean semantics of simplicial complexes [DK25].

is a tuple of the following form:

$$M = (S, \{\Sigma_i\}_{i \in Ags}, g, V)$$

1. $S \neq \emptyset$ is a set of states;
2. For every $i \in Ags$, $\Sigma_i \neq \emptyset$ is a set of i's possible actions;
3. g is a function that maps each state $w \in S$ to a game:

$$g(w) \mapsto (S^w, \{\Sigma_i^w\}_{i \in Ags}, o^w)$$

   , where $S^w \subseteq S$ is a set of possible outcomes, $\Sigma_i^w \subseteq \Sigma_i$ is a set of actions available for i at w and $o^w : \prod_{i \in Ags} \Sigma_i^w \to S^w$ is a surjective outcome function, mapping each action profile to the outcome state;
4. $V : Var \to 2^W$ is a standard evaluation function.

For any coalition $C \subseteq Ags$, the set of C's joint actions is $\Sigma_C = \prod_{i \in C} \Sigma_i$. By Σ_{Ags}, consequently, we denote the set of action profiles $\{(\sigma_1, \ldots, \sigma_n) \mid \sigma_1 \in \Sigma_1, \ldots, \sigma_n \in \Sigma_n\}$. As for the empty coalition $\emptyset$, by $\Sigma_\emptyset = \{\sigma_\emptyset\}$ we denote a singleton set of the "action" of the empty coalition: to do nothing. For any state $w \in S$ and any joint action $\sigma_C \in \Sigma_C$, the set of possible outcomes of σ_C in w is defined as:

$$out(w, \sigma_C) = \{v \in S^w \mid \exists \sigma_{Ags} \in \Sigma_{Ags} : \sigma_{Ags}|_C = \sigma_C \text{ and } o^w(\sigma_{Ags}) = v\}$$

where $\sigma_{Ags}|_C$ is σ_{Ags}'s projection on C.[6] In other words, if C does the joint action σ_C at w, then the outcome of the game $g(w)$ will be in $out(w, \sigma_C)$, no matter what everyone else does. Note that $out(w, \sigma_\emptyset) = S^w$, since for any action profile $\sigma_{Ags} \in \Sigma_{Ags}$, $\sigma_{Ags}|_\emptyset = \sigma_\emptyset$. Therefore, $out(w, \sigma_\emptyset)$ is just a set of possible outcomes of the game $g(w)$.

For any coalitions $X, Y \subseteq Ags$ and any joint actions $\sigma_X \in \Sigma_X, \sigma_Y \in \Sigma_Y$, $\sigma_X \cup \sigma_Y \in \Sigma_{X \cup Y}$ is a joint action, such that $\sigma_X \cup \sigma_Y|_X = \sigma_X$ and $\sigma_X \cup \sigma_Y|_{X \setminus Y} = \sigma_Y|_{Y \setminus X}$. In other words, $\sigma_X \cup \sigma_Y$ is a $X \cup Y$'s joint action that agrees on X's action with σ_X, and with $Y \setminus X$'s actions – with σ_Y.

Given a game model $M = (S, \{\Sigma_i\}_{i \in Ags}, g, V)$, a **play** is an infinite sequence of the form:

$$(w_0, \sigma_0, w_1, \sigma_1, \ldots),$$

where, for any $i \in \mathbb{N}$, $w_i \in W$, $\sigma_i \in \Sigma_{Ags}^{w_i}$ and $w_{i+1} \in out(w_i, \sigma_i)$. Let us denote the set of all plays in M as Pl_M. A **partial play** is a finite prefix $(w_0, \sigma_0, \ldots, \sigma_{n-1}, w_n)$

[6] Let $\vec{\sigma} = (\sigma_1, \ldots, \sigma_n) \in \prod_{i \in Ags} \Sigma_i$ be an action profile and $C \subseteq Ags$ be a subset of agents. Assume members of C are enumerated in the increasing order, i.e. $C = \{j_1, \ldots, j_k\}$, where $j_1 < \ldots < j_k$ and $|C| = k$. Then, the projection of $\vec{\sigma}$ on C is the joint action $\vec{\sigma}|_C = (\sigma_{j_1}, \ldots, \sigma_{j_k}) \in \prod_{i \in C} \Sigma_i$, where $\forall i \in C : \vec{\sigma}(i) = \vec{\sigma}|_C(i)$.

of some play $(w_0, \sigma_0, \ldots) \in Pl_M$. Let us denote the set of all partial plays in M as $PartPl_M$.

Each play can be represented as a pair of functions $\tau = (\tau_S, \tau_A)$, where $\tau_A : \mathbb{N} \to \Sigma_{Ags}$ and $\tau_S : \mathbb{N} \to S$; consequently, $\tau_S(i+1) \in out(\tau_S(i), \tau_A(i))$ for every $i \in \mathbb{N}$. Two plays τ, τ' are **state-equivalent**, denoted as $\tau \simeq_S \tau'$, iff $\tau_S(0) = \tau'_S(0)$. For any coalition $C \subseteq Ags$, two plays τ, τ' are **action-equivalent** w.r.t. C, denoted as $\tau \simeq_C \tau'$, iff $\tau \simeq_S \tau'$ and $\tau_A(0) = \sigma$, $\tau_A(0) = \sigma'$, such that $\sigma|_C = \sigma'|_C$. Informally, τ and τ' are state-equivalent iff they start from the same state. τ and τ' are action-equivalent w.r.t. C iff they start with the same action of C in the same state. By $[\tau]_S = \{\tau' \in Pl_M \mid \tau' \simeq_S \tau\}$ and $[\tau]_C = \{\tau' \in Pl_M \mid \tau \simeq_C \tau'\}$, we denote the state- and C-action equivalence classes of τ, respectively.

Tree-like game models are of technical interest. In tree-like models, there is a unique initial state, and every state has a unique predecessor. Tree-like models help to build connections between coalition logic and the logic of branching time, which will be important for us in the next section.

3.2.2. Definition (Tree-like game models). A model M is **tree-like** iff the following two conditions hold:

1. **Rootedness**: there exists a unique state $w_0 \in S$, such that for every other state $v \in S$, there is a partial play $(w_0, \sigma_0, \ldots, \sigma_n, v) \in PartPl_M$;
2. **No cycles**: for every play $(\tau_S, \tau_A) \in Pl_M$ and any $i, j \in \mathbb{N}$: $\tau_S(i) = \tau_S(j)$ iff $i = j$;
3. **Unique predecessors**: for any worlds $w, u, v \in S$, if $u \in out(w, \sigma_\emptyset)$ and $u \in out(v, \sigma_\emptyset)$, then $w = v$.

As it was proven in [GD06], Coalition Logic (as well as its more expressive counterpart, Alternating-time Temporal Logic) cannot distinguish between arbitrary and tree-like game models: for every pointed game model (M, w), there exists a tree-like pointed game model (M', v) that satisfies the same $\mathcal{L}_{CL}$-formulae.

3.2.3. Definition. Given any game model $M = (S, \{\Sigma_i\}_{i \in Ags}, g, V)$ and any state $w \in S$, we recursively define $\models$ relation as follows (clauses for atomic propositions and Boolean connectives are standard, as in Def. 3.0.3):

$$M, w \models [G]\varphi \Leftrightarrow \exists \sigma_G \in \Sigma_G (out(w, \sigma_G) \subseteq [\![\varphi]\!]_M).$$

The axiomatic system $\mathbf{L}_{CL}$, as defined in Table 3.2, was shown to be strongly sound and complete w.r.t. the class of game models [Pau02; GJT13]. ($\bot$) and ($\top$) express that no coalition can ensure $\bot$ and any coalition can trivially ensure $\top$, respectively. (N) corresponds to the fact that an outcome of each game is determined by nothing but agents' choices of actions: if there is an outcome of the game, than the grand coalition Ags can ensure that outcome. (M) expresses

(CPL)	Tautologies of classical propositional logic
($\bot$)	$\neg[G]\bot$
($\top$)	$[G]\top$
(N)	$\neg[\emptyset]\neg\varphi \to [Ags]\varphi$
(M)	$[G](\varphi \wedge \psi) \to [G]\varphi$
(S)	$([G_1]\varphi \wedge [G_2]\psi) \to [G_1 \cup G_2](\varphi \wedge \psi)$
(MP)	If $\vdash \varphi$ and $\vdash \varphi \to \psi$, then $\vdash \psi$
(RE)	If $\vdash \varphi \leftrightarrow \psi$, then $\vdash [G]\varphi \leftrightarrow [G]\psi$

Table 3.2: Logic $\mathbf{L}_{CL}$

outcome monotonicity: if the coalition G can guarantee $\varphi \wedge \psi$, then G can also guarantee a logically weaker condition φ. (S) corresponds to additivity (to be explained in detail later).

Although the semantics for $\mathcal{L}_{CL}$ appears to be way more convoluted than the general scheme we have presented in Def. 3.0.3, it follows the very same principles. Namely, every set of agents $C \subseteq Ags$ is treated as a group (or, following the terminology of a seminal paper [Pau02], a *coalition*). Coalitional power functionally depends on the members' powers, since coalitions' joint actions are nothing but aggregations of members' actions. It is not hard to see that for every coalition $C \subseteq Ags$, every joint acion $\sigma_C \in \Sigma_C$ in every pointed model (M, w):

$$out(w, \sigma_C) = \bigcap_{i \in C} out(w, \sigma_C|_{\{i\}})$$

Such a reductive approach to coalitions accounts for serious conceptual problems. First, note that coalitional powers are *additive*:

3.2.4. Proposition (Additivity in $\mathcal{L}_{CL}$). *For any disjoint coalitions $C_1, C_2 \subseteq Ags$, the following axiom schema is valid on the class of game models:*

$$\models ([C_1]\varphi \wedge [C_2]\psi) \to [C_1 \cup C_2](\varphi \wedge \psi)$$

Proof:
Let (M, w) be an arbitrary pointed game model, such that $M, w \models [C_1]\varphi \wedge [C_2]\psi$. Then, by Def.3.2.3, there exist joint actions $\sigma_{C_1} \in \Sigma_{C_1}$ and $\sigma_{C_2} \in \Sigma_{C_2}$, such that $out(w, \sigma_{C_1}) \subseteq [\![\varphi]\!]_M$ and $out(w, \sigma_{C_2}) \subseteq [\![\psi]\!]_M$. Let $\sigma_{C_1 \cup C_2} \in \Sigma_{C_1 \cup C_2}$ be a $C_1 \cup C_2$'s joint action, such that $\sigma_{C_1 \cup C_2}|_{C_1} = \sigma_{C_1}$ and $\sigma_{C_1 \cup C_2}|_{C_2} = \sigma_{C_2}$. Then, for every action profile $\sigma_{Ags} \in \Sigma_{Ags}$, $\sigma_{Ags}|_{C_1 \cup C_2} = \sigma_{C_1 \cup C_2}$ iff $\sigma_{Ags}|_{C_1} = \sigma_{C_1}$ and $\sigma_{Ags}|_{C_2} = \sigma_{C_2}$. Consequently,

$$out(w, \sigma_{C_1 \cup C_2}) = \{v \in S^w \mid \exists \sigma_{Ags} \in \Sigma_{Ags} : \sigma_{Ags}|_{C_1} = \sigma_{C_1}, \sigma_{Ags}|_{C_2} = \sigma_{C_2}, o^w(\sigma_{Ags}) = v\}$$

from which it follows that $out(w, \sigma_{C_1 \cup C_2}) = out(w, \sigma_{C_1}) \cap out(w, \sigma_{C_2})$. Since $out(w, \sigma_{C_1}) \subseteq [\![\varphi]\!]_M$ and $out(w, \sigma_{C_2}) \subseteq [\![\psi]\!]_M$, $out(w, \sigma_{C_1 \cup C_2}) \subseteq [\![\varphi]\!]_M \cap [\![\psi]\!]_M = [\![\varphi \wedge \psi]\!]_M$. Hence, there exists a joint action $\sigma_{C_1 \cup C_2} \in \Sigma_{C_1 \cup C_2}$, such that:

$$out(w, \sigma_{C_1 \cup C_2}) \subseteq [\![\varphi \wedge \psi]\!]_M.$$

Then, by Def. 3.2.3, $M, w \models [C_1 \cup C_2](\varphi \wedge \psi)$. Since (M, w) was chosen arbitrarily, we can conclude that $\models ([C_1]\varphi \wedge [C_2]\psi) \to [C_1 \cup C_2](\varphi \wedge \psi)$. □

Additivity leads us to the pragmatically questionable attributions of coalitional abilities. For example, in a possible world, where the Netherlands has not become a constitutional monarchy, the king of the Netherlands has the legal power to declare war: $[king]war$. The author of this text (being a non-Dutch citizen, thus not being eligible to conscription) has the legal power to flee the Netherlands in case of war: $[i](war \to fled)$. Then, the coalition of the king and the author (legally speaking) can ensure that the war is declared, and the author has fled: $[king, i](war \wedge fled)$. It is not wrong, but pragmatically, should we even think of the king and the author as a *coalition*, given that we speak about the author's ability to flee the war that the king can declare?

Unfortunately, additivity leads us not only to pragmatically questionable, but also to conceptually wrong attribution of coalitions' abilities, especially when it comes to the ability to coordinate.

3.2.5. Example. Let i and j be two flatmates who never talk to one another. They should pay the water and electricity bills. Each of them has enough data and money to pay either of the bills, but neither of them can pay both bills. Unless i and j cooperate and communicate, they cannot coordinate their actions to ensure that both bills are paid. For instance, i can pay either of the bills. Which one should i pay? She should pay for water iff j pays for electricity, and should pay for electricity iff j pays for water. Yet there is no way for i to know which bill j pays. The situation is analogous for j. Even if they do their best to pay both bills, there is a chance that they will end up paying one bill twice without paying the other. So i and j cannot ensure that both bills are paid, given that they do not cooperate and coordinate. If they do cooperate and coordinate (hence, act as a coalition in the common sense of this word), this ability attribution is unproblematic.

Let us model the situation where i and j do not communicate and cooperate via a game model illustrated in Figure 3.2. Let $M = (S, \Sigma_i, \Sigma_j, g, V)$, be a game model, where:

- $W = \{s, s_1, \ldots, s_4\}$ is the set of states. s is the initial state, where i and j deliberate what to do with the bills;

- $\Sigma_i = \Sigma_j = \{w, e, \epsilon\}$, where w is the action of paying the water bill, e – paying the electricity bill, ϵ – doing nothing;
- $g(s) = \{S^s, \Sigma_i^s, \Sigma_j^s, o^s\}$, where $S^s = \{s_1, \ldots, s_4\}$, $\Sigma_i^s = \Sigma_j^s = \{w, e\}$ and $o^s((w, w)) = s_1$. $o^s((w, e)) = s_2$, $o^s((e, w)) = s_3$, $o^s((e, e)) = s_4$. For any other state $v \in S \setminus \{w\}$, $g(v)$ is a trivial game, where $S^v = \{v\}$, $\Sigma_j^v = \Sigma_i^v = \{\epsilon\}$ and $o^v((\epsilon, \epsilon)) = v$. In other words, in all states, except for s, both agents can only do nothing, and nothing will happen.[7]
- Let $Var = \{water, electr\}$, where $water$ and $electr$ are the propositions meaning that the water bill is paid and the electricity bill is paid, respectively. Consequently, $V(water) = \{s_1, s_2, s_3\}$, $V(electr) = \{s_2, s_3, s_4\}$

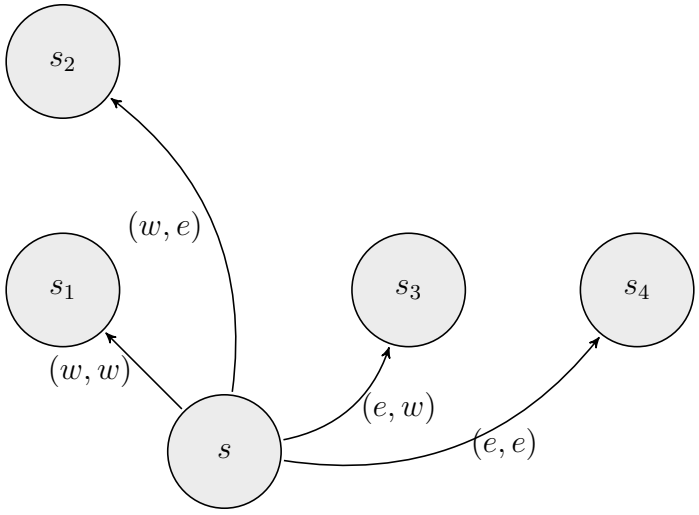

Figure 3.2: Game model M. Arrows represent transitions from the states to outcomes, labeled by the corresponding tuples (σ_i, σ_j), where the first coordinate is i's action, and the second – j's action. In the "final" states $s_1, \ldots, s_4$, every agent has only one action – to do nothing – which leads to transition to the very same state. We omit arrows for those transitions.

First, note that in (M, s), i can ensure that the water bill is paid. Since $w \in \Sigma_i$ and $out(s, w) = \{s_1, s_2\} \subseteq [\![water]\!]_M$, $M, s \models [i]water$. Analogously, j can ensure that the electricity bill is paid: $M, s \models [j]electr$. By additivity, solely from i's ability to pay the water bill and j's ability to pay the electricity bill, it follows that the coalition of $\{i, j\}$ can ensure that both bills are paid:

$$\models ([i]water \wedge [j]electr) \rightarrow [i, j](water \wedge electr)$$

Therefore, we should attribute the coalition of i and j the ability to ensure that both bills are paid: $M, w \models [i, j](water \wedge electr)$. But as we have argued, this is

[7] "Heaven is a place, a place where nothing, nothing ever happens".

conceptually wrong: since i and j do not communicate, they cannot coordinate properly to ensure that both bills are paid!

To conclude, reasoning about coalitions in games via the standard scheme of MAMLs results not only in pragmatically questionable but also in conceptually wrong attributions of coalitions' abilities. To fix it, we should stop treating non-cooperating sets of agents as coalitions.

3.2.2 Group STIT logic

Next, we move to the more expressive framework, in which we can reason not only about what agents and groups *can* do, but also what they in fact do. We work with *STIT logic*. STIT (seeing to it that) logic is a multi-agent logic of agency that originates from a series of articles by Belnap, Perloff and Xu [BPX01]. A richer version of STIT logic with groups was introduced in [Hor01]. The formalism accounts for agents' choices of action under indeterministic time, where multiple futures are possible, depending on how agents and the environment behave. For a detailed overview of STIT logics and their history, see [SMK09].

As usual, we first fix the notation we are to use, as well as the new reading of $\Box_G$-modalities. The syntax of *discrete temporal STIT logic* is defined by the following grammar:

$$\mathcal{L}^t_{STIT} \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid X\varphi \mid Y\varphi \mid G\varphi \mid H\varphi \mid \Box\varphi \mid [stit]_C\varphi$$

where $p \in Var, C \subseteq Ags$. $X\varphi$ reads as "*in the next moment in time,* φ"; $Y\varphi$ – "*in the previous moment in time,* φ"; $G\varphi$ reads as "φ *will always be true in the future*", $H\varphi$ – "φ *was always true in the past*". $\Box\varphi$ stands for "*it is historically necessary that* φ". By historic necessity, we understand that φ is true, independently of how agents and the environment behave now and in the future.

$[stit]_C\varphi$ stands for "*group* C *sees to it that* φ *holds*". By seeing to it that φ, we understand the behaviour which forces φ to be true no matter what all other agents do. We use standard abbreviations for duals of the modalities: $\Diamond := \neg\Box\neg$, $\langle stit\rangle_C := \neg[stit]_C\neg$, $P := \neg H\neg$, $F := \neg G\neg$, where $\Diamond\varphi$ reads as "*it is historically possible that* φ"; $\langle stit\rangle_C\varphi$ – "C*'s choice of action does not rule out* φ", $P\varphi$ –"φ *was true at some moment in past*", $F\varphi$ – "φ *will be true at some point in future*". As in $\mathcal{L}_{CL}$, we treat the empty group as the environment: $[stit]_\emptyset\varphi$ corresponds to the historic necessity. In other words, $\Box$ is redundant and can be expressed as follows $\Box\varphi := [stit]_\emptyset\varphi$.

We use some widely accepted acronyms in STIT logic. A group C *deliberately* sees to it that φ iff C sees to it that φ and φ is not necessary: $[dstit]_C\varphi := [stit]_C\varphi \wedge \neg\Box\varphi$. $[dstit]_C\varphi$ is a useful acronym, since $\Box\varphi$ entails $[stit]_C\varphi$. Naturally,

if φ is historically necessary (i.e., φ is true no matter what all agents will do), then whatever C does, they behave in a way that φ is true, no matter what $Ags \setminus C$ does. Via using $[dstit]_C$ as an acronym, we can distinguish between necessities and propositions at least partially caused by C. Next, we can express the groups' abilities. Namely, a group C can ensure φ iff $\Diamond[stit]_C\varphi$, i.e., C can see to it that φ. So we can express $[C]$-modality of $\mathcal{L}_{CL}$ in $\mathcal{L}^t_{STIT}$ as follows: $[C]\varphi := \Diamond[stit]_C X\varphi$ [BHT06].

The fragment of $\mathcal{L}^t_{STIT}$ without $[stit]_C$-modalities is *discrete Ockhamist branching time temporal logic*, denoted as $\mathcal{L}_{OBTL}$ [Rey02]. The fragment of $\mathcal{L}^t_{STIT}$ without H- and G-modalities is called *next time STIT logic*, denoted as $\mathcal{L}^X_{STIT}$ [Bro11b]. $\mathcal{L}^t_{STIT}$ without any temporal modalities, denoted as $\mathcal{L}_{STIT}$, is called *atemporal STIT logic* [HS08]. We use $\mathcal{L}_{OBTL}$ to reason about the temporal behaviour of the environment, independently of the agents. $\mathcal{L}^X_{STIT}$ is useful to reason about the consequences of agents' actions that are observable in finite time. $\mathcal{L}_{STIT}$ is handy to reason about agents' actions just at one time point, e.g., about one-shot or normal form games.

Standard STIT logic models agents' choices of action in indeterministic environments, *branching time structures*.

3.2.6. Definition ($BT + AC$ frames). A *branching time structure* is a tuple of the form $(T, <)$, where $T \neq \emptyset$ is a non-empty set of moments and $< \subseteq T \times T$ is an irreflexive, transitive, discrete[8] and backward-linear[9] relation on T, ordering moments by precedence/succession. In other words, $m_1 < m_2$ means that m_1 precedes m_2. For every moment m, m' is an **immediate successor of** m, denoted as $m' \in succ(m)$ iff $m < m'$, and there is no moment m'', such that $m < m'' < m'$. **Histories** are maximal $<$-ordered chains of moments: $h \subseteq T$ is a history iff for any $m \in h$, if $m' < m$ or $m < m'$, then $m' \in h$, and for any two moments $m, m' \in h$, either $m < m'$ or $m' < m$ or $m = m'$. The set of all histories is denoted as H. For any moment $m \in T$, the set of histories passing through m is defined as follows: $H_m = \{h \in H \mid m \in h\}$. **Indices** are moment/history pairs, such that the history passes through the moment. Formally, the set of indices is: $Index = \{(m, h) \in T \times H \mid m \in h\}$. Given two histories h_1, h_2 and a moment $m \in T$, h_1 and h_2 are **undivided in** m iff 1) $m \in h_1 \cap h_2$; 2) there exists a moment m', such that $m < m'$ and $m' \in h_1 \cap h_2$.

[8]By discreteness we mean that for any moment $m \in T$ there exists another moment $m' \in T$, such that $m < m'$, and there is no moment $m'' \in T$, such that $m < m'' < m'$. In the general case, discreteness is not assumed. We take this assumption throughout the thesis to stay in line with the discrete semantics of Coalition Logic. Since we are not concerned with the metaphysics of time, this assumption is conceptually harmless for our work.

[9]By backward linearity we mean that for any $m_1, m_2, m_3 \in T$, if $m_3 < m_1$ and $m_2 < m_1$, then either $m_2 < m_3$ or $m_3 < m_2$ or $m_3 = m_2$.

A $BT + AC$ frame[10] is a tuple $F = (T, <, Choice)$, where $(T, <)$ is a branching time structure, and $Choice : (Ags \times T) \to 2^{2^H}$ is a choice function, mapping each agent and each moment in time to a partition of H_m. For brevity, if two histories h_1, h_2 are in the same partition cell of $Choice(m, i)$, we denote it as $h_1 \sim_i^m h_2$.

1. **Independence of agents**.For any moment m, let $s_m : Ags \to 2^H$ be an action selection function, such that $s_m(i) \in Choice(m, i)$ for any $i \in Ags$. Then, for any action selection function s_m, $\bigcap_{i \in Ags} s_m(i) \neq \emptyset$.
2. **No choice between undivided histories**. For any agent $i \in Ags$ and any moment $m \in T$, if $h_1, h_2 \in H_m$ are undivided in m, then $h_1 \sim_i^m h_2$.

As usual, every $BT + AC$ frame F can be extended to a model $M = (F, V)$ by a valuation function $V : Var \to 2^T$.

In $BT + AC$ models, actions are idealised as choices between sets of possible futures. Intuitively, $Choice(m, i)$ represents the set of actions available for i at m (it may be seen as an analogue to Σ_i^m in game models of Def. 3.2.1). For each equivalence class $X \in Choice(m, i)$, there is an action available for i at m, such that, if performed, it guarantees that the outcome history will be in $X \subseteq H_m$. Then, *independence of agents* ensures that any choice of any agent is compatible with any other choice of everyone else. *No choice between undivided histories* guarantees that agents can only choose between histories that are about to branch: if they are to divide later in time, then no agent can distinguish them by her current choice of action.

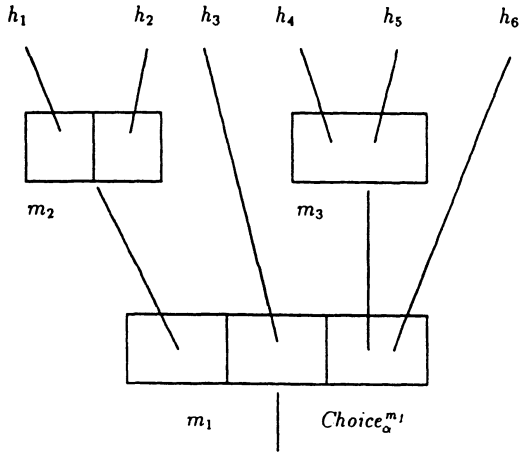

Figure 3.3: Example of a single-agent $BT + AC$ frame [HB95, Figure 2, p.589].

3.2.7. Definition (Semantic clauses for $\mathcal{L}^t_{STIT}$ over $BT + AC$). Given any $BT+AC$-model $M = (T, <, Choice, V)$ and any index $(m, h) \in Index$, we recursively

[10] $BT + AC$ stands for "Branching Time + Action Choice".

define $\models$ relation as follows:

$$
\begin{array}{lcl}
M,(m,h) \models p & \Leftrightarrow & m \in V(p) \\
M,(m,h) \models \neg\varphi & \Leftrightarrow & M,(m,h) \not\models \varphi \\
M,(m,h) \models \varphi \vee \psi & \Leftrightarrow & M,(m,h) \models \varphi \text{ or } M,(m,h) \models \psi \\
M,(m,h) \models X\varphi & \Leftrightarrow & \forall m' \in h : m' \in succ(m) \Rightarrow M,(m',h) \models \varphi \\
M,(m,h) \models Y\varphi & \Leftrightarrow & \forall m' \in h : m \in succ(m') \Rightarrow M,(m',h) \models \varphi \\
M,(m,h) \models G\varphi & \Leftrightarrow & \forall m' \in h(m < m' \Rightarrow M,(m',h) \models \varphi) \\
M,(m,h) \models H\varphi & \Leftrightarrow & \forall m' \in h(m' < m \Rightarrow M,(m',h) \models \varphi) \\
M,(m,h) \models \Box\varphi & \Leftrightarrow & \forall h' \in H_m(M,(m,h') \models \varphi) \\
M,(m,h) \models [stit]_C\varphi & \Leftrightarrow & \forall h' \in H_m(\forall i \in C(h \sim_i^m h' \Rightarrow M,(m,h') \models \varphi)
\end{array}
$$

The semantics for $\mathcal{L}^t_{STIT}$ we give in Definition 3.2.7, disregarding some arbitrary notational choices and discreteness assumption, corresponds to the one proposed in the seminal work [BPX01].

Interestingly, tree-like game models, as in Def. 3.2.2, are a special case of $BT{+}AC$ models. Let $M = (S, \{\Sigma_i\}_{i\in Ags}, g, V)$ be a tree-like game model, whose root is $w_0 \in S$. Let $<\subseteq S \times S$ be a relation, such that $w < v$ iff there exists a partial play $(w, \sigma_0, \ldots, \sigma_{n-1}, v) \in PartPl_M$. Let $H = \{(\tau_S, \tau_A) \in Pl_M \mid \tau_S(0) = w_0\}$ be the set of plays in M that start with the root state w_0. Then, given two histories $\tau = (\tau_S, \tau_A) \in H, \tau' = (\tau'_S, \tau'_A) \in H$ and a natural number $n \in \mathbb{N}$, $\tau \sim_i^n \tau'$ iff $\tau_S(n) = \tau'_S(n)$ and $\tau_A(n) = \sigma$, $\tau'_A(n) = \sigma'$, such that $\sigma(i) = \sigma'(i)$. Or, informally, two histories $\tau, \tau' \in H$ are equivalent w.r.t. i's choice at the nth moment iff τ and τ' are equal up to the nth moment, and i does the same action in both τ and τ' at the nth moment. Let $[\tau]_{\sim_i^n} = \{\tau' \in H \,|\, \tau \sim_i^n \tau'\}$. Then, $Choice : (Ags \times S) \to 2^{2^H}$ may be defined as a follows:

$$Choice(w,i) = \{[\tau]_{\sim_i^n} \mid \tau \in H, \tau_S(n) = w \text{ for some } n \in \mathbb{N}\}$$

A tuple $(S, <, Choice)$ is a $BT + AC$-model. For a detailed proof of this fact, see [BL18, Lemma 5.6, p.387].

The branching time structures provide us with both technical and philosophical complications. [Nis79; Pet95] argued against branching time structures from the philosophical point of view. Since the metaphysics of tense is not in the main scope of our inquiry, we omit the details of this critique, giving only one semi-formal counter-example, relevant to agency. As was shown in [Rey02], the following $\mathcal{L}_{OBTL}$-formula is valid on branching time (hence, also on $BT + AC$) models:

$$\models \Box G(p \to \Diamond Fp) \to \Diamond G(p \to Fp)$$

To show why this validity is undesirable, we provide an informal example that falsifies it. Think of an agent that writes an arbitrary formula of propositional logic, built of a single propositional variable q. As we know, such propositional formulae are finite, but are not bounded in length. Let p be a proposition that reads as "*the agent is not finished writing a formula*". Then, $\Box G(p \rightarrow \Diamond Fp)$ holds. In all possible future moments ($\Box G$), as soon as the agent has not yet written the formula (i.e., if p holds), she can choose to continue writing it by, e.g., adding $\wedge q$ as a suffix to the string she already has; so it is historically possible that she will not be done in the future (then, $\Diamond Fp$ holds). Yet it is historically impossible for her to keep writing the formula forever, given that the formula is finite, so it is historically necessary that she will have to write down the last symbol of the formula once: $\Box F(p \wedge \neg Fp)$. In $BT + AC$ structures, we cannot model this scenario: we will be forced to have a "limit history", where p is always true.[11] On top of such conceptual issues, branching time structures are problematic metalogically: the axiomatization of Ockhamist branching time temporal logic interpreted over branching time models remains an open problem.

To avoid both conceptual and technical problems around branching time structures, we use the alternative semantics for STIT, first introduced in [Lor13]. It is closer to the standard Kripke semantics of modal logic. In terms of the temporal component, it uses *Kamp frames* for $\mathcal{L}_{OBTL}$, which are known to avoid several problems with branching time structures.[12]

3.2.8. Definition (Kripke STIT frames). A Kripke STIT frame is a tuple $F = (W, R_{\prec}, R_{\Box}, \{R_i\}_{i \in Ags})$, where:

1. $W \neq \emptyset$ is a set of time points. For brevity, "*(time) points*" and "*(possible) worlds*" are to be used interchangeably;

2. $R_{\Box}, \{R_i\}_{i \in Ags}$ are equivalence relations on W. $wR_{\Box}v$ represents historical equivalence. wR_iv represents equivalence w.r.t. the current choice of action by agent i. $R_{\Box}, \{R_i\}_{i \in Ags}$ have the next additional properties:

   (a) $R_i \subseteq R_{\Box}$ for every $i \in Ags$: agent's choice of actions only distinguishes between historically equivalent alternatives;

   (b) *Independence of agents:* if $Card(Ags) = n$, then for any sequence of $R_{\Box}$-connected time points $(w_1, \ldots, w_n)$, $\bigcap_{i \in Ags} R_i(w_i) \neq \emptyset$;

3. $R_{\prec}$ is a serial relation on W. $wR_{\prec}v$ iff time point v is a successor of w. By $R_{<}$ we denote the transitive closure of $R_{\prec}$. $R_{\prec}$ has the next properties:

[11] In [Nis79, p.471], a counter-example, similar in spirit, is given.

[12] Kamp frames were introduced, to our knowledge, in [Tho84]. For a comparison between the two semantic frameworks for Ockhamist branching time temporal logic, see [Rey02, Section 8].

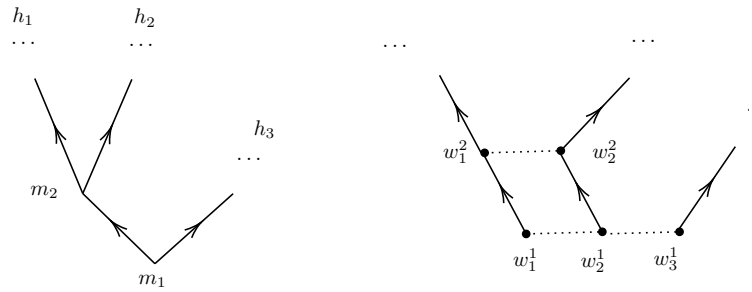

Figure 3.4: A branching time structure (on the left) and its representations as a Kamp frame (on the right). Moment/history pairs on the left correspond to time points on the right: e.g., m_1/h_1 corresponds to w_1^1, w_2^1 corresponds to m_1/h_2, etc. Arrowed lines represent $<$ and $R_<$ relations, respectively, and dotted lines represent $R_\Box$.

(a) *No branching backwards*: if $vR_\prec w$ and $uR_\prec w$, then $v = u$;

(b) *No branching forwards*: if $wR_\prec v$ and $wR_\prec u$, then $v = u$;

(c) *Irreflexivity of* $R_<$: For all $w, v \in W$: if $wR_\Box v$, then $(w, v) \notin R_\prec^n$ and $(v, w) \notin R_\prec^n$ for any $n \in \mathbb{N}$. Equivalently, $wR_\Box v$ entails that neither $wR_< v$ nor $vR_< w$. If time points w and v are historically equivalent, then neither of them succeeds the other in any number of steps;

(d) *No choice between undivided histories*: $R_\prec \circ R_\Box \subseteq \bigcap_{i \in Ags} R_i \circ R_\prec$;

4. *Group actions*: for every $G \subseteq Ags$, $R_G = \bigcap_{i \in G} R_i$;

Given a frame $F = (W, R_\prec, R_\Box, \{R_i\}_{i \in Ags})$, we denote the domain W of F as $Dom(F)$. Every Kripke STIT frame F standardly extends to a model $M = (F, V)$, where $V : Var \to 2^W$ is a valuation function.

From 3c together with reflexivity of $R_\Box$ it follows that $R_<$ is irreflexive. For every $w \in W$, $H_w = \{w\} \cup R_<(w) \cup R_<^{-1}(w)$ is the *history* of w, i.e., the set of time points that are temporally related to w, either as predecessors or as successors in any number of steps. From 3a and 3b together with irreflexivity and transitivity of $R_<$ it follows that for every history H_w, $(H_w, R_<)$ is a strict linear order. From this fact, it follows that every time point w has a unique history passing through it. $wR_\Box v$ means that w and v are time points that represent the same moment in time in different histories. In other words, a tuple $(W, R_\Box, R_<)$ represents a branching time structure, where moments are represented as $R_\Box$-equivalence classes. Given two moments $R_\Box(w)$ and $R_\Box(v)$, $R_\Box(w)$ precedes $R_\Box(v)$ iff there exist time points $x \in R_\Box(w), y \in R_\Box(v)$, such that $xR_< y$. For a graphic representation of the relation between branching time structures and their representation in Kripke STIT frames, see Figure 3.4.

R_i relation models how agent i can choose among her possible actions. Her choice of action determines which *set* of histories remains possible. Condition 3d guarantees that agents cannot choose between histories that do not branch yet. By 2b, individuals' choice of actions is independent. By 4, a group action is nothing but the simultaneous actions of group members. See Figure 3.5 for an example of a Kripke STIT frame.

3.2.9. Definition. Given a Kripke STIT model $M = (W, R_{\prec}, R_{\Box}, \{R_i\}_{i\in Ags}, V)$ and $w \in W$, the truth of a formula $\varphi \in \mathcal{L}^t_{STIT}$ is defined recursively as follows:

$$
\begin{array}{lcl}
M, w \models p & \Leftrightarrow & w \in V(p) \\
M, w \models \neg\varphi & \Leftrightarrow & M, w \not\models \varphi \\
M, w \models \varphi \vee \psi & \Leftrightarrow & M, w \models \varphi \text{ or } M, w \models \psi \\
M, w \models \Box\varphi & \Leftrightarrow & \forall v \in W(wR_{\Box}v \Rightarrow M, v \models \varphi) \\
M, w \models X\varphi & \Leftrightarrow & \forall v \in W(wR_{\prec}v \Rightarrow M, v \models \varphi) \\
M, w \models Y\varphi & \Leftrightarrow & \forall v \in W(vR_{\prec}w \Rightarrow M, v \models \varphi) \\
M, w \models G\varphi & \Leftrightarrow & \forall v \in W(wR_{<}v \Rightarrow M, v \models \varphi) \\
M, w \models H\varphi & \Leftrightarrow & \forall v \in W(vR_{<}w \Rightarrow M, v \models \varphi) \\
M, w \models [stit]_C\varphi & \Leftrightarrow & \forall v \in W(wR_C v \Rightarrow M, v \models \varphi)
\end{array}
$$

When we work with the atemporal fragment, we can simplify the semantics by dropping the relation $R_{\prec}$ from the definition of Kripke STIT frames.

3.2.10. Definition. An atemporal Kripke STIT frame (shortly, **STIT**-frame) is a tuple of the form $F = (W, R_{\Box}, \{R_i\}_{i\in Ags})$, where:

1. $W \neq \emptyset$ is a set of states;
2. $R_{\Box} \subseteq W \times W$ is an equivalence relation on W, representing historical equivalence. For every world $w \in W$, the equivalence class $[w]_{R_{\Box}} = \{v \in W \mid wR_{\Box}v\}$ represent the set of historically equivalent states.
3. For each $i \in Ags$, $R_i \subseteq W \times W$ is an equivalence relation, representing the equivalence w.r.t. i's choice of action. For every world $w \in W$, the equivalence class $[w]_{R_i} = \{v \in W | wR_i v\}$ represents the set of states that are compatible with some action of i. R_i has the following additional properties:
   (a) For any two worlds $w, v \in W$, if $wR_i v$, then $wR_{\Box}v$;
   (b) **Independence of agents:** For any worlds $w_1, \ldots, w_n \in W$, such that $w_1 R_{\Box} \ldots R_{\Box} w_n$, $\bigcap_{i\in Ags}[w_i]_{R_i} \neq \emptyset$;
   (c) **Group actions**: for every $G \subseteq Ags$, $R_G = \bigcap_{i\in G} R_i$;

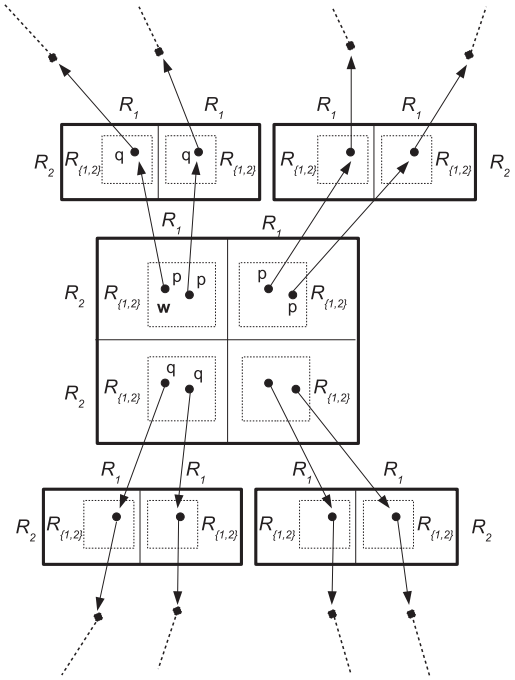

Figure 3.5: Example of a two-agent Kripke STIT model [Lor13, Figure 1, p.378].

As usual, a **STIT**-frame F can be extended to a **STIT**-model $M = (F, V)$ with a valuation function $V : Var \to 2^W$. Any equivalence relation $R \in \{R_\Box, R_1, \ldots, R_n\}$ yields the partition $[R] = \{[w]_R \mid w \in W\}$ of W. Often, we define partitions instead of equivalence relations for the sake of brevity. Such definitions are equivalent.

Group STIT logic is often used for formal models of responsibility attribution (see, e.g., [BCS20; AB22; HB95]). To hold someone morally or legally accountable for her actions, it is crucial to understand not only what she has done, but also whether she could have done otherwise [Fra18]. The expressivity of $\mathcal{L}_{STIT}$ allows us to reason about what actions the agents choose and what alternative choices they have. This expressive power was utilised to formalise the responsibility attribution to and within groups as well [Ser21; CM09]. Characteristically for MAMLs, group STIT logic treats any set of agents as a group, while any aggregation of these set members' actions is a group action. In what follows, we show which problems it creates specifically for capturing group responsibility.

In [Ser21], it is argued that the group G is causally responsible[13] for the outcome φ iff 1) G deliberately sees to it that φ, and 2) there is no proper subgroup $J \subset G$, such that J sees to it that φ. In other words, actions of all members of G is a

[13] An agent or a group of agents is held causally responsible for an outcome iff their actions have caused that outcome, regardless of intent or awareness [Hor01].

minimal sufficient condition to force φ. Formally, it is expressed in $\mathcal{L}_{STIT}$ as follows:

$$\Delta_G^{\min}\varphi := [dstit]_G\varphi \wedge \bigwedge_{J\subset G} \neg[stit]_J\varphi$$

where, from now on, $\Delta_G^{\min}\varphi$ reads as *the group G is causally responsible for φ*. The definition is in line with NESS-tests, prevalent in legal literature.[14] To challenge this definition, consider the next example.

3.2.11. Example. Consider three agents: a killer, a lookout, and a passerby (we denote them as **a**, **b**, and **c**, respectively). The killer and the lookout cooperate, while the passerby has no ties to any of them. The killer commits murder. The lookout guards the area to cover the crime in case someone happens to pass by. The passerby, being the only person who can witness the murder, takes a different route, by chance not walking past the location where the murder takes place.

Among the three agents and the groups formed by them, who is responsible for the covered-up murder?

It seems clear that the murderer and the lookout are responsible for the covered-up murder as a group. It might be plausible to attribute some form of responsibility (e.g., complicity[15]) for the murder to the lookout, and for the cover-up – to the murderer. Holding the passerby responsible for either the murder or the cover-up seems highly implausible, as well as attributing any responsibility for a "group" of, e.g., the passerby and the murderer, given that they have not cooperated.

Let us build a **STIT**-model for that case. For simplicity, we assume that each agent had the binary choice: the killer could have killed or not, the lookout could have guarded the area or not, and the passerby could have passed by or not. Each aggregation of agents' possible choices yields a unique outcome. Let $Var = \{k, l, p\}$ be a set of propositions, meaning *someone was **k**illed*, *the area was guarded by the **l**ookout*, and *someone has **p**assed by the area*, respectively. As an acronym, we introduce the following proposition: $witness = k \wedge \neg l \wedge p$, meaning that the murder was witnessed (i.e., the murder has happened, someone passed by, but no one has guarded the area to cover the murder up). The resulting **STIT**-model is a tuple $M = (W, R_\Box, R_a, R_b, R_c, V)$, where:

1. $W = \{w_1, \ldots, w_8\}$;
2. $R_\Box = W \times W$;

[14] NESS stands for "*necessary element of a sufficient set*" [Wri13]. In our context, the agent i passes NESS-test and has caused φ iff she is in a set of agents C, whose actions are sufficient for φ, but without i's contribution, C would not have ensured φ.

[15] An agent is complicit in bringing about that φ iff they have merely participated in a group action that has resulted in φ, even if their actions were neither necessary nor sufficient to achieve φ [Baz13; Kut00].

3. $[R_a] = \{\{w_1, w_2, w_3, w_4\}, \{w_5, w_6, w_7, w_8\}\}$;
4. $[R_b] = \{\{w_1, w_2, w_6, w_7\}, \{w_3, w_4, w_5, w_8\}\}$;
5. $[R_c] = \{\{w_1, w_3, w_5, w_7\}, \{w_2, w_4, w_6, w_8\}\}$;
6. $V(k) = \{w_1, w_2, w_3, w_4\}$, $V(l) = \{w_1, w_2, w_6, w_7\}$ and $V(p) = \{w_1, w_3, w_5, w_7\}$.

The real state of the Example 3.2.11 is w_2: a has $k$illed, b has $l$ooked out, and c has not $p$assed by. See Figure 3.6 for the illustration of the model.

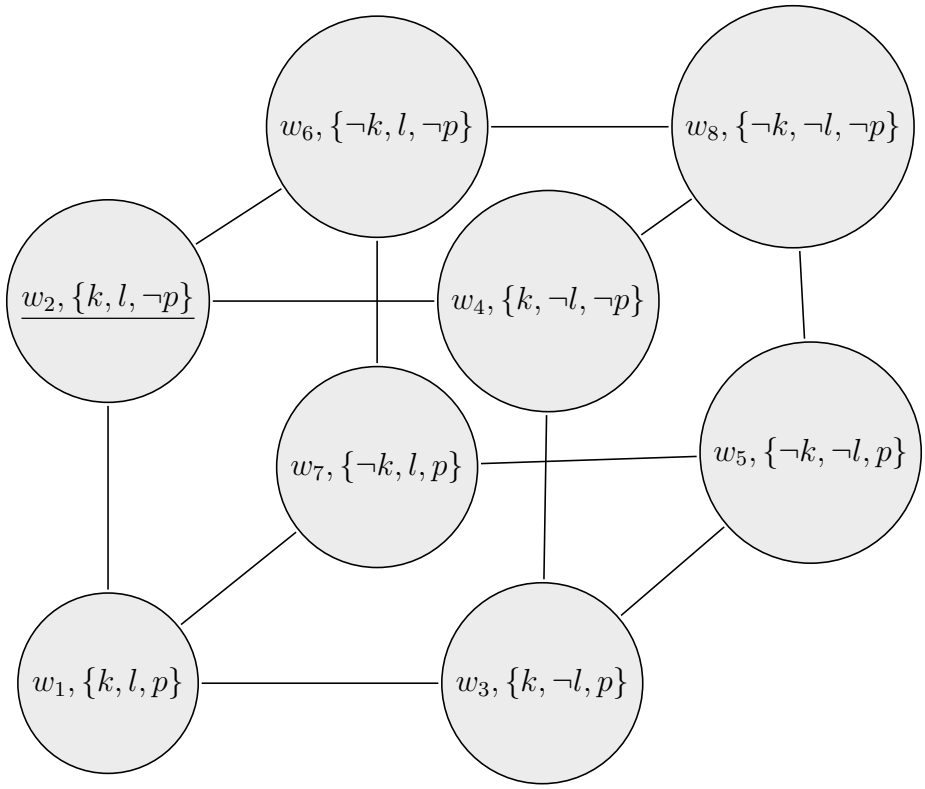

Figure 3.6: Atemporal Kripke STIT model for Example 3.2.11. Front-back faces of the cube represent $[R_a]$; left-right faces represent $[R_b]$; top-bottom faces represent $[R_c]$. We omit reflexive and transitive edges in our representation of equivalence relations. E.g., $w_1 R_a w_1$, but we omit the reflexive edge; $w_2 R_a w_3$, but we only draw the edges $w_2 R_a w_1$ and $w_1 R_a w_3$.

First, we note which propositions are true and false in the actual world. $M, w_2 \models k \wedge l \wedge \neg p \wedge \neg witness$: the murder has happened, the area was guarded, and no one passed by, so the murder was not witnessed. Second, we check what was deliberately seen to it by every agent. $M, w_2 \models [dstit]_a k \wedge [dstit]_b l \wedge [dstit]_c \neg p$, i.e., a has killed, b has guarded the area and c has not passed by. Neither of the three agents has seen to it that the unwitnessed murder has happened (i.e., $k \wedge \neg witness$): $M, w_2 \not\models [stit]_a(k \wedge \neg witness)$, since $w_2 R_a w_3$ and $M, w_3 \models witness$.

$M, w_2 \not\models [stit]_b(k \wedge witness)$, since $w_2 R_b w_7$ and $M, w_6 \models \neg k$. $M, w_2 \not\models [stit]_c(k \wedge \neg witness)$, since $w_2 R_c w_6$ and $M, w_6 \models \neg k$. Informally, the killer has not seen to it that no one witnessed the murder, while neither the lookout nor the passerby has seen to it that the murder happened.

Moving to groups, three groups have seen to it that $k \wedge \neg witness$. Unsurprisingly, $M, w_2 \models [stit]_{Ags}(k \wedge \neg witness)$. Since the group of all three agents has uniquely determined how the world is ($[w_2]_{R_{Ags}} = \{w_2\}$), they all have seen to it that the unwitnessed murder has happened. As no surprise, $M, w_2 \models [stit]_{a,b}(k \wedge \neg witness)$: the killer and the lookout have seen to it that the unwitnessed murder has happened. Since the killer had killed and the lookout had guarded the area, the murder was guaranteed to happen and remain unwitnessed, no matter what the passerby did. Ceteris paribus, had the c passed by the area (which is the case in $w_1 \in [w_2]_{R_{a,b}}$), the murder would still have happened and remained unwitnessed, given that the lookout was guarding the area and would have covered it up. The killer and the passerby have also seen to it that $k \wedge \neg witness$: $M, w_2 \models [stit]_{a,c}(k \wedge \neg witness)$. Ceteris paribus, had the b not guarded the area (which is the case in $w_4 \in [w_2]_{R_{a,c}}$), the murder would still have happened and remained unwitnessed, given that c did not pass the area, so there was no one to witness the murder even in the absence of the lookout.

So who is responsible for the covered-up murder? A group G is causally responsible for $k \wedge \neg witness$ iff G has deliberately seen to it that $k \wedge \neg witness$ and none of its proper subgroups has seen to it that $k \wedge \neg witness$. Obviously, $k \wedge \neg witness$ is not necessary, so at least someone is causally responsible for it. As we have shown:

$$M, w_2 \models [stit]_{a,b}(k \wedge \neg witness) \wedge [stit]_{a,c}(k \wedge \neg witness) \wedge [stit]_{Ags}(k \wedge \neg witness)$$

Since $\{a, b\}, \{a, c\} \subset Ags$ and both of these groups have seen to it that the covered-up murder happened, we can exclude Ags: $M, w_2 \not\models \Delta^{\min}_{Ags}(k \wedge \neg witness)$. We have previously shown that no individual agent has seen to it that the covered-up murder has happened. So no proper subgroup of $\{a, b\}$ and $\{a, c\}$ has seen to it that $k \wedge \neg witness$. We conclude that $M, w_2 \models \Delta^{\min}_{a,b}(k \wedge \neg witness) \wedge \Delta^{\min}_{a,c}(k \wedge \neg witness)$. So, surprisingly, two groups are casually responsible for the covered-up murder: the killer with the lookout and the killer with the passerby. If the former comes as no surprise, the latter is highly implausible: the passerby has no ties with the killer whatsoever!

In defense of the definition given in [Ser21], one may argue that it is still plausible to hold the killer and the passerby *causally* responsible for the covered-up murder (and the definition explicitly states to concern causal responsibility only). This attribution strikes as counterintuitive, because it ignores agents' knowledge and intention about what they have done. If we take knowledge into account, then

we can easily distinguish the killer and the lookout as the only responsible group. The story seems to imply that the killer and the lookout had common knowledge that they were causing the covered-up murder, unlike the passerby, who was clueless about what was going on. Therefore, two groups, $\{a, b\}$ and $\{a, c\}$, have caused the cover-up murder, but only the former has done it knowingly.

In reply, we say that there is still a difference between a group that unknowingly or unintentionally does something together and just a set of independent, simultaneously acting agents. It is not the case that the group $\{a, c\}$ lacks some knowledge about their group action: they lack group agency as such! To clarify the difference, we can compare it with clueless agents and non-agential indeterministic processes. For instance, think of a child who plays with a lighter in their parents' house, and it starts a fire. We hold the child causally responsible, since his actions have caused the fire. We will not hold him responsible in general, since the child was clueless about what he was doing. Compare it with a natural stochastic phenomenon such as lightning. The lightning strikes the roof of a house and, just like the child, causes a fire. Yet it seems meaningless to attribute the lightning even causal responsibility for the fire, since the lightning is not an agent at all. We argue that, in the sense of group agency, a set of independent agents relates to the unknowingly acting group in the same way that the lightning relates to the child with a lighter. The latter is an agent with a lack of knowledge about its own actions; the former is not an agent at all.

3.3 Envoi

In this chapter, we have described three different MAMLs. All three have the same characteristic feature: every set of agents is a group, while the group notion – common knowledge, coalitional power, or group action – is semantically reducible to the individual notions. In some formalisms, this feature is harmless, while in others it causes serious conceptual problems. In $\mathcal{L}_{EL}$, we get harmless yet pragmatically questionable common knowledge attributions. In $\mathcal{L}_{CL}$, we get both pragmatically questionable and conceptually wrong attributions of the coalitional abilities, especially when it comes to coordination. $\mathcal{L}_{STIT}$, being strictly more expressive, inherits the issues of $\mathcal{L}_{CL}$, but it gets further. Treating any set of agents as a group leads to conceptually wrong responsibility attributions, while modeling responsibility is one of the main use-cases of the STIT logic (if not the main one).

It is common to combine the MAMLs we have seen. Epistemic CL allows us to reason about what agents know about their own and other abilities, and what they can get to know [ÅA12]. Epistemic STIT logic provides us with tools to reason not only about what agents have and could have done, but also what they know about their own actions and actions of others [Kun+25]. With such combinations,

we get a lot of new expressivity and themes to explore: for instance, we can show the difference between having an ability ($[i]\varphi$), knowing about having the ability ($K_i[i]\varphi$), and being able to *knowingly* realise this ability ($\Diamond K_i[stit]_i\varphi$). On the other side, we get all the conceptual issues combined.

Treating any set of agents as a group is not only conceptually problematic. It also makes the formalisms computationally costly. Specifically for group STIT, the logic was proven to be finitely unaxiomatizable in cases we have at least three agents; the satisfiability problem in the same cases is undecidable [HS08]. Surprisingly, if we syntactically restrict the language $\mathcal{L}_{STIT}$ in a way, such that for some sets of agents $G \subseteq Ags$ there is no corresponding modality $[stit]_G$, we get finitely axiomatizable and decidable fragments [Sch12]. So we have a twofold motivation to "fix" the overly simplistic definition of groups in MAML: on the one hand, we will obtain conceptually better motivated formalisms; on the other hand, there is a promise that our formalisms will be computationally and metalogically better behaved.

In the next chapter, we develop logical theories of collective intention that will allow us to refine the group notions in MAMLs.

Chapter 4

Logics of collective intention: Bratmanian version[1]

As we have shown in the previous chapter, MAMLs suffer from both metalogical and conceptual shortcomings because of an overly simplistic definition of groups. As we have argued in Chapter 2, a set of agents can be treated as a group iff these agents share representational and motivational states, and act based on them. As for group representational states, we have treated common knowledge and common belief as such. We have not encountered any significant problems (or at least have abstracted away from them) in the classic formalization of these notions in Section 3.1. In the following two chapters, we focus on developing a logic of collective intention that will serve as a formal theory of group motivational states.

Conceptually, we have two conflicting theories of collective intentions. One argues that collective intention is nothing but an aggregation of individual conditional intentions that are intertwined in a specific way. The other, on the contrary, argues that collective intentions are irreducible: they are not completely independent from the individual intentions of group members, but cannot be reduced to them. We have seen sufficiently many examples of particular collective intentions being plausibly explained as either the aggregation of interdependent conditional intentions or as a holistic intention of a group. Moreover, this debate is definitely extra-logical. As Arthur Prior once beautifully said about temporal logic:

[1]Section 4.1 of this chapter is based on the paper *Daniil Khaitovich. Logic of Conditional Intention*, which is currently under review [Kha26a].

> *The logician must be rather like a lawyer* [...] *in the sense that he is there to give the metaphysician, perhaps even the physicist, the tense-logic that he wants, provided it be consistent. He must tell his client what the consequences of a given choice will be* [...] *and what alternatives are open to him; but doubt whether he can, qua logician, do more.* [Pri67, p.59]

We agree and share the same sentiment about logics of group agency. As logicians, we are not in a position to tell social ontologists what to believe about social ontology. So instead of arguing in favour of one of the theories, we suspend our philosophical judgment here and formalise both.

We begin in this chapter by formalising the theory of shared intention à la Bratman. First, we develop a general theory of conditional intentions, as they play a central role in our story (Section 4.1). Then, we extend the framework to a multi-agent modal logic and narrow our focus to agents' intentions conditional on one another's actions (Section 4.2). In this logic, we define groups' shared intentions as agglomerations of interdependent conditional intentions. Since we reduce shared intention to the aggregation of individuals' propositional attitudes, without postulating any group notions, neither semantically nor syntactically, our formalism is ontologically and conceptually conservative. Then, we prove that our postulation of rationality constraints on individuals' conditional intentions suffices to ensure that the shared intentions of these agents satisfy analogous constraints, which makes the formalism normatively conservative (Section 4.2.1).

4.1 Setting the stage: the logic of conditional intention

This section is based on [Kha26a].

Since the foundational paper of [CL90a], logicians are developing models of practical reasoning that incorporate intention as a separate, irreducible propositional attitude.[2] Philosophers of action and legal theorists lately emphasise the primary role of *conditional* intentions.[3] Broadly, conditional intention is a contingent commitment to a course of actions, which depends on a certain condition. For example, if someone *unconditionally* intends to go for a walk, then they are committed to do so no matter what; but if they intend to go for a walk *in case of* good weather, then they are committed to go for a walk as soon as they are sure that the weather is nice. Some philosophers argue that most real world intentions are conditional in nature: "*Often even an emphatic unconditional profession of intention such as 'I intend to φ at all costs' or 'I will φ no matter what' is not*

[2] See the introduction in [LH08] for a brief overview of logics of intention.

[3] See the introduction to [Lud15] for an overview of conditional intention research.

to be taken literally given that one is not really willing to φ at the price of losing one's life or making Heavens fall. Most intentions appear to be conditional in their 'deep structure' even when the conditions are not explicitly stated" [Fer09, p.700].

The neglect of conditional intention may have drastic consequences. For instance, in several legal systems, *mens rea*[4] was defined in terms of unconditional intention alone. This narrow formulation created a loophole: defendants could evade responsibility by claiming their actions were guided only by conditional intentions, which did not qualify as sufficient for *mens rea*.

> *Irwin grabbed a woman visiting another prisoner and held a knife to her throat, threatening to kill her if he was not released. A jury convicted him of assault with intent to kill and he appealed on the grounds that he did not intend to kill the woman but intended only to kill her if his demands were not met. The Court of Appeals reduced his charge to one of assault without the requisite intention* [Yaf04, p.275].[5]

Modal logics of propositional attitudes, including unconditional intention, are used to characterise *mens rea* in normative reasoning [Bro11a; LLM14; AB22]. As the mentioned legal cases indicate, conditional intentions are crucial if we want to account for *mens rea*. On top of our interest to conditional intentions as participatory, the concern around mens rea further motivates the need to develop a logical theory of conditional intentions, which, to our knowledge, has not been done yet.

4.1.1 Philosophical theories of conditional intention

We adopt the general view on intention *simpliciter* à la Bratman [Bra87]: intention is a commitment to a plan of action. Following [CL90a], the view can be sloganised as "*intention is a choice with commitment*".[6] As for *conditional* intention, we rely on theories of [Fer09; Lud15] that also inherit Bratmanian ideas. Conditional intention is a commitment to a contingent plan of action. When one says "*I intend to φ if ψ*"[7], one expresses a commitment to φ upon learning that ψ holds.

---

[4] *Mens rea* is a legal term meaning *"the mental element necessary to convict for any crime"[Say32].*

[5] Similar cases happened in the U.S. and Great Britain, where analogous arguments were successfully used by defendants that were suspected of attempted theft, car hijacking, and some other crimes. A brief overview of these cases can be found in [Yaf04; Cam82].

[6] For the detailed overview of Bratman's views on intention in practical reasoning and their expansion in other fields, the reader may consult [Bra87; CL90b; BG23].

[7] Throughout the text, we use *one intends to φ* as an abbreviation for *one intends to bring about that φ*.

While searching for a suitable theory of conditional intention, one may try to reduce the conditional intentions to the unconditional ones via material implication for the sake of parsimony. Unfortunately, such a reduction cannot be done. To clarify why it is impossible, we look into the next example borrowed from [Fer09, p.703]:

> *A nurse intends to change the patient's dressing, in case he is still alive.*

Let us try to explain what does this sentence mean using only unconditional intention. At the first glance, we have at least two possible interpretations: *de re* and *de dicto*. *De re* interpretation is "*the nurse intends that if she knows that the patient is alive, then she changes his dressing*" ($Int(K\,alive \to change)$), *de dicto* is "*if the nurse knows that the patient is alive, then she intends to change his dressing*" ($K\,alive \to Int\,change$).[8] If we take the *de re* variant, then it suffices for the bearer of intention to falsify the antecedent to vacuously satisfy the intention. In this context, the nurse can kill the patient to make the implication $K\,alive \to change$ vacuously true, which seems not to be what she really intends. In case of *de dicto*, if the bearer of the intention does not know whether the antecedent holds, nothing can be derived about her intentions from the *de dicto* statement. In our example, if the nurse does not know the patient is still alive, it does not follow that she has any intention, hence she might not be committed to any actions. On the contrary, it seems that even before the nurse gets to know that the patient is alive, she has a certain motivation to do some actions: be prepared to change the patient's dressing, make sure that it will be possible to do so, etc. In other words, even if an agent is unaware whether the antecedent holds, the proposition about the conditional intention still tells us something about the motivational state of this agent.

Therefore, neither *de re* nor *de dicto* interpretations account for the conditional intention in question, since by falsifying the antecedent of both formulations, we can trivialise the intention in undesirable ways. Since we need to distinguish between conditional intention and (necessary) material implications, from now on we will use different grammatical constructions for them. Namely, for conditional intention we will use *"one intends that φ in case of ψ"*, while *if-then* will be reserved for material implication.

Two more plausible interpretations of conditional intention go beyond *de re* and *de dicto* reductions. We adopt the terminology of [Fer09] and call them *preconditional* and *restrictive conditional* intention:

1. Preconditional intention: one preconditionally intends to φ in case of ψ iff ψ is a necessary means of bringing about that φ. Several examples: I intend to

[8] Here we use pseudo-language, where $Int\,\varphi$ stands for "*the nurse intends that φ*", $K\varphi$ stands for "*the nurse knows that φ*" and $\to$ is interpreted as material implication.

take a bus, in case the bus is coming; I intend to pay, in case I have money. In all these cases, one cannot do what is intended unless the condition of the intention holds. Obviously, it is not enough to falsify the precondition to satisfy the preconditional intention: the bearer of the intention usually wants the precondition to hold, since it is necessary to satisfy the intention.

2. Restrictive conditional intention: one has a restrictive conditional intention to φ in case of ψ iff one is committed to φ only when one is sure that ψ obtains, otherwise, one lacks reasons to φ. In other words, the condition of one's intention restricts the scope of situations under which the agent is committed to a certain course of action. Some examples: I intend to take a walk in case it does not rain; I intend to go to a library in case it is noisy at my office. In all the listed cases, the condition motivates one to fulfil the intention, while it is not a necessary means of doing so: one can go to a library even if it is perfectly quiet in one's office or go for a walk in the rain, but may not be motivated to do so. The nurse of our example has a restrictive conditional intention as well. She intends to change the patient's dressing in case he is still alive; otherwise, she still can do that, but there is no reason to do so.

Note that those are not two rival interpretations of the same phenomena, but rather two closely related concepts: *conditional intention* is an umbrella term for both of them.

Following [Lud15], we state that conditional intention can indeed be reduced[9] to an intention *simpliciter* in two cases. Namely, if a (pre)condition is intended or known. For example, if I intend to be at the train station on time with a precondition that I get up early, and I intend to get up early, then I simply intend to wake up early and be at the train station on time. The same applies to the restrictive conditional intention: going back to the nurse example, if she intends to assist the patient in case she is sure that he is alive, while having full control over his life and intending to save him, then she simply intends to save him and assist him after. As for known conditions, it seems natural to assume that there is no need to conditionalise our intentions on propositions that we know are true: "*If one knows that a condition obtains, it is not a contingency, but simply part of the background of planning. A condition is therefore a candidate for being a conditional intention's antecedent (in brief, can be an antecedent) only if one does not take oneself to know it obtains.*" [Lud15, p.38].[10]

Taking it all together, we understand the conditional intention to φ in case of ψ

[9]By reducibility of a conditional intention to the unconditional one we mean bi-conditional of the form $Int^{\psi}\varphi \leftrightarrow Int\varphi$, where $Int\varphi$ denotes an unconditional intention to φ. We will clarify and formally define the grounds for such a reduction in the following sections.

[10]In [Fer09], it is argued that one may have an intention conditional on a known proposition. For a discussion between the two authors on that topic, see [Lud15; Fer15].

as a contingent commitment of its bearer to φ in case she gets to know that ψ obtains. ψ may be a reason or a means for the agent to φ. The notion is equivalent neither to *de re* nor to *de dicto* interpretations via unconditional intention and material implication. In irreducibly conditional intention, the condition should neither be intended nor known.

4.1.2 Logical challenges

We proceed by deriving logical challenges from the philosophical theories of conditional intention. This way, we state the plausibility criteria that the modal logic of conditional intention should satisfy. First, we define a formal language we are to work with, so that it is easier to talk about logical properties we are interested in. Given some countable set of propositional variables $Var = \{p_1, p_2, \ldots\}$, we define the formal language $\mathcal{L}$ of conditional intention logic:

$$\mathcal{L} \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid Int^{\varphi}\varphi$$

where $p \in Var$. The reading of negation, disjunction, and all the derived Boolean connectives and constants ($\bot$, $\top$, $\wedge$, $\rightarrow$, $\leftrightarrow$) is standard. $\Box\varphi$ reads as "*it is necessary that* φ" or *"φ is a priori true"*. We use the following standard acronym for the a priori possibility: $\Diamond\varphi := \neg\Box\neg\varphi$.

$Int^{\varphi}\psi$ stands for *"the agent implicitly intends that ψ in case of φ"*. By implicit intention, we mean that ψ a priori follows from what the agent intends in case of φ. It might be either an intended consequence or just a side-effect. For now, every time we write "the agent intends that φ conditional on ψ", we mean the implicit intention, unless stated otherwise. We write simply $Int\varphi$ instead of $Int^{\top}\varphi$ and read it as "*the agent intends that φ*". Our language does not distinguish between preconditional and restrictive conditional intentions: both are formalised as $Int^{\psi}\varphi$.

We list some desirable and undesirable logical principles that follow from philosophical literature on (conditional) intention and the stances we have outlined. According to a generally accepted rationality constraint [Bra87; BG23; KP93; RG97], agents' intentions should be consistent. Analogously, agents' conditional intentions should be consistent w.r.t. the condition, i.e. if one intends to φ in case of ψ, then φ and ψ should be mutually consistent [Fer09; Lud15]:

$$\models Int^{\psi}\varphi \rightarrow \Diamond(\psi \wedge \varphi) \tag{4.1}$$

It is widely acknowledged that two intentions should also be consistent w.r.t. one another: if one intends that φ and ψ (under the same condition), then it should be possible that $\varphi \wedge \psi$ (under that condition) [Bra87; CL90a]:

$$\models (Int^{\psi}\varphi_1 \wedge Int^{\psi}\varphi_2) \rightarrow \Diamond(\psi \wedge \varphi_1 \wedge \varphi_2) \tag{4.2}$$

Some authors, including Bratman [Bra87], go even further and argue that intentions should not only be agglomerate*able*, but should in fact agglomerate:

$$\models (Int^{\varphi}\psi_1 \wedge Int^{\varphi}\psi_2) \rightarrow Int^{\varphi}(\psi_1 \wedge \psi_2) \tag{4.3}$$

The agglomeration principle for intention is debatable. For instance, [Sve96] contends that the unqualified version of this principle compels agents to form "*one enormous compound intention*" [p. 517] to guide their actions, which may undermine their ability to reason about such intention, given their boundedness in cognitive resources. Nonetheless, there are ways to defend the agglomeration of intentions against such a criticism; see, for example, [Zhu10] for a detailed discussion. Since we interpret $Int^{\varphi}\psi$ as " *the agent* ***implicitly*** *intends that* ψ *conditional on* φ", then agglomeration is natural: the "enormous compound intention" is a logical consequence of what the agent intends, but may not be explicitly present in her motivational attitude.

As we have pointed out in the previous section, if a condition of a conditional intention is known or intended, the conditional intention should be reducible to the unconditional one:

$$\models \Box\varphi \rightarrow (Int^{\varphi}\psi \rightarrow Int\psi) \tag{4.4}$$

$$\models Int\varphi \rightarrow (Int^{\varphi}\psi \rightarrow Int\psi) \tag{4.5}$$

Next, we move to the undesirable principles. Following insights from the previous section, the condition of a conditional intention should be either a reason or a means for an agent to commit to a certain action. Obviously, not every proposition of our language expresses a reason or a means to act. Some propositions are completely irrelevant to the agent's practical reasoning; hence, they are not rationally pressured to form a contingent commitment for every proposition. Formally speaking, the following set of formulae should be satisfiable for some φ: $\{\neg Int^{\varphi}\psi \mid \psi \in \mathcal{L}\}$ (i.e., the agent does not intend anything conditional on φ). Therefore, the *success axiom*, which is commonly accepted in logics of conditional propositional attitudes [Che75; BS06], should be falsifiable in the logic of conditional intention:

$$\not\models Int^{\varphi}\varphi \tag{4.6}$$

For example, one may intend avoiding driving the wrong lane: $Int\neg wrong\_lane$. It implies that in case one ends up driving the wrong lane, it will be done unintentionally. So it seems strange to say that such an agent is rationally pressured to intend that she drives the wrong lane in case she ends up doing so: $\neg Int^{wrong\_lane} wrong\_lane$.

Next, we face the issue of *monotonicity*. By unrestricted monotonicity we mean the following:

$$Int^{\varphi}\psi \rightarrow Int^{(\varphi \wedge \theta)}\psi \tag{Mon}$$

i.e., conditional intention is safe under logically stronger conditions. We reject unrestricted monotonicity, since agents can accept "*plan-B intentions*" without breaking any rationality constraints. For example, an agent may intend to go for a walk in case it is sunny and, as a plan-B, intend to stay home in case it is sunny but she has not gone for a walk for whatever reason. Since the second intention explicitly specifies what to do in case the first one has failed, the bearer of such intentions will never find herself in a situation where mutually inconsistent conditions of both contingent commitments in question hold. Formally:

$$Int^{(sunny \wedge \neg walk)} home, Int^{sunny} walk \not\rightsquigarrow Int^{(sunny \wedge \neg walk)} walk$$

where $\rightsquigarrow$ is a symbol to denote an intuitive notion of entailment in the ordinary language. On the other hand, monotonicity seems plausible when we do not deal with plan-A and plan-B intentions. For instance, if someone deliberates about what to do when it is sunny or windy, and intends to stay home in case it is windy, it is reasonable to assume she also intends to stay home in case it is both windy and sunny:

$$Int^{windy} home \rightsquigarrow Int^{(sunny \wedge windy)} home$$

As these examples indicate, unrestricted monotonicity is too limiting because it prevents agents from having rational plan-B intentions.

$$Int^{\varphi}\psi \not\models Int^{(\varphi \wedge \vartheta)}\psi \tag{4.7}$$

Some weaker forms of monotonicity should also fail. Even if what is intended conditional on φ is consistent with the stronger condition $\varphi \wedge \vartheta$ – and even if it is consistent with the intentions under that stronger condition – it is not sufficient to ensure that the intention is safe under that stronger condition:

$$Int^{\varphi}\psi \wedge \Diamond(\varphi \wedge \vartheta \wedge \psi) \not\models Int^{(\varphi \wedge \vartheta)}\psi \tag{Mon1}$$

$$Int^{\varphi}\psi \wedge \neg Int^{(\varphi \wedge \vartheta)}\neg\psi \not\models Int^{(\varphi \wedge \vartheta)}\psi \tag{Mon2}$$

For instance, assume an agent plans to go for a run in case of good weather. As part of that plan, she intends to wear running clothes in case the weather is good. She also has a backup plan: if the weather is good, but she has failed to go for a run, she intends to stay home instead. Obviously, she can wear running clothes even if she does not go running. Her backup plan does not specify what to wear, so she does not intend *not* to wear the running clothes in case she fails to go for a run on good weather. Still, the agent is not rationally pressured to intend to put on the sportswear in case the weather is good, but she does not go for a run. To wear the running clothes is a part of the agent's primary plan – to go for a run

conditionally on good weather – that happens to be consistent with her backup plan – to stay home conditionally on the good weather and the failure to go for a run. However, within the context of the backup plan, this part of the primary plan (to put on the sportswear) serves no purpose.

Formally, let φ, ϑ, and ψ be formulae denoting "*the weather is good*", "*one is not going for a run*", and "*one puts on the running clothes*", respectively. Then, the story we have just presented constitutes a counterexample for the two weaker forms of monotonicity described above.

So to adequately capture the failure of monotonicity, we should track which intentions are parts of which plans. Given two conditions φ and $\varphi \wedge \vartheta$, if one intends that ψ conditional on φ, whether one is rationally pressured to intend that ψ conditionally on $\varphi \wedge \vartheta$ depends on whether $\varphi \wedge \vartheta$ contradicts the whole plan the agent has conditional on φ, not just the separate conditional intention that ψ.

We can look at a closely related logic in search of parallels: the logic of conditional obligations. In normative systems, we often face requirements of the form "*in case of φ, it is obligatory that ψ*". We constantly face *contrary-to-duty obligations* that specify what we are obliged to do in case we have failed to act upon other obligations [PS96]. This sounds very similar to plan-B intentions. Contrary-to-duty obligations cause the logics of conditional obligations to falsify monotonicity, but they are not the only reason. Consider the following example:

> You ought not to act aggressively no matter what. (A)
>
> In case you meet your worst rival, you ought to act aggressively. (B)

For A and B to be consistent, unrestricted monotonicity should fail. Such cases are sorted out via the principle *lex specialis derogat legi generali*[11]: obligations with more specific conditions override those with less specific ones [Tor94; Hor94]. Here, A has no condition whatsoever (or a trivial condition $\top$), while B has the condition, which is why in case of a conflict we should follow B, not A.

The difference between obligations and intentions in this regard is that the former may be exo- and heterogeneous. Obligations may come from different sources, some of which are external forces, and this causes collisions [Owe16]. A number of examples: one may be obliged not to be rude due to general etiquette and ought to be rude to a specific person due to a custom in the given social context; one may have an obligation not to restrict anyone's freedom of speech, which yields from Universal Declaration of Human Rights, and ought to actively suppress certain kind of expressions due to domestic law of a certain country; one ought to avoid unnecessary spending due to economic rationality and ought to spend money on a specific unnecessary present as a sign of generosity due to a social norm.

[11] The specific law overrides the general law.

Unlike obligations, intentions are indo- and homogenous. They only come from the practical reasoning of the bearer of the intention, and no external force can put an intention in someone's mind [Bra87]. Therefore, collisions, analogous A-B, can be avoided for intentions. Using the very same example, it seems irrational to simultaneously intend not to be aggressive to anyone and intend to be aggressive to one's rival. One's intention is only up to oneself, so even if one is confronted with two contradictory requirements from two different sources, one needs to choose which (if any) to follow and formulate the intention accordingly.

We take an intermediate position between unrestricted monotonicity, which is too strong, and *lex specialis derogat legi generali*, which is too weak. Namely, our principle is *antecedent-consequent consistency*. Given two conditions φ and $\varphi \wedge \vartheta$, as soon as **everything** the agent intends in case of φ is consistent with $\varphi \wedge \vartheta$, the agent should intend it in case of $\varphi \wedge \vartheta$ as well. In other words, if one has an intention conditional on φ, one must save that intention under a logically stronger condition $\varphi \wedge \vartheta$, as soon as that stronger condition does not contradict one's plan for what to do in case of φ. Such a solution still allows agents to have plan-A and plan-B intentions: the condition of plan-B contradicts what is intended in plan-A, therefore, our principle will not pressure agents to merge such plans into one contradictory plan.

To motivate our principle, we provide a number of examples. Let us go back to the following pair of statements, which expresses plan-A and plan-B intentions and intuitively should be consistent:

In case it is sunny, Inne intends to go for a walk. (C)

In case it is sunny, but Inne is not going for a walk, Inne intends to stay home. (D)

Obviously, from C it does not follow that Inne intends to go for a walk in case it is sunny and she is not going for a walk: the condition $sunny \wedge \neg walk$ contradicts her intention to go for a walk, so monotonicity may fail here. Therefore, C and D do not lead us into inconsistency. On the contrary, the following pair of statements should rightfully lead us to inconsistency:

In case it is sunny, Inne intends to go for a walk. (E)

In case it is windy, Inne intends to stay home. (F)

Intuitively, E-F should not be consistent. Let us assume that it is sunny *and* windy. Should Inne still follow E or F? In case it is sunny and windy, Inne still can go for a walk or stay home; nothing rationally pressures her to abandon any of her conditional commitments, but it is impossible to fulfill them both.

E-F are indeed inconsistent, following our monotonicity principle. Everything Inne intends in case it is sunny – to go for a walk – is still possible to do in case

it is sunny and windy, therefore Inne must intend to go for a walk in case it is sunny and windy. Analogously, Inne is also required to intend to stay home in case it is sunny and windy, but it is impossible to simultaneously stay home and go for a walk. Therefore, if we adopt the proposed monotonicity principle, E-F together with 4.1 and 4.3 lead us to inconsistency, as we would intuitively expect.

To conclude this section, we have outlined two major challenges for the logic of conditional intention: cautious selection of conditions and restricted monotonicity. Both challenges pose specific requirements that are caused by our conceptual view on conditional intention, therefore, we cannot simply borrow solutions from "neighboring" logics, such as logics of conditional belief or conditional obligation. As for the selection of conditions, we will use the modification of formalisms proposed in [Che75]. Namely, as we are to see in the following section, we can make set-selection functions partial to avoid the conditionalization of intentions on any proposition whatsoever. As for restriction on monotonicity, we have argued for a principle that is stronger than *lex specialis derogat legi generali*: according to our principle, conditional intention should be monotonic, as soon as strengthening of the condition does not contradict the primary intention. In the following section, we will give a formally precise definition of this principle and show with examples why it works better than *lex specialis derogat legi generali.*

4.1.3 Formal semantics

Having pointed out the logical challenges and our views on them in Section 4.1.2, we present formal semantics for $\mathcal{L}$. We implement models with set-selection functions à la [Che75].

4.1.1. Definition (**C**-frames and models). A class of conditional frames **C** consists of tuples of the form $F = (W, f)$, where:

1. $W \neq \emptyset$ is a set of practically possible worlds. By practical possibility, we mean that worlds of W are possible with respect to actions that the agent can perform. If a world is logically possible but would not be real no matter what the agent does, it is not practically possible. For example, whilst it is logically possible that the agent in question has never been born, it is not practically possible;
2. $f : (W \times 2^W) \to 2^W$ is a *partial* set selection function, which takes a possible world, a proposition $P \subseteq W$[12] and in case the agent conditionalises her intentions on P in w, returns a set of conative P-alternatives of w, i.e. a set of possible worlds, where agent's intention in case of P is satisfied. f function has the following additional properties:

[12]Following Chellas, we will call sets of possible worlds $P \subseteq W$ *propositions* [Che75]. The reader should not confuse propositions in this sense with *propositional variables*, i.e. $p_1, p_2, \ldots \in Var$.

(a) If $f(w,P)$ is defined, then $f(w,P) \cap P \neq \emptyset$: agent's intentions in case of P are consistent with P;

(b) If $f(w,P)$ and $f(w,Q)$ are defined, while $P \cap Q \neq \emptyset$, then $f(w,P \cap Q)$ is defined: if P and Q are mutually consistent reasons or means for agent's actions, then $P \cap Q$ is one as well;

(c) If $f(w,P)$ and $f(w,P \cap Q)$ are defined and $f(w,P) \cap P \cap Q \neq \emptyset$, then $f(w,P \cap Q) \subseteq f(w,P)$: in case everything the agent intends in case of P is possible to realise in case of $P \cap Q$, her $P \cap Q$-intention should include her P-intention;

(d) Given that $f(w,P)$ is defined, if $f(w,P) \subseteq Q$ and $f(w,P \cap Q)$ is defined, then $f(w,P) = f(w,P \cap Q)$: if agent intends that Q in case of P, then P-intention and $P \cap Q$-intention are the same.

As usual, every frame $F = (W,f) \in \mathbf{C}$ can be extended to a model $M = (F,V)$ with a valuation function $V : Var \to 2^W$.

4.1.2. Definition (Semantic clauses for $\mathcal{L}$). Given a model $M = (W,f,V)$ and a possible world $w \in W$, where $[\![\varphi]\!]_M := \{w \in W \mid M,w \models \varphi\}$, the satisfiability relation $\models$ is inductively defined as follows:

$$
\begin{aligned}
&M,w \models p \Leftrightarrow w \in V(p)\\
&M,w \models \neg\varphi \Leftrightarrow M,w \not\models \varphi\\
&M,w \models \varphi \vee \psi \Leftrightarrow M,w \models \varphi \text{ or } M,w \models \psi\\
&M,w \models \Box\varphi \Leftrightarrow \forall v \in W : M,v \models \varphi\\
&M,w \models Int^{\varphi}\psi \Leftrightarrow f(w,[\![\varphi]\!]_M) \text{ is defined and } f(w,[\![\varphi]\!]_M) \subseteq [\![\psi]\!]_M
\end{aligned}
$$

$M \models \varphi$ means that $M,w \models \varphi$ for every $w \in W$. $F,w \models \varphi$ means that $(F,V),w \models \varphi$ for every valuation function V. $F \models \varphi$ means that $F,w \models \varphi$ for every $w \in W$. $\mathbf{C} \models \varphi$ means that $F \models \varphi$ for every $F \in \mathbf{C}$. When the class of frames is clear from the context, we write $\models \varphi$. In that case, we say that φ is valid.

We introduce a useful acronym: $[C]\varphi := Int^{\varphi}\top$. Intuitively, $[C]\varphi$ means that the agent conditionalises her intentions on φ. Formally, $M,w \models [C]\varphi$ iff $f(w,[\![\varphi]\!]_M)$ is defined.

Although we are using the standard set-selection functions for conditionals, there are two critical differences with the semantics proposed in [Che75]. The first difference is our interpretation of the set W. Namely, we take W to be a set of *practically* rather than logically possible worlds. We take the agent to be reasoning about some practically feasible options she chooses, so she is free not to take into account some purely logical possibilities that are irrelevant to her practical matters. The second difference is that the set-selection function f is

partial: for some proposition P and some possible world w, $f(w, P)$ may be undefined, which means that P is not a reason or a precondition for any action the agent deliberates over in w.

Despite the partiality of f, there are no truth-value gaps. From Def. 4.1.2 it follows that if $f(w, [\![\varphi]\!]_M)$ is undefined, then $M, w \models \neg Int^{\varphi}\psi$ for any ψ. If the agent does not conditionalise her intention on φ, then the agent does not intend anything at all conditional on φ. E.g., $f(w, [\![\varphi]\!]_M)$ is undefined iff $M, w \models \neg Int^{\varphi}\top$. This is an iff-statement, since if $f(w, [\![\varphi]\!]_M)$ is defined, then $f(w, [\![\varphi]\!]_M) \subseteq [\![\top]\!]_M = W$ by definition. This feature of partiality grounds our interpretation of the acronym $[C]\varphi$. If the agent has some intentions conditional on φ, then whatever is intended in case of φ, $\top$ is an a priori consequence of such an intention. $[C]\varphi := Int^{\varphi}\top$ says nothing informative about *what* is intended in case of φ, but expresses that at least something is intended in case of φ, i.e., φ is a condition.

Since there exists a pointed model (M, w), where the agent does not intend anything whatsoever conditional on φ, including tautologies, Int^{φ}-modality is not normal. Namely, the rule of necessitation fails for Int^{φ}: from $\models \psi$ it does not follow that $\models Int^{\varphi}\psi$. For instance, take any pointed model (M, w), such that $f(w, V(p))$ is not defined. Then, $M, w \not\models Int^{p}\top$, while $\models \top$.

To demonstrate how the conditional models handle non-monotonicity, we proceed with an example.

4.1.3. Example. Imagine a self-driving car Dosen with two sensors: the first detects other cars around Dosen and the second detects pedestrians who happen to be on the road. Dosen unconditionally intends to avoid road accidents no matter what: $Int\neg\texttt{crash}$. It is the strongest unconditional intention of Dosen. If the first sensor reports that there are no cars in the traffic on the left side, then Dosen intends to turn to merge in: $Int^{\texttt{free}}\texttt{turn}$.

Dosen has the following true statements in its' knowledge base. If it tries to take a turn while there is a pedestrian on the road on the same side, it will inevitably lead to an accident: $\Box((\texttt{turn} \wedge \texttt{human}) \rightarrow \texttt{crash})$. There may be no cars, but a pedestrian on the same side of the road: $\Diamond(\texttt{free} \wedge \texttt{human})$. Dozen can avoid the accident when there is a pedestrian and no cars on the same side: $\Diamond(\texttt{free} \wedge \texttt{human} \wedge \neg\texttt{crash})$.

Dosen tries to form an intention of what to do in case there are no cars but a pedestrian on the left side of the road. If he had adopted *lex specialis derogat legi generali*, then the more specific intention to merge in case there is a space free of cars would override the unconditional intention not to crash. In other words, if there was free space and a pedestrian on the road, Dosen would intend to merge even though it would lead to a road accident. Obviously, we do not want Dosen to reason in such fashion. Using our logic, can Dosen figure out what to do in case there are no cars on the left side, but there is also a pedestrian (i.e.

$\texttt{free} \wedge \texttt{human}$)?

Assume a model $M = (W, f, V)$ for this case, where $w \in W$ is the real world. For the sake of brevity, we will not fully specify the whole model. Assume that Dosen's intentions are specified for three conditions: no matter what, i.e. $f(w, W)$, there are no cars on the left, i.e. $f(w, [\![\texttt{free}]\!])$, and there are no cars but a pedestrian, i.e. $f(w, [\![\texttt{free} \wedge \texttt{human}]\!])$. We know Dosen's strongest unconditional intention, to avoid crashes, i.e. $f(w, W) = [\![\neg\texttt{crash}]\!]$; In case of free space, Dosen intends to take a turn, i.e. $f(w, [\![\texttt{free}]\!]) \subseteq [\![\texttt{turn}]\!]$. Since $f(w, [\![\texttt{free} \wedge \texttt{human}]\!])$ is defined, Dosen has specified some intention what to do in case there is a free space and a pedestrian on the left. We argue that this intention is consistent with Dosen's unconditional intention. As stated in Dosen's knowledge base, it is possible to avoid a crash while there is a pedestrian and free space to the left, i.e. $[\![\neg\texttt{crash} \wedge \texttt{free} \wedge \texttt{human}]\!] \neq \emptyset$. In other words, everything Dosen intends unconditionally is possible to realise in case there are no cars and a pedestrian on the left. Therefore, by 2(d) of Definition 4.1.1, $f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \subseteq f(w, W) \subseteq [\![\neg\texttt{crash}]\!]$. Then, $f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \subseteq [\![\neg\texttt{crash}]\!]$. By definition of $\models$, it follows that $M, w \models Int^{\texttt{free}\wedge\texttt{human}}\neg\texttt{crash}$.

We also argue that `free`-intention is not monotonic w.r.t. $\texttt{free} \wedge \texttt{human}$-intention, i.e.:

$$f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \not\subseteq f(w, [\![\texttt{free}]\!])$$

To see it, recall that, as we have just shown, $f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \subseteq [\![\neg\texttt{crash}]\!]$ and, as specified by example, $f(w, [\![\texttt{free}]\!]) \subseteq [\![\texttt{turn}]\!]$. For the sake of contradiction, assume that Dosen intends to take a turn in case there is a free space and a pedestrian on the left, i.e. $f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \subseteq [\![\texttt{turn}]\!]$. Since $f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \subseteq [\![\neg\texttt{crash}]\!]$, it follows that:

$$f(w, [\![\texttt{free} \wedge \texttt{human}]\!]) \subseteq [\![\neg\texttt{crash} \wedge \texttt{turn}]\!].$$

Then, by 2(a), it should be possible that $\texttt{free} \wedge \texttt{human} \wedge \texttt{turn} \wedge \neg\texttt{crash}$. As specified in Dosen's knowledge base, in case he turns while there is a human, it will necessarily lead to a crash, i.e. $\Box((\texttt{human} \wedge \texttt{turn}) \to \texttt{crash})$. Therefore, $[\![\texttt{free} \wedge \texttt{human} \wedge \texttt{turn} \wedge \neg\texttt{crash}]\!] = \emptyset$, what contradicts 2(a) of Definition 4.1.1. To conclude, in case there is a free space and a pedestrian on the left of Dosen, he still intends to avoid crashes and therefore does not intend to take turns:

$$M, w \models Int^{(\texttt{free}\wedge\texttt{human})}\neg\texttt{crash} \wedge \neg Int^{(\texttt{free}\wedge\texttt{human})}\texttt{turn}$$

We proceed with showing some interesting (in)validities on conditional models. First, the set of conditions is restricted, since f function is partial. In other words, given any model $M = (W, f, V)$, $f(w, [\![\varphi]\!]_M)$ is not necessarily defined for any $w \in W$ and any $\varphi \in \mathcal{L}$, hence, $\not\models [C]\varphi$. By the same reason, success

axiom (4.6) fails: $\not\models Int^{\varphi}\varphi$. Moreover, the success axiom can be falsified even if the condition is defined: $\not\models [C]\varphi \to Int^{\varphi}\varphi$, since it is not necessarily so that $f(w, P) \subseteq P$.

Second, our logic indeed validates 4.1-4.5:

4.1.4. Proposition. *The following principles are valid in the class of* **C***-frames:*

$$\models Int^{\psi}\varphi \to \Diamond(\varphi \wedge \psi) \tag{4.8}$$

$$\models (Int^{\varphi}\psi_1 \wedge Int^{\varphi}\psi_2) \to Int^{\varphi}(\psi_1 \wedge \psi_2) \tag{4.9}$$

$$\models \Box\varphi \to (Int\psi \leftrightarrow Int^{\varphi}\psi) \tag{4.10}$$

$$\models [C]\varphi \to \big(Int\varphi \to (Int\psi \leftrightarrow Int^{\varphi}\psi)\big) \tag{4.11}$$

Proof:
(4.8) holds, because of condition 2(a) of Definition 4.1.1; (4.9) holds, since if $f(w, [\![\varphi]\!]) \subseteq [\![\psi_1]\!]$ and $f(w, [\![\varphi]\!]) \subseteq [\![\psi_2]\!]$, then $f(w, [\![\varphi]\!]) \subseteq [\![\psi_1 \wedge \psi_2]\!]$. (4.10) holds, because if $M \models \Box\varphi$, then $[\![\varphi]\!]_M = [\![\top]\!]_M$, i.e. $f(w, [\![\varphi]\!]_M) = f(w, [\![\top]\!]_M)$. (4.11) holds, because of the condition 2(d) of Definition (4.1.1). □

Note that from (4.8) it follows that an agent does not conditionalise her intentions on a priori false propositions: $\models \Box\neg\varphi \to \neg Int^{\varphi}\psi$. From (4.10) it follows that if the proposition is a priori true, then any intention conditional on it is just an unconditional intention: $\models \Box\varphi \to (Int^{\varphi}\psi \leftrightarrow Int\psi)$. Therefore, our logic reflects Ludwig's thesis about conditions of intentions being "*as yet unsettled for us*" [Lud15, p.32]: in a genuine conditional intention, the condition is neither a priori true, nor a priori false.

Third, we can deduce a restricted version of a *free choice*. The free choice is a common inference pattern, where conjunctive meanings are derived from disjunctive modal sentences, which goes against the prescriptions of classical logic [Alo22]. We will demonstrate it with a deontic case. In the ordinary language, the expression "*one may smoke on the balcony* ***or*** *outside"* entails that one may smoke on the balcony **and** one may smoke outside (here, $P\varphi$ reads as "*it is permitted that* φ"):

$$P(\varphi \vee \psi) \rightsquigarrow (P\varphi \wedge P\psi) \tag{4.12}$$

It is not hard to find the analogous cases for conditional intention. For example, "*I intend to buy some apples in case they are tasty* ***or*** *cheap*" seems to entail that I intend to buy apples in case they are tasty **and** I intend to buy them in case they are cheap:

$$Int^{(\varphi\vee\psi)}\theta \rightsquigarrow (Int^{\varphi}\theta \wedge Int^{\psi}\theta)$$

In our logic, free choice inference is possible under two reasonable restrictions. First, such inference is possible if the agent conditionalises her intention on $\varphi \vee \psi$,

φ and ψ. Second, the inference is possible if there are no contradictions between φ, ψ and what is intended in case of $\varphi \vee \psi$. For example, assume that one strongly prefers walking by foot to driving a car. Therefore, in case one is either to drive or to walk, one intends to walk: $Int^{(walk \vee drive)} walk$. It is impossible to simultaneously drive and walk: $\neg\Diamond(walk \wedge drive)$. From this it seems implausible to infer that one intends to walk in case one is driving (and in our semantics, it indeed does not follow due to 2(a)):

$$\neg\Diamond(walk \wedge drive), Int^{(walk \vee drive)} walk \not\models Int^{drive} walk$$

In all other cases, i.e. when an agent has an intention conditional on disjunction $\varphi \vee \psi$ and both of the disjuncts, such that the intention in case of $\varphi \vee \psi$ is consistent with both φ and ψ, the free choice inference is valid.

4.1.5. Proposition. *The following principle is valid:*

$$\models (\neg Int^{(\varphi \vee \psi)} \neg\psi \wedge \neg Int^{(\varphi \vee \psi)} \neg\varphi \wedge [C]\varphi \wedge [C]\psi \wedge [C](\varphi \vee \psi)) \rightarrow (Int^{(\varphi \vee \psi)} \vartheta \rightarrow (Int^{\varphi} \vartheta \wedge Int^{\psi} \vartheta))$$

Proof:
See Theorem 4.1.6 of the following section. □

4.1.4 Logic, soundness, completeness

In what follows, we define an axiomatic system **L** (see Table 4.1 for the definition of **L**) and prove it to be sound and complete w.r.t. the class of **C**-models.

(CPL)	Tautologies of classical propositional logic
(S5)	S5 axioms and rules for $\Box$
(C)	$Int^{\varphi}\psi \rightarrow [C]\varphi$
(K)	$Int^{\varphi}(\psi \rightarrow \theta) \rightarrow (Int^{\varphi}\psi \rightarrow Int^{\varphi}\theta)$
(U)	$(\Box\psi \wedge [C]\varphi) \rightarrow Int^{\varphi}\psi$
(D+)	$Int^{\varphi}\psi \rightarrow \Diamond(\varphi \wedge \psi)$
(Cut)	$\big([C]\alpha \wedge [C](\alpha \wedge \psi)\big) \rightarrow \big(Int^{\alpha}\psi \rightarrow (Int^{\alpha}\varphi \leftrightarrow Int^{(\alpha \wedge \psi)}\varphi)\big)$
(R-Mon)	$([C]\alpha \wedge [C](\alpha \wedge \beta)) \rightarrow \big((Int^{\alpha}\varphi \wedge \neg Int^{\alpha}(\alpha \rightarrow \neg\beta)) \rightarrow Int^{(\alpha \wedge \beta)}\varphi\big)$
(C-inter)	$([C]\varphi \wedge [C]\psi \wedge \Diamond(\varphi \wedge \psi)) \rightarrow [C](\varphi \wedge \psi)$
(MP)	If $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$, then $\vdash \psi$
(RN′)	If $\vdash \psi$, then $\vdash [C]\varphi \rightarrow Int^{\varphi}\psi$
(RE)	If $\vdash \varphi \leftrightarrow \psi$, then $\vdash Int^{\varphi}\vartheta \leftrightarrow Int^{\psi}\vartheta$

Table 4.1: Logic **L**

Axiom (C) ensures that if the agent intends anything conditional on φ, then φ should be a condition. (C-Inter) guarantees that whenever φ and ψ are consistent conditions, $\varphi \wedge \psi$ is a condition as well; this principle allows merging intentions with consistent conditions. Axiom (K) guarantees that intentions agglomerate and are closed under conjuncts, i.e. $Int^{\varphi}(\psi_1 \wedge \psi_2) \leftrightarrow (Int^{\varphi}\psi_1 \wedge Int^{\varphi}\psi_2)$. (U) ensures that $\Box$ behaves as a universal modality. (D+) ensures that conditional intention is consistent. (Cut) is a generalization of a schema $([C]\varphi \wedge Int\varphi) \rightarrow (Int\psi \leftrightarrow Int^{\varphi}\psi)$, i.e., if the condition of one intention is (conditionally) intended, such a conditional intention can be reduced.[13] (R-Mon) corresponds to the restricted monotonicity principle that we have argued for.

4.1.6. Theorem (Theorems and admissible rules of inference). *The following are derivable in* **L***:*

1. $Int^{\varphi}\psi \rightarrow \Diamond\varphi$
2. $Int^{\varphi}\psi \rightarrow \neg Int^{\varphi}\neg\psi$
3. *If* $\vdash \psi_1 \leftrightarrow \psi_2$, *then* $\vdash Int^{\varphi}\psi_1 \leftrightarrow Int^{\varphi}\psi_2$
4. *If* $\vdash \alpha \rightarrow \beta$, *then* $\vdash Int^{\varphi}\alpha \rightarrow Int^{\varphi}\beta$
5. *From* $\vdash ([C]\varphi \wedge [C]\psi \wedge [C](\varphi \vee \psi))$ *and* $\vdash (\neg Int^{(\varphi\vee\psi)}\neg\varphi \wedge \neg Int^{(\varphi\vee\psi)}\neg\psi)$ *it follows that* $\vdash (Int^{(\varphi\vee\psi)}\theta \rightarrow (Int^{\varphi}\theta \wedge Int^{\psi}\theta)$

Proof:

We derive only the last statement, since others are either easily derivable using D+ and K axioms or known to be derivable by standard methods in normal modal logics.[14]

[13] We call this axiom schema (Cut) as a reference to the cut rule for the logic of conditionals: from $\vdash \alpha > \psi \wedge (\alpha \wedge \psi) > \varphi$ infer $\vdash \alpha > \varphi$, where $\varphi > \psi$ reads as "*conditional on* φ, ψ" [KLM90]. If we replace $>$ with Int, we can rewrite it as an axiom in the following way: $Int^{\alpha}\psi \rightarrow (Int^{(\alpha\wedge\psi)}\varphi \rightarrow Int^{\alpha}\varphi)$, which is one of the directions of the bi-implication in the consequent of (Cut). The other direction – $Int^{\alpha}\psi \rightarrow (Int^{\alpha}\varphi \rightarrow Int^{(\alpha\wedge\psi)}\varphi)$ – follows from (R-Mon), since from $Int^{\alpha}\psi$, using (D+), it follows that $\neg Int^{\alpha}(\alpha \rightarrow \neg\psi)$. Hence, we could have written the (Cut) axiom in a weaker form, exactly as it is in the conditional logics, without loss of deductive power. We choose to write it in that redundant form just for the sake of intuitive understanding of what the axiom means.

[14] See different formulations of normal modal logics K and D in [BRV01].

1. $\neg Int^{(\varphi\vee\psi)}\neg\varphi \wedge \neg Int^{(\varphi\vee\psi)}\neg\psi$	
2. $Int^{(\varphi\vee\psi)}\theta$	Assumptions
3. $[C]\varphi \wedge [C]\psi \wedge [C](\varphi \vee \psi)$	
4. $((\varphi \vee \psi) \to \neg\varphi) \leftrightarrow \neg\varphi$	CPL
5. $(Int^{(\varphi\vee\psi)}\theta \wedge \neg Int^{(\varphi\vee\psi)}((\varphi \vee \psi) \to \neg\psi)) \to Int^{((\varphi\vee\psi)\wedge\psi)}\theta$	R-Mon, 4
6. $(\varphi \wedge (\varphi \vee \psi)) \leftrightarrow \varphi$	CPL
7. $(Int^{(\varphi\vee\psi)}\theta \wedge \neg Int^{(\varphi\vee\psi)}\neg\psi) \to Int^{\psi}\theta$	4-6, RE
8. $Int^{\psi}\theta$	1,2, 7. MP
9. $Int^{\varphi}\theta$	Analogously to 1-8
10. $Int^{\varphi}\theta \wedge Int^{\psi}\theta$	8,9
11. $Int^{(\varphi\vee\psi)}\theta \to (Int^{\varphi}\theta \wedge Int^{\psi}\theta)$	2-11, CPL

□

We move to proving the completeness of **L**. For any formula $\phi \in \mathcal{L}$, we will construct a special finitary language $\mathcal{L}_\phi$ and build a canonical model in a standard way, using maximal **L**-consistent subsets of $\mathcal{L}_\phi$ as possible worlds of the canonical model. As a consequence, the canonical model for every formula will be finite; hence, it follows that our logic has a finite model property.

4.1.7. Definition (Restricted languages). Take any formula $\phi \in \mathcal{L}$. By a *modal depth* of ϕ, $md(\phi)$, we mean the deepest degree of nesting of modalities $\Box$ and Int in ϕ. In other words, $md(\phi)$ is recursively defined as follows:

$$
\begin{aligned}
&md(p) = 0\\
&md(\neg\phi) = md(\phi)\\
&md(\phi_1 \vee \phi_2) = max(md(\phi_1), md(\phi_2))\\
&md(\Box\phi) = 1 + md(\phi)\\
&md(Int^{\phi_1}\phi_2) = 1 + max(md(\phi_1), md(\phi_2))
\end{aligned}
$$

By $Var(\phi)$ we mean the set of propositional variables that occur in ϕ.

For every formula $\phi \in \mathcal{L}$, we define a special restricted language $\mathcal{L}_\phi$ as follows:

$$\mathcal{L}_\phi = \{\psi \in \mathcal{L} \mid md(\psi) \leq md(\phi), Var(\psi) \subseteq Var(\phi)\}$$

Given that (RE) rule is admissible for $\Box$ and Int modalities[15] and that there are finitely many logically non-equivalent modal-free[16] formulae that contain finitely many propositional variables, $\mathcal{L}_\phi$ is *finite modulo* **L**[17] for any ϕ.

[15] If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \Box\varphi \leftrightarrow \Box\psi$, $\vdash Int^{\varphi}\theta \leftrightarrow Int^{\psi}\theta$, $\vdash Int^{\theta}\varphi \leftrightarrow Int^{\theta}\psi$.

[16] A formula is modal-free iff its modal depth is 0.

[17] A set of formulae S is finite modulo **L** iff there are finitely many classes of logically equivalent formulae in S. In other words, let $[\psi] = \{\varphi \in S \mid \vdash_{\mathbf{L}} \psi \leftrightarrow \varphi\}$. Then, S is finite modulo **L** iff the set $\{[\psi] \mid \psi \in S\}$ is finite.

4.1.8. Definition (MCSs and binary relations on them). Given any formula $\phi \in \mathcal{L}$, let $MCS_\phi \subseteq 2^{\mathcal{L}_\phi}$ be a collection of maximal **L**-consistent sets of formulae (shortly, mcs) of $\mathcal{L}_\phi$. Note that MCS_ϕ for any ϕ is finite: $\mathcal{L}_\phi$ is finite modulo **L** and for every $\Delta \in MCS_\phi$ and $\varphi \in \mathcal{L}_\phi$, $\varphi \in \Delta$ iff $\{\psi \in \mathcal{L}_\phi \mid \vdash_{\mathbf{L}} \varphi \leftrightarrow \psi\} \subseteq \Delta$.

Given any mcs $\Gamma \in MCS_\phi$, let $\Box\Gamma = \{\Box\psi \in \mathcal{L}_\phi \mid \Box\psi \in \Gamma\}$ be a $\Box$-theory of Γ, i.e. all the $\Box$-formulae that are in Γ. Then, $[\Gamma]_\Box = \{\Delta \in MCS_\phi \mid \Box\Gamma = \Box\Delta\}$ be a collection of maximal **L**-consistent sets of formulae that agree with Γ on all $\Box$-formulae. For any $\varphi \in \mathcal{L}_\phi$ and any mcs Γ, $Int^\varphi\Gamma = \{\psi \in \mathcal{L}_\phi \mid Int^\varphi\psi \in \Gamma\}$.

4.1.9. Definition (Canonical model). Having some $\Gamma \in MCS_\phi$ fixed, we define a canonical model $M_\Gamma = (W_\Gamma, f_\Gamma, V_\Gamma)$, where:

- $W_\Gamma = [\Gamma]_\Box$;
- For any $\varphi \in \mathcal{L}_\phi$, $\widehat{\varphi} = \{\Delta \in W_\Gamma \mid \varphi \in \Delta\}$;
- $f_\Gamma(\Delta, \widehat{\varphi}) = \begin{cases} \{\Sigma \in W_\Gamma \mid Int^\varphi\Delta \subseteq \Sigma\} & \text{if } [C]\varphi \in \Delta \\ \text{undefined} & \text{otherwise} \end{cases}$
- $V_\Gamma(p) = \widehat{p}$ for any $p \in Var$.

Note that, since $W_\Gamma \subseteq MCS_\phi$ is finite modulo **L**, every $\Delta \in W_\Gamma$ contains finitely many logically non-equivalent formulae. Hence, every canonical world Δ has its corresponding unique finite set of logically non-equivalent formulae S_Δ. Then, any set $P \subseteq W_\Gamma$ corresponds to $\widehat{\bigvee_{\Delta\in P}\bigwedge_{\psi\in S_\Delta}\psi}$. Therefore, any subset of possible worlds $P \subseteq W$ may be identified with some formula φ_P, such that $\widehat{\varphi_P} = P$.

First, we are to check that a canonical model is in **C**.

4.1.10. Lemma (Any canonical frame is a conditional frame). *For any* $\Gamma \in MCS_\phi$, $(W_\Gamma, f_\Gamma) \in \mathbf{C}$.

Proof:
Since **L** is consistent, $W_\Gamma \neq \emptyset$ for any $\Gamma \in MCS_\phi$. It is a routine to check that $f_\Gamma : (W_\Gamma \times 2^{W_\Gamma}) \to 2^{W_\Gamma}$ is a well-defined partial function of the corresponding type.

2a: Assume that $f_\Gamma(\Delta, P)$ is defined for some $\Delta \in W_\Gamma, P \subseteq W_\Gamma$. Then, $P = \widehat{\varphi}$ for some $\varphi \in \mathcal{L}_\phi$ and $[C]\varphi \in \Delta$. For contradiction, assume that $f_\Gamma(\Delta, \widehat{\varphi}) \cap \widehat{\varphi} = \emptyset$. Then, there is no canonical world $\Sigma \in \widehat{\varphi}$, such that $Int^\varphi\Delta \subseteq \Sigma$. Then, $\{\varphi\} \cup \Box\Gamma \cup Int^\varphi\Delta \vdash \bot$. Then, (since $\Box\Gamma$ is consistent) there exists some $\psi_1, \ldots, \psi_n \in Int^\varphi\Delta$, such that $\Box\Gamma \vdash \neg(\varphi \wedge \psi_1 \wedge \ldots \wedge \psi_n)$. From S5 for $\Box$, $\Box\Gamma \vdash \neg\Diamond(\varphi \wedge \psi_1 \wedge \ldots \wedge \psi_n)$. Then, since $\psi_1, \ldots, \psi_n \in Int^\varphi\Delta$, $Int^\varphi\psi_1, \ldots, Int^\varphi\psi_n \in \Delta$. By (K) and (RN'), $Int^\varphi(\psi_1 \wedge \ldots \wedge \psi_n) \in \Delta$. Since $\neg\Diamond(\varphi \wedge \psi_1, \ldots, \psi_n) \in \Box\Gamma \subseteq \Delta$, $\neg\Diamond(\varphi \wedge \psi_1, \ldots, \psi_n) \in \Delta$. Hence, $Int^\varphi(\psi_1 \wedge \ldots \wedge \psi_n) \wedge \neg\Diamond(\varphi \wedge \psi_1 \wedge \ldots \wedge \psi_n) \in \Delta$, which contradicts (D+) axiom, which cannot happen, given that Δ by definition is **L**-consistent.

Therefore, $f_\Gamma(\Delta, P) \cap P \neq \emptyset$ for any Δ and P, s.t. $f_\Gamma(\Delta, P)$ is defined.

2b: If $f_\Gamma(\Delta, P)$ and $f_\Gamma(\Delta, Q)$ are defined, while $P \cap Q \neq \emptyset$, then $f_\Gamma(P \cap Q)$ is defined. Follows directly from (C-Inter) axiom.

2c: If $f_\Gamma(\Delta, P)$ and $f_\Gamma(\Delta, P \cap Q)$ are defined and $f_\Gamma(\Delta, P) \cap P \cap Q \neq \emptyset$, then $f_\Gamma(\Delta, P \cap Q) \subseteq f_\Gamma(\Delta, P)$. Assume that $f_\Gamma(\Delta, P)$ and $f_\Gamma(\Delta, P \cap Q)$ are defined. Then, $P = \widehat{\varphi_1}$, $P \cap Q = \widehat{\varphi_2}$ for some φ_1, φ_2, such that $\widehat{\varphi_2} \subseteq \widehat{\varphi_1}$ (i.e. $\widehat{\varphi_2} = \widehat{\varphi_1 \wedge \varphi_2}$), and $[C]\varphi_1, [C]\varphi_2 \in \Delta$. Assume that $f_\Gamma(\Delta, \widehat{\varphi_1}) \cap \widehat{\varphi_1 \wedge \varphi_2} \neq \emptyset$. It means that there exists some $\Sigma \in W_\Gamma$, such that $Int^{\varphi_1}\Delta \cup \{\varphi_1 \wedge \varphi_2\} \subseteq \Sigma$. We claim that $\neg Int^{\varphi_1}(\varphi_1 \to \neg(\varphi_1 \wedge \varphi_2)) \in \Delta$. For contradiction, assume otherwise. Then, since Δ is mcs, $Int^{\varphi_1}(\varphi_1 \to \neg(\varphi_1 \wedge \varphi_2)) \in \Delta$. Then, by definition of f_Γ, if $\Sigma \in f_\Gamma(\Delta, \widehat{\varphi_1})$, then $\varphi_1 \to \neg(\varphi_1 \wedge \varphi_2) \in \Sigma$. Take any $\Sigma \in f_\Gamma(\Delta, \widehat{\varphi_1}) \cap \widehat{\varphi_1 \wedge \varphi_2}$. Then, by our assumptions, $\varphi_1 \to \neg(\varphi_1 \wedge \varphi_2), \varphi_1 \wedge \varphi_2 \in \Sigma$. Then, $\Sigma \vdash \bot$, which contradicts the fact that Σ is mcs. Then, $f_\Gamma(\Delta, \widehat{\varphi_1}) \cap \widehat{\varphi_1 \wedge \varphi_2} = \emptyset$, which contradicts our initial assumption. Hence, $\neg Int^{\varphi_1}(\varphi_1 \to \neg(\varphi_1 \wedge \varphi_2)) \in \Delta$. Then, given that $[C]\varphi_1, [C](\varphi_1 \wedge \varphi_2) \in \Delta$, by (R-Mon) axiom we get that $Int^{\varphi_1}\psi \to Int^{(\varphi_1 \wedge \varphi_2)}\psi \in \Delta$ for any $\psi \in \mathcal{L}_\phi$. Then, $Int^{\varphi_1}\Delta \subseteq Int^{(\varphi_1 \wedge \varphi_2)}\Delta$, i.e. $f_\Gamma(\Delta, \widehat{\varphi_1 \wedge \varphi_2}) \subseteq f_\Gamma(\Delta, \widehat{\varphi_1})$, i.e. $f_\Gamma(\Delta, P \cap Q) \subseteq f_\Gamma(\Delta, P)$ (given that $P \cap Q = \widehat{\varphi_2} \subseteq \widehat{\varphi_1} = P$).

2d: If $f_\Gamma(\Delta, P) \subseteq Q$ and $f_\Gamma(\Delta, P \cap Q)$ is defined, then $f_\Gamma(\Delta, P) = f_\Gamma(\Delta, P \cap Q)$. Again, assume that $f_\Gamma(\Delta, P), f_\Gamma(\Delta, P \cap Q)$ are defined and $f_\Gamma(\Delta, P) \subseteq Q$. Then, $P = \widehat{\varphi}$, $P \cap Q = \widehat{\varphi \wedge \psi}$ for some $\varphi, \psi \in \mathcal{L}_\phi$, s.t. $[C]\varphi, [C](\varphi \wedge \psi) \in \Delta$, where $P = \widehat{\varphi}, Q = \widehat{\psi}$. By our assumption, $f_\Gamma(\Delta, \widehat{\varphi}) \subseteq \widehat{\psi}$. In other words, if $Int^\varphi\Delta \subseteq \Sigma$, then $\psi \in \Sigma$ for any $\Sigma \in W_\Gamma$. We claim that $Int^\varphi\psi \in \Delta$. Assume otherwise. Since Δ is mcs, $\neg Int^\varphi\psi \in \Delta$. The set $\Box\Gamma \cup Int^\varphi\Delta \cup \{\neg\psi\}$ is consistent. To prove this consistency claim, assume otherwise. Then, $\Box\Gamma \cup Int^\varphi\Delta \vdash \psi$. In other words, there exists $\vartheta_1, \ldots, \vartheta_n \in Int^\varphi\Delta$, such that $\Box\Gamma \cup \{\vartheta_1, \ldots, \vartheta_n\} \vdash \psi$. I.e., $Int^\varphi\vartheta_1 \wedge \ldots \wedge Int^\varphi\vartheta_n \in \Delta$ and $\Box\Gamma \vdash (\vartheta_1 \wedge \ldots \wedge \vartheta_n) \to \psi$. Then, by (K), $Int^\varphi(\vartheta_1 \wedge \ldots \wedge \vartheta_n) \in \Delta$ and $\Box\Gamma \vdash (\vartheta_1 \wedge \ldots \wedge \vartheta_n) \to \psi$. Since $\Box\Gamma \subseteq \Delta$, by the derivable rule $\vdash (\vartheta_1 \wedge \ldots \wedge \vartheta_n) \to \psi \Rightarrow \vdash Int^\varphi(\vartheta_1 \wedge \ldots \wedge \vartheta_n) \to Int^\varphi\psi$, $Int^\varphi\psi \in \Delta$, contradicting the fact that $\neg Int^\varphi\psi \in \Delta$ and Δ is consistent. Hence, $\Box\Gamma \cup Int^\varphi\Delta \cup \{\neg\psi\}$ is consistent. Then, there exists some mcs $\Sigma \in W_\Gamma$, such that $\Box\Gamma \cup Int^\varphi\Delta \cup \{\neg\psi\} \subseteq \Sigma$. Then, since $Int^\varphi\Delta \subseteq \Sigma$, $\Sigma \in f_\Gamma(\Delta, \widehat{\varphi})$. But $\neg\psi \in \Sigma$, while by our initial assumption $f_\Gamma(\Delta, \widehat{\varphi}) \subseteq \widehat{\psi}$. Contradiction. So we conclude that $Int^\varphi\psi \in \Delta$, from which it follows that $[C]\varphi \in \Delta$. By (Cut) axiom, $[C](\varphi \wedge \psi) \to (Int^\varphi\theta \leftrightarrow Int^{(\varphi \wedge \psi)}\theta) \in \Delta$. Since $[C](\varphi \wedge \psi) \in \Delta$, $Int^\varphi\theta \leftrightarrow Int^{(\varphi \wedge \psi)}\theta \in \Delta$ for any $\theta \in \mathcal{L}_\phi$, from which it follows that $f_\Gamma(\Delta, \widehat{\varphi}) = f_\Gamma(\Delta, \widehat{\varphi \wedge \psi})$. □

4.1.11. Lemma (Existence). *For any $\Delta \in W_\Gamma$, if $\neg Int^\varphi\psi \in \Delta$, then there exists $\Sigma \in f_\Gamma(\Delta, \widehat{\varphi})$, s.t. $\neg\psi \in \Sigma$.*

Proof:
Assume that $\neg Int^\varphi\psi \in \Delta$. Then, as we have shown in Lemma 4.1.10, $\Box\Gamma \cup$

$Int^{\varphi}\Delta \cup \{\neg\psi\}$ is consistent. In this case, there exists a mcs $\Sigma \in W_\Gamma$, such that $\Box\Gamma \cup Int^{\varphi}\Delta \cup \{\neg\psi\} \subseteq \Sigma$. Since $Int^{\varphi}\Delta \subseteq \Sigma$, $\Sigma \in f_\Gamma(\Delta, \widehat{\varphi})$ and $\neg\psi \in \Sigma$. □

4.1.12. Lemma (Truth lemma). *Given any canonical model $M_\Gamma = (W_\Gamma, f_\Gamma, V_\Gamma)$, $\Delta \in W_\Gamma$ and $\varphi \in \mathcal{L}_\phi$, $M_\Gamma, \Delta \models \varphi$ iff $\varphi \in \Delta$.*

Proof:
By induction on the complexity of φ. Cases of φ being atomic or modal-free are trivial. Let $\varphi := \Box\psi$.

($\Rightarrow$) Assume $M_\Gamma, \Delta \models \Box\psi$. Then, $W_\Gamma = [\![\psi]\!]_{M_\Gamma}$, i.e. by IH $W_\Gamma = \widehat{\psi}$. Then, $\Box\Gamma \vdash \psi$, from which it follows that $\Box\psi \in \Delta$, since $\Box\Gamma = \Box\Delta$.

($\Leftarrow$) Assume $\Box\psi \in \Delta$. Then, since $\Box\Delta = \Box\Gamma = \Box\Sigma$ for any $\Sigma \in W_\Gamma$, $\widehat{\Box\psi} = W_\Gamma$. Then, by axiom T for $\Box$, $\widehat{\psi} = W_\Gamma$, i.e. by IH $[\![\psi]\!]_{M_\Gamma} = W_\Gamma$, i.e. $M_\Gamma, \Delta \models \Box\psi$.

Let $\varphi := Int^{\alpha}\psi$.

($\Rightarrow$) Assume $M_\Gamma, \Delta \models Int^{\alpha}\psi$. Then, $f_\Gamma(\Delta, [\![\alpha]\!]_{M_\Gamma}) \subseteq [\![\psi]\!]_{M_\Gamma}$. By IH, $f_\Gamma(\Delta, \widehat{\alpha}) \subseteq \widehat{\psi}$. For contradiction, assume that $Int^{\alpha}\psi \notin \Delta$. Then, $\neg Int^{\alpha}\psi \in \Delta$. By Lemma 4.1.11, there exists $\Sigma \in f_\Gamma(\Delta, \widehat{\alpha})$, such that $\neg\psi \in \Sigma$, which contradicts that $f_\Gamma(\Delta, \widehat{\alpha}) \subseteq \widehat{\psi}$. Hence, $Int^{\alpha}\psi \in \Delta$.

($\Leftarrow$) Assume that $Int^{\alpha}\psi \in \Delta$. Then, $f_\Gamma(\Delta, \widehat{\alpha}) \subseteq \widehat{\psi}$. By IH, $f_\Gamma(\Delta, [\![\alpha]\!]_{M_\Gamma}) \subseteq [\![\psi]\!]_{M_\Gamma}$, i.e. $M_\Gamma, \Delta \models Int^{\alpha}\psi$. □

4.1.13. Theorem (Soundness and completeness). *Given any formula $\varphi \in \mathcal{L}$, the following holds:*

$$\mathbf{C} \models \varphi \Leftrightarrow \mathbf{L} \vdash \varphi$$

Proof:

($\Leftarrow$) By the routine check of axioms' validity.

($\Rightarrow$) By contraposition. Assume that $\mathbf{L} \nvdash \varphi$. Then, $\{\neg\varphi\}$ is $\mathbf{L}$-consistent. Take a language $\mathcal{L}_{\neg\varphi}$, pick a mcs $\Gamma \ni \neg\varphi$ and build a model $M_\Gamma = (W_\Gamma, f_\Gamma, V_\Gamma)$. Then, $\neg\varphi \in \Gamma \in W_\Gamma$, from which by Lemma 4.1.12 it follows that $M_\Gamma, \Gamma \models \neg\varphi$. By Lemma 4.1.10, $(W_\Gamma, f_\Gamma) \in \mathbf{C}$, hence, $\mathbf{C} \not\models \varphi$. □

4.1.5 Time, action, and and side-effects

We have developed a formalism that captures key logical properties of conditional intention: restricted monotonicity and free choice inferences together with a lim-

ited set of conditions. In this section, we briefly discuss ways to refine our analysis with respect to other aspects: temporality, actions, and the side-effects problem. We have not addressed these aspects from the start, since they may be redundant depending on the goals of the logic's user analysis and the varying philosophical stances she might hold.

Explicit time and actions

So far, we have implicitly treated the object of intention as a proposition about outcomes of an action the agent ruminates on doing. Such a view presupposes a certain connection with actions and time. E.g., one cannot intend to change the past or to see to it that the future will be in a certain way if it is out of one's control. In this subsection, we will make the temporal structure and the agent's actions explicit, both in the language and semantics.

We analyse conditional intentions as *present-directed*: by intending that ψ conditional on φ, the agent contingently commits herself to a currently available action that brings about that ψ in case of φ. We do not deal with the *future-directed* conditional intentions, by which agents contingently commit themselves to some distant future action. For example, the author's intention to finish typing this sentence in case their laptop keeps working is in the scope of our inquiry, while the future parents' intention to pay for their children's education in case they have children is not.

We concentrate on present-directed intention because of its connection to group agency and *mens rea*. Present-directed intentions are intentions *with which* agents act, and exactly present-directed intentions are important when we want to understand the mental component of the agents' responsibility for their actions, or whether the agent acted as a member of a group. Future-directed intentions, on the other hand, may or may not bind agents to any actions whatsoever at the moment of deliberation. As Cohen and Levesque concisely put it, "[future-directed intentions] *guide agents' planning and constrain their adoption of other intentions, whereas* [present-directed intentions] *function causally in producing behavior*" [CL90a, p.216].

For instance, it is important to understand which present-directed intentions Irvin had when he grabbed a woman and held a knife to her throat. Did he intend to kill her, conditional on not being released? Or did he intend to make an empty threat? It seems irrelevant whether he had malicious or merciful future-directed intentions towards what to do when and if he is released, since those did not bind him to assaulting the woman at the moment (at least not directly, but via the dependent present-directed intention). Analogously, to understand whether two agents have done something together, we need to determine whether their actions were caused by some shared motivational state composed of their

conditional intentions; whether they were planning to do something together in the distant future is irrelevant.

Present-directedness distinguishes our analysis from that of Cohen and Levesque. In their seminal work, they explicitly state that they analyse future-directed intention and every time they use the word "intention", they mean the future-directed one [CL90a, p.216]. Therefore, some Cohen and Levesque's core issues – e.g., persistence of intention over time – are irrelevant to us. Yet the principles that are non-specific to future-directed intention are to be satisfied: rational present-directed intentions should still be consistent; a rational intender should (believe to) be able to satisfy her intention; and rational agents may not intend all the expected side-effects of their intention.

Having clarified the essential relation between intention, actions, and time that we have in mind, we continue with the refinement of our formalism to make these points explicit. First, we extend the language $\mathcal{L}$ to $\mathcal{L}_t$ in the following manner:

$$\mathcal{L}_t \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid X\varphi \mid \varphi U \varphi \mid [stit]\varphi \mid Int^{\varphi} X\varphi$$

where $X\varphi$ reads as *in the next moment in time, φ will hold.* $\varphi U \psi$ reads as *ψ is true now or will be true at some moment in the future, and φ holds until ψ.* We will use the widely accepted acronyms: $F\varphi := X(\top U \varphi)$ (*in some moment in future, φ holds*); $G\varphi := \neg F \neg \varphi$ (*from now on, it will always be true that φ*).

$[stit]\varphi$ reads as *the agent sees to it that φ holds.* $\Box\varphi$ reads as *at the given moment, it is practically necessary that φ.* We still have the same understanding of practical necessity: φ is practically necessary iff φ is true independently of how all agents and the environment behave. As we are to see later, practical necessity coincides with historical necessity.

$Int^{\varphi} X\psi$ reads as *the agent intends that* ***in the next moment in time*** *ψ will hold, conditional on φ.* We adopt the acronym from Section 4.1.2: $[C]\varphi := Int^{\varphi} X \top$. We explicitly take the object of one's intention to be a proposition about the immediate consequences of one's currently available actions she commits herself to. Our language does not allow meaningless expressions about intentions towards propositions in the past (e.g., *I intend that I have finished writing this subsection yesterday*) or in the moment of deliberation (e.g., *I intend that this subsection is already written now, as I am writing it*). No matter what the agent is going to do, propositions about the past or present are either true or false independently of her or anyone's choice of action; so propositions about the past and present are either practically necessary (if they are true) or practically impossible (if they are false).

We still can express the intention towards the long-term consequences of one's current actions: e.g., $Int^{\varphi} X(F\psi)$ reads as the agent intends that ψ will hold eventually after the next moment, conditional on φ. Yet this statement should not be

confused with the expression of the future-directed intention! $Int^{\varphi}X(F\psi)$ means that the agent intends to do some currently available action, such that, conditional on φ, the action has *as its immediate consequence* that ψ will eventually be true. It does not mean that the agent intends to do some action in the future that guarantees ψ conditional on φ. For instance, I might have a present-directed intention that my colleague will eventually die from poisoning, conditional on her drinking her coffee. This intention causes me to poison her coffee. Given that my colleague drinks the poisoned coffee, she will survive the next couple of hours or even days, but will inevitably die sooner or later. Her inevitable death in the future is an immediate consequence of the action I am deliberating upon now. On the contrary, I may have a future-directed intention to poison my colleague. In that case, I still intend that my colleague will eventually die from poisoning, but my future-directed intention does not necessarily cause any action of mine at the given moment: I might commit to the act of poisoning in, e.g., the next year.

As for the semantics, we will adapt game models (as in Def. 3.2.1) to interpret a single-agent STIT logic over them, and add the set-selection functions as in Def. 4.1.1.

4.1.14. Definition (Game models with intentions). Let $Ags = \{i, e\}$ be a set of agents, where i is the agent whose intentions we model, and e is the environment. Then, a tuple $M = (S, \Sigma_i, \Sigma_e, g, f, V)$ is a game model with intentions iff:

1. $(S, \Sigma_i, \Sigma_e, g, V)$ is a game model, as in Def. 3.2.1. To remind the reader, by Pl_M we denote the set of infinite plays $(w_0, \sigma_0, \ldots)$ in the game model. Two plays $\tau = (w_0, \sigma_0, \ldots) \in Pl_M$, $\kappa = (v_0, \sigma'_o, \ldots) \in Pl_M$ are state equivalent, $\tau \simeq_S \kappa$ iff they begin with the same state: $w_0 = v_0$. τ and κ are action-equivalent w.r.t. i, $\tau \simeq_{\{i\}} \kappa$, iff $\tau \simeq_S \kappa$ and τ agree on the first action of i: $\sigma_0(i) = \sigma'_0(i)$;
2. $f : Pl_M \to (2^{Pl_M} \to 2^{Pl_M})$ is a partial set-selection function with the following properties:
   (a) If $f(\tau, P)$ is defined, then $[\tau]_S \cap f(\tau, P) \cap P \neq \emptyset$: agent's intentions in case of P are consistent with P at the given moment;
   (b) If $f(\tau, P)$ is defined, then $f(\tau, P) \subseteq [\tau]_S$: the object of the agent's present-directed conditional intention is a proposition about what the agent can bring about *now*, conditional on her and the environment's behaviour;
   (c) If $f(\tau, P)$ and $f(\tau, Q)$ are defined, while $P \cap Q \cap [\tau]_S \neq \emptyset$, then $f(\tau, P \cap Q)$ is defined;
   (d) If $f(\tau, P)$ and $f(\tau, P \cap Q)$ are defined and $f(\tau, P) \cap P \cap Q \neq \emptyset$, then $f(\tau, P \cap Q) \subseteq f(\tau, P)$;

(e) Given that $f(\tau, P)$ is defined, if $f(\tau, P) \subseteq Q$ and $f(\tau, P \cap Q)$ is defined, then $f(\tau, P) = f(\tau, P \cap Q)$;

(f) *Control constraint*: if $f(\tau, P) = Q$, then there exists an action $\sigma \in \Sigma_i$, such that: there is a play $\kappa \in P$, such that $\kappa \equiv_S \tau$ and $\kappa_A(0)(i) = \sigma_i$, and for any play $\lambda \in [\kappa]_{\{i\}}$, if $\lambda \in P$, then $\lambda \in Q$;

(g) *Intentions are up to the agent*: if $\tau \equiv_{\{i\}} \kappa$, then $f(\tau, P) = f(\kappa, P)$.

As we see, f function has all the properties as in Def. 4.1.1. Additionally, we require (2a) to treat state-equivalent plays as practically possible worlds and (2b) to make sure we model present-directed (conditional) intention. Control constraint expressed in (2f) requires that if the agent intends Q conditional on P, then the agent must be able to bring about Q conditional on P. To be more precise, if $f(\tau, P) = Q$, then there exists an action $\sigma_i \in \Sigma_i$, which is compatible with P and which guarantees that Q in case of P. We require this action to be compatible with the condition to avoid the situations where the agent has a conditional intention, but can "satisfy" it only by preventing the condition from happening. Condition (2g) expresses that the agent performs her atomic actions with the same intention, independently of its outcome.

4.1.15. Definition (Semantics for $\mathcal{L}_t$). Given a model $M = (W, \Sigma_i, \Sigma_e, g, f, V)$ and a play $\tau = (w_0, \sigma_0, \ldots) \in Pl_M$, the satisfiability relation $\models$ is inductively defined as follows:

$$
\begin{array}{lcl}
M, \tau \models p & \Leftrightarrow & w_0 \in V(p) \\
M, \tau \models \neg\varphi & \Leftrightarrow & M, \tau \not\models \varphi \\
M, \tau \models \varphi \vee \psi & \Leftrightarrow & M, \tau \models \varphi \text{ or } M, \tau \models \psi \\
M, \tau \models \Box\varphi & \Leftrightarrow & \forall \kappa \in Pl_M : \tau \simeq_S \kappa \Rightarrow M, \kappa \models \varphi \\
M, \tau \models X\varphi & \Leftrightarrow & M, (w_1, \sigma_1, \ldots) \models \varphi \\
M, \tau \models \varphi U \psi & \Leftrightarrow & \exists n \in \mathbb{N} : M, (w_n, \sigma_n, \ldots,) \models \psi \text{ and} \\
 & & \forall m \in \mathbb{N}(0 \leq m \leq n \Rightarrow M, (w_m, \sigma_m, \ldots) \models \varphi) \\
M, \tau \models [stit]\varphi & \Leftrightarrow & \forall \kappa \in Pl_M : \tau \simeq_{\{i\}} \kappa \Rightarrow M, \kappa \models \varphi \\
M, \tau \models Int^{\varphi} X\psi & \Leftrightarrow & f(\tau, [\![\varphi]\!]_M) \text{ is defined and } f(\tau, [\![\varphi]\!]_M) \subseteq [\![X\psi]\!]_M
\end{array}
$$

where $[\![\varphi]\!]_M = \{\kappa \in Pl_M \mid M, \kappa \models \varphi\}$.

4.1.16. Example (The rent contract). Anne is homeless and looking for a place to rent in Amsterdam. Since she struggles to find anything long-term, she considers *antikraak*.[18] Anne gets a proposition via antikraak to rent a place from a

[18] Antikraak is a property-management practice in the Netherlands where people rent spaces in temporally vacant buildings to prevent squatting. The rent is typically cheap and the contract is easier to get, but the tenant has to move out on a short notice as soon as the owner needs to use the building again, so renting via antikraak presupposes unpredictability of terms.

house owner who has temporarily left Amsterdam for Brussels. If Anne agrees, as soon as the house owner gets bored in Brussels and wants to come back to Amsterdam, Anne will have to leave the house.

Anne constantly gets information about the housing market from the insiders, so she can act conditionally on the updates she might receive. Anne intends to agree to rent via antikraak, in case no indefinite-term lease offers are coming on the market for the next month. Anne also intends to deny the antikraak proposal and wait for a lease for an indefinite term, if such an offer appears on the market in the next month.

Let $Var = \{m, h, p\}$, where m stands for *Anne has to **m**ove out*, h – *Anne has a **h**ouse to stay in*, p – *renting **p**roperty for an indefinite term appears on the market*.

We construct the following model for Example 4.1.16. Let $\Sigma_i = \{wait, accept, \epsilon\}$ be the agent's actions, where *wait* is to reject the antikraak and wait for the indefinite term proposal; *accept* – to accept the antikraak, ϵ – to do nothing. $\Sigma_e = \{contr, no, prolong, back, \epsilon\}$ are the possible behaviours of the environment, where: *contr* – the indefinite-term contract appears on the market, *no* – no such contract appears on the market, *prolong* – the owner of the antikraak property stays in Brussel for one more month, *back* – the owner gets back to Amsterdam, ϵ – do nothing.

Let w_0 be a world, where Anne receives the antikraak proposal and may either accept it or reject it and wait for the indefinite term proposal; w_1 – the world where Anne accepts antikraak, while no indefinite term proposals arrive the market; w_2 – the world where Anne rejected antikraak and kept waiting for the indefinite term proposal, and the proposal appears on the market; w_3 – the world were Anne has to move out, because the home owner got bored in Brussel; w_4 – the world where Anne accepts antikraak, while the indefinite term proposal arrives on the market. The game model for the example is as in Fig. 4.1.

Let τ be any play, such that $\tau_S(0) = w_0$. Let $P = [\![Xp]\!]_M$, and $Q = [\![X\neg p]\!]_M$. Then, $f(\tau, P) = \{\kappa \in Pl_M \mid \kappa \equiv_S \tau, \kappa_S(1) = w_2\}$; $f(\tau, Q) = \{\kappa \in Pl_M \mid \kappa \equiv_S \tau; \kappa_S(1) = w_1\}$. Then, Anne intends to rent the house indefinitely in case she gets an undetermined offer, and she intends to accept the antikraak proposal and stay there until she has to move out in case there is no indefinite term offer: $M, \tau \models Int^{Xp}X(Gh) \wedge Int^{X\neg p}X(hUXm)$.

Note that because of the control constraint, Anne cannot, e.g., unconditionally intend to rent a house for the next 6 months – $IntX(X^6h)$, where $X^1\varphi := X\varphi$, $X^{n+1}\varphi := X(X^n\varphi) \wedge X^n\varphi$ – since there is no action of Anne to guarantee that. Anne has two actions at w_0: *accept* and *wait*. If Anne chooses *accept* at w_0, depending on how long the homeowner will stay in Brussels, she will or will not have a house for 6 months. If she chooses *wait* at w_0, she might or might

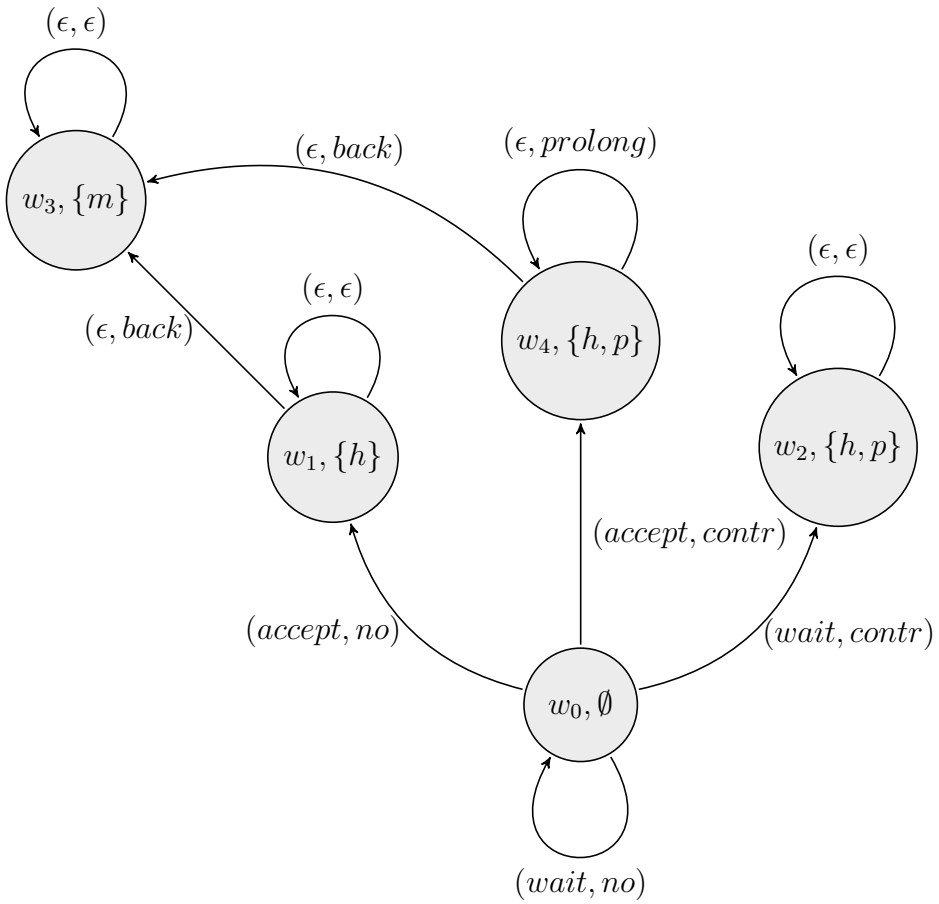

Figure 4.1: Game model for Example 4.1.16.

not get a house at all, depending on whether there will be an undetermined offer on the market. Moreover, Anne cannot intend to rent a house for the next 6 months, even conditionally on her getting an undetermined offer or the homeowner staying in Brussels for 6 months. Let the condition be expressed as follows: $\varphi := X^7(\neg m) \vee Xp$, i.e. either the homeowner is not going back to Amsterdam for the next 7 months, or a permanent contract becomes available during this month. Still, $Int^{\varphi}X(X^6h)$ cannot be satisfied on the game model in Figure 4.1 with any well-defined function f. No single action of Anne guarantees $X(X^6h)$ in case of φ. If there will be no permanent contract but the homeowner is not going back to Amsterdam, then Anne should choose *accept*; if a permanent contract becomes available but the homeowner is coming back to Amsterdam, Anne should choose *wait*. Yet no action will guarantee $X(X^6h)$ in any case where φ holds.

Next, we list some interesting (in)validities on the class of game models with intentions. First, note that axioms and rules of **L**, adequately translated from $\mathcal{L}$ to $\mathcal{L}_t$, are still valid:

4.1.17. Proposition. *The following formulae are valid in the class of concurrent game structures with intentions:*

$\models Int^{\varphi}X\psi \to [C]\varphi$
$\models Int^{\varphi}X(\psi \to \theta) \to (Int^{\varphi}X\psi \to Int^{\varphi}X\theta)$
$\models (\Box X\psi \wedge [C]\varphi) \to Int^{\varphi}X\psi$
$\models Int^{\varphi}X\psi \to \Diamond(\varphi \wedge X\psi)$
$\models ([C]X\alpha \wedge [C](X\alpha \wedge X\psi)) \to (Int^{X\alpha}X\psi \to (Int^{X\alpha}X\varphi \leftrightarrow Int^{(X\alpha \wedge X\psi)}X\varphi))$
$\models ([C]X\alpha \wedge [C](X\alpha \wedge X\beta)) \to \big((Int^{X\alpha}X\varphi \wedge \neg Int^{X\alpha}X(\alpha \to \neg\beta)) \to Int^{(X\alpha \wedge X\beta)}X\varphi\big)$
$\models \big([C]\varphi \wedge [C]\psi \wedge \Diamond(\varphi \wedge \psi)\big) \to [C](\varphi \wedge \psi)$

Proof:
Follows from (2a-2e) of Def. 4.1.14, analogously to the proof of **L**'s soundness in Theorem 4.1.13. □

The following validities concern the interaction between agents' intentions, actions, and abilities:

4.1.18. Proposition. *The following formulae are valid in the class of concurrent game structures with intentions:*

$$\models Int^{\varphi}X\psi \leftrightarrow [stit]Int^{\varphi}X\psi \tag{4.13}$$
$$\models Int^{\varphi}X\psi \to \Diamond\big([stit](\varphi \to X\psi) \wedge \langle stit\rangle\varphi\big) \tag{4.14}$$

Proof:
(4.13) corresponds to (2g), and (4.14) – to (2f) of Def. 4.1.14. □

Informally, (4.14) expresses the control constraint: if the agent intends that $X\psi$ conditional on φ, then she has an action to guarantee that $X\psi$ if φ that is compatible with φ. (4.13) expresses the fact that intentions are up to the agent: she intends something iff she sees to it that she so intends.

In our formalism, we do not presuppose the persistence of intentions, since they are understood as present-directed. For instance, in [CL90a], it is argued that if the idealised rational agent intends that φ, then she must continue to intend so until she reaches the state where either φ is true or no longer possible for her to achieve. In $\mathcal{L}_t$, this can be expressed by the following formula: $IntX(F\varphi) \rightarrow (IntX(F\varphi))U(\varphi \vee \Box G\neg\varphi)$. However, this formula is not valid in the class of CGSs with intentions. Given that we interpret $IntX(F\varphi)$ as a present-directed intention, this is intuitively plausible. For example, assume I have a present-directed intention that my colleague will eventually die from poisoning. Then, according to (4.14), I should have an available action that would immediately satisfy my intention, e.g., I can poison her coffee. After I poisoned her coffee, I would have known that she would eventually die from poisoning, so there would be no need for me to keep my intention until she actually died. I would have achieved what I aimed for, but I would not have seen the desired consequences yet. If I, despite my intention, failed to poison my colleague's coffee, I still might drop that intention: it could be that I had reasons to poison my colleague only at that past moment, but in the present, there is no reason for me to do so. On the contrary, if I have a future-directed intention that my colleague will eventually die from poisoning, then my intention should persist until I either reach my goal or realise that the goal is no longer possible to reach.

We proceed with showing which important notions and distinctions can be expressed in $\mathcal{L}_t$ as opposed to $\mathcal{L}$. In $\mathcal{L}_t$, we can track whether the agent has control over the conditions of her intentions. This is important, since, as Ludwig argues, whenever the agent has control over the condition and exercises it, the corresponding conditional intention should be reducible to the unconditional one:

> "*If conditions are in an agent's control, they can be antecedents only if the agent has sufficient reason not to exercise control (negative or positive as the case may be) over their obtaining*" [Lud15, p.43].

The positive control over φ means the ability to ensure that φ holds: $\Diamond[stit]\varphi$; the negative control over φ means the ability to prevent φ: $\Diamond[stit]\neg\varphi$.

Often it makes sense to have an intention conditional on φ, while having only negative or only positive control over φ and refusing to exercise it. For instance, a nurse has only positive control over her patient's death and only negative control

over his survival (i.e., she can kill him, but cannot ensure that he will survive). She may, e.g., intend to change his dressing in case he survives and intend to notify the doctor in case he dies. In both cases, whether the patient dies or survives is a contingency for her, given that she does not intend to kill the patient. If the nurse has both positive and negative control over the patient's life, it seems rational for her to have such conditional intentions if she refuses to exercise her control over the patient's life, i.e., if she intends neither to kill nor to save him. If there are no reasons not to decide the patient's fate, intentions conditional on the decision she intends to make are reducible to her unconditional intentions via the (Cut) axiom.

Notably, we can distinguish *pre*conditional intentions in $\mathcal{L}_t$. $Int^{\varphi}X\psi \wedge \neg\Diamond[stit]X\psi$ stands for "*the agent intends that ψ conditional on φ and is unable to ensure ψ unconditionally*". As we have noted before, $Int^{\varphi}X\psi$ entails $\Diamond\big([stit](\varphi \to X\psi) \wedge \langle stit\rangle\varphi\big)$. Hence, the agent intends that ψ conditional on φ, can ensure ψ conditionally on φ, but cannot ensure it unconditionally. So φ entails a necessary precondition for that agent to ensure ψ. Moreover, sometimes we can separate reasons from preconditions when they are mixed. For instance, assume one intends to go to a movie theatre, in case a good film is screened, and she has enough money to buy a ticket. Then, the condition contains a reason (the good film is screened) and a precondition (one has enough money to get a ticket) at the same time. In $\mathcal{L}_t$, we can express the difference. Let φ, ϑ and ψ stand for "*the good film is screened*", "*one has enough money to get a ticket*" and "*one watches the film in the movie theatre*", respectively. Then, the intention in question is expressible as follows: $Int^{(\varphi\wedge\vartheta)}X\psi$. One can watch a movie in the movie theatre in case she has enough money, regardless of whether the film is good or not:

$$\Diamond\big([stit](\vartheta \to X\psi) \wedge \langle stit\rangle\vartheta\big)$$

Yet she cannot watch a film unconditionally, since she needs money to buy a ticket: $\neg\Diamond[stit]X\psi$. Hence, in the condition $(\varphi \wedge \vartheta)$, φ is a reason, while ϑ is a precondition.

To conclude this subsection, we have provided a richer formalism with the explicit treatment of action and time, where we understand conditional intention as present-directed. In this formalism, we can clearly see with which conditional intention the agent acts, which is exactly what we need to capture participatory intentions. By taking intentions to be present-directed, we have parted ways with the seminal Cohen and Levesque's work. We can potentially develop a similar formalism for future-directed intentions: we would need to drop the constraint (2b); in the formulation of control constraint ((2f) of Def. 4.1.14), we would need to replace the existence of an atomic action with the existence of a long-term *strategy*, as well as add new constraints to account for persistence of conditional intentions. We leave this investigation for future work.

Side-effects problem

The logic **L** suffers from the problem of *logical omniscience*: conditional intention is closed under necessary entailment.

$$Int^{\varphi}\psi \wedge \Box(\psi \to \vartheta) \models Int^{\varphi}\vartheta \tag{4.15}$$

Hence, our formalism cannot handle the problem of *side-effects*. By a side effect, we mean an outcome of the intended action, which the agent does not intend to achieve. As philosophical [Bra87; Fin88], psychological and experimental [Kno06; Kno03] investigations show, agents do not intend all the consequences of their intended actions. They are not rationally pressured to do so, while the difference between intended and unintended consequences of intentional actions is crucial, especially in normative reasoning.

In general, an agent can fail to intend some a priori consequence of her intention for two reasons [KÖ25]:

1. She lacks cognitive resources to grasp the consequence or compute the proof that the consequence indeed follows from what she intends. For instance, one may intend to write down Peano axioms without intending to write down an axiomatic system in which the theorem of prime numbers is provable, given that one does not know how to derive the theorem from Peano axioms.

2. The consequence in question is irrelevant to the agent's practical matters. For instance, while on a trip to France, one might intend to buy a bottle of wine in a legally operating store to avoid buying uncertified products. Buying goods from a legally operating store in France practically entails paying some VAT in the budget of France. Even knowing that entailment, one still might not intend to pay some money to the French budget, simply because it is completely irrelevant to her practical matters.

We can abstract away from the first kind of failures of the closure under necessary entailment by assuming that the agent whose intentions we model is a perfect logician: she is unbounded in her cognitive capacities when it comes to logical reasoning. The second kind of closure's failures is more interesting. We can demonstrate how our logic suffers from the inability to model them by reformulating the famous gentle murder paradox [For84] in terms of intention. The following statements seem to be consistent with one another:

I intend not to kill anyone. (S1)

It is not true that I intend to kill in case I am killing. (S2)

In case I am killing someone, I intend to do it gently. (S3)

If I kill someone gently, it necessarily follows that I kill. (S4)

If we formalise (S1)-(S4) in our logic, we will get the next set of formulae:

$$\{Int\neg kill, \neg Int^{kill}kill, Int^{kill}gentle, \Box(gentle \to kill)\}$$

By (4.15), $Int^{kill}gentle, \Box(gentle \to kill) \models Int^{kill}kill$, hence,

$$\{Int\neg kill, \neg Int^{kill}kill, Int^{kill}gentle, \Box(gentle \to kill)\} \models \neg Int^{kill}kill \wedge Int^{kill}kill$$

Therefore, in our logic we cannot handle the gentle murder paradox: we cannot model situations where one is unintentionally committing a murder, while intentionally doing it as gently as possible.

Since the problem of side effects is not the primary concern of this section, we have abstracted away from it to keep our framework simple. We can alleviate the problem by either implementing some raw syntactical restrictions of the closure, using machinery of awareness functions [FH87] as it was done in [RG97; LHM96], or taking a more subtle solution, using a question-sensitive framework [BG23; KÖ25].[19] To keep things simple, we show a toy solution in the spirit of [FH87]. Namely, we do not explicitly model why some consequences of one's intentions are unintended and simply restrict the closure via the analogue of awareness functions.

First, we expand our language with *explicit intention* modality:

$$\mathcal{L}_I \ni \varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid Int^{\varphi}\psi \mid \mathcal{I}^{\varphi}\psi$$

where $Int^{\varphi}\psi$ reads as "*ψ is entailed by what the agent intends in case of φ*" (from now on, will be abbreviated to "*agent implicitly intends that ψ in case of φ*"), $\mathcal{I}^{\varphi}\psi$ – "*agent explicitly intends that ψ in case of φ*". Explicit unconditional intentions are expressible via conditional ones standardly: $\mathcal{I}\varphi := \mathcal{I}^{\top}\varphi$.

As for the semantics, we will extend conditional frames with *availability functions*.

4.1.19. Definition (Hyperintensional conditional models). A class of hyperintensional conditional frames **H** consists of tuples of the form $F = (W, f, a)$, where:

1. (W, f) is a conditional frame as in Definition 4.1.1;
2. $a : W \times 2^W \to 2^{\mathcal{L}_I}$ is a partial availability function[20] that takes a world w, a proposition P and maps it to the set of formulae that are available for the agent in w to intend in case of P. a has the following properties:

[19] The hyperintensional framework of [KÖ25] is presented in detail in Section 5.1.

[20] In [FH87], such functions are called *awareness* functions, and $\varphi \in a(w)$ is supposed to indicate that the agent is aware of φ in w. Here, we purposefully avoid this name, since an agent being aware of the proposition is not enough for that proposition to be available for the agent to intend. See discussions around cases of foreseeing without intending, where agents foresee (and hence are aware of) some consequences of their intentions, but still do not intend them [Bra87; BG23; Gil99].

(a) Definedness: $a(w, P)$ is defined iff $f(w, P)$ is defined;

(b) Restricted monotonicity: given that $a(w, P)$ and $a(w, P \cap Q)$ are defined, if $f(w, P) \cap P \cap Q \neq \emptyset$, then $a(w, P) \subseteq a(w, P \cap Q)$;

(c) Cut: given that $a(w, P)$ and $a(w, P \cap Q)$ are defined, if $f(w, P) \subseteq Q$, $a(w, P) = a(w, P \cap Q)$.

As usual, any frame $F = (W, f, a)$ can be extended to a model $M = (F, V)$, where V is a standard valuation function.

Availability function a specifies what formulae the agent can conditionally intend. It takes a world and a proposition $P \subseteq W$, since the very same formulae may or may not be available, depending on the condition of the intention. For example, it is irrational to unconditionally intend to win the lottery, since it is out of the player's control, but one can intend to win the lottery in case one gets to know which lottery ticket is the winning one. Property 2a means that one can intend something conditional on the proposition $P \subseteq W$ iff P is a condition for one's intention. Properties 2b and 2c ensure that the monotonicity principles we have argued for in the previous sections are valid for explicit intentions.

4.1.20. Definition (Semantic clauses for $\mathcal{L}_I$). For any hyperintensional conditional model $M = (W, f, a, V)$ and any world $w \in W$, $\models$ relation inductively defined as follows (all cases, except of $\mathcal{I}^{\varphi}\psi$, are identical to Def. 4.1.2):

$$M, w \models \mathcal{I}^{\varphi}\psi \Leftrightarrow M, w \models Int^{\varphi}\psi \text{ and } \psi \in a(w, [\![\varphi]\!])$$

The notions of $M \models \varphi$, $F, w \models \varphi$, $F \models \varphi$ and $\mathbf{H} \models \varphi$ are defined standardly.

Using hyperintensional models, we can easily avoid the closure under entailment and hence alleviate the side-effects problem. Going back to the gentle murderer paradox, it is simple to handle. Take a model $M = (W, f, a, V)$, where:

- $W = \{w_1, w_2, w_3\}$, where in w_1 one is not killing anyone, in w_2 one is murdering someone gently and in w_3 one is murdering someone cruelly. Consequently, $V(kill) = \{w_2, w_3\}$, $V(gentle) = \{w_2\}$. Let the real world be w_1.
- $f(w_1, W) = \{w_1\}$, i.e. one intends not to kill anyone no matter what.
- $f(w_1, \{w_2, w_3\}) = \{w_2\}$, i.e. in case one is killing someone, one intends to do it gently.
- $a(w_1, W) = \{\neg kill, kill\}$, $a(w_1, \{w_2, w_3\}) = \{gentle, \neg gentle\}$ i.e. the agent can unconditionally intend either to kill or not to kill, but in case he kills, he can only intend either to do it gently or not: given that $f(w_1, W) \subseteq W \setminus \{w_2, w_3\}$, the intention in case he kills is a plan-B intention, hence, the

agent finds himself in a situation of unintentional killing, and committing a murder is not his choice and cannot be intended.

Clearly, $M, w_1 \models \mathcal{I}\,\neg kill \wedge \mathcal{I}^{kill} gentle \wedge \neg\mathcal{I}^{kill} kill$, despite the fact that $M, w_1 \models \Box(gentle \to kill)$, since $kill \notin a(w_1, [\![kill]\!])$. Interestingly, if we changed the story and imagined that the agent in question unconditionally intends to kill, then he would intend to kill in case he kills as well: $\mathcal{I}\,kill, [C]kill \models \mathcal{I}^{kill} kill$. This would happen due to the properties of availability functions we have established. This is desirable, since then the plan for the case he commits a murder is not a plan-B for him, but the very same plan as the unconditional one, hence, he should explicitly intend the same things.

Since we are abstracting away from the exact reasons why some formulae cannot be explicitly intended by the agent, we do not impose any further conditions on availability functions. Here we list what conditions should be added to satisfy apparently plausible closure principles. The closure principles are on the right, while the conditions of availability functions are on the left.[21] In what follows, $\circ \in \{\wedge, \vee\}$:

$\psi_1 \circ \psi_2 \in a(w, [\![\varphi]\!]) \Leftrightarrow \psi_2 \circ \psi_1 \in a(w, [\![\varphi]\!])$	iff	$\mathcal{I}^{\varphi}(\psi_1 \circ \psi_2) \leftrightarrow \mathcal{I}^{\varphi}(\psi_2 \circ \psi_1)$
$\psi_1 \wedge \psi_2 \in a(w, [\![\varphi]\!]) \Leftrightarrow \psi_1, \psi_2 \in a(w, [\![\varphi]\!])$	iff	$\mathcal{I}^{\varphi}(\psi_1 \wedge \psi_2) \leftrightarrow (\mathcal{I}^{\varphi}\psi_1 \wedge \mathcal{I}^{\varphi}\psi_2)$
$\psi_1, \psi_2 \in a(w, [\![\varphi]\!]) \Rightarrow \psi_1 \vee \psi_2 \in a(w, [\![\varphi]\!])$	iff	$(\mathcal{I}^{\varphi}\psi_1 \wedge \mathcal{I}^{\varphi}\psi_2) \to \mathcal{I}^{\varphi}(\psi_1 \vee \psi_2)$
$\neg\neg\psi \in a(w, [\![\varphi]\!]) \Leftrightarrow \psi \in a(w, [\![\varphi]\!])$	iff	$\mathcal{I}^{\varphi}\neg\neg\psi \leftrightarrow \mathcal{I}^{\varphi}\psi$
If $\models \psi$, then $\psi \notin a(w, [\![\varphi]\!])$	iff	If $\vdash \psi$, then $\vdash \neg\mathcal{I}^{\varphi}\psi$

Using availability functions, we can also adjust our logic for an opponent of propositionalism about intention. Namely, assume we have a set of propositions $Var = \{p_1, p_2, \ldots\}$ with a distinguished subset of *agentives* $Ag \subseteq Var$, i.e. propositions about the agent's actions. Restricting the range of a function to modal-free formulae built from agentives, we can avoid situations where the object of the agent's intention is a proposition about anything but her own actions.

If we do not impose any of those additional properties on a, the logic $\mathbf{L}_I = \{\varphi \in \mathcal{L}_I \mid \mathbf{H} \models \varphi\}$ of hyperintensional conditional frames $\mathbf{H}$ can be axiomatised as follows: see Table 4.2.

The proof of the fact that $\mathbf{L}_I$ is sound and complete w.r.t. $\mathbf{H}$ strongly resembles the proof of Theorem 4.1.13 with the slight modification of canonical model

[21] We emphasise the meaning of the last condition on the availability functions. Namely, it prevents the agent from explicitly intending logical truths. This condition was advocated, among others, by Cohen and Levesque, who viewed intentions as a certain kind of *achievement goals* that are believed to be currently false [CL90a, Def.3.24]. A rational agent should not perceive logical truths as false and aim at making them true.

(**L**)	Axioms and rules of **L**
(Inc)	$\mathcal{I}^{\varphi}\psi \to Int^{\varphi}\psi$
(Ex-Cut)	$([C]\varphi \wedge [C](\varphi \wedge \psi)) \to (Int^{\varphi}\psi \to (\mathcal{I}^{\varphi}\theta \leftrightarrow \mathcal{I}^{(\varphi\wedge\psi)}\theta))$
(R-ExMon)	$([C]\alpha \wedge [C](\alpha \wedge \beta)) \to (\neg Int^{\alpha}(\alpha \to \neg\beta) \to (\mathcal{I}^{\alpha}\varphi \to \mathcal{I}^{(\alpha\wedge\beta)}\varphi))$
(RE)	If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \mathcal{I}^{\varphi}\theta \leftrightarrow \mathcal{I}^{\psi}\theta$

Table 4.2: Logic $\mathbf{L}_I$

construction. From that point, by following the steps of completeness proof of **L** w.r.t. **C**, one may get the analogous result for $\mathbf{L}_I$ and **H**.

We still want to construct a finite canonical model for every formula ϕ, but we cannot use the same strategy with restricted languages anymore, since $\mathcal{I}$ modality does not validate the rule of equivalence. From $\vdash \varphi \leftrightarrow \psi$ it does not follow that $\vdash \mathcal{I}^{\theta}\varphi \leftrightarrow \mathcal{I}^{\theta}\psi$, therefore, if we construct the restricted language for ϕ the same way we did it in Section 4.1.4, we will have infinite mcs in that language. Hence, we will redefine the restricted language to avoid it.

4.1.21. Definition (Restricted sub-languages of $\mathcal{L}_I$). Given any formula $\phi \in \mathcal{L}_I$, let $sub(\phi)$ denote the set of subformulae of ϕ. Then, the set of formulae Σ_I^{ϕ} is defined as follows:

$$\Sigma_I^{\phi} \ni \psi ::= \pi \mid \neg\psi \mid (\psi \vee \psi) \mid \Box\psi \mid Int^{\psi}\psi \mid \mathcal{I}^{\psi}\pi$$

where $\pi \in sub(\phi)$. Note that for any $\psi \in \Sigma_I^{\phi}$, $Var(\psi) \subseteq Var(\phi)$ and for any $\psi \in \Sigma_I^{\phi}$ there are finitely many $\mathcal{I}^{\psi}$-formulae. Then, we define a restricted language $\mathcal{L}_I^{\phi}$ as follows:

$$\mathcal{L}_I^{\phi} = \{\psi \in \Sigma_I^{\phi} \mid md(\psi) \leq md(\phi)\}$$

where md function is defined analogously to Def. 4.1.7, i.e. all the induction steps for Boolean connectives, $\Box$ and Int are the same and

$$md(\mathcal{I}^{\varphi_1}\varphi_2) = 1 + max(md(\varphi_1), md(\varphi_2))$$

4.1.22. Definition (Canonical model for $\mathbf{L}_I$). Let $MCS_{\phi} \subseteq 2^{\mathcal{L}_I^{\phi}}$ be a collection of mcs of the restricted language $\mathcal{L}_I^{\phi}$. Then, given any $\Gamma \subseteq MCS$, a canonical model is a tuple $M_{\Gamma}^c = (W_{\Gamma}^c, f_{\Gamma}^c, a_{\Gamma}^c, V_{\Gamma}^c)$, where:

- $(W_{\Gamma}^c, f_{\Gamma}^c, V_{\Gamma}^c)$ are defined just like in Def. 4.1.9;
- $a_{\Gamma}^c(\Delta, P) = \begin{cases} \{\psi \in \mathcal{L} \mid \mathcal{I}^{\varphi}\psi \in \Delta\} & \text{if } P = \widehat{\varphi}, [C]\varphi \in \Delta \\ \text{undefined} & \text{otherwise} \end{cases}$

4.1.23. LEMMA (Truth lemma for M^c). *Given any restricted language $\mathcal{L}_I^\phi$ and a canonical model $M_\Gamma^c = (W_\Gamma^c, f_\Gamma^c, a_\Gamma^c, V_\Gamma^c)$, where $\Gamma \in MCS_\phi$, any mcs $\Delta \in W_\Gamma$ and any formula $\varphi \in \mathcal{L}_I^\phi$, $M_\Gamma^c, \Delta \models \varphi$ iff $\varphi \in \Delta$.*

Proof:
By induction on the complexity of φ. All cases, except of $\mathcal{I}$-formulae, were proven in Lemma 4.1.12. Assume that $\varphi := \mathcal{I}^{\psi_1}\psi_2$. Left-to-right. $M_\Gamma^c, \Delta \models \mathcal{I}^{\psi_1}\psi_2$. I.e. $f_\Gamma^c(\Delta, [\![\psi_1]\!]) \subseteq [\![\psi_2]\!]$ and $\psi_2 \in a_\Gamma^c(\Delta, [\![\psi_1]\!])$. By induction hypothesis, $f_\Gamma^c(\Delta, \widehat{\psi_1}) \subseteq \widehat{\psi_2}$ and $\psi_2 \in a_\Gamma^c(\Delta, \widehat{\psi_1})$. Then, there exists some formula θ, such that $\widehat{\psi_1} = \widehat{\theta}$ (i.e. $\Box\Gamma \vdash \psi_1 \leftrightarrow \theta$), and $[C]\theta, \mathcal{I}^\theta\psi_2 \in \Delta$. By (RE) it follows that $\mathcal{I}^{\psi_1}\psi_2 \in \Delta$. Right-to-left: Assume that $\mathcal{I}^{\psi_1}\psi_2 \in \Delta$. Then, by definition of a_Γ^c, $\psi_2 \in a_\Gamma^c(\Delta, \widehat{\psi_1})$, i.e., by induction hypothesis, $\psi_2 \in a_\Gamma^c(\Delta, [\![\psi_1]\!])$. Since $\mathcal{I}^{\psi_1}\psi_2 \in \Delta$, by (Inc), $Int^{\psi_1}\psi_2 \in \Delta$, from which by Lemma 4.1.12 it follows that $f_\Gamma^c(\Delta, [\![\psi_1]\!]) \subseteq [\![\psi_2]\!]$. Taking it all together, $M_\Gamma^c, \Delta \models \mathcal{I}^{\psi_1}\psi_2$. □

4.1.24. LEMMA. *Take any formula $\phi \in \mathcal{L}_I$. Let $M_\Gamma^c = (W_\Gamma^c, f_\Gamma^c, a_\Gamma^c, V_\Gamma^c)$ be a canonical model for some $\Gamma \in MCS_{\mathcal{L}_I^\phi}$. Then, M_Γ^c is a* **H***-model.*

Proof:

From Lemma 4.1.10 we know that $(W_\Gamma^c, f_\Gamma^c, V_\Gamma^c)$ is a **C**-model. Hence, it is left for us to prove that $a_\Gamma^c : W_\Gamma \times 2^{W_\Gamma} \to 2^{\mathcal{L}_I^\phi}$ is a partial availability function with the listed properties in Def. 4.1.19. It is a routine to check that a_Γ is a well-defined partial function of the corresponding type. As for properties:

1. $a_\Gamma^c(\Delta, P)$ is defined iff $f_\Gamma^c(\Delta, P)$ is defined. Follows directly from definitions of f_Γ^c and a_Γ^c.

2. Given that $a_\Gamma^c(\Delta, P)$ and $a_\Gamma^c(\Delta, P \cap Q)$ are defined, if $f_\Gamma^c(\Delta, P) \cap P \cap Q \neq \emptyset$, then $a_\Gamma^c(\Delta, P) \subseteq a_\Gamma^c(\Delta, P \cap Q)$. Assume that $a_\Gamma^c(\Delta, P)$ and $a_\Gamma^c(\Delta, P \cap Q)$ are defined and $f_\Gamma^c(\Delta, P) \cap P \cap Q \neq \emptyset$. Then, by definition of a_Γ^c, $P = \widehat{\varphi}$, $P \cap Q = \widehat{\varphi \wedge \psi}$ for some formulae $\varphi, \psi \in \mathcal{L}_I^\varphi$, such that $[C]\phi \in \Delta$ and $[C](\phi \wedge \psi) \in \Delta$. Then, $f_\Gamma^c(\Delta, P) \cap P \cap Q \neq \emptyset$ means that $f_\Gamma^c(\Delta, [\![\varphi]\!]) \cap [\![\varphi \wedge \psi]\!] \neq \emptyset$, from which it follows that $M_\Gamma^c, \Delta \models [C]\varphi \wedge [C](\varphi \wedge \psi) \wedge \neg Int^\varphi(\varphi \to \neg(\varphi \wedge \neg\psi))$. From Lemma 4.1.23 it follows that $[C]\varphi \wedge [C](\varphi \wedge \psi) \wedge \neg Int^\varphi(\varphi \to \neg(\varphi \wedge \neg\psi)) \in \Delta$. Then, by (R-ExMon) axiom, $\mathcal{I}^\varphi\vartheta \to \mathcal{I}^{(\varphi\wedge\psi)}\vartheta \in \Delta$ for any ϑ, from which it directly follows that $a_\Gamma^c(\Delta, \widehat{\varphi}) \subseteq a_\Gamma^c(\Delta, \widehat{\varphi \wedge \psi})$.

3. Given that $a_\Gamma^c(\Delta, \widehat{\varphi})$ and $a_\Gamma^c(\Delta, \widehat{\varphi \wedge \psi})$ are defined, if $f_\Gamma^c(\Delta, \widehat{\varphi}) \subseteq \widehat{\psi}$, then $a_\Gamma^c(\Delta, \widehat{\varphi}) = a_\Gamma^c(\Delta, \widehat{\varphi})$. Assume that $a_\Gamma^c(\Delta, \widehat{\varphi})$ and $a_\Gamma^c(\Delta, \widehat{\varphi \wedge \psi})$ are defined and $f_\Gamma^c(\Delta, \widehat{\varphi}) \subseteq \widehat{\psi}$. Then, by Definition 4.1.22, $M_\Gamma^c, \Delta \models [C]\varphi \wedge [C](\varphi \wedge \psi) \wedge Int^\varphi\psi$. From Lemma 4.1.12 it follows that $[C]\varphi \wedge [C](\varphi \wedge \psi) \wedge Int^\varphi\psi \in \Delta$. Then, by (Ex-Cut) axiom, $\mathcal{I}^\varphi\theta \leftrightarrow \mathcal{I}^{(\varphi\wedge\psi)}\theta$ for any $\theta \in \mathcal{L}_I^\phi$, from which it

directly follows that $a^c_\Gamma(\Delta, \widehat{\varphi}) = a^c_\Gamma(\Delta, \widehat{\varphi \wedge \psi})$.

□

4.1.25. THEOREM. *Given any $\varphi \in \mathcal{L}_I$, the following holds:*

$$\mathbf{H} \models \varphi \Leftrightarrow \mathbf{L}_I \vdash \varphi$$

Proof:
($\Leftarrow$) By a routine check of the validity of axioms and the validity preservation of rules of inference.

($\Rightarrow$) By contraposition. Assume that $\mathbf{L}_I \nvdash \varphi$. Then, $\{\neg\varphi\}$ is $\mathbf{L}_I$-consistent. Take a language $\mathcal{L}_I^{\neg\varphi}$ and take any mcs $\Gamma \ni \neg\varphi$. Construct a canonical model $M^c_\Gamma = (W^c_\Gamma, f^c_\Gamma, a^c_\Gamma, V^c_\Gamma)$. Then, by Lemma 4.1.23, $M^c_\Gamma, \Gamma \models \neg\varphi$, while M^c_Γ is a $\mathbf{H}$-model, as we know from Lemma 4.1.24. Hence, $\mathbf{H} \not\models \varphi$. □

4.2 Simpler multi-agent framework

In the previous section, we developed a general theory of conditional intention. For our purposes, we can significantly simplify it. We only need individual agents' intentions conditionalised on the actions of others. So there is no need to deal with the plan-B intentions and related monotonicity issues anymore. Consequently, the kind of conditional ability we need to account for the control constraint is also somewhat simpler, so that we will only need (an extension of) Coalition Logic instead of STIT logic.

Generally, we will keep working with game models. We heavily rely on [GJ22], where the logic ConStR is presented. ConStR extends $\mathcal{L}_{CL}$ with three different kinds of conditional modalities.[22] Namely, given two arbitrary sets of agents $A, B \subseteq Ags$:

- $[A]_c(\varphi; \langle B\rangle\psi)$ reads as *A has a joint action σ_A which, when applied, guarantees the truth of φ and enables B to apply a joint action σ_B that is consistent with σ_A and guarantees ψ when additionally applied by B.*
- $[A]_\alpha(\varphi; \langle B\rangle\psi)$ reads as *the coalition $B \setminus A$ has an action $\sigma_{B\setminus A}$ such that if A applies any action that guarantees the truth of φ, then $B \setminus A$ can guarantee the truth of ψ by applying additionally the action $\sigma_{B\setminus A}$*. Informally, $[A]_\alpha(\varphi; \langle B\rangle\psi)$ means that $B \setminus A$ has a *proactive* ability to ensure ψ conditional on A acting towards φ.

[22] Authors denote both the logic and the formal language as ConStR, so we follow their notation.

- $[A]_\beta(\varphi; \langle B\rangle\psi)$ read as *for any joint action σ_A of A that guarantees the truth of φ, when applied by A, there is an action σ_B that is consistent with σ_A and guarantees ψ when additionally applied by B.*. Informally, $[A]_\beta(\varphi; \langle B\rangle\psi)$ means that $B \setminus A$ has a *reactive* ability to ensure ψ conditional on A acting towards φ.

The first kind of conditional ability is required to satisfy the control constraint in the general case. When i intends that ψ conditional on φ, independently of whether i or someone else has control over φ, i should be able to ensure that ψ if φ, without preventing φ: $[i]_c(\varphi \to \psi; \langle Ags \setminus \{i\}\rangle\varphi)$. For the case when the condition is someone else's action, we can concentrate on proactive abilities. Assume that an agent $i \in Ags$ intends that φ, conditional on the other agent, $j \in Ags$, acting towards ψ. Then, i contingently commits herself to an action that guarantees φ in case j acts towards ψ. The commitment is contingent on i making sure that j acts towards ψ. Hence, it should suffice for i to get to know that j acts towards ψ, without knowing the j's exact strategy by which she ensures ψ, to satisfy her conditional intention. So i should be able to ensure φ conditional on j ensuring ψ without knowing j's strategy to do so, which is a *pro*active rather than *re*active ability.

For instance, assume i plays the rock-paper-scissors game against j and intends to win, conditional on j choosing either the rock or the scissors. i has a reactive ability to do so. For every action of j that is consistent with the condition, i has an action to guarantee the win: i can reply with paper for j's rock and with rock – to j's scissors. Assume that i is assured that j is going to choose either rock or scissors. Then, the condition of i's contingent commitment is satisfied, and i should choose an action to satisfy her conditional intention. Which one should i choose? There is still no action to guarantee the winning, i needs to know j's exact move to choose accordingly. On the contrary, assume that they play a rock-paper-scissors-well: the game is exactly like the rock-paper-scissors, but each player has an additional move, to choose a well. The well wins against the rock and the scissors (since both can fall into it), but loses against the paper (since you can cover the well with the paper sheet). In such a scenario, i can rationally intend to win, conditional on j choosing either the rock or the well. i has a proactive ability to win in such a case: if i is sure that j is choosing the rock or the well, i can reply with the paper, and no matter which of the two figures j chooses, i will win.

In what follows, we combine the coalition logic with proactive abilities with a simplified multi-agent version of the formalism for conditional intentions we have developed in Section 4.1.5.

Having a finite set of agents $Ags = \{1, \ldots, n\}$ and a countable set of propositional variables $Var = \{p_1, p_2, \ldots\}$ fixed, we recursively define a language of conditional intentions and abilities $\mathcal{L}^I_{CL_\alpha}$ as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid [A](\varphi \mid_B \varphi) \mid Int_i(\varphi \mid_B \varphi)$$

where $i \in Ags$. $A, B \subseteq Ags$. $[A](\varphi \mid_B \psi)$ reads as "*A has a proactive ability to ensure φ conditional on B ensuring ψ*. $Int_i(\varphi \mid_B \psi)$ reads as "*i intends that φ conditional on B ensuring ψ*". We abbreviate unconditional intention of i as follows: $Int_i\varphi := Int_i(\varphi \mid_{Ags\setminus\{i\}} \top)$. Analogously, unconditional ability of C to ensure that φ is abbreviated as $[C]\varphi := [C](\varphi \mid_{Ags\setminus C} \top)$. $[\emptyset]\varphi$ stands for "*φ is true in all possible outcomes of the current game*". $\langle\emptyset\rangle\varphi$ – "*there is an outcome of the current game, in which φ is true*".

We have conditional intention operators only for individual agents, i.e., $Int_A(\varphi|_B\gamma)$ is not a well-formed formula if $Card(A) \neq 1$. Following the conservativity of shared intentions over individual ones, we do not need group intentions in our language, since we should be able to express them via the individual intentions of group members.

Semantically, we extend game models with a set-selection function f that has all the good properties we have argued for in Section 4.1.

4.2.1. Definition (Intentional game models). An *intentional game model* is a tuple of the form:

$$M = (S, \{\Sigma_i\}_{i\in Ags}, g, f, V)$$

1. $(S, \{\Sigma_i\}_{i\in Ags}, g, V)$ is a game model;
2. f is a function that takes an agent $i \in Ags$, a possible world $w \in W$ and returns a partial function $f^i_w : \mathcal{P}(\Sigma^w_{\bar{i}}) \to \mathcal{P}(\Sigma^w_i)$, such that:
   (a) If $f^i_w(P)$ is defined, then $f^i_w(P) \neq \emptyset$;
   (b) If $f^i_w(P)$ is defined, then $P \neq \emptyset$;
   (c) If $f^i_w(P)$ and $f^i_w(Q)$ are defined and $P \cap Q \neq \emptyset$, then $f^i_w(P \cap Q)$ is defined;
   (d) If $f^i_w(P)$ and $f^i_w(P \cap Q)$ are defined, $f^i_w(P \cap Q) \subseteq f^i_w(P)$;

f^i_w function takes a condition – a set of $\bar{i} = Ags \setminus \{i\}$'s joint actions – and returns a set of actions to which i restricts her choice in case $\bar{i}$ does any of the joint actions in the condition. In other words, the conditional intention is modelled as a commitment to a partial choice of action, conditional on the way others are going to behave. The unconditional intention of the agent i then is represented as $f^i_w(\Sigma^w_{\bar{i}})$, i.e., actions that i commits to no matter what everyone else does. We will abbreviate $f^i_w(\Sigma^w_{\bar{i}})$ simply as $f^i_w(\top)$. Condition (2a) expresses that the agent's intentions are consistent with at least one of her available actions. (2b) expresses that agents conditionalise their intentions only on consistent conditions. (2c) expresses that conditions are closed under non-empty intersections. (2d) expresses

monotonicity. Here, we model only intentions that are conditionalised on what *others* do. We abstract away from agents' plan-B intentions, since participatory intentions, as we have argued, are either unconditional or conditional on actions of others. So we can safely impose (2d) in such a strong form, without running into counterexamples we have discussed in Section 4.1.

For every state $w \in S$, set of states $X \subseteq S$ and coalition $C \subseteq Ags$, let $\|X\|^C_w$ be a set of C's joint actions that ensure that the outcome of the game played in w will be in X:

$$\|X\|^C_w = \{\sigma_C \in \Sigma^w_C \mid out(w, \sigma_C) \subseteq X\}$$

For any sets $X, Y \subseteq S$ and coalitions $A, B \subseteq Ags$, let $\|X \mid_B Y\|^A_w$ be a set of A's actions to guarantee X, conditional on B ensuring Y at w:

$$\|X \mid_B Y\|^A_w = \{\sigma_A \in \Sigma^w_A \mid \forall \sigma_B \in \|Y\|^B_w : \sigma_B \cup \sigma_A \in \|X\|^{A\cup B}_w\}$$

For every $A \subseteq B \subseteq Ags$, let $\|X\|^{B|A}_w = \{\sigma_B \in \Sigma^w_B \mid \sigma_B|_A \in \|X\|^A_w\}$. In other words, $\|X\|^{B|A}_w$ is a set of all B's joint actions in which A guarantees X, without any restrictions on $B \setminus A$'s joint actions. Alternatively, for every $\sigma_A \in \|X\|^A_w$ and every $\sigma_{B\setminus A} \in \Sigma^w_{B\setminus A}$, $\sigma_A \cup \sigma_{B\setminus A} \in \|X\|^{B|A}_w$.

4.2.2. Definition (Semantic clauses for $\mathcal{L}^I_{CL_\alpha}$). Given any formula $\varphi \in \mathcal{L}^I_{CL_\alpha}$, any intentional game model $M = (S, \{\Sigma_i\}_{i\in Ags}, g, f, V)$ and any state $w \in S$, we recursively define $\models$ relation as follows:

$$\begin{array}{lcl}
M, w \models [A](\varphi \mid_B \psi) & \Leftrightarrow & \|\varphi \mid_B \psi\|^A_w \neq \emptyset \\
M, w \models Int_i(\varphi \mid_B \psi) & \Leftrightarrow & f^i_w(\|\psi\|^{Ags\setminus\{i\}|B}_w) \text{ is defined and} \\
 & & f^i_w(\|\psi\|^{Ags\setminus\{i\}|B}_w) \subseteq \|\varphi \mid_B \psi\|^i_w
\end{array}$$

where:

$$\begin{array}{lcl}
\|\varphi\|^A_w & = & \|\,[\![\varphi]\!]_M\|^A_w = \{\sigma_A \in \Sigma^w_A \mid out(w, \sigma_A) \subseteq [\![\varphi]\!]_M\} \\
\|\varphi \mid_B \psi\|^A_w & = & \{\sigma_A \in \Sigma^w_A \mid \forall \sigma_B \in \|\psi\|^B_w : \sigma_B \cup \sigma_A \in \|\varphi\|^{A\cup B}_w\} \\
[\![\varphi]\!]_M & = & \{s \in S \mid M, s \models \varphi\}
\end{array}$$

And the rest of the cases are standard.

According to Def. 4.2.2, i intends that φ conditional on B seeing to it that ψ iff i contingently commits herself to actions that will lead to φ in case B sees to it that ψ, conditional on B actually seeing to it that ψ.

To clarify the machinery of the formalism, let us formalise a toy example of a protest cascade from Section 2.1.

4.2.3. Example (Protest cascade). There are three agents: Anne, Boris, and Chloe. All of them share the same political goals but differ in their beliefs about the most effective means of achieving them. Anne believes that solo protests make sense, so she is ready to join a protest no matter what. Boris believes that only protests of at least two participants make sense. Chloe is even more sceptical than Boris: she believes that a protest is pointless unless at least three people participate in it. Everyone is also aware of others' views. So Anne unconditionally intends to go on a protest to ensure that there is at least someone protesting. Boris intends that there will be at least two protesters, in case Anne joins the protest. Chloe intends that there will be at least three protesters, conditional on Boris and Anne joining the protest.

To rationally hold such intentions, agents should have the corresponding abilities. Anne should be able to go on a protest no matter what; Boris should be able to ensure that there are at least two protesters in case Anne is protesting; Chloe should be able to ensure that there are at least three protesters, conditional on Anne and Boris joining the protest. As the story goes, it seems to be the case.

Let $Ags = \{a, b, c\}$, standing for Anne, Boris, and Chloe, respectively. Let $Var = \{p_n \mid n \in \mathbb{N}\}$, where p_n denotes the proposition "*there are at least n-many protesters participating*".

We start with a pointed game model $M = (S, \{\Sigma_i\}_{i \in Ags}, g, f, V)$. Let $s_0 \in S$ be a state, where Anne, Boris, and Chloe deliberate whether to go on the protest. For any $i \in Ags$, let $\Sigma_i = \Sigma_i^{s_0} = \{1_i, 0_i\}$, where 1_i is to go and 0_i – not to go. Then, the game form at s_0 is as follows:

$$g(s_0) = \{\{s_1, \dots, s_8\}, \Sigma_a, \Sigma_b, \Sigma_c, o^{s_0}\}$$

,

where, in the lexicographical order:

$$\begin{aligned} &o^{s_0}((0_a, 0_b, 0_c)) = s_1 \\ &\dots \\ &o^{s_0}((1_a, 1_b, 1_c)) = s_8 \end{aligned}$$

Consequently, $V(p_0) = S$, $V(p_1) = \{s_2, \dots, s_8\}$, $V(p_2) = \{s_4, s_6, s_7, s_8\}$ and $V(p_3) = \{s_8\}$. See Figure 4.2 for an illustration of $g(s_0)$.

Clearly, each agent has the unconditional ability to ensure that at least 1 protester will be present – she can go to the protest. Formally, $M, s_0 \models \bigwedge_{i \in Ags}[i]p_1$, since $\|p_1\|^i_{s_0} = \{1_i\} \neq \emptyset$. Moreover, each agent can ensure that there will be at least two protesters, conditional on someone else ensuring that there will be at least

	0_b	1_b
0_a	$s_1 : p_0$	$s_3 : p_0, p_1$
1_a	$s_5 : p_0, p_1$	$s_7 : p_0, p_1, p_2$

0_c

	0_b	1_b
0_a	$s_2 : p_0, p_1$	$s_4 : p_0, p_1, p_2$
1_a	$s_6 : p_0, p_1, p_2$	$s_8 : p_0, \ldots, p_3$

1_c

Figure 4.2: Normal game form for $g(s_0)$

someone: $M, s_0 \models \bigwedge_{i,j \in Ags}^{i \neq j} [i](p_2 \mid_j p_1)$, since:

$$\begin{aligned}
&\|p_2 \mid_j p_1\|^i_{s_0} \\
&= \{\sigma_i \in \Sigma_i \mid \forall \sigma_j \in \|p_1\|^j_{s_0} : out(s_0, \sigma_i \cup \sigma_j) \subseteq V(p_2)\} \\
&= \{\sigma_i \in \Sigma_i \mid out(s_0, \sigma_i \cup 1_j) \subseteq V(p_2)\} \\
&= \{1_i\}
\end{aligned}$$

Analogously, everyone can ensure that there are at least three protesters, conditional on everyone else joining:

$$\begin{aligned}
&\|p_3 \mid_{\{j,k\}} p_2\|^i_{s_0} \\
&= \{\sigma_i \in \Sigma_i \mid \forall \sigma_{\{j,k\}} \in \|p_2\|^{\{j,k\}}_{s_0} : out(s_0, \sigma_i \cup \sigma_{\{j,k\}}) \subseteq V(p_3)\} \\
&= \{\sigma_i \in \Sigma_i \mid out\big(s_0, \sigma_i \cup (1_j, 1_k)\big) \subseteq V(p_3)\} \\
&= \{1_i\}
\end{aligned}$$

Then, let us specify agents' conditional intentions. Anne unconditionally commits herself to going: $f^a_{s_0}(\top) = \{1_a\} = \|p_1\|^a_{s_0}$. Boris is committed to go, conditional on Anne going: $f^b_{s_0}(\|p_1\|^{\bar{b}|a}_{s_0}) = \{1_b\} = \|p_2 \mid_a p_1\|^b_{s_0}$. Finally, Chloe is committed to go, conditionally on Boris and Anne going:

$$f^c_{s_0}(\|p_2\|^{\{a,b\}}_{s_0}) = \|p_3 \mid_{\{a,b\}} p_2\|^c_{s_0} = \{1_c\}$$

Then, by Def. 4.2.2: $M, s_0 \models Int_a p_1 \wedge Int_b(p_2 \mid_a p_1) \wedge Int_c(p_3 \mid_{\{a,b\}} p_2)$.

So in our model, every agent's intention satisfies the control constraint: if i intends that φ, conditional on j seeing to it that ψ, then i can ensure that φ, conditional on j ensuring ψ. In general, our formalism imposes this rationality constraint on all agents in all models:

4.2.4. Proposition. *The following is valid in the class of intentional game models:*

$$\models Int_i(\varphi \mid_A \psi) \rightarrow [i](\varphi \mid_A \psi)$$

Proof:
Assume that $M, w \models Int_i(\varphi|_A\psi)$ for some pointed intentional game model (M, w). For contradiction, assume that $M, w \not\models [i](\varphi \mid_A \psi)$. By Def. 4.2.2, $M, w \models Int_i(\varphi|_A\psi)$ iff $f^i_w(\|\psi\|^{\bar{i}|A}_w) \subseteq \|(\varphi|_A\psi)\|^i_w$. Then, by (2a-2b) of Def. 4.2.2, $\|\psi\|^{\bar{i}|A}_w \neq \emptyset$ (and consequently $\|\psi\|^A_w \neq \emptyset$) and $\|\varphi \mid_A \psi\|^i_w \neq \emptyset$. But from $M, w \not\models [i](\varphi \mid_A \psi)$ it follows that $\|\varphi \mid_A \psi\|^i_w = \emptyset$. Contradiction. Hence, $M, w \models [i](\varphi \mid_A \psi)$. □

4.2.1 Formalizing participatory intentions in $\mathcal{L}^I_{CL_\alpha}$

In this subsection, we approximate participatory intentions in our language $\mathcal{L}^I_{CL_\alpha}$. Throughout the section, we make some idealizations about the agents' knowledge and behaviour. First, we assume that agents are *ex-ante omniscient*: at each state, they commonly know which actions are available to them and to others, as well as which outcomes are yielded by every action profile. Agents are uncertain only about their own and others' choice of action. For instance, two players of the rock-paper-scissors game are ex-ante omniscient before they play. They commonly know that each player has three available moves – rock, paper, or scissors – as well as what outcome is yielded by every action profile. The only thing the agents do not know is who is going to choose which figure. Second, we assume that individual agents always act intentionally and never fall under the weakness of will. Namely, if any agent sees to it that φ, then she does so intentionally; if the agent intends that φ, then she will see to it that φ. From the last assumption, it follows that there is no difference between, e.g., the intention of j, conditional on the intention of i that φ, and the intention of j, conditional on i seeing to it that φ.

Having these idealizations in mind, we proceed with formalizing shared intention à la Bratman in $\mathcal{L}^I_{CL_\alpha}$. Since shared intention has several possible realizations, we cannot simply define a translation function that maps the expression "the group G shares the intention that φ" to some well-formed formula in $\mathcal{L}^I_{CL_\alpha}$. In what follows, we show which formulations of the shared intention are plausible and which are not. For intuitive clarity, we will use the following simple two-agent example.

Let $\varphi_i, \varphi_j, \varphi \in \mathcal{L}^I_{CL_\alpha}$ be formulae, such that φ stands for *i and j have met at the library at 2 o'clock*, φ_i – *i went to the library at 2 o'clock*, φ_j – *j went to the library at 2 o'clock*. The fact that $\{i, j\}$'s joint action consists of i and j doing their parts can be expressed via the formula $[\emptyset]((\varphi_i \wedge \varphi_j) \rightarrow \varphi)$ (i.e., in all possible outcomes, if i and j both go to the library at 2 o'clock, then they will meet one another). See Figure 4.3 for the illustration of the corresponding game model.

Let (M, w_0) be a pointed intentional game model depicted in Figure 4.3. First, we start with formulations of participatory intentions that we do not find plausible

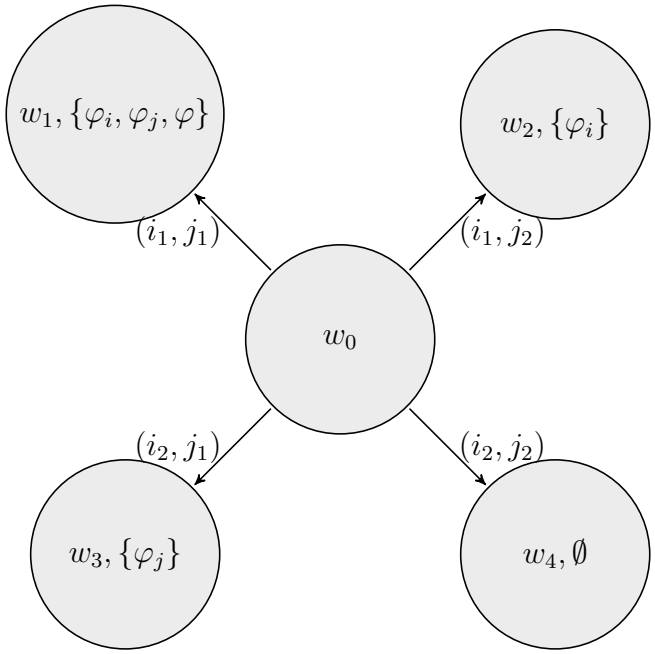

Figure 4.3: The game model $M = (S, \Sigma_i, \Sigma_j, g, V)$. $S = \{w_0, \ldots, w_4\}$. $A_i = \{i_1, i_2\}$, $A_j = \{j_1, j_2\}$. Arrows represent o^{w_0} function. For any other state $s \in S \setminus \{w_0\}$, $o^s(\sigma_{Ags}) = s$ for any $\sigma_{Ags} \in \Sigma_{Ags}$. $[\![\varphi_i]\!] = \{w_1, w_2\}$, $[\![\varphi_j]\!] = \{w_1, w_3\}$, $[\![\varphi]\!] = [\![\varphi_i \wedge \varphi_j]\!]$. f function will be defined later in the text.

and explain why. Agents may simply intend to do their parts: $Int_i\varphi_i \wedge Int_j\varphi_j$. The formula is satisfiable at (M, w_0): if $f^i_{w_0}(\top) = \{i_1\}$ and $f^j_{w_0}(\top) = \{j_1\}$, then $M, w_0 \models Int_i\varphi_i \wedge Int_j\varphi_j$. If i acts towards φ_i and implements i_1, while j acts towards φ_j and implements j_1, then the outcome will be $w_1 = o^{w_0}((i_1, j_1))$, and $M, w_1 \models \varphi$. The problem with such an approach is that φ – i.e., joint action – is absent in the content of both i and j's intentions. To hold such intentions, there is no need for i and j to recognise one another as co-participants of the joint action. For example, i may intend to visit a library and j may intend to visit the library at the same time. If they follow their intentions, they will meet one another at the library, while neither of them thought of the other as a co-participant; maybe both of them would rather avoid this encounter. It might even be commonly known among i and j that they both intend to go to the library at the same time. For instance, i and j might be publicly invited to visit the library at the same time, while it is commonly known among them that they always accept invitations and act accordingly. Both i and j may despise one another and prefer to avoid the encounters, but still both intend to do actions that will result in them stumbling upon one another. Obviously, in such a case, i and j do not share the intention to meet one another in the library: their meeting is rather an undesirable side-effect of what each of them intends by their own.

We can blatantly read Bratman's proposal and presuppose that both i and j simply intend that φ: $Int_i\varphi \wedge Int_j\varphi$. But this cannot be satisfied at (M, w_0), since the control constraint is broken! Clearly, $M, w_0 \not\models [i]\varphi \vee [j]\varphi$: neither i nor j has a strategy to ensure that φ. As we have proven in Prop. 4.2.4, $\models Int_i(\varphi \,|_j\, \gamma) \to [i](\varphi|_j\gamma)$, i.e, by contraposition, $\models \neg[i]\varphi \to \neg Int_i\varphi$. Since $M, w_0 \models \neg[i]\varphi \wedge \neg[j]\varphi$, we can conclude that $M, w_0 \models \neg Int_i\varphi \wedge \neg Int_j\varphi$, whatever the definition of f is. Moreover, neither i nor j can intend that φ conditional on the other ensuring that φ: since $\|\varphi\|^i_{w_0} = \|\varphi\|^j_{w_0} = \emptyset$, by (2b) of Def. 4.2.2, $f_i(\|\varphi\|^j_{w_0})$ and $f_j(\|\varphi\|^i_{w_0})$ are undefined, hence, $M, w_0 \not\models Int_i(\varphi \,|_j\, \varphi) \vee Int_j(\varphi \,|_i\, \varphi)$.

In what follows, we list plausible formulations. We start with a case where both i and j are initiators:

$$Int_i\varphi_i \wedge Int_i(\varphi \,|_j\, \varphi_j) \qquad (\alpha)$$
$$Int_j\varphi_j \wedge Int_j(\varphi \,|_i\, \varphi_i) \qquad (\beta)$$

Informally, α means that i intends to do her part – to go to the library – no matter what, and i also intends that i and j will meet there, conditional on j going to the library. β expresses the mirroring intention for j. First of all, $\alpha \wedge \beta$ is satisfiable: given that $f^i_{w_0}(\top) = f^i_{w_0}(\{j_1\}) = \{i_1\}$ and $f^j_{w_0}(\top) = f^j_{w_0}(\{i_1\}) = \{j_1\}$, $M, w_0 \models \alpha \wedge \beta$. This time, φ is in the content of both i and j's intention. Yet it apparently might not be enough. Namely, i and j conditionalise their intention

that φ only on the other doing her part, maybe not having their joint activity in mind. To clarify our concern and how we deal with it, we can look into the following example:

> *Suppose I have as an end that we dig a tunnel under the English Channel. You are in France and will dig from Calais, and I am in England and will dig from Dover. The tunnels will connect in the middle of the English Channel. I don't care whether you have as an end that we dig the tunnel, or whether you are simply digging a tunnel from Calais to the middle of the English Channel for a bet.* [...] *Your sentiments mirror mine.* [Mil01, p.75]

In this example, both agents intend that they dig the tunnel, conditional on the other digging their half of the tunnel. Neither of them cares if the other does her part to jointly dig the tunnel under the English Channel, or just to dig the tunnel to the middle of the channel alone, for whatever reason. Do such agents share the intention to dig the tunnel together in that case?

If we add the assumption that it is commonly known among i and j that both of them "*have as an end*" that they dig the tunnel under the English Channel, neither of them even considers it possible that the other does not have their joint action as an object of intention. Therefore, under the common knowledge assumption, the question boils down to the preference. "*I don't care whether you have as an end that we dig the tunnel*" cannot be read epistemically, since I know that you do intend that we dig the tunnel. I just would not care if you did your part for whatever reason other than having the participatory intention. As highlighted by Bratman in response to this example, such preferences are irrelevant to the shared intention:

> *After all, I am not on the lookout for ways to ensure that you instead participate because of an alternative intention to win a bet by digging halfway. Interlocking is a relation between our actual intentions, and is compatible with the absence of a preference that favours the intention of the other over other possible intentions that might suffice for achieving the joint activity. So Miller's example need not pose a problem for* [(2) of the definition]. [Bra13, p.52]

To conclude with the case of initiators only, $\alpha \wedge \beta$ is a satisfiable formalisation of the participatory intentions to constitute a shared intention of i and j to φ. We can generalise this formulation beyond the two-agent case:

4.2.5. Definition (Shared intention, the strongest formulation). A set of agents $G \subseteq Ags$, each of whom is an initiator, share the intention that φ iff for some

$\{\psi_i\}_{i\in G} \subseteq \mathcal{L}^I_{CL_\alpha}$:

$$\bigwedge_{i\in G} \big(Int_i\psi_i \wedge Int_i(\varphi \mid_{G\setminus\{i\}} \bigwedge_{j\in G\setminus\{i\}} \psi_j)\big) \wedge [\emptyset]((\bigwedge_{i\in G} \psi_i) \to \varphi)$$

Or, informally, iff every group member unconditionally intends to do her part of some joint action that entails φ and intends that this joint action will be done, conditionally on all the other members doing their parts.

The shared intention as specified in Def. 4.2.5 is satisfiable.

4.2.6. Proposition. *There exists a game model with intentions:*

$$M = (S, \{\Sigma_i\}_{i\in Ags}, g, f, V)$$

and a state $w \in S$, such that, for some $\{\psi_i\}_{i\in Ags}, \varphi \subseteq \mathcal{L}^I_{CL_\alpha}$:

$$M, w \models \bigwedge_{i\in G} \big(Int_i\psi_i \wedge Int_i(\varphi \mid_{G\setminus\{i\}} \bigwedge_{j\in G\setminus\{i\}} \psi_j)\big) \wedge [\emptyset]((\bigwedge_{i\in G} \psi_i) \to \varphi)$$

Proof:
Let $G = \{1, \ldots, k\} \subseteq Ags$. Take any pointed game model with intentions (M, w), such that $\Sigma^w_i = \{\sigma_i, \sigma'_i\}$ for any $i \in Ags$. Let $[\![\varphi]\!]_M = out(w, (\sigma_1, \ldots, \sigma_k))$ and $[\![\psi_i]\!]_M = out(w, \sigma_i)$ for any $i \in \{1, \ldots, k\}$. For any $i \in \{1, \ldots, k\}$, $f^i_w(\top) = \{\sigma_i\}$ and:

$$f^i_w(\| \bigwedge_{j\in G\setminus\{i\}} \psi_j \|^{\bar{i}|G\setminus\{i\}}) = f^i_w(\{\sigma_{\bar{i}} \in \Sigma_{\bar{i}} | \sigma_{\bar{i}}|_{G\setminus\{i\}} = (\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_k)\}) = \{\sigma_i\}$$

Then, $M, w \models \bigwedge_{i\in G} \big(Int_i\psi_i \wedge Int_i(\varphi \mid_{G\setminus\{i\}} \bigwedge_{j\in G\setminus\{i\}} \psi_j)\big) \wedge [\emptyset]((\bigwedge_{i\in G} \psi_i) \to \varphi)$, which can be seen by a tedious but straightforward model-checking. □

This way of sharing intentions, however, is not unique and quite demanding: it requires all agents to intend to do their parts no matter what. E.g., both i and j intend to show up at the library even if the other will not. Such a strong commitment from all agents is not necessary to share the intention. To be more precise, at least some members may have only *restrictive* conditional intentions to do their parts. Although everyone can do her part regardless of what others do, some members may have a reason to do so only if at least someone initiates and does their part no matter what. E.g., i, as an initiator, may intend to go to the library no matter what (even if j will not show up) and intend to meet with j there, conditional on j showing up; then, j may intend to go to the library and meet i there conditionally on i initiating by going there no matter what. Since at

least one of them initiates, this should be sufficient for the successful formation of a shared intention to meet at the library. Formally, this is expressible as follows:

$$Int_i\varphi_i \wedge Int_i(\varphi \mid_j \varphi_j) \qquad (\alpha')$$
$$Int_j(\varphi_j \mid_i \varphi_i) \wedge Int_j(\varphi \mid_i \varphi_i) \qquad (\beta')$$

$\alpha' \wedge \beta'$ is satisfiable: given that $f^j_{w_0}(\{i_1\}) = f^j_{w_0}(\|\varphi_i\|^i_{w_0}) = \{j_1\} = \|\varphi_j \mid_i \varphi_i\|^j_{w_0} = \|\varphi \mid_i \varphi_i\|^j_{w_0}$ and $f^i_{w_0}(\top) = \{i_1\} = \|\varphi_i\|^i_{w_0}$, $f^i_{w_0}(\{j_1\}) = f^i_{w_0}(\|\varphi_j\|^j_{w_0}) = \{i_1\} = \|\varphi \mid_j \varphi_j\|^i_{w_0}$, $M, w_0 \models \alpha' \wedge \beta'$.

Generalizing beyond the two-agent case, there is a variety of ways the group G can implicitly share the intention that φ in weaker forms rather than having everyone as initiators. Namely, for every subset $\emptyset \neq J \subseteq G$, there is at least one way in which only members of J are initiators.

4.2.7. Definition (Shared intention in $\mathcal{L}^I_{CL_\alpha}$, general scheme). Given a set of agents $G = \{1, \ldots, k\} \subseteq Ags$, of which $\{1, \ldots, m\} \subseteq G$ are initiators, G shares the implicit intention that φ iff:

$$\bigwedge_{i\in\{1,\ldots m\}} \big(Int_i\psi_i \wedge Int_i(\varphi \mid_{G\setminus\{i\}} \bigwedge_{j\in G\setminus\{i\}} \psi_j)\big) \wedge$$
$$\bigwedge_{j\in\{m+1,\ldots,k\}} \big(Int_j(\psi_j \mid_{\{1,\ldots,m\}} \psi_1 \wedge \ldots \wedge \psi_m) \wedge Int_j(\varphi \mid_{G\setminus\{j\}} \bigwedge_{l\in G\setminus\{j\}} \psi_l)\big) \wedge$$
$$[\emptyset]((\bigwedge_{i\in G} \psi_i) \rightarrow \varphi)$$

or, informally, iff every member intends that the joint action that entails φ will be done, conditional on others doing their parts of the joint action. The initiators intend to do their parts no matter what, and the rest intend to do their parts conditionally on the initiators.

4.2.8. Proposition. *There exists a game model with intentions*

$$M = (S, \{\Sigma_i\}_{i\in Ags}, g, f, V)$$

and a state $w \in S$, such that, for some $\{\psi_i\}_{i\in G}, \varphi \in Ags$:

$$M, w \models \bigwedge_{i\in\{1,\ldots m\}} \big(Int_i\psi_i \wedge Int_i(\varphi \mid_{G\setminus\{i\}} \bigwedge_{j\in G\setminus\{i\}} \psi_j)\big) \wedge$$
$$\bigwedge_{j\in\{m+1,\ldots,k\}} \big(Int_j(\psi_j \mid_{\{1,\ldots,m\}} \psi_1 \wedge \ldots \psi_m) \wedge Int_j(\varphi \mid_{G\setminus\{j\}} \bigwedge_{l\in G\setminus\{j\}} \psi_l)\big) \wedge$$
$$[\emptyset]((\bigwedge_{i\in G} \psi_i) \rightarrow \varphi)$$

Proof:
Let $G = \{1, \ldots, k\} \subseteq Ags$ and $m \leq k$. Take any pointed game model with intentions (M, w), such that $\Sigma_i^w = \{\sigma_i, \sigma_i'\}$ for any $i \in Ags$. Let $[\![\varphi]\!]_M = out(w, (\sigma_1, \ldots, \sigma_k))$ and $[\![\psi_i]\!]_M = out(w, \sigma_i)$ for any $i \in \{1, \ldots, k\}$. For any $i \in \{1, \ldots, m\}$, $f_w^i(\top) = \{\sigma_i\}$ and for any $i \in \{1, \ldots, k\}$:

$$f_w^i(\| \bigwedge_{j \in G\setminus\{i\}} \psi_j \|^{\bar{i}|G\setminus\{i\}}) = f_w^i(\{\sigma_{\bar{i}} \in \Sigma_{\bar{i}} | \sigma_{\bar{i}}|_{G\setminus\{i\}} = (\sigma_1, \ldots, \sigma_{i-1}, \sigma_{i+1}, \ldots, \sigma_k)\}) = \{\sigma_i\}$$

For any $j \in \{m+1, \ldots, k\}$, $f_w^j(\|\psi_1 \wedge \ldots \wedge \psi_m\|_w^{\bar{i}|\{1,\ldots,m\}}) = f_w^j(\{\sigma_{\bar{i}} \in \Sigma_{\bar{i}} | \sigma_{\bar{i}}|_{\{1,\ldots,m\}} = (\sigma_1, \ldots, \sigma_m)\}) = \{\sigma_j\}$

Then, by a tedious but straightforward model-checking:

$$\begin{aligned} M, w \models & \bigwedge_{i \in \{1,\ldots m\}} \big(Int_i\psi_i \wedge Int_i(\varphi\,|_{G\setminus\{i\}} \bigwedge_{j \in G\setminus\{i\}} \psi_j)\big) \wedge \\ & \bigwedge_{j \in \{m+1,\ldots,k\}} \big(Int_j(\psi_j\,|_{\{1,\ldots,m\}}\, \psi_1 \wedge \ldots \wedge \psi_m) \wedge Int_j(\varphi\,|_{G\setminus\{j\}} \bigwedge_{l \in G\setminus\{j\}} \psi_l)\big) \wedge \\ & [\emptyset]((\bigwedge_{i \in G} \psi_i) \rightarrow \varphi) \end{aligned}$$

□

Note that in both Def. 4.2.5 and 4.2.7, there is no vicious circle: the condition of non-initiators' intentions coincides with the unconditional intentions of initiators.

So far, we have proven that: 1) the accounts of shared intention that we have informally argued against in Section 2.1 are unsatisfiable in our formalism; 2) our own account with initiators is satisfiable: agents who satisfy all the rationality constraints we have argued for in Sec. 4.1 may have the corresponding participatory intentions, without breaking the control constraint or falling into vicious circle of conditionalisation. Moreover, our account is ontologically individualistic: we have not postulated the existence of anything but individuals and their mental states. It is left for us to show that shared intentions are normatively conservative. Namely, given that members' participatory intentions are rational, so is their shared intention. In [Bra13], the following norms guiding intentions are given:

(Cons) Consistency: intentions should be consistent;

(Aggl) Agglomeration: it should be possible to agglomerate one's intentions into a larger intention;

(ME) Means-ends coherence: if one intends that φ and believes ψ to be a necessary means to bring about that φ, then one should intend that ψ;

(Stable) Stability: although intentions can be revised, one should keep one's intention as long as there is no reason to abandon it (i.e., one should keep intending that φ as long as φ is not achieved yet, but it is still possible to achieve).

We abstract away from the stability and means-ends coherence. As for the former, it concerns future-directed intentions and plans, whereas our formalisation so far deals only with agents' present-directed intentions, i.e., what they are to do in the given state. As for the latter, $\mathcal{L}^I_{CL_\alpha}$ contains only expressions about implicit intentions, so any consequence of implicit intention is implicitly intended, whether it is a necessary means or a side-effect, trivially satisfying (**ME**). We will address the problem of side effects for group intention in detail in the following Chapter. As for the rest, we demonstrate that Bratman's thesis about conservativity holds for (**Cons**) and (**Aggl**). If individuals' conditional intentions do not break any rationality constraints, the intentions they share (if any) are consistent, and they agglomerate.

For simplicity, we limit ourselves to the two-agent case, where both agents are initiators. All proofs for more agents and weaker forms of shared intentions are analogous. So we abbreviate a shared intention of $\{i,j\}$ that φ as follows:

$$Int_{i,j}\varphi := Int_i\varphi_i \wedge Int_i(\varphi \,|_j\, \varphi_j) \wedge Int_j\varphi_j \wedge Int_j(\varphi \,|_i\, \varphi_i) \wedge [\emptyset]((\varphi_i \wedge \varphi_j) \to \varphi) \quad \text{for some } \varphi_i, \varphi_j.$$

There are natural formulations of consistency and agglomeration norms in $\mathcal{L}^I_{CL_\alpha}$:

$$Int_{i,j}\varphi \to \langle\emptyset\rangle\varphi \tag{Cons}$$

$$(Int_{i,j}\varphi \wedge Int_{i,j}\psi) \to Int_{i,j}(\varphi \wedge \psi) \tag{Aggl}$$

Moving to the proofs of validity, we show that the principle that is stronger than (Cons) holds for shared intention: namely, the group analogue of the control constraint. If i and j share the intention to φ, not only φ is possible, but i and j have a joint strategy to ensure that φ:

4.2.9. Proposition. *For any $\varphi, \varphi_i, \varphi_j \in \mathcal{L}^I_{CL_\alpha}$, the following holds: $\models Int_{i,j}\varphi \to [\{i,j\}]\varphi$*

Proof:
Assume that $M, w \models Int_{i,j}\varphi$ for some intentional game model M and some state w. In other words:

$$M, w \models Int_i\varphi_i \wedge Int_i(\varphi \,|_j\, \varphi_j) \wedge Int_j\varphi_j \wedge Int_j(\varphi \,|_i\, \varphi_i) \wedge [\emptyset]((\varphi_i \wedge \varphi_j) \to \varphi)$$

Then, from $M, w \models Int_i\varphi_i \wedge Int_j\varphi_j$ it follows that $f^i_w(\top) \subseteq \|\varphi_i\|^i_w$ and $f^j_w(\top) \subseteq \|\varphi_j\|^j_w$. Then, there exist some strategy $\sigma_i \in \Sigma_i$, such that $out(\sigma_i, w) \subseteq [\![\varphi_i]\!]$, and

there exists some strategy $\sigma_j \in \Sigma_j$, such that $out(w, \sigma_j) \subseteq [\![\varphi_j]\!]$. Then, by def. of *out*, $out(w, \sigma_i \cup \sigma_j) \subseteq [\![\varphi_i \wedge \varphi_j]\!]$.

From $M, w \models [\emptyset]((\varphi_i \wedge \varphi_j) \to \varphi)$ it follows that there exists a strategy $\sigma_\emptyset \in \Sigma_\emptyset$, such that $out(w, \sigma_\emptyset) \subseteq [\![(\varphi_i \wedge \varphi_j) \to \varphi]\!]$. Since $\Sigma_\emptyset = \{\sigma_\emptyset\}$ for any game model M, it simply means that for any outcome state $s \in S^w$ of w, $M, s \models (\varphi_i \wedge \varphi_j) \to \varphi$. Take any outcome state $s \in S^w$, such that $s \in out(w, \sigma_i \cup \sigma_j)$. As we have shown before, $M, s \models \varphi_i \wedge \varphi_j$. Then, since s is an outcome state for w, $M, s \models (\varphi_i \wedge \varphi_j) \to \varphi$. Therefore, $M, s \models \varphi$. Therefore, there exists a state $s \in out(w, \sigma_\emptyset)$, such that $M, s \models \varphi$, hence, $M, w \models \langle\emptyset\rangle\varphi$.

Since s was chosen arbitrary, for any $s \in out(w, \sigma_i \cup \sigma_j)$, $M, s \models \varphi$. Then, $out(w, \sigma_i \cup \sigma_j) \subseteq [\![\varphi]\!]$, i.e., $(\sigma_i \cup \sigma_j) \in \|\varphi\|_w^{\{i,j\}}$. Hence, $\|\varphi\|_w^{\{i,j\}} \neq \emptyset$, from which it follows that $M, w \models [\{i, j\}]\varphi$. □

Before moving to the proof of agglomeration, we demonstrate a lemma that is interesting in its own right. Given that we do not deal with plan-B intentions and the associated restrictions on monotonicity, we can agglomerate agents' intentions with consistent conditions.

4.2.10. Lemma (Agglomeration of conditional intentions). *Given any two-agent game-model with intentions $M = (S, \Sigma_i, \Sigma_j, g, f, V)$ and any state $w \in S$, the following holds:*

$$M, w \models (Int_i(\varphi \mid_j \varphi_j) \wedge Int_i(\psi \mid_j \psi_j)) \to \big([j](\varphi_j \wedge \psi_j) \to Int_i(\varphi \wedge \psi \mid_j \varphi_j \wedge \psi_j)$$

Proof:
Assume $M, w \models Int_i(\varphi \mid_j \varphi_j) \wedge Int_i(\psi \mid_j \psi_j) \wedge [j](\varphi_j \wedge \psi_j)$. Then, $f_w^i(\|\varphi_j\|_w^j) \subseteq \|\varphi|_j \varphi_j\|_w^i$ and $f_w^i(\|\psi_j\|_w^j) \subseteq \|\psi|_j \psi_j\|_w^i$. Since $M, w \models [j](\varphi_j \wedge \psi_j)$, $\|\varphi_j \wedge \psi_j\|_w^j \neq \emptyset$. Obviously, for any strategy $\sigma_j \in \Sigma_j^w$, $\sigma_j \in \|\varphi_j\|_w^j \cap \|\psi_j\|_w^j$ iff $\sigma_j \in \|\varphi \wedge \psi\|_w^j$. Hence, $\|\varphi_j\|_w^j \cap \|\psi_j\|_w^j \neq \emptyset$. Then, by (2c) of Def. 4.2.1, $f_w^i(\|\varphi_j\|_w^j \cap \|\psi_j\|_w^j) = f_w^i(\|\varphi_j \wedge \psi_j\|_w^j)$ is defined. Then, by (2d), $f_w^i(\|\varphi_j \wedge \psi_j\|_w^j) \subseteq f_w^i(\|\varphi_j\|_w^j)$ and $f_w^i(\|\varphi_j \wedge \psi_j\|_w^j) \subseteq f_w^i(\|\psi_j\|_w^j)$. Then, $f_w^i(\|\varphi_j \wedge \psi_j\|_w^j) \subseteq \|\varphi \mid_j \varphi_j\|_w^i \cap \|\psi \mid_j \psi_j\|_w^i$.

Take any $\sigma_i \in \|\varphi|_j \varphi_j\|_w^i \cap \|\psi|_j \psi_j\|_w^i$. Take any $\sigma_j \in \|\varphi_j \wedge \psi_j\|_w^j$. Since $\sigma_j \in \|\varphi_j\|_w^j$ and $\sigma_i \in \|\varphi|_j \varphi_j\|_w^i$, $out(w, \sigma_i \cup \sigma_j) \subseteq [\![\varphi]\!]_M$. Since $\sigma_j \in \|\psi_j\|_w^j$ and $\sigma_i \in \|\psi|_j \psi_j\|_w^i$, $out(w, \sigma_i \cup \sigma_j) \subseteq [\![\psi]\!]_M$. So for any $\sigma_j \in \|\varphi_j \wedge \psi_j\|_w^j$, $out(w, \sigma_i \cup \sigma_j) \subseteq [\![\varphi \wedge \psi]\!]_M$. Then, it follows that $\sigma_i \in \|\varphi \wedge \psi|_j \varphi_j \wedge \psi_j\|_w^i$. Since $\sigma_i \in \|\varphi|_j \varphi_j\|_w^i \cap \|\psi|_j \psi_j\|_w^i$ was chosen arbitrarily, we conclude that $\|\varphi \mid_j \varphi_j\|_w^i \cap \|\psi \mid_j \psi_j\|_w^i \subseteq \|\varphi \wedge \psi \mid_j \varphi_j \wedge \psi_j\|_w^j$.

So we have that $f_w^i(\|\varphi_i \wedge \psi_j\|_w^j) \subseteq \|\varphi \mid_j \varphi_j\|_w^i \cap \|\psi \mid_j \psi_j\|_w^i \subseteq \|\varphi \wedge \psi \mid_j \varphi_j \wedge \psi_j\|_w^j$, from which it follows that $M, w \models Int_i(\varphi \wedge \psi \mid_j \varphi_j \wedge \psi_j)$. □

4.2.11. Proposition (Shared intentions agglomerate). *For any* $\varphi_i, \varphi_j, \psi_i, \psi_j, \varphi, \psi \in \mathcal{L}^I_{CL_\alpha}$*, the following holds:*

$$\models (Int_{i,j}\varphi \wedge Int_{i,j}\psi) \rightarrow Int_{i,j}(\varphi \wedge \psi)$$

Proof:
Assume that $M, w \models Int_{i,j}\varphi \wedge Int_{i,j}\psi$, where:

$$Int_{i,j}\varphi := Int_i\varphi_i \wedge Int_i(\varphi|_j\varphi_j) \wedge Int_j\varphi_j \wedge Int_j(\varphi|_i\varphi_i) \wedge [\emptyset]((\varphi_i \wedge \varphi_j) \rightarrow \varphi)$$
$$Int_{i,j}\psi := Int_i\psi_i \wedge Int_i(\psi \,|_j\, \psi_j) \wedge Int_j\psi_j \wedge Int_j(\psi \,|_i\, \psi_i) \wedge [\emptyset]((\psi_i \wedge \psi_j) \rightarrow \psi)$$

for some $\varphi_i, \varphi_j, \psi_i, \psi_j$. We show that $M, w \models Int_{i,j}(\varphi \wedge \psi)$, i.e.:

$$M, w \models Int_i(\varphi_i \wedge \psi_j) \wedge Int_i(\varphi \wedge \psi|_j \varphi_j \wedge \psi_j) \wedge Int_j(\varphi_j \wedge \psi_j) \wedge Int_j(\varphi \wedge \psi|_i \varphi_i \wedge \psi_i) \wedge [\emptyset]((\varphi_i \wedge \psi_i \wedge \varphi_j \wedge \psi_j) \rightarrow (\varphi \wedge \psi))$$

By Lemma 4.2.10, from $M, w \models Int_i\varphi_i \wedge Int_i\psi_i$ it follows that $M, w \models Int_i(\varphi_i \wedge \psi_i)$. Analogously, from $M, w \models Int_j\varphi_j \wedge Int_j\psi_j$ it follows that $M, w \models Int_j(\varphi_j \wedge \psi_j)$. Then, by (2a) of Def. 4.2.1, we also have that $M, w \models [i](\varphi_i \wedge \psi_i)$ and $[j](\varphi_j \wedge \psi_j)$. By Lemma 4.2.10, from $M, w \models Int_i(\varphi \,|_j\, \varphi_j) \wedge Int_i(\psi \,|_j\, \psi_j)$ and $M, w \models [j](\varphi_j \wedge \psi_j)$ it follows that $M, w \models Int_i(\varphi \wedge \psi \,|_j\, \varphi_j \wedge \psi_j)$. Analogously, $M, w \models Int_j(\varphi \wedge \psi|_i \varphi_i \wedge \psi_i)$. From $M, w \models [\emptyset]((\varphi_i \wedge \varphi_j) \rightarrow \varphi) \wedge [\emptyset]((\psi_i \wedge \psi_j) \rightarrow \psi)$ it follows that $M, w \models [\emptyset]((\varphi_i \wedge \varphi_j \wedge \psi_i \wedge \psi_j) \rightarrow (\varphi \wedge \psi))$. Taking it all together, $M, w \models Int_{i,j}(\varphi \wedge \psi)$. □

4.2.2 Envoi

We have introduced a formal language $\mathcal{L}^I_{CL_\alpha}$, in which we have formalised our reading of Bratman's theory of shared intention. We have demonstrated that our reading is consistent (Proposition 4.2.8), as well as ontologically, conceptually, and normatively conservative (Propositions 4.2.9, 4.2.11) over the theory of individual conditional intention. So if we share Bratman's philosophical stance on group agency, we can distinguish random sets of agents from groups in $\mathcal{L}^I_{CL_\alpha}$: the latter should share some non-trivial intentions, unlike the former.

Our work here can be continued, both conceptually and formally. From the conceptual side, we have only addressed present-directed shared intentions, such as to dance together, to meet at the library, to applaud the concert, etc. Present-directed intentions suffice to distinguish who acts as a group at any given moment. So we have ignored the future-directed intentions: commitments to long-term strategies. The examples of a shared future-directed intention we abstracted away

from are: to dance together whenever the opportunity presents itself; to regularly meet at the library every Sunday for the next year; to love and to cherish one another till death do us part. Although future-directed shared intentions are not necessary for our line of work, they are worth exploring in their own right.

Formally, we have not explored any meta-logical properties of the formalisms we have introduced beyond the single-agent logic of conditional intention. To axiomatise the set of validities $\{\varphi \in \mathcal{L}^I_{CL_\alpha} \mid \models \varphi\}$, we will need to axiomatise the fragment of ConStR with $[\,.\,]_\alpha(\,.\,;\langle\,.\,\rangle\,.\,)$-modality, which is still an open problem. It might be easier to embed the corresponding fragment of ConStR in some fragment of Strategy Logic [Mog13] or First-Order Coalition Logic [CGM25] first, which we reserve for future work.

Chapter 5

Collective intentions and decision problems[1]

In this chapter, we formally approximate team reasoning as a theory of collective intention. In short, the main take-home message from Bacharach's work, discussed in Section 2.2 is as follows. The object of a collective intention is a (proposition about) a solution to the group's decision problem. The solutions to group decision problems are joint actions, understood as aggregations of group members' actions. The group decision problems, on the contrary, cannot be reduced to the aggregations of group members' individual decision problems: the group problem might be guided by scarcer, irreducible distinctions, such as common interest, collective obligation, etc. So we have the following desiderata for the formalism to model the theory: it should have an explicit representation of individual and group decision problems, in the context of which intentions are formulated. The difference between individual and collective decision problems should ground the irreducibility of collective intention to the aggregation of individual ones. As we are to see, making the intention problem-sensitive also helps us to alleviate the problem of *side-effects*. Given that intentions exist in the context of a decision problem, whenever the agent intends that φ and φ entails ψ, whether the agent is rationally pressured to intend ψ depends on whether ψ has something to do with her decision problem.

The chapter is organised as follows. In Section 5.1, we introduce a single-agent problem-sensitive logic of intention. In this formalism, an agent's intention de-

[1]Section 5.1 of this chapter is based on the paper *Daniil Khaitovich and Aybüke Özgun. Hyperintensional Intention. In: Proceedings of TARK 2025. Edited by Adam Bjorndahl. EPTCS. Waterloo, Australia: Open Publishing Association, 2025, pages 44–60.* [KÖ25].

pends on the decision problem the agent has. Introducing decision problems allows to distinguish between practically relevant and irrelevant propositions, which provides us with a more precise solution to the side-effects problem, addressed in Section 4.1.5. In Section 5.2, we extend out framework to multi-agent case with groups, holistically defining groups' decision problems as irreducible.

5.1 Problem-sensitive intention: single-agent case

Intentions do not obey many of the *closure principles* logicians bake in their formalisms. While we are rationally committed to intend some of the consequences of what we intend, we are not rationally committed to intend, e.g., all logical, known, or believed consequences of our intentions. To cite a classic example [CL90a, p. 218], one intends to go to the dentists to get one's tooth filled, believing or knowing that filling tooth will lead to being in pain (maybe because the agent doesn't know about anesthetics or they have high tolerance to anesthetics or they cannot take much anesthetics due to other health issues), while not intending to be in pain. Agents not only do not intend the unforeseen side-effects of their actions, but they also do not intend all of their foreseen side-effects [Bra87; CL90a].

This has already been well documented in the literature on logics of intentions [CL90a; KP93], and much ink has been spilled to meet the challenge of finding a logical formalism for intentions that neither over- nor undergenerates the *intended* consequences of one's intentions such that the overgeneration problem is solved in a principled, well-motivated way [CL99; BG23; Roy08].

The concept of the so-called *issue-relevance* [Ber18; Kro24] or *question-sensitivity* [Min11; Yal18; Hoe22; BBS18] of propositional mental states may be helpful to achieve this goal. Question-sensitivity of epistemic/doxastic attitudes can explain the failure of some closure principles by modeling these attitudes as dependent on inquiry relevant to an agent's epistemic/doxastic agenda, alleviating the infamous problem of logical omniscience [Sta91; Var86; HÖB20] or accounting for the effects of inquiry on a rational, idealised agent's knowledge and belief [BBS18; Kro24]. In a similar vein, observing the parallel between the problems of logical omniscience and side-effects, recent work [BG23] developed a question-sensitive theory of rational intentions, modeling intentions as dependent on the agent's practical question *what to do?*, namely, their decision problem. We take our cue from this theory but argue that it still overgenerates validities: it is still subject to the problem of side-effects. To briefly explain (and to be further elaborated below), the purely possible-worlds, partition-based mechanism employed in [BG23] avoids closure under logical, necessary or known/believed entailments but still validates closure under (undesired instances of) logical and necessary *equivalents* for intentions. We provide compelling examples that challenge closure under log-

ical, necessary, and known/believed equivalents and, in turn, argue that a logic of intentions should also avoid these closure principles of equivalents. In particular, it should be *hyperintesional*.

Our proposal employs tools from topic- or subject matter-sensitive semantics for knowledge and belief [ÖB21; Ber22; HÖB20] and use them to model *decision problems* in a finer, hyperintensional way. We fully develop this formalism based on a bi-modal language with a global and an intention modality, and provide a sound and strongly complete axiomatization for the proposed hyperintensional logic of intention.

5.1.1 Logics of Intension and The Problem of Side-Effects

Formal epistemologists and philosophical logicians often turn to Kripke semantics to model propositional mental attitudes such as knowledge, belief and, of particular interest for us, *intention*. As well known, however, this mainstream approach validates certain closure principles for these attitudes that are fit to model only highly idealised knowers and believers. As also emphasised, e.g., in [KP93], these closure principles are a more serious treat for intention, as they cannot be defended for intentions even as idealizations. In this paper, we will focus on the following closure principles, formulated for intentions via operator I:

1. **Closure under logical entailment:** $\models \varphi \to \psi$ then $\models Int\varphi \to Int\psi$
2. **Necessary entailment:** $\models \Box(\varphi \to \psi) \to (Int\varphi \to Int\psi)$
3. **Closure under logical equivalence:** $\models \varphi \leftrightarrow \psi$ then $Int\varphi \leftrightarrow Int\psi$
4. **Necessary equivalence:** $\models \Box(\varphi \leftrightarrow \psi) \to (Int\varphi \leftrightarrow Int\psi)$

A few notes on notation is warranted. In the formulation of principles 1-4, $\models$ represents the logical consequence relation of the relevant logic, $Int\varphi$ is read as "the agent intends that φ". $\Box\varphi$ is read as "it is apriori that φ" or "the agents knows that φ", depending on the context and examples we target. The reason why we do not introduce distinct operators for knowledge and apriori necessity is merely practical and the lack of such distinct operators does not bear on any substantive conceptual points we want to make in this work. In our formalism, we interpret $\Box$ as the global modality, closer to the standard interpretation of apriori necessity. Introducing an additional knowledge operator K interpreted on a subset of the logical space, representing epistemically accessible worlds of the agent (as in, e.g., the Kripke semantics of epistemic logic) does not make any difference for the kinds of logical principles we focus on in this work.[2] This much

[2]This claim assumes that the set of so-called conative alternatives wrt which the operator *Int* is interpreted is a subset of the epistemically accessible worlds. Moreover, one can also read

formalism should be sufficient for our exposition. All these notions and notation will be properly introduced later.

We say that a logic *suffers from the problem of side-effects* if it validates at least one of 1-4 (even when φ and ψ are propositional variables). Just as an epistemic agent can be non-omniscient for diverse reasons [Ber22; HÖB20], the problem of side-effects can be explained by different factors. Since the early stages of intention logic research, principles 1 and 2 have been considered undesirably strong [KP93; CL99]. There are at least two factors explaining the failure of these principles:

1. *Epistemic flaws*: One can intend that φ only if one has the epistemic resources to reason about φ. Agents cannot derive and do not always know/believe all the consequences of what they intend, thus, they do not always intend all necessary or logical consequences of what they intend.

For example, one may intend to maintain a vegan diet without intending not eating anything that contains cysteine simply because one has no idea what cysteine is or does not know that it is made out of animal products.[3]

2. *Control constraint*: One can intend that φ only if one believes that one has control over φ. So that, if one intends that φ, φ entails ψ, but one lacks control over ψ, one will not intend ψ. [BG23, p.357].

Epistemic flaws are not specific to intention and are widely studied in formal theories of knowledge and belief (see, e.g., [Var86; Sta84; HÖB20]). On the contrary, control constraint is a unique feature of intention and similar motivational attitudes. Cases of the failure of closure under entailment due to control constraint have been discussed in the existing literature [Roy08; BG23]. Authors often point to *foreseeing without intending*: one may intend to do some action, foresee some necessary consequences of that action (hence know that the intended action entails such consequences), yet still not intend these consequences [Aun66]. We focus on the less discussed cases, where even the weaker principle of closure under equivalence fails.[4]

Below we present a few examples motivating the failure of 3-4 exactly due to control constraint. In all such cases, the bearers of intention intends that φ, knows that φ is (necessarily or materially) equivalent to ψ, but does not intend

---

$\Box$ as belief. In this section, we often refer to knowledge to keep the presentation more concise.

[3]This example is inspired by Stalnaker's famous William III example for belief [Sta84], also used in [BG23] for intention to make a similar point.

[4]A single paper known to us that addresses the closure under equivalence as well is [CL99]. Authors approach the issue from the perspective of resource-bounded agents. As we are to see in the following sections, the presented semantics there is overly simplistic to capture the control constraint and too syntactic, relying solely on how the formulas are built from the atomic propositions in order to evaluate which side-effects they force.

that ψ. Since the equivalences are known by the agents in question, such failures cannot be explained by epistemic flaws.

5.1.1. Example. Imagine a patient who suffers from a severe dental disease. The doctor tells the patient that without any treatment, he will certainly lose his teeth, but the treatment decreases the chance only by 50%. By agreeing to the treatment, the patient intends to save his teeth with 50% chance, but he doesn't intend to lose his teeth with 50% chance.[5]

5.1.2. Example. In some monarchy, it is obligatory to pay taxes. A citizen of the monarchy does not want to face any legal consequences of tax avoidance, hence, she intends to pay taxes. The monarch necessarily spends tax money on official ceremonies. In this context, the propositions "*one's taxes are paid*" and "*one's paid taxes are spent on official ceremonies*" are necessarily equivalent (in all the worlds that comply with the customs of the monarchy): if one pays taxes, then the monarch spend them on ceremonies, while the spending one's paid taxes on ceremonies entails that one's taxes are paid. Being familiar with the customs of the monarch, this equivalence is known by the citizen. Nevertheless, the citizen may not intend that her paid taxes are spent on official ceremonies, despite intending to pay taxes.

We cannot explain these failures of the closure under equivalence by epistemic flaws. Instead, they can be explained via control constraint.[6] Intuitively, the patient has control over the decrease of the chance of teeth loss by 50%, since there is an option to take the treatment, but has no control over the fact that there still will be 50% chance of losing the teeth. The taxpayer controls whether her taxes are paid or not, but it is not in her control to prevent the monarch from spending her tax money not to her liking.

Therefore, the control constraint poses a challenge not only to the principles of closure under entailment (1 & 2), but also to the principles of closure under

[5]Similar examples can be found in the literature on framing effects [KT84].

[6]We can give a stronger philosophical explanation as to why one may have control over some proposition, but lack control over the equivalent one. While the early theories of agency [Chi64] define control strictly in terms of causation – the agent has control over φ iff she can force φ by her actions – the more recent accounts refine these definitions, connecting control with explanatory relations. In [Kelng], it is argued that an agent has control over φ only if the agent can act in a way that will lead to φ and the explanation why φ will obtain is centered around the actions of the agent. This leads to the failure of the closure under equivalents: it is possible that φ holds iff so does ψ, but explanations why those propositions are true differ. Explanation is often seen as an example of hyperintensional phenomena. All mathematical truths are necessary true and hence logically and apriori equivalent, nevertheless, they have different explanations. Such phenomena occur with contingent propositions as well: "*it is true that grass is green*" is explained by "*the grass is green*", but not vise versa, while two propositions are obviously necessarily equivalent. Since the metaphysics and logic of explanation are out of the scope of the present paper, we do not elaborate further on this argument. See, e.g., [Nol14, p.157] and [Sch11] for the further discussion.

equivalences (3-4). A logic of rational intention should avoid them. Yet, there are a number of simple and intuitive logical principles that rational intention should validate; that is, a logic of intention is still possible. For example, the restricted versions of the above mentioned closure principles, constrained to the cases where the agent has control over both propositions in question and there are no epistemic obstacles, should be validated. Moreover, as we have argued before, rational intentions are consistent (if one intends that φ, one doesn't intend that $\neg\varphi$) [CL90a; KP93; CL99; Roy08] and they agglomerate (if one intends that φ and one intends that ψ, then one intends that $\varphi \wedge \psi$) [Zhu10; BG23].

5.1.2 Formalising decision problems as partitions

One of the recent solutions to the problem of side-effects borrows another tool from formal epistemology – question-sensitivity. Some have argued that our knowledge and beliefs exist in the context of *questions* [Sta84; Yal18; BBS18; Kro24]. Likewise, one can argue that our intentions exist in the context of *decision problems*. Just like any belief we hold is an answer to some question, any intention we have is a solution to a problem "*what to do?*" [BG23]. Question-sensivity explains why we may not intend some consequences of our intentions, even when we know what follows from them: some propositions might be *undefined* in the context of our decision problems; they might not constitute (partial) solutions to them.

How to formally model a decision problem? So far, authors take a decision problem to be a question and formalise it in terms of a partition of a logical space [BG23; Hoe22; BBS18; BM12; Min11]:

5.1.3. Definition (Decision problem as partition). Given a non-empty set of possible worlds W, a decision problem Π is a partition of W.[7]

5.1.4. Definition (Complete and partial solutions). Every cell $X \in \Pi$ is a complete solution to the decision problem Π. A union of any set of cells $A = \bigcup_{i\in I} X_i$ for some $\{X_i\}_{i\in I} \subseteq \Pi$ is a partial solution to Π.

It is easy to see that every complete solution $X \in \Pi$ is a partial solution to Π: $X = \bigcup\{X\}$ with $\{X\} \subseteq \Pi$.

Intuitively, we associate a decision problem with a set of its mutually exhaustive complete solutions that the agent considers implementing. *Partial* solutions may be viewed as indeterministic choices between a set of complete solutions: agents may intend only to partially solve their decision problem, given that sometimes the difference between some complete solutions is of no importance to them. For example, one may be dealing with the quest of getting to the railway station and

[7] A partition Π of W is a subset $\Pi \subseteq 2^W$ such that $\emptyset \notin \Pi$, $\bigcup \Pi = W$, and for any $X, Y \in \Pi$ such that $X \neq Y$, $X \cap Y = \emptyset$.

consider three possible ways to do that: by bus, by tram or by foot. Both the bus and the tram depart from the same place and take roughly equal amount of time to get to the destination point, so that, one may intend to simply go to the bus/tram station and take a bus or a tram without settling upon any of these two complete solutions.

A decision problem restricts what propositions the agent can intend. Namely, if one intends that φ, then φ should be *defined on the agent's decision problem*. Beddor & Goldstein [BG23] has two proposals to formalise the latter notion of definedness on a decision problem. To recap briefly, the first one takes it that φ is defined on the agent's decision problem Π iff φ is a partial solution to Π (in the sense described in Definition 5.1.4). The corresponding logic of intention invalidate closure under entailment and validate agglomeration, but still forces closure under equivalents in full generality. The second proposal is closer to what we aim at, so we introduce its formal components in more detail and argue that it still overgenerates validities. Below, we assume that the object of intention is a sentence of a language of classical propositional logic, i.e., a formula that is built from atomic propositions using classical Boolean connectives ($\neg, \wedge, \vee$. etc). The semantics for such expressions is the standard possible worlds semantics, where $[\![\varphi]\!]$ denotes the set of possible worlds that make φ true.

The second proposal (see [BG23, Section 6]) takes it that a proposition φ is defined on a decision problem iff the *subject matter* of φ is included in the decision problem, where the notion of subject matter is also defined as partitions, in line with Lewisian theory of subject matters [Lew88a; Lew88b].[8] Here, we briefly restate the relevant definitions in [BG23].

5.1.5. Definition (Subject matters). Given a logical space W and any formula φ, $sm(\varphi) \subseteq 2^{2^W}$ is the subject matter of φ, defined recursively as follows (where p is atomic):

$$\begin{array}{lcl} sm(\top) & = & \{W\} \\ sm(\neg\varphi) & = & sm(\varphi) \\ sm(p) & = & \{[\![p]\!], W \setminus [\![p]\!]\} \setminus \{\emptyset\} \\ sm(\varphi \wedge \psi) = sm(\varphi \vee \psi) & = & \{X \cap Y \mid X \in sm(\varphi), Y \in sm(\psi)\} \setminus \{\emptyset\} \end{array}$$

Note that for any formula φ, $sm(\varphi)$ is a partition on W[9]; Moreover, $\exists P \subseteq sm(\varphi) : \bigcup P = [\![\varphi]\!]$ and $\exists N \subseteq sm(\varphi) : \bigcup N = [\![\neg\varphi]\!]$.

[8] A detailed presentation of theories of subject matter is beyond the scope of this work. We refer the reader to [Lew88b; Haw18; PS21; Lew88a].

[9] Our definition of *sm* function differs from the one given in [BG23]. The only differences are that we define $sm(\top)$ as a primitive notion and make sure that for any $\varphi \in \mathcal{L}$, $\emptyset \notin sm(\varphi)$. It affects the framework neither conceptually nor technically; and is done for the sake of consistency with our notation and definitions.

5.1.6. Definition (Parthood on subject matter). Given a logical space W and two propositions φ, ψ, a subject matter of φ is a part of subject matter of ψ, $sm(\varphi) \sqsubseteq sm(\psi)$, iff for any $X \in sm(\varphi)$ there exists a subset $S \subseteq sm(\psi)$, such that $\bigcup S = X$.

We are now set up to define the formal framework. Given some logical space W, we can represent intention via (1) the set of conative alternatives, i.e. the non-empty set of possible worlds $Con \subseteq W$, such that the agent's intention is satisfied in Con-worlds; (2) the decision problem Π , which is a partition of the logical space W.

5.1.7. Definition (Question-sensitive intention). Given a logical space W, a set of conative alternatives Con and a decision problem Π, an agent intends that φ iff (1) $Con \subseteq [\![\varphi]\!]$ and (2) $sm(\varphi) \sqsubseteq \Pi$, i.e. for any $X \in sm(\varphi)$ there exists a partial solution $S \subseteq \Pi$, s.t. $X = \bigcup S$.

The intuition behind Definition 5.1.7 is the following: agent intends that φ iff (1) the agent intends to solve her decision problem in a way that will force φ and (2) φ is defined on the agent's decision problem. Agent has control over φ, since φing is considered a partial solution to her decision problem. By the very same reason, agent is aware that she brings about that φ by the solution she is committed to implement. Hence, both control constraint and epistemic flaws are accounted for in this semantics.

Definition 5.1.7 validates consistency and agglomeration for intention. The closure under apriori entailment (2) and equivalence (4) fail, since the scope of the intention is restricted by the decision problem. Nevertheless, the usage of subject matters does not rule out all cases when the closure under apriori or known equivalences may fail. The closure cannot be falsified for atomic propositions: it not hard to see that under Definition 5.1.7, for any atomic propositions p, q, if $[\![p]\!] = [\![q]\!]$, then $sm(p) = sm(q)$ and hence one intends that p iff one intends that q. Examples 5.1.1 and 5.1.2 falsify this principle.

Yet another, more conceptual, argument against formalizing decision problems as partitions comes from the way the notion of parthood on questions is defined. Intuitively, one decision problem is a part of another iff every complete solution to the former is a partial solution to the latter (a complete solution to a coarser problem is a partial solution to a finer problem). More formally, Π_1 is a part of Π_2, denoted by $\Pi_1 \sqsubseteq \Pi_2$, iff for all $A \in \Pi_1$ there is $\{X_i\}_{i \in I} \subseteq \Pi_2$ such that $A = \bigcup_{i \in I} X_i$. We refer to $\sqsubseteq$ relation as *intensional parthood*, i.e. $\Pi_1 \sqsubseteq \Pi_2$ reads as "Π_1 is an intensional part of Π_2". Another useful acronym: $\Pi_1 \simeq \Pi_2 := \Pi_1 \sqsubseteq \Pi_2 \& \Pi_2 \sqsubseteq \Pi_1$, referred to as *intensional equivalence* (Π_1 is intensionally equivalent to Π_2). The claim according to which there is nothing more to parthood than an intensional parthood, is going to be referred as *intensional parthood claim*. Modeling decision problems as partitions supports the intensional parthood claim,

affecting both proposal of question-sensitivity in [BG23]. On the contrary, we reject the intensional parthood claim, since there are decision problems Π_1, Π_2, where $\Pi_1 \sqsubseteq \Pi_2$, but the agent can intentionally solve Π_2 without any intention to solve Π_1. Consider the next example.

5.1.8. Example. Ticket office clerk works at a train station with two platforms. At every platform, a train arrives once in an hour, at j:00 sharp. Every train has its unique code of the form $(i; j)$: the train $(i; j)$ arrives at the ith platform at j:00 sharp. Let r^i_j be a proposition "*someone has a ticket to a train service number* $(i; j)$", t_j – "*someone has a ticket to a train service at* j*:00 sharp*", p_i – "*someone has a ticket to a train service that departs from* i*th platform*". Note that every atomic proposition r^i_j is equivalent to the corresponding conjunction $p_i \wedge t_j$: every service has its unique pair of departure platform and time.

Someone comes to the train station and asks the clerk to sell him a ticket for a train number $(2; 13)$: the customer only knows the train service number, since he was asked to buy that ticket with no further details provided. The clerk knows the system well and is aware that the train service $(2; 13)$ departs from platform 2 at 13:00 sharp. There is a pile of 48 tickets to every train: $(1; 1), (1; 2), \ldots, (2; 24)$. The clerk needs to choose which ticket he should take from the pile and intends that the customer gets the ticket to $(2; 13)$. The clerk intends neither that the customer gets a ticket for a train that departs from the second platform nor that the customer gets a ticket for a train that departs at 13:00, since this is just out of the focus of the clerk.

Intuitively, Example 5.1.8 is just another instance of foreseeing without intending. The clerk intends to give the customer the ticket the customer has asked for. The clerk also foresees that as a consequence of that, the customer will receive a ticket for a train that departs at 13:00 from platform number 2. Nevertheless, the clerk does not intend to give the customer a ticket for a train with a specific departure time or place, since it is just irrelevant to the decision problem the clerk is facing. Unfortunately, formalization via subject matters of solutions cannot capture this nuance. The clerk needs to choose the ticket from the pile, i.e., he considers all the tickets, and his decision problem is $\Pi = \{[\![r^i_j]\!] \mid i \in \{1, 2\}, 0 \leq j \leq 24\} \cup \{W \setminus \bigcup\limits_{0\leq j\leq 24}^{i\in\{1,2\}} [\![r^i_j]\!]\}$. In other words, $\Pi = sm(\bigvee\limits_{0\leq j\leq 24}^{i\in\{1,2\}} r^i_j)$. Under Definition 5.1.6, $sm(\bigvee\limits_{0\leq j\leq 24}^{i\in\{1,2\}} r^i_j) = sm(\bigvee\limits_{0\leq j\leq 24}^{i\in\{1,2\}} (p_i \wedge t_j))$. So that, r^i_j is defined on Π iff $p_i \wedge t_j$ is defined on Π. Under Definition 5.1.7, the clerk cannot intend that r^i_j without intending that p_i and that t_j.

We may come up with a number of similar examples: one may look for an exact building without minding the street and the house number of that building or wonder about one's age without thinking about neither the day nor the month

nor the year one was born, etc.: in all such cases, an agent solves a problem without minding the intensional parts of the problem.

Summarizing, modeling decision problems as partitions does not seem to be sufficiently fine-grained, since it endorses intensional parthood claim and, as a consequence, the closure under equivalence 4 is satisfied in contexts where it should fail.

5.1.3 Hyperintensional logic of intention

As we have just seen, modeling decision problems as partitions overgenerates validities: it validates undesired, atomic instaces of 4. Luckily, there is a concurrent way to represent decision problems: as atomic objects. Namely, we can simply state that there is a set P of problems, where no problem $a \in P$ has internal structure. Then, we can define a solution function s, which maps every formula of our language φ to a set of problems: if $a \in s(\varphi)$, then bringing it about that φ is a partial solution to a. Definedness on a decision problem and control constraint are expressible via s. If the agent is solving a decision problem $a \in P$ and $a \in s(\varphi)$, then bringing about that φ is a partial solution to a the agent considers. Hence, φ is defined on a, which means that the agent has control over φ and is aware that some solution to the decision problem forces φ. To have an adequate notion of parthood on decision problems, we define the corresponding partial order $\leq$ on P. Then, following our intuition about parthood, s function should be $\leq$-upward closed: if $a \leq b$, then for any φ, $a \in s(\varphi)$ entails $b \in s(\varphi)$. Informally, if a is a part of b, then any partial solution to a is a partial solution to b. As we are to see, this approach steps on the hyperintensional terrain, where the closure under logical and necessary equivalences fail. Moreover, it will allow us to differentiate even between the necessarily equivalent atomic propositions and adequately model cases such as Examples 5.1.1-5.1.8.

Formalizing decision problems as atomic objects

First, we define the formal language we are to work with. Let $\mathcal{L}_{CPL}$ be the language of classical propositional logic with a countable set of atomic propositions Prop, defined recursively as follows:

$$\alpha := \top \mid p \mid \neg\alpha \mid (\alpha \vee \alpha) \mid (\alpha \wedge \alpha)$$

where $p \in \mathsf{Prop}$. Note that we have $\top$, $\neg$, $\vee$ and $\wedge$ as primitives and do not define one in terms of another. We use the interdefinability of $\neg$, $\vee$, $\bot$ and $\wedge$ in terms of each other as our language cannot discern, e.g. $\varphi \wedge \psi$ from $\neg\varphi \vee \neg\psi$. However, we have specifically defined $\top$ as a primitive: it is pure tautology without non-logical context. The reader should not confuse $\top$ with another tautology expressed using the elements of Prop, such as $p \vee \neg p$ and $p \to p$. They will be logically equivalent

but they won't be substitutable with each other under the scope of our intention modality. For $\bot$, we set $\bot := \neg\top$. Based on $\mathcal{L}_{CPL}$, the well-formed formulas of the language $\mathcal{L}_I$ of our logic of intention are given by the following grammar:

$$\alpha \mid \neg\varphi \mid (\varphi \vee \varphi) \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid Int\alpha$$

where $\alpha \in \mathcal{L}_{CPL}$. $Int\alpha$ reads as "*agent intends that* α". Notice that the maximum modal depth of a formula wrt Int is one, that is, I takes only the elements of $\mathcal{L}_{CPL}$ in its scope. It is done so to exclude expressions about higher order intentions, such as intention to intend. The reason for that restriction is to be discussed later. $\Box\varphi$ read as "*it is necessarily true that* φ". We employ the usual abbreviations for propositional connectives $\rightarrow$ and $\leftrightarrow$ as $\varphi \rightarrow \psi := \neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi := (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)$. We interpret this language on the so-called *problem-sensitive Kripke models*, obtained by endowing Kripke models with a join-semilattice that represents decision problems and their parthood relation. The former is familiar in modal logic and requires no elaboration [BBW06]. The component for the representation of decision problems, called *problems model*, is a variation of topic-sensitive models [Ber22] and warrants further explanation.

5.1.9. Definition (Problems model). A problems model $\mathcal{P}$ is a tuple $(P, \oplus, s)$, where:

1. $P \neq \emptyset$ is a non-empty set of decision problems;
2. $\oplus : P \times P \rightarrow P$ is a binary idempotent, commutative, associative operation: problem fusion. We assume the unrestricted fusion, that is, $\oplus$ is always defined on P: $\forall A \subseteq P \exists a \in P : \bigoplus A = a$. The parthood relation $\leq$ is defined in terms of $\oplus$: $\forall a, b \in P : a \leq b$ iff $a \oplus b = b$;
3. $s : \mathsf{Prop} \rightarrow 2^P$ is a function assigning a set of decision problems to each element in Prop. Namely, if $a \in s(p)$, then bringing it about that p is a suitable (partial) solution to a. s extends to the whole propositional language $\mathcal{L}_{CPL}$ as follows:

$$\begin{array}{ll} s(\top) = P & s(\varphi \vee \psi) = s(\varphi) \cap s(\psi) \\ s(\neg\varphi) = s(\varphi) & s(\varphi \wedge \psi) = \{a \oplus b \mid a \in s(\varphi), b \in s(\psi)\} \end{array}$$

   Given any $p \in Prop$, $s(p)$ is upward closed: $a \leq b \Rightarrow (a \in s(p) \Rightarrow b \in s(p))$.

The extension of s to the whole propositional language is motivated as follows. Given any problem $a \in P$, forcing φ is a partial solution to the problem iff to decide if φ or not to φ is a part of a, hence, $s(\varphi) = s(\neg\varphi)$. $\varphi \vee \psi$ partially solves a iff both disjuncts – to bring about that φ, to bring about that ψ – solve a: this corresponds to our intuition about partial solutions as indeterministic choice between a set of solutions. Finally, $\varphi \wedge \psi$ is a partial solution to a iff a can be splitted into two problems, i.e. $a = a_1 \oplus a_2$, such that φing partially solve a_1 and

ψing partially solve a_2. The following lemma shows that s is upward closed over the whole language $\mathcal{L}_{CPL}$.

5.1.10. Lemma. *Given a problem model $\mathcal{P} = (P, \oplus, s)$, $a, b \in P$ and $\varphi \in \mathcal{L}_{CPL}$, the following holds:*

$$a \leq b \Rightarrow (a \in s(\varphi) \Rightarrow b \in s(\varphi)).$$

Proof:
The proof follows by induction on the structure of φ. The case for atomic propositions holds by the definition of s. Case $\varphi := \top$ is trivial as $s(T) = P$ and case $\varphi := \neg\psi$ follows from by induction hypothesis (IH) and the fact that $s(\neg\psi) = s(\psi)$.

Case $\varphi := \psi \vee \chi$: Suppose that $a \leq b$ and $a \in s(\psi \vee \chi)$. By the defn. of s, this means that $a \in s(\psi)$ and $a \in s(\chi)$. Then, by the IH, we have $b \in s(\psi)$ and $b \in s(\chi)$, thus, $b \in s(\psi) \cap s(\chi) = s(\psi \vee \chi)$.

Case $\varphi := \psi \wedge \chi$: Suppose that $a \leq b$ and $a \in s(\psi \wedge \chi)$. By the defn. of s, this means that $a = d \oplus c$ for some $d \in s(\psi)$ and $c \in s(\chi)$. Observe that $a = d \oplus c$ implies that $d \leq a$ and $c \leq a$, and, in turn, $d \leq b$ and $c \leq b$. Then, by IH, we have $b \in s(\psi)$ and $b \in s(\chi)$. As $b = b \oplus b$, by defn. of s, we have $b \in s(\psi \wedge \chi)$. □

Semantics, axiomatization, and completeness

5.1.11. Definition (Problem-sensitive frames and models). A problem-sensitive frame is a tuple $F = (W, R, \mathcal{P}, f)$, where:

- $W \neq \emptyset$ is a non-empty set of possible worlds;
- $R \subseteq W \times W$ is a serial binary relation on W, such that $R(w)$ is the set of conative alternatives to w, i.e. possible worlds, where agent's intention in w is satisfied;
- $\mathcal{P} = (P, \oplus, s)$ is a problems model;
- $f : W \to P$ is a function, which takes a world and returns a decision problem that the agent is solving in the given world;

As usual, any frame F may be extended to a model $\mathcal{M} = (F, V)$, where $V : \mathsf{Prop} \to 2^W$ is a standardly defined valuation function.

5.1.12. Lemma. *Given a problems model $\mathcal{P} = (P, \oplus, s)$, a decision problem $a \in P$, and formulas $\varphi, \psi \in \mathcal{L}_{CPL}$, $a \in s(\varphi \vee \psi)$ iff $a \in s(\varphi \wedge \psi)$.*

Proof:

($\Rightarrow$) Assume $a \in s(\varphi \vee \psi)$. By Defn. 5.1.9, this means that $a \in s(\varphi) \cap s(\psi)$, i.e. $a \in s(\varphi)$ and $a \in s(\psi)$. Then, since $a = a \oplus a$, we obtain that $a \in s(\varphi \wedge \psi)$.

($\Leftarrow$) Assume $a \in s(\varphi \wedge \psi)$. By Defn. 5.1.9, this means that there is $b, c \in P$ such that $a = b \oplus c, b \in s(\varphi), a \in s(\psi)$. The fact that $a = b \oplus c$ implies that $b \leq a$ and $c \leq a$. Then, by Lemma 5.1.10, we obtain that $a \in s(\varphi)$ and $a \in s(\psi)$, that is, $a \in s(\varphi) \cap t(\psi)$. By Defn. 5.1.9, we conclude $a \in s(\varphi \vee \psi)$. □

We show the connection between the decision problems as atomic objects and as partitions. Namely, given a problem-sensitive model $\mathcal{M}$, every atomic decision problem $a \in P$, as it is given in Definition 5.1.9, can be transformed into a partition $\lceil a \rceil_{\mathcal{M}}$. Intuitively, $\lceil a \rceil_{\mathcal{M}}$ is an *intensional representation* of a: if we accepted the intensional parthood claim, then $\lceil a \rceil_{\mathcal{M}}$ would be identified with a and treated as in Definition 5.1.7. Moreover, we can show that if $a \leq b$, then $\lceil a \rceil_{\mathcal{M}} \sqsubseteq \lceil b \rceil_{\mathcal{M}}$. In other words, $\leq$ relation is not arbitrary, but meaningful w.r.t solutions: if $a \leq b$, then every partial solution to a is a partial solution to b, hence, $a \sqsubseteq b$. Nevertheless, the converse does not hold: we may have decision problems $a, b \in P$, such that $\lceil a \rceil_{\mathcal{M}} \sqsubseteq \lceil b \rceil_{\mathcal{M}}$, but $a \not\leq b$. This corresponds to our denial of the intensional parthood claim: intensional parthood does not entail parthood in general.

5.1.13. Definition. Given a problem-sensitive model $\mathcal{M} = (W, R, \mathcal{P}, f, V)$ and a problem $a \in P$, let $s^{-1}(a) = \{\varphi \in \mathcal{L}_{CPL} \mid a \in s(\varphi)\}$; that is, $\varphi \in s^{-1}(a)$ iff bringing it about that φ partially solves a. Let $S^{-1}(a)$ be a collection of maximal satisfiable subsets (shortly, mss) of $s^{-1}(a)$, i.e. $\Gamma \in S^{-1}(a)$ iff 1) $\Gamma \subseteq s^{-1}(a)$; 2) there exists a world $w \in W$, such that $\mathcal{M}, w \models \varphi$ for all $\varphi \in \Gamma$ (from now on, abbreviated as $\mathcal{M}, w \models \Gamma$) and 3) for any $\Delta \subseteq s^{-1}(a)$, such that $\Gamma \subset \Delta$, there is no world $w \in W$, such that $\mathcal{M}, w \models \Delta$.

For every $\Gamma \in S^{-1}(a)$ we define a set $[\![\Gamma]\!] = \{w \in W \mid \mathcal{M}, w \models \varphi \text{ for all } \varphi \in \Gamma\}$. Then, clearly, the set $\lceil a \rceil_{\mathcal{M}} = \{[\![\Gamma]\!] | \Gamma \in S^{-1}(a)\}$ is a partition of W. We call $\lceil a \rceil_{\mathcal{M}}$ an intensional representation of a. When the model is clear from the context, we omit the subscript and write simply $\lceil a \rceil$.

5.1.14. Proposition (Intensional representation preserves solutions). *Given any problem-sensitive model $\mathcal{M} = (W, R, \mathcal{P}, f, V)$ and any decision problem $a \in P$, if $a \in s(\varphi)$, then φ is defined on $\lceil a \rceil$, i.e. $sm(\varphi) \sqsubseteq \lceil a \rceil$.*

Proof:
We show that for any $\varphi \in \mathcal{L}_{CPL}$, if $a \in s(\varphi)$ then φ is defined on $\lceil a \rceil$.

By induction on φ. Let $\varphi := p$. Assume $a \in s(p)$ for some proposition $p \in$ Prop. W.l.o.g., let $[\![p]\!] \neq W$. By Definition 5.1.5, $sm(p) = \{[\![p]\!], W \setminus [\![p]\!]\}$. Let $\underline{p} = \{\bigcup [\![\Gamma]\!] \mid \Gamma \in S^{-1}(a), p \in \Gamma\}$. Clearly, $\underline{p} = [\![p]\!]$: since $p \in s^{-1}(a)$, for every $\Gamma \in S^{-1}(a)$ either $p \in \Gamma$ or $\neg p \in \Gamma$. Hence, for any possible world $w \in W$, if

$\mathcal{M}, w \models p$, then there exists $\Gamma \in S^{-1}(a)$, such that $\mathcal{M}, w \models \Gamma$, from which it follows that $w \in \underline{p}$. On the other hand, if $w \in \underline{p}$, then there exists a maximal satisfiable set $\Gamma \in S^{-1}(a)$, such that $p \in \Gamma$. Since $\mathcal{M}, w \models \Gamma$, $\mathcal{M}, w \models p$, from which by definition it follows that $w \in [\![p]\!]$. Then, $[\![p]\!] = \bigcup_{i \in I} X_i$ for some $\{X_i\}_{i \in I} \subseteq \lceil a \rceil$, and $W \setminus [\![p]\!] = \bigcup_{j \in J} X_j$, where $\{X_j\}_{j \in J} = \lceil a \rceil \setminus \{X_i\}_{i \in I}$ i.e. $sm(p) \sqsubseteq \lceil a \rceil$, i.e. p is defined on $\lceil a \rceil$.

Let $\varphi := \neg\psi$. By IH, if $a \in sm(\psi)$, then ψ is defined on $\lceil a \rceil$. Since $sm(\psi) = sm(\neg\psi) = sm(\varphi)$, if $a \in sm(\varphi)$, then $sm(\varphi) \sqsubseteq \lceil a \rceil$.

Let $\varphi := \psi_1 \vee \psi_2$ and $a \in s(\psi_1 \vee \psi_2)$. Since $s(\psi_1 \vee \psi_2) = s(\psi_1) \cap s(\psi_2)$, $a \in s(\psi_1)$ and $a \in s(\psi_2)$. By IH, if $a \in s(\psi_1), s(\psi_2)$, then ψ_1 and ψ_2 are defined on $\lceil a \rceil$, i.e. $sm(\psi_1) \sqsubseteq \lceil a \rceil$ and $sm(\psi_2) \sqsubseteq \lceil a \rceil$. By Definition 5.1.6, it means that $\forall A \in sm(\psi_1) \exists \{X_i\}_{i \in I} \subseteq \lceil a \rceil : \bigcup_{i \in I} X_i = A$ and $\forall B \in sm(\psi_1) \exists \{X_j\}_{j \in J} \subseteq \lceil a \rceil : \bigcup_{i \in J} X_j = B$. By Definition 5.1.5, $sm(\psi_1 \vee \psi_2) = \{A \cap B \mid A \in sm(\psi_1), B \in sm(\psi_2)\} \setminus \{\emptyset\}$. Take any $A \cap B \in sm(\psi_1 \vee \psi_2)$. As we have shown before, $A = \bigcup_{i \in I} X_i$ and $B = \bigcup_{j \in J} X_j$ for some $\{X_k\}_{k \in I \cup J} \subseteq \lceil a \rceil$. Hence, given that all X_i's and X_j's are disjoint[10], $A \cap B = \bigcup_{l \in I \cap J} X_l$. Since $A \cap B$ was an arbitrary element of $sm(\psi_1 \vee \psi_2)$, we can conclude that $\forall O \in sm(\psi_1 \vee \psi_2) \exists \{X_l\}_{l \in L} \subseteq \lceil a \rceil : \bigcup_{l \in L} X_l = O$, so that, $sm(\psi_1 \vee \psi_2) \sqsubseteq \lceil a \rceil$, which means that $\psi_1 \vee \psi_2$ is defined on $\lceil a \rceil$.

Let $\varphi := \psi_1 \wedge \psi_2$. Analogously to the previous case (cases are analogous in the light of Lemma 5.1.12). □

5.1.15. Proposition. *Given any problem-sensitive model $\mathcal{M}$ and any two problems a, b, if $a \leq b$, then $\lceil a \rceil \sqsubseteq \lceil b \rceil$.*

Proof:
Let $a \leq b$ for some arbitrary decision problems $a, b \in P$. Then, by Lemma 5.1.10, if $a \in s(\varphi)$, then $b \in s(\varphi)$ for any $\varphi \in \mathcal{L}_{CPL}$. In other words, $s^{-1}(a) \subseteq s^{-1}(b)$.

Let $X \in \lceil a \rceil$ be an arbitrary complete solution to $\lceil a \rceil$. By the definition of $\lceil a \rceil$, $X = [\![\Gamma]\!]$, where Γ is a maximal satisfiable subset of $s^{-1}(a)$. Since $s^{-1}(a) \subseteq s^{-1}(b)$, there exists some maximal satisfiable subsets $\Sigma \subseteq s^{-1}(b)$ such that $\Gamma \subseteq \Sigma$. Let $[\Gamma]_b = \{\Sigma \in S^{-1}(b) \mid \Gamma \subseteq \Sigma\}$ be a collection of such maximal satisfiable subsets. As we have just shown, $[\Gamma]_b \neq \emptyset$. Moreover, as we are to show, $\bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!] = [\![\Gamma]\!]$. From left-to-right: assume $w \in \bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!]$. Then, $w \in [\![\Sigma_i]\!]$ for some $\Sigma_i \in [\Gamma_b]$. Then, $\mathcal{M}, w \models \Sigma_i$. Since $\Gamma \subseteq \Sigma_i$, $\mathcal{M}, w \models \Gamma$, i.e. $w \in [\![\Gamma]\!]$. Right-to-left: assume $w \in [\![\Gamma]\!]$. Then, there exists some $\Sigma_i \in [\Gamma_b]$, such that $\mathcal{M}, w \models \Sigma_i$.[11] Then,

[10] For the contradiction, assume they are not disjoint. Then, there exists $X_i = [\![\Gamma_i]\!]$, $X_j = [\![\Gamma_j]\!]$, such that $X_i \cap X_i \neq \emptyset$. Let $w \in X_i \cap X_j$. Then, $\mathcal{M}, w \models \Gamma_i \cup \Gamma_j$, which contradicts the fact that Γ_i and Γ_j are *maximal satisfiable* subsets of $s^{-1}(a)$.

[11] In order to see why it holds, note that for any problem $x \in P$, $s^{-1}(x)$ is closed under

$w \in [\![\Sigma_i]\!]$ and since $[\![\Sigma_i]\!] \subseteq \bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!]$, $w \in \bigcup_{\Sigma \in [\Gamma]_b} [\![\Sigma]\!]$.

Moreover, by definition of $\lceil b \rceil$, $\{[\![\Sigma]\!] \mid \Sigma \in [\Gamma]_b\} \subseteq \lceil b \rceil$. Hence, X, a complete solution to $\lceil a \rceil$, is a partial solution to $\lceil b \rceil$. Since X was arbitrary, we can conclude that any complete solution to $\lceil a \rceil$ is a partial solution to $\lceil b \rceil$, from which by Definition 5.1.6 it follows that $\lceil a \rceil \sqsubseteq \lceil b \rceil$. □

5.1.16. Proposition ($\lceil a \rceil \sqsubseteq \lceil b \rceil \not\Rightarrow a \leq b$). *Given any problem-sensitive model $\mathcal{M} = (W, R, \mathcal{P}, f, V)$ and two problems $a, b \in P$, the fact that $\lceil a \rceil \sqsubseteq \lceil b \rceil$ does not entail $a \leq b$.*

Proof:
We prove this statement by counterexample. Let $\mathcal{M} = (W, R, \mathcal{P}, f, V)$ be a problem-sensitive model, such that for two propositions $p, q \in \mathsf{Prop}$, $V(p) = V(q)$. Then, let a be a problem, such that $a \in s(p)$ and for any formula $\psi \in \mathcal{L}_{CPL}$, if ψ contains any propositions other that p, then $a \notin s(\psi)$. Consequently, $a \notin s(q)$. Let b be a decision problem, such that $b \in s(q)$ and for any formula $\psi \in \mathcal{L}_{CPL}$, if ψ contains any propositions other that q, then $b \notin s(\psi)$. Consequently, $b \notin s(p)$. Obviously, $\lceil a \rceil = \lceil b \rceil$, hence, $\lceil a \rceil \sqsubseteq \lceil b \rceil$. At the same time, $a \not\leq b$, since it is not the case that for any $\varphi \in \mathcal{L}_{CPL}$, if $a \in s(\varphi)$, then $b \in s(\varphi)$: $a \in s(p)$, $b \notin s(p)$. □

We interpret $\mathcal{L}_I$ on problem-sensitive models as in the following definition.

5.1.17. Definition (Semantic for $\mathcal{L}_I$). Given a problem-sensitive model $\mathcal{M} = (W, R, \mathcal{P}, f, V)$ and a possible world $w \in W$, the satisfiability relation $\models$ is defined as follows:

$$\begin{array}{lcl}
\mathcal{M}, w \models p & \Leftrightarrow & w \in V(p) \\
\mathcal{M}, w \models \neg\varphi & \Leftrightarrow & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \vee \psi & \Leftrightarrow & \mathcal{M}, w \models \varphi \text{ or } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \Leftrightarrow & \forall v \in W : \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models Int\alpha & \Leftrightarrow & R(w) \subseteq [\![\alpha]\!]_{\mathcal{M}} \text{ and } f(w) \in s(\alpha)
\end{array}$$

where $[\![\alpha]\!]_{\mathcal{M}} = \{w \in W \mid \mathcal{M}, w \models \varphi\}$.

The notions of logical consequence and validity are defined standardly, as in Def. 3.0.3. While the semantic clauses for the Booleans and $\Box$ are standard, the intention operator $Int\varphi$ is interpreted in a problem-sensitive way: the agent intends that α at w iff (1) α is true in all conative alternatives at w and (2) α is a partial solution to the agent's decision problem at w. (2) is what invalidates 1-4, making the resulting logic hyperintensional.

negation, i.e.for any formula $\varphi \in \mathcal{L}_{CPL}$, $\varphi \in s^{-1}(x)$ iff $\neg\varphi \in s^{-1}(x)$ (since $s(\varphi) = s(\neg\varphi)$). Given that and the fact that $s^{-1}(a) \subseteq s^{-1}(b)$, we can complete any maximal satisfiable set $\Gamma \in S^{-1}(a)$ to some maximal satisfiable set $\Sigma \in S^{-1}(b)$ by the analogue of Lindenbaum construction.

5.1.18. THEOREM. *Principles 1-4 are invalid (also for atomic propositions).*

Proof:
We provide a counterexample against 4. The same counterexample can be used to prove the invalidity of the others. Consider the problem-sensitive model $\mathcal{M} = (\{w\}, \{(w,w)\}, (\{a,b,c\}, \oplus, s), f, V)$ such that $a \oplus b = c$, $a \not\leq b$ and $b \not\leq a$, $s(p) = \{a,c\}$, $s(q) = \{b,c\}$, $f(w) = a$, and $V(p) = V(q) = \{w\}$. We then have that $\mathcal{M} \models \Box(p \leftrightarrow q) \wedge Intp$, but $\mathcal{M}, w \not\models Intq$, since $f(w) = a \notin s(q)$. □

To state some of the axioms and other relevant principles in the complete system, we use the abbreviation '$\overline{\varphi}$' to denote the tautology $\bigwedge_{p \in Var(\varphi)}(p \vee \neg p)$[12], following a similar idea in [Gio19]. It is easy to see that $Int\overline{\varphi}$ simply expresses that φ *is a partial solution to the agent's decision problem* ($\mathcal{M}, w \models Int\overline{\varphi}$ iff $f(w) \in s(\varphi)$). We list the axioms and rules of the logic of problem-sensitive models in Table 5.1.

(CPL)all classical propositional tautologies and Modus Ponens
($\mathsf{S5}_\Box$) S5 axioms and rules for $\Box$
(Ax1) $Int\top$
(Ax2) $Int\varphi \rightarrow Int\overline{\varphi}$
(Ax3) $Int\varphi \rightarrow \neg I\neg\varphi$
(Ax4) $(Int\varphi \wedge Int\psi) \leftrightarrow Int(\varphi \wedge \psi)$
(Ax5) $\Box(\psi \rightarrow \varphi) \rightarrow ((Int\psi \wedge Int\overline{\varphi}) \rightarrow Int\varphi)$

Table 5.1: Axiomatization Log for the logic of problem-sensitive models.

We briefly comment on the axioms for the intention operator. Ax1 means that one always has at least the weakest intention possible: to bring it about that $\top$. Ax2 states that one intends that φ only if φ is a partial solution to the one's decision problem, reflecting problem-sensitivity of intention. Ax3 says that intentions are consistent and Ax4 states that intentions agglomerate. Finally Ax5 is our restricted closure principle: one intends those a priori consequences of one's intensions which also constitute partial solutions to one's decision problem. We conclude the section by stating our main technical result.[13]

5.1.19. THEOREM. Log *is a sound and strongly complete axiomatization of* $\mathcal{L}_I$ *with respect to the class of all problem-sensitive models.*

Proof:

[12] In order to have a unique definition of each $\overline{\varphi}$, we set the convention that elements of $Var(\varphi)$ occur in $\bigwedge_{p \in Var(\varphi)}(p \vee \neg p)$ from left-to-right in the order they are enumerated in $\mathsf{Prop} = \{p_1, p_2, \dots\}$, as in, e.g., [ÖB21].

[13] Unsurprisingly, the axiomatization is similar to the axiomatization of the logic of simple hyperintensional belief in [ÖB21]. We do not have an intention counterpart of their axiom $B\varphi \rightarrow \Box B\varphi$, since the semantics of I depends on the actual world.

For the detailed completeness proof, see Appendix 5.2.3. □

5.2 Problem-sensitive collective intentions

We proceed by modeling many agents' intentions. There are multiple agents, each having her own decision problem and intention. Some of these agents form groups and share intentions. We treat collective decision problems in line with the theory of team reasoning, presented in Section 2.2.

We begin with a motivating example of a *stag hunt game*.

5.2.1. Example. Two hunters are going for a hunt and face a dilemma. Each of them can either chase a hare or go for a stag. Hunting the stag is way more beneficial, since the stag is more valuable. Nevertheless, as it is commonly known among the hunters, they can only hunt a stag if both of them are chasing it: if one of the hunters went for a stag, while the other did not, the former would be left with nothing. On the other hand, each of the hunters can hunt a hare by themselves, but will not benefit as much. The hunters cannot communicate with one another to deliberate on the choice of actions, and cannot get to know what their peer has decided. Intuitively, both hunters will be better off if they just go for a stag.

The stag hunt game poses a challenge for standard game theory. Despite it being intuitively clear that the best option for both hunters is to go for a stag, standard solution concepts do not advise doing so. Let us formalise the scenario as a normal form game (as in Def. 2.2.1), in the following way. $G = (N, \{A_i, u_i\}_{i \in N})$ is a game, where:

- $N = \{a, b\}$ is the set of agents, i.e. two hunters;
- For any $i \in N$, $A_i = \{s_i, h_i\}$, i.e. both hunters have two strategies: to go for a stag, s_i, or for a hare, h_i;
- $u_i : \prod_{i \in N} A_i \to \mathbb{R}$ is a utility function that ascribes i's payoff for every outcome. See Figure 5.1.

	s_b	h_b
s_a	$3, 3$	$0, 1$
h_a	$1, 0$	$1, 1$

Figure 5.1: Stag hunt game.

A classically rational agent should frame her decision problem as the choice between two strategies characterised solely in terms of expected utility. Hence, if both players are best-reply reasoners, each should have a decision problem with

two complete solutions: to go for a stag or to go for a hare. Since neither of these two strategies dominates the other, individual decision problems cannot be framed as a choice between a dominating and a dominated strategy.

A team-reasoner should first choose the solution to the collective decision problem: *what should* ***we*** do? If a team-reasoner thinks about the collective decision problem in terms of payoff-structure, she can distinguish the unique Pareto-optimal outcome (s_a, s_b) from all the others. Hence, the group decision problem, where at least one joint action is distinguished from all the others as dominating, is not a fusion of individual decision problems, where nothing can be said about the dominance in terms of payoffs at all.

Generalizing this and other examples we have seen in Section 2.2, we want to outline the interesting relation between the collective and individual decision problems. Intensionally, solutions to the former are reducible to the aggregations of solutions to the latter for all members: joint actions are nothing but tuples of individual actions. Hyperintensionally, we use the irreducible concepts of group agency to describe a collective decision problem. In Example 5.2.1, the irreducible notion in the collective decision problem is payoff-dominance: we cannot redefine the payoff-dominance of the action profile (s_a, s_b) in terms of strategic dominance of s_a over h_a and s_b over h_b. Moving from quantitative to qualitative notions, we have a variety of impossibility results, showing that collective obligations [TH20], collective responsibility [HIH22], collective belief [HZ79], judgment [LP02], and preference [Arr63] cannot be reduced by any plausible forms of aggregation.

For individual intention, we have clearly distinguished its intensional component – a set of conative alternatives – from its hyperintensional component – the decision problem. In what follows, we use the same idea to model collective decision problems as intensionally additive and hyperintensionally super-additive. To concentrate on this primary goal, we abstract away from several irrelevant nuances via the following idealizations. First, we assume that members of all groups group-identify accordingly, and it is commonly known among group members. In other words, we assume there is no uncertainty about which members have which group identities. The group members commonly know their group decision problem. In that sense, our notion of collective intention is *thick*[14]: we do not tackle collective intentions of loose groups, whose communication channels are too unreliable to establish the common knowledge about group decision problems and the chosen (partial) solutions to them.

We move to the formalization of the outlined conceptual notions. To reason about the intentions of multiple agents and groups, we need to extend $\mathcal{L}$ to its multi-

[14]In [RS19], the distinction between thin and thick shared intention is made: *"Let us call 'thick' an account of shared intentions that makes use of a common knowledge condition, and call the account 'thin' otherwise."* The reader should avoid the confusion with the distinction of thick and thin propositions in the realm of aboutness [Yab14].

agent version. Given a finite set of agents $Ags = \{1, \ldots, n\}$ and countable set of propositional variables Prop, we recursively define $\mathcal{L}_G$ as follows:

$$\alpha ::= \top \mid p \mid \neg\alpha \mid (\alpha \vee \alpha) \mid (\alpha \wedge \alpha)$$

$$\mathcal{L}_G \ni \varphi ::= \alpha \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid Int_G\alpha$$

where $p \in \mathsf{Prop}$, $G \subseteq Ags$. The meaning of Boolean connectives and $\Box$ modality is the same as in $\mathcal{L}$. $Int_G\varphi$ reads as *"group G collectively intends that φ"*. For every agent $i \in Ags$, we will abbreviate $Int_{\{i\}}\varphi$ to I_i.

We need to introduce some minor changes in our definition of problems models in order to capture the absence of decision problems for some groups. So far, the problems models are atomless Boolean algebras (i.e., they might lack minimal elements). Here we introduce such a minimal element, representing the "empty" problem. Such a problem will be assigned to sets of agents that do not constitute a group.

5.2.2. Definition (Problems model, revised). A problems model is a tuple $\mathcal{P} = (P, \epsilon, \oplus, s)$, where $(P, \oplus)$ is as in Def. 5.1.9. $\epsilon \in P$ is a problem, such that $\forall a \in P \setminus \{\epsilon\} : \epsilon < a$ (i.e. $a \oplus \epsilon = a$, but $\epsilon \oplus a \neq \epsilon$). $s : \mathsf{Prop} \to 2^{P\setminus\{\epsilon\}}$ is a solution function assigning a set of decision problems to each element of Prop. $a \in s(p)$ still means that bringing about that p partially solves a. s extends to the whole propositional language $\mathcal{L}_{CPL}$ as follows:

$$\begin{array}{ll} s(\top) = P \setminus \{\epsilon\} & s(\varphi \vee \psi) = s(\varphi) \cap s(\psi) \\ s(\neg\varphi) = s(\varphi) & s(\varphi \wedge \psi) = \{a \oplus b \mid a \in s(\varphi), b \in s(\psi)\} \end{array}$$

Given any $p \in \mathsf{Prop}$, $s(p)$ is upward closed: $a \leq b \Rightarrow (a \in s(p) \Rightarrow b \in s(p))$.

Note that from Def. 5.2.2 it follows that for any $\varphi \in \mathcal{L}_{CPL}$, $\epsilon \notin s(\varphi)$: there are no solutions to the non-existent problem.

5.2.3. Definition (Multi-agent problem-sensitive frames and models). Given a finite set of agents $Ags = \{1, \ldots, n\}$, a multi-agent problem-sensitive model is a tuple $\mathcal{M} = (W, \{R_i\}_{i\in Ags}, \mathcal{P}, f, V)$, where:

- $W \neq \emptyset$ is a set of possible worlds;
- $R_i \subseteq W \times W$ is a serial binary relation on W, representing conative alternatives of agent i;
- $\mathcal{P} = (P, \epsilon, \oplus, s)$ is a problems model, where $\epsilon \in P$ is an empty problem: $s^{-1}(\epsilon) = \emptyset$;

- $f : (2^{Ags} \times W) \to P$ is a function that takes a group, a world and returns a decision problem that the group is solving in the given world. For the sake of brevity, $f_G(w) := f(G)(w)$, $f_i(w) := f_{\{i\}}(w)$. f has the next additional properties. For every $w \in W$ and for any $G \subseteq Ags$, such that $Card(G) \geq 2$:

$$f_i(w) \neq \epsilon \text{ for any } i \in Ags \tag{Ags}$$

$$f_G(w) = \epsilon \text{ or } \bigoplus_{i \in G} f_i(w) \leq f_G(w) \tag{Super-Add}$$

$$\text{If } \bigcap_{i \in G} R_i(w) = \emptyset, \text{ then } f_G(w) = \epsilon \tag{Triv}$$

$$\text{If } f_{G_1}(w) \neq \epsilon, \ldots, f_{G_n}(w) \neq \epsilon \text{ and } \bigcap_{1 \leq i \leq n} G_i \neq \emptyset, \text{ then } \bigcap_{i \in G_1 \cup \ldots \cup G_n} R_i(w) \neq \emptyset \tag{G-Cons}$$

$$\text{If } f_G(w) \neq \epsilon, \text{ then } \lceil f_G(w) \rceil = \lceil \bigoplus_{i \in G} f_i(w) \rceil \tag{Add}$$

- $V : Var \to 2^W$ is a standard valuation function.

The class of multi-agent problem-sensitive models is denoted as $\mathbf{M}_G$.

The idea behind this semantics is the following. If a set of agents does not constitute a group and does not have any collective intention, then their "collective decision problem" is empty, i.e. ϵ. Property (Ags) means that individuals have agency, hence non-trivial decision problems. (Super-Add) corresponds to superadditivity of group decision problems: a group problem (if non-trivial) may be bigger than the fusion of members' decision problems. (Triv) corresponds to the fact that agents with contradictory intentions do not form a group. (G-Cons) expresses the consistency of collective intentions among groups with common members. Informally, if $G_1, \ldots, G_n \subseteq Ags$ constitute groups and share some common members, then the collective intentions of $G_1, \ldots, G_n$ are consistent. Since we idealise group members as commonly knowing group decision problems and goals, we do not model situations where one agent is a member of a number of groups with contradictory goals. (Add) means that a nontrivial group problem, intensionally represented, is nothing but the fusion of individual decision problems. E.g., joint actions, intensionally speaking, are just aggregations of individuals' actions.

Our model allows agents to have contradictory intentions: there is a pointed model $(\mathcal{M}, w)$, such that two agents $i, j \in Ags$ have inconsistent sets of conative alternatives, i.e. $R_i(w) \cap R_j(w) = \emptyset$. One may object against it on the grounds of control constraint: if one agent intends that φ and other intends that $\neg\varphi$, then the one has control over φ and can bring about that φ, while the other can

bring about that $\neg\varphi$, which is contradictory, since they cannot bring about φ and $\neg\varphi$ simultaneously. For that matter, it is common for logics of group agency to impose *independence of agents* constraint: if one agent can bring about that φ and the other can bring about that ψ, then φ and ψ should be consistent [Hor01].

As a reply, we can point out that the control constraint is formulated in terms of agents' *beliefs* about control. Simply saying, if one agent intends that φ and another intends that $\neg\varphi$, then at least one of them has false belief about her abilities. For example, two athletes may compete against one another, while both of them have the intention to win, given that both athletes believe that a certain strategy of hers will secure her a decisive advantage. Nevertheless, we do not allow contradictory intentions among members of groups: group members communicate with one another in order to plan and coordinate their joint actions, hence they can revise their false beliefs about their abilities, in case some contradictions about members' abilities manifest themselves in the process of deliberation. For example, if the same two athletes are involved in match-fixing and were bribed to show the predetermined result of a competition, in the process of deliberating how they are to stage the match, they will resolve their false beliefs about their abilities to deal with certain moves or strategies of one another.

We distinguish individual decision problems from the group ones, which might seem to contradict the very idea of team reasoning. Shouldn't group members just share the very same decision problem, given that each deliberates what *they* should do? The group decision problem provides group members with reasons to choose solutions to their own decision problems: first, the team reasoner selects the best solution to the group decision problem; then, she chooses her own share as the solution to her individual decision problem, and commits to it. A group member can grasp the group decision problem, but cannot adopt it as her own, since that would break the control constraint. The solutions to group decision problems are up to all members, so each individual, lacking control over the actions of others, cannot commit to those solutions. Nevertheless, the solution to the group decision problem grounds the choice of their own actions.

Having explained our motivations behind Def. 5.2.3, we move to define semantics clauses for $\mathcal{L}_G$.

5.2.4. Definition. Given a multi-agent problem-sensitive model

$$\mathcal{M} = (W, \{R_i\}_{i\in Ags}, \mathcal{P}, f, V)$$

and any world $w \in W$, $\models$ relation is inductively defined as follows:

$$
\begin{array}{lcl}
\mathcal{M}, w \models p & \Leftrightarrow & w \in V(p) \\
\mathcal{M}, w \models \neg\varphi & \Leftrightarrow & \mathcal{M}, w \not\models \varphi \\
\mathcal{M}, w \models \varphi \vee \psi & \Leftrightarrow & \mathcal{M}, w \models \varphi \text{ or } \mathcal{M}, w \models \psi \\
\mathcal{M}, w \models \Box\varphi & \Leftrightarrow & \forall v \in W : \mathcal{M}, v \models \varphi \\
\mathcal{M}, w \models Int_G\alpha & \Leftrightarrow & \forall v \in W((w, v) \in \bigcap\limits_{i \in G} R_i \Rightarrow \mathcal{M}, v \models \alpha) \text{ and } f_G(w) \in s(\alpha)
\end{array}
$$

The satisfiability in a model ($\mathcal{M} \models \varphi$), a (pointed) frame ($F, w \models \varphi$, $F \models \varphi$) and the whole class of frames ($\mathbf{F}_G \models \varphi$) are defined as usual.

The intuition behind the semantic clause for $I_G\varphi$ reflects our ideas about the relation between intensional and hyperintensional characterizations of group actions. Intensionally, joint action is an aggregation of individual actions; therefore, the set of groups' conative alternatives is nothing but the intersection of members' sets of conative alternatives. Hyperintensionally, a group's decision problem may properly include the fusion of individual decision problems. Note that if some set of agents $G \subseteq Ags$ do not constitute a group, i.e. $f_G(w) = \epsilon$, then for any $\alpha \in \mathcal{L}_{CPL}$, $\mathcal{M}, w \not\models Int_G\alpha$, since $\epsilon \notin s(\alpha)$ for any formula.

Having the formal tools to model superadditive collective intentions, we proceed with the practical use-case. The *many hands problem* refers to the situation where we struggle to hold any individual responsible for a negative outcome that was reached as a consequence of the actions of multiple agents. The problem is easy to solve if we deal with the group who share the intention, rather than with independent individuals who happen to simultaneously act in a certain way. In the following example, we model the situation where group members' individual intentions and actions are insufficient to hold them responsible for the negative outcome, while their collective intention provides a solid justification to ascribe them some form of collective responsibility.

5.2.5. Example (Toxic waste and many hands). A company needs to cut its operating costs. It has two operating plants. Both plants utilise their toxic waste in a specialised disposal, which is very expensive. The most effective cost-cutting option involves switching to utilizing the toxic waste in a cheaper but environmentally dangerous way. If only one of the plants starts utilizing the waste in such a way, then there is no danger, but if both of them start doing it, then the ecological damage is almost inevitable, which is commonly known. Switching to that questionable practice is the only easy way to cut the company's operating costs by 50%: there are other possible reorganizations to optimise the production, but they are extremely hard to develop and implement.

After negotiating about it, the company's plant managers and the CEO agree that it is in the company's best interest to intentionally take the ecological risk

for the sake of economic prosperity. So they proceed with the new policy. Namely, the CEO approves the 50% budget cut for the operating costs. Managers of both plants instruct their workers to utilise the toxic waste in the same unregulated location. As a result, an ecological disaster happens. The investigation reveals that the incident would not have happened had the waste of at least one of the two plants been utilised elsewhere. Being accused of causing the disaster, the CEO argues that he only intended to cut the operating costs and has never directly instructed plant managers on implementing dangerous policies to reach the goal. Each of the plant managers, being confronted about causing the disaster, argues that he only intended to utilise his share of the toxic waste, which is relatively insignificant and could not have caused the incident by itself.

Let $Ags = \{ceo, m_1, m_2\}$ be the set of agents, representing the CEO and the two managers, respectively. Let $\mathsf{Prop} = \{t_1, t_2, c, d\}$, where t_1 and t_2 mean that the toxic waste from the first resp. the second plant is utilised in a risky way; c means that the aggressive cost cuts are approved, and d means that the high risk of ecological disaster is created. We represent the scenario of Example 5.2.5 as a pointed model $\mathcal{M} = (W, R_{ceo}, R_{m_1}, R_{m_2}, \mathcal{P}, f, V)$, where:

- $W = \{w_1, \dots, w_4\}$ is a set of possible worlds. $V(c) = \{w_1, w_2, w_3\}$, $V(t_1) = \{w_1, w_2\}$, $V(t_2) = \{w_1, w_3\}$, $V(d) = \{w_1\}$. The real world is w_1;
- $R_{ceo}(w_1) = \{w_1, w_2, w_3\} = V(c)$, $R_{m_1} = \{w_1, w_2\} = V(t_1)$, $R_{m_2} = \{w_1, w_3\} = V(t_2)$;
- $\mathcal{P} = (P, \epsilon, \oplus, s)$. $(P, \oplus)$ is as illustrated in Figure 5.2. Informally, p_{ceo}, p_{m_1} and p_{m_2} are the decision problems of the CEO, the first and the second manager, respectively. p_{Ags} is a collective decision problem of the company. $s^{-1}(p_{ceo}) = \{\varphi \in \mathcal{L}_{CPL} \mid Var(\varphi) \subseteq \{c\}\}$, $s^{-1}(p_{m_1}) = \{\varphi \in \mathcal{L}_{CPL} \mid Var(\varphi) \subseteq \{t_1\}\}$, $s^{-1}(p_{m_2}) = \{\varphi \in \mathcal{L}_{CPL} \mid Var(\varphi) \subseteq \{t_2\}\}$, $s^{-1}(p_{Ags}) = \mathcal{L}_{CPL}$.
- $f_{ceo}(w_1) = p_{ceo}, f_{m_1}(w_1) = p_{m_1}, f_{m_2}(w_1) = p_{m_2}, f_{Ags}(w_1) = p_{Ags}$.

It is not hard to see that no one has intended to cause the disaster individually and was not even anticipating it in their decision problems:

$$M, w_1 \models Int_{ceo}c \wedge Int_{m_1}t_1 \wedge Int_{m_2}t_2 \wedge \bigwedge_{i \in Ags} \neg Int_i\overline{d}$$

As a group, they collectively intend to take the environmental risk: $\mathcal{M}, w \models Int_{Ags}d$. Given that the story is consistent with our idealizations, the group's decision problem, as well as the adopted solution, was commonly known among the group members, and their actions were perfectly consistent with the solution. Hence, one may argue that at least each of them is complicit in causing the ecological disaster.

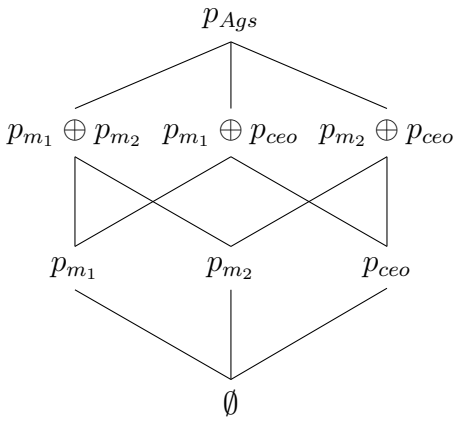

Figure 5.2: Haase diagram for a problems-model $\mathcal{P}$, where elements of P are partially ordered by $\leq$.

5.2.1 Logic

Here, we provide a sound and complete axiomatization for the logic of $\mathbf{M}_G$. We work with a fragment of $\mathcal{L}_G$, where Prop is finite. Assume that $Card(\mathsf{Prop}) = n$ for some $n \in \mathbb{N}$. Let $\mathcal{L}_G^k$ be a fragment of $\mathcal{L}_G$, built from such a finite set of propositional variables Prop. Then, let $\mathcal{L}_{CPL}^k$ be a set of all logically inequivalent propositional formulae of $\mathcal{L}_G^k$. Note that $\mathcal{L}_{CPL}^k$ is finite: $Card(\mathcal{L}_{CPL}^k) = 2^{2^n} = k$ for the corresponding $k \in \mathbb{N}$.

Throughout the section, we use the following acronym: $\bigcirc_G := Int_G\top$. $M, w \models \bigcirc_G \Leftrightarrow f_G(w) \neq \epsilon$. Intuitively, $\bigcirc_G$ means that the set of agents G constitutes a group and has a non-trivial collective decision problem.

We show that the logic Log_G (on Table 5.2) is a sound and complete axiomatization of the set of validities $\{\varphi \in \mathcal{L}_G^k \mid \mathbf{M}_G \models \varphi\}$.

Axioms $\mathsf{Ax1} - \mathsf{Ax5}$ bear the same meaning as in Log, expressing the very same properties of collective intentions. Triv means that unless $G \subseteq Ags$ constitutes a group, G does not collectively intend anything at all. Formally, Triv ensures that the set of formulae $\{\neg Int_G\varphi \mid \varphi \in \mathcal{L}_G^k\}$ is Log_G-consistent. $\mathsf{s} - \mathsf{Add}$ expresses the super-additivity of collective intentions. Namely, if a group G intends that φ, then its supergroup G' intends that φ as well. Given our understanding of groups as thick – i.e., groups are capable of establishing common knowledge – it seems natural that a subgroup's intentions are included in a supergroup's intention. Con expresses the consistency of groups' intentions that share common members. AxAdd corresponds to the intensional additivity of the group's decision problems. If α is a partial solution to the group's decision problem, then, intensionally, α is nothing but an indeterministic choice (syntactically, disjunction) between aggregations (syntactically, conjunctions) of solutions to the members' decision

(CPL)	All classical propositional tautologies and Modus Ponens
($\mathsf{S5}_\Box$)	$\mathsf{S5}$ axioms and rules for $\Box$
($\mathsf{Ax1}$)	$Int_i\top$ for any $i \in Ags$
($\mathsf{Ax2}$)	$Int_G\varphi \to Int_G\overline{\varphi}$
($\mathsf{Ax3}$)	$Int_G\varphi \to \neg Int_G\neg\varphi$
($\mathsf{Ax4}$)	$(Int_G\varphi \wedge Int_G\psi) \leftrightarrow Int_G(\varphi \wedge \psi)$
($\mathsf{Ax5}$)	$\Box(\psi \to \varphi) \to ((Int_G\psi \wedge Int_G\overline{\varphi}) \to Int_G\varphi)$
(Triv)	$\neg\bigcirc_G \to \neg Int_G\varphi$
($\mathsf{s-Add}$)	$(\bigcirc_G \wedge \bigcirc_{G'}) \to (Int_G\varphi \to Int_{G'}\varphi)$ for any $G \subseteq G' \subseteq Ags$
(Con)	$(\bigcirc_{G_1} \wedge \ldots \wedge \bigcirc_{G_n}) \to$ $((Int_{G_1}\varphi_1 \wedge \ldots \wedge Int_{G_n}\varphi_n) \to \neg\Box\neg(\varphi_1 \wedge \ldots \wedge \varphi_n))$ if $G_1 \cap \ldots \cap G_n \neq \emptyset$
(AxAdd)	$Int_G\overline{\alpha} \to$ $\bigvee_{\{\alpha_j^i\}_{1\le j\le k}^{i\in G} \subseteq \mathcal{L}^k_{CPL}} \big(\bigwedge_{i\in G} (Int_i\overline{\alpha_1^i} \wedge \ldots \wedge Int_i\overline{\alpha_k^i}) \wedge \Box(\alpha \leftrightarrow (\bigvee_{1\le j\le k} \bigwedge_{i\in G} \alpha_j^i)) \big)$

Table 5.2: Axiomatization Log_G for the logic of multi-agent problem-sensitive models

problems. Given that,e.g., $G = \{1, \ldots, n\}$, if $Int_1\overline{\alpha_1^j}, \ldots, Int_n\overline{\alpha_n^j}$ (i.e., α_1^j is the solution of $1 \in G$, ..., α_n^j is the solution of $n \in G$), then, $\alpha_1^j \wedge \ldots \wedge \alpha_n^j$ is the aggregation of members' solutions. Given the restriction of our language, there might be at most k such aggregations. Consequently, $\bigvee_{1\le j\le k} \bigwedge_{i\in G} \alpha_j^i$ is an indeterministic choice between k-many (not necessarily unique) aggregations of individual solutions.

Note that, analogously to Lemma 5.2.7, $\vdash (Int_G\overline{\varphi} \wedge Int_G\overline{\psi}) \to Int_G(\overline{\varphi} \vee \overline{\psi})$. Informally, it means that the groups are able to coordinate. If φ and ψ are solutions to the group's decision problem (e.g., $(stag, stag)$ and $(hare, hare)$ in Example 5.2.1), then $\varphi \vee \psi$ is also a partial solution that the group considers, hence, has control over it. This theorem aligns with our understanding of groups as thick: given that the group members can communicate to establish common knowledge among themselves, coordination puzzles, such as the stag hunt game, should not be puzzling for them.

Our proof strategy for completeness is as follows. First, we define a class of pseudo-models (Def. 5.2.20) and show it to be $\mathcal{L}_G$-invariant to $\mathbf{M}_G$ (Lemma 5.2.21). Then, we show that every Log_G-consistent set of $\mathcal{L}^k_G$-formulae is satisfiable on a pseudo-model, implementing the standard canonical model construction, analogous to the one we have implemented for Log.

5.2.6. Theorem (Soundness and completeness). *Given any set of formulae $\Delta \subseteq \mathcal{L}^k_G$ and any formula $\varphi \in \mathcal{L}^k_G$, the following holds:*

$$\Gamma \vdash_{\mathsf{Log}_G} \varphi \Leftrightarrow \Gamma \models_{\mathbf{M}_G} \varphi$$

Proof:
See Section 5.2.4 in Appendices. □

5.2.2 Envoi

In this section, we have developed a hyperintensional logic of collective intention. The problem-sensivity accounts for the failure of closure under logical and necessary equivalence and entailment, alleviating the side-effects problem. Any group decision problem is intensionally equivalent to the fusion of group members' individual problems, while there might be hyperintensional differences that make the group problem irreducible. Our analysis is again incomplete: unlike in Chapter 4, we have abstracted away from agents' actions and abilities, although they are implicitly present in our framework as solutions to individual decision problems. Explicit representation of agents' actions, as in, e.g., game models, would do a better job in expressing the intensional side of the control constraint.

Another interesting extension of our framework is conditional intentions. Namely, we would need to account for hyperintensional differences not only between propositions that are intended, but also between those that express conditions. The differences get especially subtle in the case of restrictive conditional intentions. The condition of a restrictive conditional intention is a reason for the intender to act. We can find two necessary equivalent propositions, such that one constitutes a reason for the agent to act, while the other is not. For instance, assume a gambling mathematician believes even numbers to be luck-promoting. She can bet on either 4 or 7 in some numbers game. The gambler might bet on 4 for the reason that 4 is an even number. "4 is an even number" is a mathematical, hence a priori truth, so it is logically equivalent to any other mathematical truth, e.g., the Knaster-Tarski theorem. The gambling mathematician is aware of the theorem, as well as the fact that all mathematical truths are necessarily equivalent. Yet, the Knaster-Tarski theorem is not the reason why the gambler bets on 4. It seems like the difference also lies in the relation of the propositions to the gambler's decision problem, but the propositions in question are not solutions! What it means for a condition to be relevant to the agent's decision problem is an interesting question we leave for further investigations.

5.2.3 Appendix: soundness and completeness of Log

We prove completeness via a canonical model construction. Our proof closely resembles the completeness proof in [Gio19] except for the construction of the problem-sensitive components. We follow [Sie21] for the construction of P^c and $\oplus^c$, and the construction of f^c is similar to the construction of awareness sets in awareness logics [FH87].

Consistency, maximal consistency, and derivability for Log are defined standardly [BRV01].

5.2.7. Lemma. *The following are derivable in* Log*:*

1. $Int\overline{\varphi} \leftrightarrow Int(\bigwedge_{p \in Var(\varphi)} \overline{p})$
2. $Int\overline{\varphi} \rightarrow Int\overline{\psi}$, *if* $Var(\psi) \subseteq Var(\varphi)$
3. $(Int\varphi \wedge Int\psi) \rightarrow Int(\varphi \vee \psi)$

Proof:

(1): Follows from Ax4. (2): Follows from Theorem 5.2.7.1. (3): □

1.	$\Box(\varphi \rightarrow (\varphi \vee \psi)) \rightarrow ((Int\varphi \wedge Int(\overline{\varphi \vee \psi})) \rightarrow Int(\varphi \vee \psi))$	Ax5
2.	$\varphi \rightarrow (\varphi \vee \psi)$	CPL
3.	$\Box((\varphi \wedge \psi) \rightarrow (\varphi \vee \psi))$	CPL, $\mathsf{S5}_\Box$
4.	$(Int\varphi \wedge Int\psi) \rightarrow Int(\varphi \wedge \psi)$	Ax4
5.	$Int(\varphi \wedge \psi) \rightarrow Int(\overline{\varphi \vee \psi})$	Ax2
6.	$(Int\varphi \wedge Int\psi) \rightarrow Int(\overline{\varphi \vee \psi})$	4, 5, MP
7.	$(Int\varphi \wedge Int\psi) \rightarrow Int\varphi$	CPL
8.	$(Int\varphi \wedge Int\psi) \rightarrow (\Box(\varphi \rightarrow (\varphi \vee \psi)) \wedge Int\varphi \wedge Int(\overline{\varphi \vee \psi}))$	(3, 6, 8, CPL)
9.	$(Int\varphi \wedge Int\psi) \rightarrow Int(\varphi \vee \psi)$	1, 8, MP

5.2.8. Lemma. *For every maximally consistent set (mcs)* Γ *of* Log *and* $\varphi, \psi \in \mathcal{L}_I$*, the following hold:*

1. $\Gamma \vdash \varphi$ *iff* $\varphi \in \Gamma$*,*
2. *if* $\varphi \in \Gamma$ *and* $\varphi \rightarrow \psi \in \Gamma$ *then* $\psi \in \Gamma$*,*
3. *if* $\vdash \varphi$ *then* $\varphi \in \Gamma$*,*
4. $\varphi \in \Gamma$ *and* $\psi \in \Gamma$ *iff* $\varphi \wedge \psi \in \Gamma$*,*
5. $\varphi \in \Gamma$ *iff* $\neg\varphi \notin \Gamma$*.*

Proof:
Standard. □

We sometimes use Lemma 5.2.8 without explicitly mentioning it.

5.2.9. Lemma (Lindenbaum's Lemma). *Every* Log*-consistent set can be extended to a maximally* Log*-consistent one.*

Proof:
Standard. □

Let W^c be the set of all maximally Log-consistent sets. For each $\Gamma \in W^c$, define

$$\begin{aligned}\Gamma[\Box] &:= \{\varphi \in \mathcal{L} : \Box\varphi \in \Gamma\},\\ \Gamma[I] &:= \{\varphi \in \mathcal{L}_I : I\psi \wedge \Box(\psi \to \varphi) \in \Gamma \text{ for some } \psi \in \mathcal{L}_{CPL}\}\end{aligned}$$

Moreover, we define $\sim_\Box$ and $\to^c$ on W^c, respectively, as

$$\begin{aligned}\Gamma \sim_\Box \Delta &\text{ iff } \Gamma[\Box] \subseteq \Delta,\\ \Gamma \to^c \Delta &\text{ iff } \Gamma[Int] \subseteq \Delta.\end{aligned}$$

Since $\Box$ is an S5 modality, $\sim_\Box$ is an equivalence relation [BRV01]. For any maximally Log-consistent set Γ, we denote by $[\Gamma]_\Box$ the equivalence class of Γ induced by $\sim_\Box$, i.e., $[\Gamma]_\Box = \{\Delta \in W^c : \Gamma \sim_\Box \Delta\}$.

The following lemma shows that $\to^c \subseteq \sim_\Box$.

5.2.10. Lemma. *For all $\Gamma, \Delta \in W^c$, if $\Gamma \to^c \Delta$, then $\Gamma \sim_\Box \Delta$.*

Proof:
Let $\Gamma, \Delta \in W^c$ such that $\Gamma \to^c \Delta$, i.e., that $\Gamma[Int] \subseteq \Delta$. Let $\varphi \in \Gamma[\Box]$. This means that $\Box\varphi \in \Gamma$. Then, by $\mathsf{S5}_\Box$ (since $\vdash \Box\varphi \leftrightarrow \Box(\top \to \varphi)$) and Lemma 5.2.8, we have that $\Box(\top \to \varphi) \in \Gamma$. Moreover, by Ax1, we have that $Int\top \in \Gamma$. Then, by the definition of $\Gamma[Int]$, we conclude that $\varphi \in \Gamma[Int]$. Then, by the first assumption that $\Gamma[Int] \subseteq \Delta$, we obtain $\varphi \in \Delta$. Therefore, $\Gamma \sim_\Box \Delta$. □

5.2.11. Lemma. *Given a mcs Γ, for all finite $\Phi \subseteq \Gamma[Int]$, we have $\bigwedge \Phi \in \Gamma[Int]$.*

Proof:
Let $\Phi = \{\varphi_1, \ldots, \varphi_n\} \subseteq \Gamma[Int]$. This means that, for each φ_j with $1 \leq j \leq n$, there is a $\psi_j \in \mathcal{L}_{CPL}$ such that $Int\psi_j \wedge \Box(\psi_j \to \varphi_j) \in \Gamma$. Thus, $\bigwedge_{1\leq j\leq n} Int\psi_j \wedge \bigwedge_{1\leq j\leq n} \Box(\psi_j \to \varphi_j) \in \Gamma$. Then, by Ax4, we obtain that $Int(\bigwedge_{j\leq n} \psi_j) \in \Gamma$. By $\mathsf{S5}_\Box$, we also have $\Box(\bigwedge_{j\leq n} \psi_j \to \bigwedge_{j\leq n} \varphi_j) \in \Gamma$. Therefore, $\bigwedge \Phi \in \Gamma[Int]$. □

5.2.12. Lemma. *Given a mcs Γ of* Log*, $\Gamma[I]$ is consistent.*

Proof:
Assume, toward contradiction, that $\Gamma[I]$ is not consistent, i.e., $\Gamma[I] \vdash \bot$. This means that there is a finite subset $\Phi = \{\varphi_1, \ldots, \varphi_n\} \subseteq \Gamma[I]$ such that $\vdash \bigwedge \Phi \supset \neg\varphi_j$ for some $j \leq n$. By Lemma 5.2.11, we have that $\bigwedge \Phi \in \Gamma[I]$, thus, there is a $\psi \in \mathcal{L}_{CPL}$ such that $Int\psi \in \Gamma$ and $\Box(\psi \to \bigwedge \Phi) \in \Gamma$. Since $\vdash \bigwedge \Phi \to \neg\varphi_j$, by $\mathsf{S5}_\Box$, we also have $\Box(\psi \to \neg\varphi_j) \in \Gamma$. Hence, $\neg\varphi_j \in \Gamma[Int]$ too. As $\varphi_j \in \Gamma[Int]$, we also

have a $\psi' \in \mathcal{L}_{CPL}$ with $Int\psi' \in \Gamma$ and $\Box(\psi' \to \varphi_j) \in \Gamma$. From $\Box(\psi \to \neg\varphi_j) \in \Gamma$ and $\Box(\psi' \to \varphi_j) \in \Gamma$, by $\mathsf{S5}_\Box$, we obtain that $\Box(\psi \to \neg\overline{\psi'}) \in \Gamma$. As $Int\psi' \in \Gamma$, by Ax2 and Lemma 5.2.7.2, $Int\neg\overline{\psi'} \in \Gamma$. Therefore, $Int\neg\overline{\psi'} \in \Gamma$, $\Box(\psi \supset \neg\psi') \in \Gamma$, $I\psi \in \Gamma$, by Ax5, imply that $Int\neg\psi' \in \Gamma$, contradicting the consistency of Γ: $Int\psi' \in \Gamma$ implies $\neg Int\neg\psi' \in \Gamma$, by Lemmma 5.2.7.1. Therefore, $\Gamma[Int]$ is consistent. □

Given a mcs Γ_0 of Log, the canonical model for Γ_0 is a tuple $\mathcal{M}^c = \langle [\Gamma_0]_\Box, R^c, \mathcal{P}^c, f^c, V^c\rangle$ where

- $[\Gamma_0]_\Box$ is as described above.
- $R^c = \to^c \cap ([\Gamma_0]_\Box \times [\Gamma_0]_\Box)$
- $\mathcal{P}^c = (P^c, \oplus^c, s^c)$, where
  - $P^c = \mathcal{P}(\mathsf{Prop})$
  - $\oplus^c : P^c \times P^c \to P^c$ such that for all $A, B \in \mathcal{P}(\mathsf{Prop})$, $A \oplus^c B = A \cup B$.
  - $s^c : \mathcal{L} \to \mathcal{P}(P^c)$ such that $s^c(p) = \{A \in P^c \mid p \in A\}$ for all $p \in \mathsf{Prop}$. s extends to the language $\mathcal{L}_{CPL}$ as in Definition 5.1.9.
- $f^c : [\Gamma_0]_\Box \to P^c$ such that $f^c(\Gamma) = \{p \in \mathsf{Prop} \mid I\overline{p} \in \Gamma\}$.
- $V^c : \mathsf{Prop} \to \mathcal{P}([\Gamma_0]_\Box)$ such that $V^c(p) = \{\Gamma \in [\Gamma_0]_\Box : p \in \Gamma\}$.

5.2.13. Lemma. *For all $\varphi \in \mathcal{L}_{CPL}$, $s^c(\varphi) = \{A \subseteq \mathsf{Prop} \mid Var(\varphi) \subseteq A\}$.*

Proof:
The proof follows by induction on the structure of φ. The case of constant $\top$ is trivial: $Var(\top) = \emptyset$, hence, $Var(\top) \subseteq A$ for any $A \in P^c$, i.e. $\{A \subseteq \mathsf{Prop} \mid Var(\top) \subseteq A\} = \mathcal{P}(\mathsf{Prop}) = P^c$, which by Definition 5.1.9 is equal to $s^c(\top)$. The case for atomic propositions follows directly by the defn. of s^c. Case $\varphi := \neg\psi$ follows from induction hypothesis (IH) and the fact that $s^c(\neg\psi) = s^c(\psi)$ and $Var(\neg\psi) = Var(\psi)$.

Case: $\varphi := \psi \vee \chi$:

$$\begin{aligned}
s^c(\psi \vee \chi) &= s^c(\psi) \cap s^c(\chi) && \text{(by the defn. of } s^c\text{)}\\
&= \{A \subseteq \mathsf{Prop} \mid Var(\psi) \subseteq A\} \cap \{A \subseteq \mathsf{Prop} \mid Var(\chi) \subseteq A\} && \text{(by IH)}\\
&= \{A \subseteq \mathsf{Prop} \mid Var(\psi) \cup Var(\chi) \subseteq A\} && \text{(simple set theory)}\\
&= \{A \subseteq \mathsf{Prop} \mid Var(\psi \vee \chi) \subseteq A\} && (Var(\psi \vee \chi) = Var(\psi) \cup Var(\chi))
\end{aligned}$$

Case: $\varphi := \psi \wedge \chi$:

$$\begin{aligned} s^c(\psi \wedge \chi) &= \{A \cup B \subseteq \mathsf{Prop} \mid A \in s^c(\psi), B \in s^c(\chi)\} && \text{(by the defn. of } s^c\text{)} \\ &= \{A \cup B \subseteq \mathsf{Prop} \mid Var(\psi) \subseteq A, Var(\chi) \subseteq B\} && \text{(by IH)} \\ &= \{A \subseteq \mathsf{Prop} \mid Var(\psi) \cup Var(\chi) \subseteq A\} && \text{(simple set theory)} \\ &= \{A \subseteq \mathsf{Prop} \mid Var(\psi \wedge \chi) \subseteq A\} && (Var(\psi \wedge \chi) = Var(\psi) \cup Var(\chi)) \end{aligned}$$

□

5.2.14. Corollary. *For any $A, B \in P^c$ and $\varphi \in \mathcal{L}_{CPL}$, if $A \subseteq B$ and $A \in s^c(\varphi)$, then $B \in s^c(\varphi)$.*

5.2.15. Lemma. *Given a mcs Γ_0, the canonical model $\mathcal{M}^c = \langle [\Gamma_0]_\Box, R^c, \mathcal{P}^c, f^c, V^c \rangle$ for Γ_0 is a problem-sensitive model.*

Proof:
Obviously functions and operations $s^c, \oplus^c, f^c$ are well-defined. Since $\oplus^c$ is defined as $\cup$ on $\mathcal{P}(\mathsf{Prop})$, it is idempotent, commutative, associative, and it satisfies unrestricted fusion. By Corollary 5.2.14, we know that s^c is upward closed. Seriality of R^c follows from Lemmas 5.2.12, 5.2.9, 5.2.10.

□

5.2.16. Lemma (Existence lemma). *Given a mcs Γ_0, the canonical model $\mathcal{M}^c = \langle [\Gamma_0]_\Box, R^c, \mathcal{P}^c, f^c, V^c \rangle$ for Γ_0, a world $\Delta \in [\Gamma_0]_\Box$ and a formula $\varphi \in \mathcal{L}_{CPL}$, if $\neg Int\varphi \in \Delta$, then either $Int\overline{\varphi} \notin \Delta$ or there exists a world $\Gamma \in [\Gamma_0]_\Box$, such that $\Delta R^c \Gamma$ and $\neg\varphi \in \Gamma$.*

Proof:
Suppose $\neg Int\varphi, Int\overline{\varphi} \in \Delta$. Since Δ is maximal and consistent, $Int\varphi \notin \Delta$. Then, consider the set $\Delta' = \{\neg\varphi\} \cup \Delta[Int]$. This set is consistent. To show this claim, assume otherwise. Then, $\Delta[Int] \vdash \varphi$. In other words, there exists a finitely many formulae $\psi_1, \ldots, \psi_n$ in $\Delta[Int]$ such that $\psi_1, \ldots, \psi_n \vdash \varphi$. This implies that $\Box(\bigwedge_{1 \leq i \leq n} \psi_i \to \varphi) \in \Delta$. Moreover, by Lemma 5.2.11, we also have that $\bigwedge_{1 \leq i \leq n} \psi_i \in \Delta[I]$. Thus, there is $\chi \in \mathcal{L}_{CPL}$ such that $Int\chi \wedge \Box(\chi \to \bigwedge_{1 \leq i \leq n} \psi_i) \in \Delta$. Then, by $\mathsf{S5}_\Box$, we have $Int\chi \wedge \Box(\chi \to \varphi) \in \Delta$. This, together with $Int\overline{\varphi}$ and $\mathsf{Ax5}$, implies that $Int\varphi \in \Delta$, contradicting consistency of Δ. Therefore, $\Delta' = \{\neg\varphi\} \cup \Delta[Int]$ is consistent. By Lemma 5.2.9, it can be extended to a msc Γ. Obviously $\neg\varphi \in \Gamma$. Since $\Delta[Int] \subseteq \Gamma$ and $\Gamma \in [\Gamma_0]_\Box$ (by Lemma 5.2.10), we also have $\Delta R^c \Gamma$. □

5.2.17. Lemma. *Given a mcs Γ_0, the canonical model $\mathcal{M}^c = \langle [\Gamma_0]_\Box, R^c, \mathcal{P}^c, f^c, V^c \rangle$ for Γ_0, a world $\Delta \in [\Gamma_0]_\Box$ and a formula $\varphi \in \mathcal{L}_{CPL}$, $Int\overline{\varphi} \in \Delta$ iff $Var(\varphi) \subseteq f^c(\Delta)$.*

Proof:
$Int\overline{\varphi} \in \Delta$ iff I $(\bigwedge_{p \in Var(\varphi)} \overline{p}) \in \Delta$ (by Lemma 5.2.7.1) iff $Int\overline{p} \in \Delta$ for all $p \in Var(\varphi)$ (by $\mathsf{Ax4}$) iff $p \in f^c(\Delta)$ for all $p \in Var(\varphi)$ (by the defn. of f^c) iff $Var(\varphi) \subseteq f^c(\Delta)$.

□

5.2.18. Lemma (Truth lemma). *Let Γ_0 be a mcs of Log and $\mathcal{M}^c = \langle [\Gamma_0]_\Box, R^c, \mathcal{P}^c, f^c, V^c \rangle$ be the canonical model for Γ_0. Then, for all $\varphi \in \mathcal{L}_I$ and $\Delta \in [\Gamma_0]_\Box$, we have $\mathcal{M}^c, \Delta \models \varphi$ iff $\varphi \in \Delta$.*

Proof:
By induction on φ. Cases of propositional atoms, Boolean connectives and $\Box$ are proven as usual.

Case of $\varphi := Int\psi$.

($\Leftarrow$) Assume $Int\psi \in \Delta$. Then, by $\mathsf{Ax2}$, $Int\overline{\psi} \in \Delta$. This means, by Lemma 5.2.17 that $Var(\psi) \subseteq f^c(\Delta)$. By Lemma 5.2.13, this means that $f^c(\Delta) \in s^c(\varphi)$. Now let $\Gamma \in [\Gamma_0]_\Box$ such that $\Delta R^c \Gamma$, i.e., that $\Delta[Int] \subseteq \Gamma$. Observe that $Int\psi \in \Delta$ implies that $\psi \in \Delta[Int]$ (since $\Box(\psi \to \psi) \in \Delta$, by $\mathsf{S5}_\Box$). By $\Delta[Int] \subseteq \Gamma$, this implies that $\psi \in \Gamma$. Then, by IH, we obtain that $\mathcal{M}^c, \Gamma \models \psi$. As Γ has been chosen arbitrarily, we obtain that $R^c(\Delta) \subseteq [\![\psi]\!]_{\mathcal{M}^c}$. Putting everything together, we conclude $\mathcal{M}^c, \Delta \models Int\psi$.

($\Rightarrow$) Assume $\mathcal{M}^c, \Delta \models Int\psi$. Then, by the semantics, (1) $f^c(\Delta) \in s^c(\psi)$ and (2) for all $\Gamma \in [\Gamma_0]_\Box$ such that $\Delta R^c \Gamma$, $\mathcal{M}^c, \Gamma \models \psi$. By (1) and Lemma 5.2.13, we know that $Var(\psi) \subseteq f^c(\Delta)$. This implies, by Lemma 5.2.17, that $Int\overline{\psi} \in \Delta$. Now, toward contradiction, assume $Int\psi \notin \Delta$, i.e. $\neg Int\psi \in \Delta$. By Lemma 5.2.16, there exists a world $\Gamma \in [\Gamma_0]_\Box$, such that $\Delta R^c \Gamma$ and $\neg\psi \in \Gamma$. By IH, we have $\mathcal{M}^c, \Gamma \models \neg\psi$, implying that $\mathcal{M}^c, \Delta \not\models Int\psi$, which contradicts our initial assumption. Hence, $Int\psi \in \Delta$. □

5.2.19. Theorem (Soundness and completeness). *Given any set of formulae $\Delta \subseteq \mathcal{L}$ and any formula $\varphi \in \mathcal{L}$, the following holds:*

$$\Gamma \vdash \varphi \text{ iff } \Gamma \models \varphi$$

Proof:
($\Rightarrow$) By a routine check of axiom validities and validity preservation under the rules of inference of Log.

($\Leftarrow$) By contraposition. Assume $\Gamma \nvdash \varphi$. Then, $\Gamma \cup \{\neg\varphi\}$ is Log-consistent. By Lemma 5.2.9, there exists a maximal consistent set $\Gamma_0 \subseteq \mathcal{L}$, such that $\Gamma \cup \{\neg\varphi\} \subseteq$

Γ_0. Then, we can build a canonical model $\mathcal{M}^c = \langle [\Gamma_0]_\Box, R^c, \mathcal{P}^c, f^c, V^c \rangle$. Since $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma_0$, by Lemma 5.2.18, $\mathcal{M}^c, \Gamma_0 \models \Gamma \cup \{\neg\varphi\}$. By Lemma 5.2.15, $\mathcal{M}^c$ is a problem-sensitive model. Hence, there exists a pointed problem-sensitive model, where Γ is true and φ is false, i.e. $\Gamma \not\models \varphi$. □

5.2.4 Appendix: soundness and completeness of Log_G

5.2.20. Definition (Pseudo-models). A pseudo-model is a tuple:

$$M = (W, \{R_G\}_{G \subseteq Ags}, \mathcal{P}, f, V),$$

such that:

1. $W \neq \emptyset$ is a set of possible worlds;
2. $\{R_G\}_{G \subseteq Ags}$ is a collection of binary relations, such that for every $i \in Ags$, $R_{\{i\}}$ is serial and for any $G \subseteq G' \subseteq Ags$, $R_{G'} \subseteq R_G$;
3. $\mathcal{P} = (P, \epsilon, \oplus, s)$ is a problems model as in Def. 5.2.2;
4. $f : (2^{Ags} \times W) \to P$ is a function that takes a group, a world and returns a decision problem that the group is solving in the world. f satisfies the properties (Ags-Add) from Def. 5.2.3;
5. $V : \mathsf{Prop} \to 2^W$ is a valuation function.

$\mathcal{L}_G$ is interpreted over pseudo-models in a slightly different manner. All clauses are as in Def. 5.2.4, except of Int_G. Namely, for any pointed pseudo-model (M, w):

$$M, w \models Int_G\alpha \Leftrightarrow \forall v \in W(wR_Gv \Rightarrow M, v \models \alpha) \text{ and } f_G(w) \in s(\alpha)$$

i.e., Int_G is interpreted via R_G, not $\bigcap_{i \in G} R_i$.

In other words, the only difference between multi-agent problem-sensitive models and pseudo-models lies in the relations between singleton relations R_i and group relations R_G. According to Def. 5.2.3, multi-agent problem sensitive models are pseudo-models, such that $R_G = \bigcap_{i \in G} R_i$, while in pseudo-models in general, $R_G \subseteq \bigcap_{i \in G} R_i$. In other words, in pseudo-models group relations may be *super-additive*, while in "real" models they are *additive*.

5.2.21. Lemma. *Every pseudo-model as in Def 5.2.20 is $\mathcal{L}_G$-invariant to some multi-agent problem-sensitive model as in Def 5.2.3.*

Proof:
Take any pseudo-model $M = (W, \{R_G\}_{G \subseteq Ags}, \mathcal{P}, f, V)$. By a path on M we mean a finite non-empty string (potentially consisting of only one symbol) of the form

$\vec{w} = (w_1, G_1, \ldots, G_{n-1}, w_n)$, such that $w_1, \ldots, w_n \in W$, $G_1, \ldots, G_{n-1} \subseteq Ags$ and for every pair w_i, w_{i+1}: $(w_i, w_{i+1}) \in R_{G_i}$. By $tail(\vec{w})$ we denote the last world that appears in $\vec{w}$ (i.e. $tail((w_1, G_1, \ldots, G_{n-1}, w_n)) = w_n$). Note that for any world $w \in W$, (w) is also a path.

Then, we define an *unraveling* $\mathfrak{M} = (\mathfrak{W}, \{\mathfrak{R}_i\}_{i \in Ags}, \mathcal{P}, \mathfrak{f}, \mathfrak{V})$ of the pseudo-model M, where:

- $\mathfrak{W} = \{\vec{w} \mid \vec{w} \text{ is a path on } M\}$;
- $\mathfrak{R}_i \subseteq \mathfrak{W} \times \mathfrak{W}$ is a binary relation, s.t. $\vec{w}\mathfrak{R}_i\vec{v}$ iff $\vec{v}$ extends $\vec{w}$ with one i-step, i.e., if $\vec{v} = (\vec{w}, J, v)$ for some $J \subseteq Ags$, $v \in W$, such that $i \in J$;
- $\mathcal{P}$ is the same problems model as in M;
- $(\mathfrak{f} : 2^{Ags} \times \mathfrak{W}) \to P$ is a function, such that $\mathfrak{f}_G(\vec{w}) = f_G(tail(\vec{w}))$;
- $\mathfrak{V} : \mathsf{Prop} \to 2^{\mathfrak{W}}$, such that for any $p \in \mathsf{Prop}$, $\mathfrak{V}(p) = \{\vec{w} \in \mathfrak{W} \mid tail(\vec{w}) \in V(p)\}$;

It is straightforward to check that $\mathfrak{M}$ is a multi-agent problem-sensitive model.

Next, we are to show that M and $\mathfrak{M}$ are $\mathcal{L}_G$-invariant. Namely, for any possible world $w \in W$, any path $\vec{w} \in \mathfrak{W}$, such that $tail(\vec{w}) = w$ and any formula $\varphi \in \mathcal{L}_G$, $M, w \models \varphi$ iff $\mathfrak{M}, \vec{w} \models \varphi$. Cases for the Boolean and $\Box$-formulae are trivial, so we explicitly show it only for the case when $\varphi = Int_G\psi$.

($\Rightarrow$) Assume $M, w \models Int_G\psi$. Then, $R_G(w) \subseteq [\![\psi]\!]_M$ and $f_G(w) \in s(\psi)$. Then, since $\mathfrak{f}_G(\vec{w}) = f_G(w)$, $\mathfrak{f}_G(\vec{w}) \in s(\psi)$. For contradiction, assume that $\bigcap_{i \in G} \mathfrak{R}_i(\vec{w}) \not\subseteq [\![\psi]\!]_{\mathfrak{M}}$. Then, there exists a path $\vec{v}$, such that $\vec{w}\mathfrak{R}_i\vec{v}$ for every $i \in G$ and $\mathfrak{M}, \vec{v} \not\models \psi$. Let $tail(\vec{v}) = v$. By IH, $M, v \not\models \psi$. Since $\vec{w}\mathfrak{R}_i\vec{v}$ for al $i \in G$, from definition of $\mathfrak{R}_i$ it follows that $\vec{v} = (\vec{w}, J, v)$, where $G \subseteq J \subseteq Ags$. Then, from the definition of a path, $tail(\vec{w})R_Jv$, i.e., wR_Jv. Since $G \subseteq J$, by Def. 5.2.20, $R_J \subseteq R_G$. Then, wR_Gv. But $M, v \not\models \psi$, while $R_G(w) \subseteq [\![\psi]\!]_M$. Contradiction. Hence, $\bigcap_{i \in G} \mathfrak{R}_i(\vec{w}) \subseteq [\![\psi]\!]_{\mathfrak{M}}$, and, as we have already shown, $\mathfrak{f}_G(\vec{w}) \in s(\psi)$. Then, by Def. 5.2.4, $\mathfrak{M}, \vec{w} \models Int_G\psi$.

($\Leftarrow$) Assume that $\mathfrak{M}, \vec{w} \models Int_G\psi$. Then, $\bigcap_{i \in G} \mathfrak{R}_i(\vec{w}) \subseteq [\![\psi]\!]_{\mathfrak{M}}$ and $\mathfrak{f}_G(\vec{w}) \in s(\psi)$. By def. of $\mathfrak{f}$, $\mathfrak{f}_G(\vec{w}) = f_G(w)$, hence, $f_G(w) \in s(\psi)$. For contradiction, assume that $R_G(w) \not\subseteq [\![\psi]\!]_M$. Then, there exists a world $v \in W$, such that wR_Gv and $M, v \not\models \psi$. Then, let $\vec{v}$ be a path, such that $\vec{v} = (\vec{w}, G, v)$. By IH, $\mathfrak{M}, \vec{v} \not\models \psi$. Since $\vec{v}$ extends $\vec{w}$ with one G-step, $\vec{w}\mathfrak{R}_i\vec{v}$ for all $i \in G$. Therefore, $(\vec{w}, \vec{v}) \in \bigcap_{i \in G} \mathfrak{R}_i$ But $\bigcap_{i \in G} \mathfrak{R}_i(\vec{w}) \subseteq [\![\psi]\!]_{\mathfrak{M}}$. Contradiction. Hence, $R_G \subseteq [\![\psi]\!]_M$, and, as we have shown, $f_G(w) \in s(\psi)$. Then, by Def. 5.2.4, $M, w \models Int_G\psi$. $\Box$

Next, we proceed with the canonical model construction.

Let W^c be the set of all maximally Log_G-consistent sets. For each $\Gamma \in W^c$, define

$$\begin{aligned}\Gamma[\Box] &:= \{\varphi \in \mathcal{L}_G^k : \Box\varphi \in \Gamma\},\\ \Gamma[Int_G] &:= \{\varphi \in \mathcal{L}_G^k : Int_G\psi \wedge \Box(\psi \to \varphi) \in \Gamma \text{ for some } \psi \in \mathcal{L}_{CPL}\}.\end{aligned}$$

Moreover, we define $\sim_\Box$ and $\to_G^c$ on W^c, respectively, as

$$\begin{aligned}\Gamma \sim_\Box \Delta &\text{ iff } \Gamma[\Box] \subseteq \Delta,\\ \Gamma \to_G^c \Delta &\text{ iff } \Gamma[Int_G] \neq \emptyset \text{ and } \Gamma[Int_G] \subseteq \Delta.\end{aligned}$$

Since $\Box$ is an $\mathsf{S5}$ modality, $\sim_\Box$ is an equivalence relation [BRV01]. For any maximally Log_G-consistent set Γ, we denote by $[\Gamma]_\Box$ the equivalence class of Γ induced by $\sim_\Box$, i.e., $[\Gamma]_\Box = \{\Delta \in W^c : \Gamma \sim_\Box \Delta\}$. Lemmas 5.2.10, 5.2.11 and 5.2.12 are proved analogously for $\sim_\Box$ and $\to_G^c$.

Given a mcs Γ_0 of Log_G^k, the canonical model for Γ_0 is a tuple:

$$\mathcal{M}^c = ([\Gamma_0]_\Box, \{R_G^c\}_{G\subseteq Ags}, \mathcal{P}^c, f^c, V^c),$$

where:

- $[\Gamma_0]_\Box$ is as described above.
- $R_G^c = \to_G^c \cap ([\Gamma_0]_\Box \times [\Gamma_0]_\Box)$
- $\mathcal{P}^c = (P^c, \epsilon, \oplus^c, s^c)$, where
  - $P^c = \{A \cup \{\top\} \mid A \subseteq \mathsf{Prop}\} \cup \{\emptyset\}$, where $\epsilon = \emptyset$;
  - $\oplus^c : (P^c \times P^c) \to P^c$ such that for all $A, B \in P^c$, $A \oplus^c B = A \cup B$.
  - $s^c : \mathcal{L}_{CPL}^k \to \mathcal{P}(P^c)$ such that $s^c(x) = \{A \mid x \in A\}$ for $x \in \mathsf{Prop} \cup \{\top\}$. s extends to the language $\mathcal{L}_{CPL}^k$ as in Definition 5.1.9.
- For any $G \subseteq Ags$, $f_G^c : [\Gamma_0]_\Box \to P^c$ such that $f_G^c(\Gamma) = \{x \in \mathsf{Prop} \cup \{\top\} \mid I_G\overline{x} \in \Gamma\}$.
- $V^c : \mathsf{Prop} \to \mathcal{P}([\Gamma_0]_\Box)$ such that $V^c(p) = \{\Gamma \in [\Gamma_0]_\Box : p \in \Gamma\}$.

Lemmas analogous to Lemmas 5.2.13,5.2.16, and 5.2.18 hold for this canonical model. Their proofs are also analogous.

5.2.22. Lemma. *Given a mcs Γ_0, the canonical model:*

$$\mathcal{M}^c = ([\Gamma_0]_\Box, \{R_G^c\}_{G\subseteq Ags}, \mathcal{P}^c, f^c, V^c)$$

is a pseudo-model in accordance with Def. 5.2.20.

Proof:
Obviously functions and operations $s^c, \oplus^c, f^c$, are well-defined. Since $\oplus^c$ is defined as $\cup$, it is idempotent, commutative, associative, and it satisfies unrestricted fusion. Since for every $A \in P^c \setminus \{\epsilon\}$, $\top \in A$ and as $\epsilon = \emptyset$, $\epsilon < A$ for every $A \in P^c \setminus \{\epsilon\}$. By Corollary 5.2.14, we know that s^c is upward closed. Seriality of R_i^c for any $i \in Ags$ follows from Lemmas 5.2.9, 5.2.10, 5.2.12 and axiom $\mathsf{Ax1}$. Super-additivity of group relations, i.e. $R_J^c \subseteq R_G^c$ for any $G \subseteq J \subseteq Ags$ follows from $\mathsf{s-Add}$ axiom and the definition of $\{R_G^c\}_{G \subseteq Ags}$. Property (Ags) follows by def. of f_G^c from $\mathsf{Ax1}$ and the fact that $\top \notin \epsilon$. (Super-Add) follows from the following theorem:

$$\vdash \bigwedge_{i \in G} \bigcirc_G \rightarrow (I_i \overline{\varphi} \rightarrow I_G \overline{\varphi})$$

which is easily derivable as an instance of the axioms $\mathsf{s-Add}$.

We show that (Triv) holds for any two-agent group $\{a, b\}$: the proof can be easily extrapolated for a group of any cardinality. Assume that $R_a^c(\Delta) \cap R_b^c(\Delta) = \emptyset$ for some $\Delta \in [\Gamma_0]_\Box$. By def. of R_a^c, R_b^c, $\Delta[Int_a] \cup \Delta[Int_b] \vdash \bot$, while $\Delta[Int_a]$ and $\Delta[Int_b]$ are non-empty. Then, there exists some $Int_a\alpha_1, \ldots, Int_a\alpha_n, Int_b\beta \in \Delta$, such that $\alpha_1 \wedge \ldots \wedge \alpha_n \vdash \neg\beta$. For the sake of contradiction, assume that $Int_{\{a,b\}}\gamma \in \Delta$ for some $\gamma \in \mathcal{L}_{CPL}^k$. Then, by $\mathsf{Ax1}$, $\bigcirc_{\{a,b\}} \in \Delta$. Given that $Int_a\alpha_1, \ldots, Int_a\alpha_n \in \Delta$, by $\mathsf{Ax4}$ $Int_a(\alpha_1 \wedge \ldots \wedge \alpha_n) \in \Delta$. Since $Int_b\beta \in \Delta$, by $(\mathsf{s-Add})$ we have that $Int_{\{a,b\}}(\alpha_1 \wedge \ldots \wedge \alpha_n) \wedge Int_{\{a,b\}}\beta \in \Delta$. Since $\alpha_1 \wedge \ldots \wedge \alpha_n \vdash \neg\beta$, by rule of necessitation for $\Box$ we have that $\Box((\alpha_1 \wedge \ldots \wedge \alpha_n) \rightarrow \neg\beta)$. Then, by $\mathsf{Ax2}$, $Int_{\{a,b\}}\overline{\beta} \in \Delta$, which means that $Int_{\{a,b\}}\overline{\neg\beta} \in \Delta$. By $\mathsf{Ax5}$:

$$\Box((\alpha_1 \wedge \ldots \wedge \alpha_n) \rightarrow \neg\beta) \rightarrow ((Int_{a,b}(\alpha_1 \wedge \ldots \wedge \alpha_n) \wedge Int_{\{a,b\}}\overline{\neg\beta}) \rightarrow Int_{\{a,b\}}\neg\beta) \in \Delta$$

Then, by Modus Ponens we get that $Int_{\{a,b\}}\neg\beta \in \Delta$. But as we have shown earlier, $Int_{\{a,b\}}\beta \in \Delta$, hence, Δ contradicts $\mathsf{Ax3}$, while Δ by definition is Log_G-consistent. Contradiction. Hence, $Int_{\{a,b\}}\gamma \notin \Delta$ for any $\gamma \in \mathcal{L}_{CPL}$, from which by definition of f^c follows that $f_{\{a,b\}}^c(\Delta) = \emptyset = \epsilon$.

Finally, (G-Cons) property directly follows from Con axiom. For (Add), see Lemma 5.2.23. Therefore, we can conclude that the canonical model M^c for any mcs Γ_0 is a pseudo-model. $\Box$

5.2.23. Lemma. *(Add) holds for any canonical model M^c.*

Proof:
Before proceeding, let us fix some acronyms to be used throughout the proof. Let $f_G^c(\Delta) = A_G \subseteq \mathsf{Prop} \cup \{\top\}$, $f_i^c(\Delta) = A_i \subseteq \mathsf{Prop} \cup \{\top\}$ for every $i \in G$. For brevity, we denote $\bigcup_{i \in G} A_i$ simply as $\bigcup A_i$. Let $\mathcal{L}_{CPL}^{A_G} = \{\alpha \in \mathcal{L}_{CPL}^k \mid Var(\alpha) \subseteq A_G\}$ be a fragment of propositional language, where Prop is restricted to A_G.

Then, by Lemma 5.2.13, $s^{-1}(A_G) = \mathcal{L}_{CPL}^{A_G}$, i.e., every propositional formula built from the variables of A_G is a partial solution to A_G. Then, by definition of $\lceil\ .\ \rceil$, $\lceil A_G \rceil_{M^c} = \{[\![\Gamma]\!]_{M^c} \mid \Gamma \text{ is a mcs of } \mathcal{L}_{CPL}^{A_G}\}$.[15] Analogously, $\lceil A_i \rceil = \{[\![\Gamma]\!]_{M^c} | \Gamma \text{ is a mcs of } \mathcal{L}_{CPL}^{A_i}\}$. Since $A_i = f_i^c(\Delta)$ and $\oplus^c$ is defined as $\cup$, $\bigoplus_{i\in G} f_i^c(\Delta) = \bigcup A_i$. Consequently, $s^{-1}(\bigoplus_{i\in G} f_i^c(\Delta)) = \mathcal{L}_{CPL}^{\bigcup A_i}$ and:

$$\lceil \bigoplus f_i^c(\Delta) \rceil = \{[\![\Gamma]\!]_{M^c} \mid \Gamma \text{ is a mcs of } \mathcal{L}_{CPL}^{\bigcup A_i}\}.$$

We are now ready to proceed with the proof. Assume $A_G \neq \emptyset$. First, we show that $\lceil \bigoplus f_i^c(\Delta) \rceil \sqsubseteq \lceil f_G^c(\Delta) \rceil$. Note that from $(\mathsf{s-Add})$ axiom it follows that $\bigcirc_G \to (Int_i\overline{p} \to Int_G\overline{p}) \in \Delta$ for every $p \in \mathsf{Prop} \cup \{\top\}$. By our assumption, $A_G \neq \emptyset$. Then, at least $\top \in A_G$. Therefore, $Int_G\overline{\top} = Int_G\top = \bigcirc_G \in \Delta$. So, if $Int_i\overline{p} \in \Delta$, then $Int_G\overline{p} \in \Delta$ for every $i \in G$ and every $p \in \mathsf{Prop}$. Therefore:

$$\begin{aligned} A_G &= \{p \in \mathsf{Prop} \cup \{\top\} \mid Int_G\overline{p} \in \Delta\} \\ &\supseteq \bigcup_{i\in G} \{p \in Prop \cup \{\top\} \mid Int_i\overline{p} \in \Delta\} \\ &= \bigcup_{i\in G} f_i^c(\Delta) \\ &= \bigoplus_{i\in G} f_i^c(\Delta) \end{aligned}$$

Then, by Proposition 5.1.15, since $\bigoplus_{i\in G} f_i^c(\Delta) \leq f_G^c(\Delta)$, $\lceil \bigoplus_{i\in G} f_i^c(\Delta) \rceil \sqsubseteq \lceil f_G^c(\Delta) \rceil$.

It is left for us to prove the other direction: $\lceil f_G^c(\Delta) \rceil \sqsubseteq \lceil \bigoplus f_i^c(\Delta) \rceil$. Note that, by (AxAdd), if $Int_G\overline{\alpha} \in \Delta$, then, for some propositional formulae $\{\alpha_j^i\}_{1\leq j\leq k}^{i\in G} \subseteq \mathcal{L}_{CPL}^k$:

$$\bigwedge_{1\leq j\leq k}^{i\in G} Int_i\overline{\alpha_j^i} \wedge \Box(\alpha \leftrightarrow (\bigvee_{1\leq j\leq k} \bigwedge_{i\in G} \alpha_j^i)) \in \Delta$$

Then, for every $\alpha \in \mathcal{L}_{CPL}^{A_G}$, there are some $\alpha_1^i, \ldots, \alpha_k^i \in \mathcal{L}_{CPL}^{A_i}$ for every $i \in G$, such that $\Box(\alpha \leftrightarrow \bigvee_{1\leq j\leq k} \bigwedge_{i\in G} \alpha_j^i) \in \Delta$. If $\alpha_1^i, \ldots, \alpha_k^i \in \mathcal{L}_{CPL}^{A_i}$ for every $i \in G$, then $\bigvee_{1\leq j\leq k} \bigwedge_{i\in G} \alpha_j^i \in \mathcal{L}_{CPL}^{\bigcup A_i}$. So we can define a translation function $g : \mathcal{L}_{CPL}^{A_G} \to \mathcal{L}_{CPL}^{\bigcup A_i}$ as follows:

$$g(\alpha) = \bigvee_{1\leq j\leq k} \bigwedge_{i\in G} \alpha_j^i \text{ s.t. } \Box(\alpha \leftrightarrow \bigvee_{1\leq j\leq k} \bigwedge_{i\in G} \alpha_j^i) \in \Delta$$

[15] In Def. 5.1.13, $\lceil a \rceil = \{\Gamma \subseteq s^{-1}(a)\} \mid \Gamma$ is a maximal satisfiable subset of $s^{-1}(a)\}$. By the analogue of Lemma 5.2.18 (Truth lemma), the notions of maximal satisfiable and maximal consistent sets coincide.

Then, by the Truth lemma, for any $\alpha \in \mathcal{L}_{CPL}^{A_G}$, $[\![\alpha]\!]_{M^c} = [\![g(\alpha)]\!]_{M^c}$.

Finally, to prove that $\lceil f_G^c(\Delta)\rceil \sqsubseteq \lceil \bigoplus_{i\in G} f_i^c(\Delta)\rceil$, we need to show that for every mcs $\Gamma \subseteq S^{-1}(f_G^c(\Delta)) = \mathcal{L}_{CPL}^{A_G}$, $[\![\Gamma]\!]_{M^c} = \bigcup\{[\![\Delta_i]\!]_{M^c}\}_{i\in I}$ for some $\{[\![\Delta_i]\!]\}_{i\in I} \subseteq S^{-1}(\bigoplus f_i^c(\Delta)) = \mathcal{L}_{CPL}^{\bigcup A_i}$, where I is some arbitrary index-set. As we have shown, for every propositional formula $\alpha \in \mathcal{L}_{CPL}^{A_G}$, $[\![\alpha]\!]_{M^c} = [\![g(\alpha)]\!]_{M^c}$, where $g(\alpha) \in \mathcal{L}_{CPL}^{\bigcup A_i}$. Hence, for every mcs $\Gamma \subseteq \mathcal{L}_{CPL}^{A_G}$:

$$[\![\Gamma]\!]_{M^c} = \bigcup\{[\![\Delta]\!]_{M^c} \in \lceil \bigoplus f_i^c(\Delta)\rceil \mid \forall \alpha \in \Gamma : g(\alpha) \in \Delta\}$$

To conclude, $\lceil f_G^c(\Delta)\rceil \sqsubseteq \lceil \bigoplus f_i^c(\Delta)\rceil$ and $\lceil f_G^c(\Delta)\rceil \sqsupseteq \lceil \bigoplus f_i^c(\Delta)\rceil$, from which it follows that $\lceil f_G^c(\Delta)\rceil = \lceil \bigoplus f_i^c(\Delta)\rceil$. □

5.2.24. THEOREM (Soundness and completeness). *Given any set of formulae $\Delta \subseteq \mathcal{L}_G^k$ and any formula $\varphi \in \mathcal{L}_G^k$, the following holds:*

$$\Gamma \vdash_{\mathsf{Log}_G} \varphi \Leftrightarrow \Gamma \models_{\mathbf{M}_G} \varphi$$

Proof:
($\Rightarrow$) By a routine check of axiom validities and validity preservation under the rules of inference of Log_G.

($\Leftarrow$) By contraposition. Assume $\Gamma \nvdash_{\mathsf{Log}_G} \varphi$. Then, $\Gamma \cup \{\neg\varphi\}$ is Log_G-consistent. By Lemma 5.2.9, there exists an mcs Γ_0, such that $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma_0$. Then, we can build a canonical pseudo-model $M^c = ([\Gamma_0]_\Box, \{R_G^c\}_{G\subseteq Ags}, \mathcal{P}^c, f^c, V^c)$. Since $\Gamma \cup \{\neg\varphi\} \subseteq \Gamma_0$, by Truth Lemma it follows that $M^c, \Gamma_0 \models \Gamma \cup \{\neg\varphi\}$. By Lemma 5.2.22, M^c is a pseudo-model. By Lemma 5.2.21, there exists a multi-agent problem-sensitive model $\mathcal{M}$, such that M^c is $\mathcal{L}_G^k$-invariant to $\mathcal{M}$. In other words, there exists a world w of $\mathcal{M}$, such that $\mathcal{M}, w \models \psi$ iff $M^c, \Gamma_0 \models \psi$. Then, $\mathcal{M}, w \models \Gamma \cup \{\neg\varphi\}$. Hence, there exists a pointed multi-agent problem-sensitive model, where all formulae in Γ are true and φ is false, i.e. $\Gamma \not\models_{\mathbf{M}_G} \varphi$. □

Chapter 6

Fixing MAMLs: the case of STIT logic[1]

So far, we have demonstrated that an overly simplistic definition of groups causes both conceptual and technical problems for MAMLs (Chapter 3). We have argued that real groups should share motivational states (Chapter 2); afterwards, we have developed tools to formally define these states (Chapters 4, 5). In this chapter, we show how these tools may be integrated into MAMLs to alleviate the problems.

Standard multi-agent epistemic logics do not suffer as much from the weak definition of groups. Common knowledge ascription to a random set of agents does not imply ascribing these agents with group agency. On the contrary, in multi-agent modal logics of agency, the distinction between groups and random sets of agents is crucial. Therefore, we concentrate on MAMLs of actions and abilities. Specifically, we will work with STIT logics, because most popular multi-agent modal logics of agency can be embedded in variations of STIT logic [Ram23; BHT06].

The main problem we have with STIT logic is an aggregative definition of group actions. Given any arbitrary set of agents $G = \{i_1, \ldots, i_n\}$ and any actions σ_1 of i_1,...,σ_n of i_n, the tuple $(\sigma_1, \ldots, \sigma_n)$ is regarded as a joint action of G. Ascribing group agency to any set of agents entails holding them responsible for their "joint actions". This is philosophically problematic. Moreover, the aggregative definition overgenerates joint actions, while all of them are intertwined in a sophisticated way, which makes it overly complicated to axiomatically capture their

[1]Sections 6.2, 6.3.1 and 6.3.3 are partly based on the paper *Daniil Khaitovich. Plan-restricted group STIT logic*, which is currently under review [Kha26b]. Section 6.4 is based on the extended abstract of the talk *Daniil Khaitovich, Hein Duijf, Jan Broersen. Solving coordination games with minimal communication*, which was presented at the 16th Conference on Logic and the Foundations of Game and Decision Theory [KDB26].

properties. So we have both philosophical and technical reasons to strengthen the aggregative definition of group actions. To do so in a philosophically informed way, we take our clue from the following idea, concisely put by Raimo Tuomela:

> *In the strongest sense of acting together we require acting on a joint, agreed-upon plan.*[...] *This kind of joint action is a collective social action in its most central sense, acting as a team.* [Tuo05, p.83]

This idea – groups act upon their collective motivational states – is shared by most of the philosophers of group agency, and may be found in slightly different terms in other essential texts on the topic. The difference lies in the conceptualisations of the group motivational state: for Tuomela, it is a *joint, agreed-upon plan*; for Margaret Gilbert, it is a *joint commitment* [Gil92]; for Michael Bratman, it is an *intention in favour of joint activity* [Bra92]; for John Searle, it is *we-intention* [Sea95]. The landscape of such conceptualisations is rich, and we did our best in Chapter 2 to map it.

The main goal of this chapter is to integrate the philosophically informed definition of group actions into a standard group STIT logic and show how it alleviates both conceptual and metalogical shortcomings. We start with a detailed overview of metalogical properties of group STIT logic, going beyond the short introduction of Section 3.2.2. Then, we introduce our refinement of it, named *plan-restricted group STIT logic*, and show what conceptual and formal advantages it brings.

6.1 Group STIT logic and its technical problems

As we have briefly mentioned in Section 3.2.2, the set of validities $\mathbf{L}^t_{STIT} = \{\varphi \in \mathcal{L}^t_{STIT} \mid \models \varphi\}$ is notoriously hard (and as we are to show, actually impossible) to axiomatise. The first problem arises from the temporal component of the language. $\mathcal{L}^t_{STIT}$ embeds $\mathcal{L}_{OBTL}$, whose axiomatization, when interpreted over branching time structures, remains an open problem. If we use Kamp frames as in Def. 3.2.8, the known axiomatization of $\mathcal{L}_{OBTL}$-validities over this class of models requires us to use the non-standard irreflexivity rule [Rey02, Chapter 7]:

$$\frac{(p \wedge H\neg p) \rightarrow \varphi}{\varphi} \text{ , where } p \notin sub(\varphi).$$

Even with the atemporal fragment $\mathcal{L}_{STIT}$ interpreted over the simplified models as in Def.3.2.10, the set $\mathbf{L}_{STIT} = \{\varphi \in \mathcal{L}_{STIT} \mid \models \varphi\}$ still cannot be finitely axiomatised. We proof-sketch this fact and draw connections with mathematically similar formalisms.

First, we introduce a *product universal modal logic*, since it is going to play an important role in our story. Given a final set of indices $I = \{1, \ldots, n\}$, we define

a language $\mathcal{L}_{S5^n}$ by a following grammar:

$$\varphi := p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box_i\varphi$$

where $p \in Var$, $i \in I$. There are different possible readings of $\Box_i$-modalities, depending on how we understand the index-set I. Three different interpretations will be provided later. Next, we look into a class of *universal product frames*, over which $\mathcal{L}_{S5^n}$ is interpreted.

6.1.1. Definition ($\mathbf{F}_{S5^n}$-frames). A universal product frame is a tuple $F = (X, \{R_i\}_{i\in I})$, where:

1. $X = X_1 \times \ldots \times X_n$ for some non-empty sets $X_1, \ldots, X_n$.
2. For each $i \in I$, $R_i \subseteq X \times X$ is a binary relation, such that:

$$(x_1, \ldots, x_n)R_i(y_1, \ldots, y_n) \text{ iff } x_j = y_j \text{ for all } j \in I \setminus \{i\}$$

Standardly, every universal product frame F can be extended to a model $M = (F, V)$ with a standard evaluation function $V : Var \to 2^X$.

$\mathcal{L}_{S5^n}$ is interpreted over universal product models in a standard way, as in Def. 3.0.3. That is, for any pointed $\mathbf{F}_{S5^n}$-model (M, w), $M, w \models \Box_i\varphi$ iff $\forall v \in X(wR_iv \Rightarrow M, v \models \varphi)$.

Universal product frames were shown to correspond to *cylindric set algebras*.

6.1.2. Definition (Cylindric set algebra). A cylindric set algebra of dimension n is a tuple

$$\mathcal{A} = (A, \cup, \cap, \setminus, \emptyset, X^n, \{c_i\}_{i\in\{1,\ldots,n}\}),$$

where:

1. (A, X^n) is the field of sets. In other words, $A \subseteq 2^{X^n}$, such that $X^n \in A$ and A is closed under taking unions and complements with respect to X^n.
2. $(A, \cup, \cap, \setminus, \emptyset, X^n)$ is a Boolean algebra on sets, where $\emptyset$ is 0, X^n is 1, $A \subseteq 2^{X^n}$ is the domain, and $\cup, \cap, \setminus$ have their standard set-theoretic meaning;
3. For every $i \in \{1, \ldots, n\}$, $c_i : 2^{X^n} \to 2^{X^n}$ is an unary operation, such that:

$$c_i(U) = \{(x_1, \ldots, x_n) \in X^n \mid \exists(y_1, \ldots, y_n) \in U : x_j = y_j \text{ for all } j \in \{1, \ldots, n\} \setminus \{i\}\}$$

Think of X in Def.6.1.1 as an n-dimensional space, where for every $i \in I$, X_i is a set of coordinates of ith dimension. Then, every formula φ denotes some region of this space $[\![\varphi]\!]_M \subseteq X$. $\Diamond_i$ then may be interpreted as the *cylindrification* operator

c_i: given a region $[\![\varphi]\!]_M$, we compute $[\![\Diamond_i\varphi]\!]_M$ by taking all the points of X that differ from some point in $[\![\varphi]\!]_M$ in at most ith coordinate:

$$[\![\Diamond_i\varphi]\!]_M = \{(x_1, \ldots, x_n) \in X \,|\, \forall j \in I \setminus \{i\} : x_j = y_j \text{ for some } (y_1, \ldots, y_n) \in [\![\varphi]\!]_M\}$$

For an illustration of what cylindrification does, see the Figure 6.1.

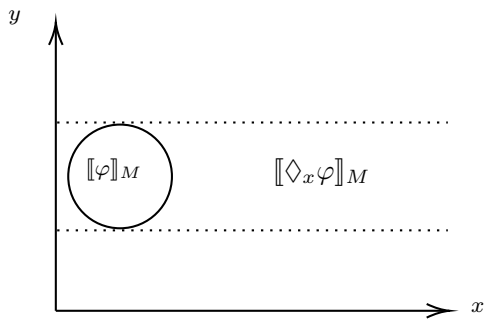

Figure 6.1: x-cylindrification of the region $[\![\varphi]\!]_M$ of the two-dimensional space. Doing the same operation in three-dimensional space may provide a good explanation as to why it is called *cylindrification*.

The equational theory of cylindric set algebras is not finitely axiomatizable [Kur13]. Moreover, "cylindric algebras stand in the same relationship to first-order logic as Boolean algebras stand to sentential logic"[Mon76, p.212]. To give an intuition about the connection between cylindric set algebras (and hence, modal logic of universal product frames) and first-order logic, consider a first-order language $\mathcal{L}_{FO}$ with n-many variables $I = \{x_1, \ldots, x_n\}$. Take any standard Tarskian model M for it, where D is the domain of M. For any $\varphi \in \mathcal{L}_{FO}$, let $[\![\varphi]\!]_M = \{\vec{x} \in D^I | M, \vec{x} \models \varphi\}$ be a set of variable assignments that satisfy φ. Let $A = \{[\![\varphi]\!]_M \,|\, \varphi \in \mathcal{L}_{FO}\}$. Then, observe that (A, D^I) is the field of sets, where for every $[\![\varphi]\!]_M \in A$ and any $x \in I$, $c_x([\![\varphi]\!]_M) = [\![\exists x\varphi]\!]_M$.

The connections between the modal logic of universal product frames, cylindric set algebras, and first-order logic that we have demonstrated in a hand-wavy way have been formally proven. The algebraization of first-order logic via cylindric algebras was established by Henkin, Monk and Tarski [HMT71]. The correspondence between restricted first-order logic with finitely many variables and the modal logic of universal product frames is due to [Ven95], and between cylindric set algebras and the modal logic of universal product frames – due to [Kur13]. Having these correspondences at hand, one can transfer the known metatheoretical results about cylindric set algebras and first-order logic to the modal logic of universal product frames. Namely, we know that if $Card(I) \geq 3$, neither the logic $\{\varphi \in \mathcal{L}_{S5^n} \mid \models \varphi\}$ is axiomatizable, nor the satisfiability problem of $\mathcal{L}_{S5^n}$ is decidable [Kur+03].

There are two ways in which we can mitigate the negative results for cylindric set algebras and first-order logic. We can either generalise the semantics or restrict the language, obtaining better-behaved fragments. Both paths were fruitfully explored [Grä99; ANB98; Ven95]. As for semantic generalisations, the technique of relativisation allows us to recover decidability and finite axiomatisability. Namely, instead of taking the field of sets (A, X^n) for all possible tuples X^n, we can take an arbitrary subset of such tuples $V \subseteq X^n$.

6.1.3. Definition (Cylindric-relativised set algebras). Let $\emptyset \neq V \subseteq X^n$ for some $X \neq \emptyset$, and let (A, V) be a field of sets. Then, *cylindric-relativised set algebra* of dimension n is the algebra

$$\mathfrak{A} = (A, \cup, \cap, \setminus, \emptyset, V, \{c_i^V\}_{i \in \{1,\ldots,n\}}),$$

where $\cup, \cap, \setminus, \emptyset$ have their standard set-theoretical meaning, and for any $i \in \{1, \ldots, n\}$, $U \subseteq V$:

$$c_i^V(U) = \{(x_1, \ldots, x_n) \in V | \exists (y_1, \ldots, y_n) \in U : x_j = y_j \text{ for all } j \in \{1, \ldots, n\} \setminus \{i\}\}$$

The equational theory of cylindric-relativised set algebras is strictly weaker than that of cylindric set algebras. Moreover, it is finitely axiomatisable and decidable [HMT71, Chapter 2.2]. It also has a natural connection with *general assignment models*, where we restrict the set of all possible variable assignments D^I to some arbitrary non-empty subset $V \subseteq D^I$. This allows us to model dependencies between variables [Abr+16], as well as restrictions in quantifier domains as they often occur in natural language [SG00]. To give an intuition about both cases, assume we have two variables $I = \{x, y\}$, where x denotes a day and y – a month. Possible interpretations of x and y range over the sets $Day = \{1, \ldots, 31\}$ and $Month = \{January, \ldots, December\}$, respectively. Not all pairs in $Day \times Month$ make sense: e.g., 30th of February should never occur. In the natural language, we take it as a background of our reasoning and omit explicit mentions of the restriction: e.g., one can say "At any day of any month, I am happy to help you", obviously meaning only existing dates. To formalise this, we can build a model with a restricted domain $V \subseteq Day \times Month$. Alternatively, we can define a relation $Date \subseteq Day \times Month$, specifying in the definition of a language that every quantified sentence should be *guarded by Date*: e.g., given an atomic formula φ, $\forall x \forall y(Date(x, y) \rightarrow \varphi)$ would be a well-formed formula of such a restricted language, while $\forall x \forall y \varphi$ would not. For a detailed exposition of the connection between cylindric-relativised set algebras, generalised assignment models and the guarded fragment of first-order logic, see [Ben13].

But what does all of this have to do with STIT logic? $\mathcal{L}_{S5^n}$ can be embedded into $\mathcal{L}_{STIT}$. In [HS08], this correspondence between $\mathcal{L}_{S5^n}$ and $\mathcal{L}_{STIT}$ is formally

demonstrated. From this correspondence, one can reduce decidability and axiomatizability of $\mathbf{L}_{S5^n}$ to the same notions for group STIT logic, obtaining negative results: for $Card(Ags) \geq 3$, $\mathbf{L}_{STIT} = \{\varphi \in \mathcal{L}_{STIT} \mid \models \varphi\}$ is not finitely axiomatizable, and the satisfiability problem for $\mathcal{L}_{STIT}$ is undecidable [HS08, Theorems 22, 23].

The following correspondence will be the most important for us:

$$CSA_n \simeq \mathcal{L}_{S5^n} \simeq \mathcal{L}_{STIT}$$

where CSA_n is the variety of n-dimensional cylindric set algebras. Essentially, it gives us a clue that some techniques applied for cylindric set algebras to obtain a better metatheoretical behaviour may also work for group STIT logic. As we are to show, the relativisation has a surprisingly clear connection with Tuomela's definition of joint action.

6.2 Plan-restricted group STIT logic

This section is based on [Kha26b].

In this chapter, we develop a *plan-restricted group STIT logic* – a variation of group STIT logic in which we can explicitly distinguish between genuine group actions and aggregations of actions by independent individuals. After introducing its syntax and semantics, we show its conceptual advantages by revising definitions of causal responsibility for individuals and groups.

For a finite set of agents $Ags = \{1, \ldots, n\}$, by $Gr = \{G \subseteq Ags \mid Card(G) \geq 2\}$ we denote the collection of non-empty non-singleton sets of agents. Then, we recursively define the formal language $\mathcal{L}^t_{pSTIT}$ by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid X\varphi \mid Y\varphi \mid G\varphi \mid H\varphi \mid \Box\varphi \mid [stit]_i\varphi \mid [stit]_C\varphi \mid \varrho_C$$

where $p \in Var$, $i \in Ags$, $C \in Gr$. Propositional variables, Boolean connectives, X, Y, G, H, $\Box$, $[stit]_i$ and $[stit]_C$-modalities have their standard meaning, as in $\mathcal{L}^t_{STIT}$. ϱ_C reads as "*C acts upon their plan*". We use the following acronyms. $[pstit]_C\varphi := \varrho_C \wedge [stit]_C\varphi$. $[pstit]_C\varphi$ reads as "*C acts upon their plan and sees to it that φ*", or, for brevity, "*C jointly sees to it that φ*". Jointly seeing to it that φ corresponds to Tuomela's understanding of seeing to it that φ as a "collective social action in its most central sense". As with $\mathcal{L}^t_{STIT}$, the atemporal fragment of $\mathcal{L}^t_{pSTIT}$ is denoted as $\mathcal{L}_{pSTIT}$.

Semantically, to interpret $\mathcal{L}^t_{pSTIT}$, we need to equip temporal Kripke STIT frames with some representation of group plans. We understand plans in line with the following idea of Bratman:

> [Plans] *constrain solutions to* [group decision] *problems, providing a filter of admissibility for options.* [...] *They narrow the scope of the deliberation to a limited set of options.* [Bra87, p.33]

To provide a "filter of admissibility", a group plan should limit the group's choice of present and future joint actions. Therefore, for any non-empty non-singleton set of agents $C \in Gr$, we can abstractly represent C's plan as a subset of C's possible current and future actions that adhere to that plan. In case this set is empty, C has no plan and does not constitute a group.[2]

To introduce this idea in STIT framework, we will associate every non-singleton subset of agents $C \in Gr$ with some possibly empty set of time points $U_C \subseteq W$, where members of C act in accordance with their joint, agreed-upon plan. If members of C are not a group, i.e., they do not coordinate their actions in any way, then the corresponding set of worlds will be empty. In this case, any behaviour of C's members is just the simultaneous acting of individuals rather than the acting of a group.

We will impose a number of constraints on these sets of points. First, U_C should be $\bigcap_{i \in C} R_i$-closed: if $w \in U_C$, wR_iv for all $i \in C$, then $v \in U_C$. Informally, the group plan is a subset of members' available actions: whether a group's plan will be fulfilled is only up to the group members' choice of actions. Second, as in Section 5.2, we assume that groups with shared members have non-contradictory plans.

6.2.1. Definition (**pSTIT**-models). A tuple $F = (W, R_\prec, R_\Box, \{R_i\}_{i \in Ags}, \{U_C\}_{C \in Gr})$ is a **pSTIT**-frame iff:

1. $(W, R_\prec, R_\Box, \{R_i\}_{i \in Ags})$ is a Kripke STIT frame in accordance with Definition 3.2.8. Standardly, $R_G = \bigcap_{i \in G} R_i$ for every $\emptyset \neq G \subseteq Ags$;
2. For every $C \in Gr$, $U_C \subseteq W$ is a possibly empty set of time points, where C acts in accordance with their plan. $\{U_C\}_{C \in Gr}$ have the next properties:
   (a) Every U_C is a R_C-closed subset of worlds.[3]
   (b) For any collection of non-empty non-singleton sets $C_1, \ldots, C_n \in Gr$, such that $C_1 \cap \ldots \cap C_n \neq \emptyset$, if $R_\Box(w) \cap U_{C_1} \neq \emptyset, \ldots, R_\Box(w) \cap U_{C_n} \neq \emptyset$, then $R_\Box(w) \cap U_{C_1} \cap \ldots \cap U_{C_n} \neq \emptyset$.

As usual, a **pSTIT**-frame F is extended to a model $M = (F, V)$ with a standard evaluation function $V : Var \to 2^W$.

Condition 2a expresses that C's plan is a subset of C's shared actions. Condition

[2] A similar way to formalise group plans in a game-theoretic setting was used in [Dui+21].

[3] The subset of worlds $U_C \subseteq W$ is R_C-closed iff the following holds: $\forall w \in U_C \forall v \in W(wR_Cv \Rightarrow v \in U_C)$.

2b expresses that groups with shared members have consistent plans. As with $\mathcal{L}^t_{STIT}$, when working with the atemporal fragment $\mathcal{L}_{pSTIT}$, we can simplify the model by dropping the relation $R_{\prec}$.

A natural question arises: why do we have the definition of $U_C \subseteq W$ only for non-empty non-singleton sets $C \in Gr$? Individual agents can also have plans and act either in accordance with or against them. The key difference lies in the relation between causal agency and plans for individuals and groups. If the agent acts against her plan, e.g., as an instance of her weak will, it is still *her* action. On the contrary, when at least some members of a group act against the plan of the group, it undermines the group agency. For example, think of two friends who decided to organise a running club and agreed to go on weekly runs every Sunday at 10:00 at a local park. Every time they show up at the given time and place and go for a run, we can count it as a group action of the running club. But if they do not show up for a run and stay home without notifying one another, it seems strange to regard it as a group action of their club: at this point, those are just two individuals who defy their mutual agreement. On the contrary, if a single agent planned to go for a run but ended up staying home, it is still the action of that agent.

Another way to look at the difference between individual and group agency is through *independence*. Individual agents are indeed independent: for every agent $i \in Ags$ and every available action of this agent σ_i, no matter what others are going to do, σ_i is still available for i. As group members, agents are not necessarily independent: restricting group members' choices by plan may force dependence. For example, the members of the running club may agree that everyone should show up, unless the organizer notifies about the cancellation of the event. Each member as an individual can either show up for a run or not, independently of what organizer is going to do. But to act as a group, everyone should show up for a run unless the organizer chooses to cancel the event.

6.2.2. Definition. For a **pSTIT**-model $M = (W, R_{\prec}, R_{\Box}, \{R_i\}_{i\in Ags}, \{U_C\}_{C\in Gr}, V)$ and a point $w \in W$, we recursively define $\models$ relation for $\mathcal{L}^t_{pSTIT}$ as follows:

$$M, w \models \varrho_C \Leftrightarrow w \in U_C$$

and all other clauses are the same as in Def. 3.2.9.

From Def. 6.2.2 it follows that $M, w \models [pstit]_C\varphi$ iff $M, w \models [stit]_C\varphi$ and $w \in U_C$. Consequently, if $U_C = \emptyset$, then $M \models \neg[pstit]_C\varphi$ for any $\varphi \in \mathcal{L}^t_{pSTIT}$. If C does not constitute a group, then C does not perform group actions and does not jointly see to it that anything holds whatsoever.

To show what we gain conceptually by distinguishing group actions from aggregations of individual actions, we revise the definition of causal responsibility. For

individuals, the definition of causal responsibility stays the same as in [Ser21]: an agent $i \in Ags$ is causally responsible for φ iff $[stit]_i\varphi \wedge \neg\Box\varphi$. Or, following the previously accepted acronym, the agent is causally responsible for φ iff the agent deliberately sees to it that φ:

$$[dstit]_i\varphi := \neg\Box\varphi \wedge [stit]_i\varphi$$

For non-empty non-singleton sets, the definition is strengthened. To hold a set of agents $C \in Gr$ causally responsible for φ, we require not only that actions of C is a minimal sufficient condition for φ to obtain, but also that members of C act upon some common group plan. For every $C \in Gr$, let $\Sigma^{min}_C\varphi$ stand for " C *is causally responsible for* φ". Then, formally:

$$\Sigma^{min}_C\varphi := \neg\Box\varphi \wedge [stit]_C\varphi \wedge \bigwedge_{C' \subset C} \neg[stit]_{C'}\varphi \wedge \bigvee_{C \subseteq C'} \varrho_{C'}$$

Or, informally: $\Sigma^{min}_C\varphi$ iff φ is not necessary, C sees to it that φ, no proper subgroup of C sees to it that φ, and agents in C act upon some joint agreed-upon plan, hence, coordinate their actions.

Consequently, the agent i *bears collective causal responsibility for* φ iff i is a member of some set C that is causally responsible for φ. In other words, i acts as a member of some group that ensures φ, while actions of i are part of the minimal sufficient condition for that group to ensure φ. Let $colresp_i\varphi$ stands for "i bears collective responsibility for φ". Then, formally:

$$colresp_i\varphi := \bigvee_{C \in \uparrow\{i\}} \Sigma^{min}_C\varphi$$

.

where $\uparrow\{i\} = \{C \subseteq Ags \mid i \in C\}$. By such a strengthening of causal collective responsibility's definition, we get one obvious improvement: to bear collective responsibility for φ, an agent should be a member of a group that jointly sees to it that φ.

Further, we define the weaker form of collective responsibility: *complicity*. In line with [Baz13], the agent is complicit in bringing about that φ iff the agent acts as a member of a group that jointly sees to it that φ, regardless of whether actions of that agent are necessary to ensure φ. For every agent $i \in Ags$, let $compl_i\varphi$ stands for "i *is complicit in* φ". Then:

$$compl_i\varphi := \neg\Box\varphi \wedge \bigvee_{C \in \uparrow\{i\}} [pstit]_C\varphi$$

It is not hard to see that if i bears collective causal responsibility for φ, then i is complicit in φ, but not the other way around. To distinguish between those

who bear collective causal responsibility and those who are only complicit, we say that the agent is *merely complicit in* φ, denoted as $mcompl_i\varphi$, iff she is complicit in φ, but does not bear collective responsibility for it:

$$mcompl_i\varphi := compl_i\varphi \wedge \neg colresp_i\varphi$$

In the existing literature on STIT-based models of responsibility, complicity is overlooked, as merely participating agents are labeled as *free-riders* and escape any form of responsibility [Ben12, Chapter 5], [Ser21]. This comes as no surprise, given that the authors rely on the aggregative definition of group actions. In standard group STIT logic, the following is valid: $\models [stit]_{C_1}\varphi \to [stit]_{C_1 \cup C_2}\varphi$ for any $C_1, C_2 \subseteq Ags$. If some set of agents sees to it that φ, then any of its supersets sees to it that φ as well. Then, as soon as at least someone forces φ, everybody else merely participates in the aggregative "joint action" that forces φ, and we cannot distinguish whether there is even a real joint action to participate in.

Finally, the agent is not responsible for φ, denoted as $notresp_i\varphi$, iff she is neither causally responsible for φ, nor bears collective responsibility for φ, nor merely complicit in φ:

$$notresp_i\varphi := \neg[dstit]_i\varphi \wedge \neg colresp_i\varphi \wedge \neg mcompl_i\varphi$$

For brevity, while talking about different forms of causal responsibility, we will omit "causal" in the corresponding terms, since we do not talk about any non-causal forms of responsibility so far.

To see these definitions at work, let us go back to Example 3.2.11 and build an atemporal **pSTIT**-model for it. Let $Ags = \{a, b, c\}$ be the same set of agents: the killer, the lookout, and the passerby, respectively. Consequently, we have four non-empty non-singleton sets of agents: $Gr = \{\{a, b\}, \{a, c\}, \{b, c\}, Ags\}$. We use the same set of propositional variables: $Var = \{k, l, p\}$, meaning *someone was **k**illed*, *the area was guarded by the **l**ookout*, and *someone has **p**assed by the area*, respectively. Then, let $M = (W, R_\Box, R_a, R_b, R_c, \{U_C\}_{C \in Gr}, V)$, be an atemporal **pSTIT**-model, where:

1. $W = \{w_1, \ldots, w_8\}$;
2. $R_\Box = W \times W$;
3. $[R_a] = \{\{w_1, w_2, w_3, w_4\}, \{w_5, w_6, w_7, w_8\}\}$;
4. $[R_b] = \{\{w_1, w_2, w_6, w_7\}, \{w_3, w_4, w_5, w_8\}\}$;
5. $[R_c] = \{\{w_1, w_3, w_5, w_7\}, \{w_2, w_4, w_6, w_8\}\}$;
6. $V(k) = \{w_1, w_2, w_3, w_4\}$, $V(l) = \{w_1, w_2, w_6, w_7\}$ and $V(p) = \{w_1, w_3, w_5, w_7\}$.

7. $U_{\{a,b\}} = V(k) \cap V(l) = \{w_1, w_2\}$. For any other $C \in Gr$, $U_C = \emptyset$.

In this model, as the story of Example 3.2.11 goes, the killer and the lookout constitute a group. According to their plan, the killer is to kill, and the lookout is to guard the area. The passerby does not share any motivational states with either the killer or the lookout. As before, the real state of the Example 3.2.11 is w_2: a has $k$illed, b has $l$ooked out, and c has not $p$assed by. See Figure 6.2 for the illustration of the model.

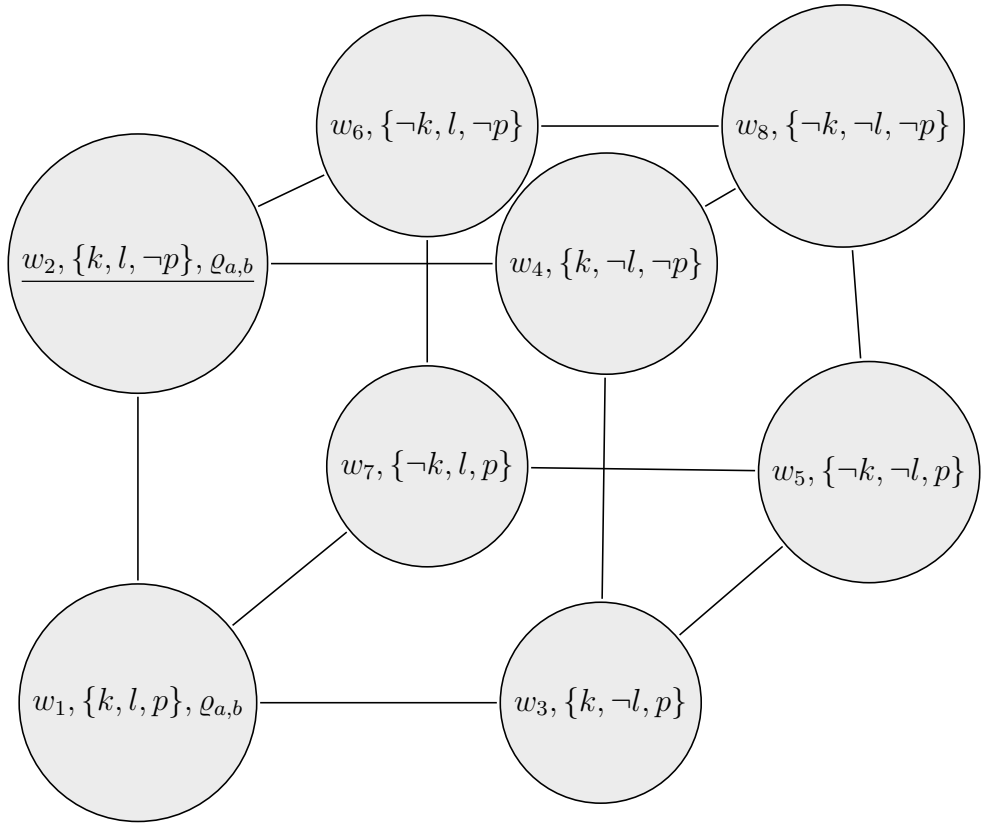

Figure 6.2: Atemporal Kripke **pSTIT**-model for Example 3.2.11. Front-back faces of the cube represent $[R_a]$; left-right faces represent $[R_b]$; top-bottom faces represent $[R_c]$. We omit reflexive and transitive edges in our representation of equivalence relations. E.g., $w_1 R_a w_1$, but we omit the reflexive edge; $w_2 R_a w_3$, but we only draw the edges $w_2 R_a w_1$ and $w_1 R_a w_3$. Worlds in $U_{\{a,b\}}$ are labeled with $\varrho_{a,b}$.

Let $\varphi := k \wedge (l \vee \neg p)$ be a formula, standing for "*the unwitnessed murder has happened*" (i.e., the killer has murdered someone, and either the lookout has covered it or the only possible witness has not passed by the murder scene). According to our definitions, the killer and the passerby are not causally responsible for the covered-up murder, since they have not acted upon any common group plan:

$M, w_2 \models \neg\Sigma^{min}_{a,c}\varphi$, since $M, w_2 \models \bigwedge_{C\supseteq\{a,c\}} \neg\varrho_C$. Moreover, the passerby is not responsible for the covered-up murder at all, since she neither deliberately sees to it, nor acts as a member of any group that jointly sees to it: $M, w_2 \models noresp_c\varphi$.

On the contrary, the group of the killer and the lookout are causally responsible for the covered-up murder: $M, w \models \Sigma^{min}_{a,b}\varphi$, since $M, w \models \neg\Box\varphi \wedge [stit]_{a,b}\varphi \wedge \bigwedge_{C\subset\{a,b\}} \neg[stit]_C\varphi \wedge \bigvee_{C\supseteq\{a,b\}} \varrho_C$. Therefore, both the killer and the lookout bear collective responsibility for the covered-up murder: $M, w_2 \models colresp_k\varphi \wedge colresp_l\varphi$. As for the responsibility for the murder alone, independently of its cover-up, the killer is causally responsible for it, while the lookout is merely complicit: $M, w_2 \models [dstit]_a k \wedge mcompl_b k$.

Such responsibility attributions are intuitively plausible and stand in line with the philosophical ideas around collective responsibility in [Baz13; Kut00]. On the one hand, we attribute collective responsibility only to members of groups, not random sets of agents. On the other hand, we hold agents complicit in group action if they participate in the group action, even if their contribution is not necessary.

To conclude, introducing plans as shared motivational states in group STIT logic allows us to strengthen the aggregative definition of joint actions in the standard STIT logic. As a result, we alleviate conceptual problems around it and improve models of causal responsibility attribution. Not only do we avoid holding random sets of agents collectively responsible, but we also capture subtler notions, such as complicity. In the following section, we will investigate the technical advantages we obtain.

6.3 Metalogical results

In this section, we investigate two metalogical properties of plan-restricted group STIT logic and its fragments: decidability and finite axiomatizability. $\mathcal{L}^t_{pSTIT}$ contains $\mathcal{L}^t_{STIT}$ as its fragment. $\mathcal{L}^t_{STIT}$, as we know, is not finitely axiomatizable, and its satisfiability problem has no decision procedure. Therefore, these negative results also hold for $\mathcal{L}^t_{pSTIT}$. So we have two paths to take. We can either generalise the semantics for $\mathcal{L}^t_{pSTIT}$ or restrict its language in search of well-behaved fragments. In this section, we will do both.

So our contributions here are twofold. First, we show how results about relativisations in cylindric algebras can be translated to STIT logic, and why having an explicit representation of plans makes relativisation conceptually justified. Second, we show that several fragments of $\mathcal{L}^t_{STIT}$ are finitely axiomatisable and decidable, going beyond the known results of [Sch12; Lor13]. These results easily transfer to the analogous fragments of $\mathcal{L}^t_{pSTIT}$.

6.3.1 Relativisations and plans

For atemporal **pSTIT**-models, we can weaken the semantic definitions without unjustifiably abstracting away from the crucial properties of what we model. The important property of most multi-modal logics of agency, including STIT, is the independence of agents. As we have discussed before, independence is not necessary if we assume that agents cooperate and act in accordance with their group plans. We often make such an assumption in our everyday practical reasoning.

For example, assume I have agreed to meet my friend at Amsterdam Centraal in two hours, unless he notifies me that he is not going to come. Even if I am notified that he is going and waiting for me, I am still causally able to refrain from going. In my everyday practical reasoning, I will be inclined to ignore this ability, as it is purely metaphysical: even though I potentially can do that, I have no reason to ever actualise this ability. So I might say in this situation: *"I cannot refrain from going to Amsterdam Centraal, unless my friend declines our meeting, as I have promised to show up".*

So in natural language, we seem to quantify over our abilities in a way that is surprisingly similar to how general assignment models work! We often say that we can or cannot do something, while we have these (in)abilities only conditional on following some plan. We take such restrictions on our own and others' behaviour as a background for our reasoning. Such restrictions can be naturally modelled in plan-restricted group STIT logic, via *plan-based submodels*.

6.3.1. Definition (Plan-based subframes of **pSTIT**). Given an atemporal **pSTIT**-model

$$M = (W, R_\Box, \{R_i\}_{i\in Ags}, \{U_C\}_{C\in Gr}, V)$$

and a group $G \in Gr$, a *U_G-based submodel* of M is a tuple:

$$M_G = (\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i\in Ags}, \{\mathcal{U}_C\}_{C\in Gr}, \mathcal{V}),$$

such that:

- $\mathcal{W} = U_G$;
- $\mathcal{R}_\Box = R_\Box \cap (U_G \times U_G)$;
- For any $i \in Ags$: $\mathcal{R}_i = R_i \cap (U_G \times U_G)$. Consequently, $\mathcal{R}_C = \bigcap_{i\in C} \mathcal{R}_i$ for any $C \subseteq Ags$;
- For all $C \in Gr$: $\mathcal{U}_C = U_C \cap U_G$;
- For any $p \in Var$: $\mathcal{V}(p) = V(p) \cap U_G$.

A notion of a U_G-based subframe $F_G = (\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i\in Ags}, \{\mathcal{U}_C\}_{C\in Gr})$ is defined accordingly. We denote a class of non-empty plan-based subframes of **pSTIT** as

$SF(\mathbf{pSTIT})$, i.e. $F \in SF(\mathbf{pSTIT})$ iff F is a non-empty plan-based subframe of some **pSTIT**-frame for some group $G \in Gr$. For every non-empty non-singleton set $G \in Gr$, by $SF_G(\mathbf{pSTIT})$ we denote a class of non-empty U_G-based subframes.

Intuitively, given some **pSTIT**-model and a non-empty non-singleton group $G \in Gr$, a U_G-submodel of M simulates agents' possible choices of action under the assumption that members of G will behave as a group and follow their plan. Such models are especially fruitful for reasoning the point of view of G's members: group members naturally expect themselves and their peers to follow the mutually agreed-upon plan, so they might not take into account possible actions of themselves and their peers that go against the plan.

In this subsection, we will work with the atemporal language $\mathcal{L}_{pSTIT}$. The satisfiability relation on plan-based subframes is defined standardly:

6.3.2. Definition. Given any plan-based model:

$$M_G = (\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i \in Ags}, \{\mathcal{U}_C\}_{C \in Gr}, \mathcal{V})$$

and any formula $\varphi \in \mathcal{L}_{pSTIT}$, we recursively define $\models$ relation as follows:

$$\begin{array}{lcl}
M_G, w \models p & \Leftrightarrow & w \in \mathcal{V}(p) \\
M_G, w \models \neg\varphi & \Leftrightarrow & \text{not } M_G, w \models \varphi \\
M_G, w \models \varphi \vee \psi & \Leftrightarrow & M_G, w \models \varphi \text{ or } M_G, v \models \psi \\
M_G, w \models \Box\varphi & \Leftrightarrow & \forall v \in \mathcal{W}(w\mathcal{R}_\Box v \Rightarrow M_G, v \models \varphi) \\
M_G, w \models [stit]_C\varphi & \Leftrightarrow & \forall v \in \mathcal{W}(w\mathcal{R}_C v \Rightarrow M_G, v \models \varphi) \\
M_G, w \models \varrho_C & \Leftrightarrow & w \in \mathcal{U}_C
\end{array}$$

Alternatively, if we want to take it for granted that members of some fixed group $C \in Gr$ always act in accordance with their plan, we could have restricted the language, using ϱ_C as a guard:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box(\varrho_C \to \varphi) \mid [stit]_J(\varrho_C \to \varphi) \mid \varrho_{C'}$$

where $J \subseteq Ags$, $C' \in Gr$. This approach brings nothing conceptually new to the table in comparison with plan-based submodels, and choosing between the two is in a sense a matter of taste: whether we prefer to restrict the language or generalise the semantics. We are more inclined to proceed with the plan-based submodels, not to multiply formal language definitions.

Interestingly, the logic of the U_{Ags}-based subframes can be finitely axiomatised. See the definition of $\mathbf{L}^p_G$ in Table 6.1.

(PL)	Axioms for classical propositional logic
(S5)	**S5** axioms and rules for $\Box$, $[stit]_i$, $[stit]_C$
(Incl)	$\Box\varphi \to [stit]_i\varphi$ for any $i \in Ags$
(S-Add)	$[stit]_C\varphi \to [stit]_{C'}\varphi$ for every $C \subseteq C' \subseteq Ags$
(Contr)	$\varrho_C \leftrightarrow [stit]_C\varrho_C$ for every $C \subseteq Ags$
(Pl)	$\Box\varrho_{Ags}$
(Cons)	$(\Diamond\varrho_{C_1} \wedge \ldots \wedge \Diamond\varrho_{C_j}) \to \Diamond(\varrho_{C_1} \wedge \ldots \wedge \varrho_{C_j})$ for every $C_1, \ldots, C_j \in Gr$, s.t. $\bigcap_{1\leq i\leq j} C_i \neq \emptyset$
(MP)	If $\vdash \varphi \to \psi$ and $\vdash \varphi$, then $\vdash \psi$

Table 6.1: Logic $\mathbf{L}^p_G$

(S5) expresses that $\mathcal{R}_\Box$, $\{\mathcal{R}_G\}_{G\subseteq Ags}$ are equivalence relations. (Incl) corresponds to $\mathcal{R}_i \subseteq \mathcal{R}_\Box$ for any $i \in Ags$. (S-Add) expresses that $\mathcal{R}_C \subseteq \mathcal{R}_{C'}$ for any $C' \subseteq C \subseteq Ags$. (Contr) axiomatically expresses that each $\mathcal{U}_C$ is $\mathcal{R}_C$-closed ((2a) of Def. 6.2.1), and (Cons) corresponds to the consistency of group plans whenever groups share some members ((2b) of Def. 2a). The axiom (Pl) expresses that, in U_{Ags}-based subframes, Ags necessarily acts in accordance with its plan.

We proceed with proving that $\mathbf{L}^p_G$ is sound and strongly complete w.r.t $SF_{Ags}(\mathbf{pSTIT})$-models. First, we show that $\mathbf{L}^p_G$ is sound and strongly complete w.r.t. a class of Kripke frames $\mathbf{C_2}$. Then we define a subclass of $\mathbf{C}_1 \subseteq \mathbf{C}_2$ and show that every $\mathbf{C}_1$-frame is a bounded morphic image of some $\mathbf{C}_2$-frame, which guarantees that $\mathbf{C_2} \models \varphi$ iff $\mathbf{C_1} \models \varphi$ for any $\varphi \in \mathcal{L}_{pSTIT}$. After that, we show that every $\mathbf{C_1}$-frame is a U_{Ags}-based subframe of some **pSTIT**-frame.

6.3.3. Definition ($\mathbf{C_2}$-frames). A $\mathbf{C_2}$-frame is a tuple of the form:

$$F = (W, R_\Box, \{R_G\}_{G\subseteq Ags}, \{U_G\}_{G\in Gr})$$

1. $\{R_i\}_{i\in Ags}$, $R_\Box$ are equivalence relations;
2. $R_i \subseteq R_\Box$ for every $i \in Ags$;
3. $\{R_G\}_{G\subseteq Ags}$ are equivalence relations, such that $R_{G'} \subseteq R_G$ for every $G \subseteq G' \subseteq Ags$;
4. For every $G \in Gr$, $U_G \subseteq W$ is a possibly empty subset of worlds that is R_G-closed.
5. For every world $w \in W$, $R_\Box(w) \subseteq U_{Ags}$.
6. For every world $w \in W$ and every $C_1, \ldots, C_j \in Gr$, such that $C_1 \cap \ldots \cap C_j \neq \emptyset$: if $wR_\Box v_1, \ldots, wR_\Box v_j$ and $v_1 \in U_{C_1}, \ldots, v_j \in U_{C_j}$, then there exists $u \in W$, such that $wR_\Box u$ and $u \in U_{C_1} \cap \ldots \cap U_{C_j}$.

6.3.4. Definition ($\mathbf{C_1}$-frames). A tuple $F = (W, R_\Box, \{R_G\}_{G\subseteq Ags}, \{U_G\}_{G\in Gr})$ is a $\mathbf{C_1}$-frame iff F is a $\mathbf{C_2}$-frame, such that for any group $G \subseteq Ags$, $R_G = \bigcap_{i\in G} R_i$.

From Definitions 6.3.4 and 6.3.3 it follows that $\mathbf{C_1} \subseteq \mathbf{C_2}$. The difference is only in the definition of $\{R_G\}_{G\subseteq Ags}$ relations: $\mathbf{C}_1$ requires further that $R_G = \bigcap_{i\in G} R_i$.

6.3.5. Lemma. *$\mathbf{L}^p_G$ is sound and strongly complete w.r.t. $\mathbf{C_2}$-frames, i.e. for any set of formulae $\Gamma \subseteq \mathcal{L}_{pSTIT}$ and any formula $\varphi \in \mathcal{L}_{pSTIT}$:*

$$\Gamma \vdash_{\mathbf{L}^p_G} \varphi \Leftrightarrow \Gamma \models_{\mathbf{C_2}} \varphi.$$

Proof:
All $\mathbf{L}^p_G$ axioms are Sahlqvist formulae whose first-order translations directly correspond to (1)-(6) of Definition 6.3.3. By the Sahlqvist theorem [Sah75], $\mathbf{L}^p_G$ is canonical and hence sound and strongly complete w.r.t. $\mathbf{C_2}$. □

Next, we show that $\mathcal{L}_{pSTIT}$ cannot distinguish between $\mathbf{C_2}$ and $\mathbf{C_1}$-frames. We do it by a fairly standard unravelling technique first introduced (to the best of our knowledge) in [FHV92] and later simplified in [WÅ20].

6.3.6. Definition (Paths, initial segments). Take any $\mathbf{C_2}$-frame:

$$F = (W, R_\Box, \{R_G\}_{G\subseteq Ags}, \{U_C\}_{C\in Gr})$$

By a path on F we mean a finite, non-empty sequence $\vec{w} = (w_1, G_1, \ldots, G_{n-1}, w_n)$, such that $w_1, \ldots, w_n \in W$, $G_1, \ldots, G_{n-1} \subseteq Ags$ and for every pair w_i, w_{i+1}: $(w_i, w_{i+1}) \in R_{G_i}$. If for some i, $G_i = \emptyset$, then $(w_i, w_{i+1}) \in R_\Box$. By $head(\vec{w})$ we mean the first world that appears in the sequence (i.e. $head((w_1, G_1, \ldots, G_{n-1}, w_n)) = w_1$); $tail(\vec{w})$ is the last world that appears in $\vec{w}$ (i.e. $tail((w_1, G_1, \ldots, G_{n-1}, w_n)) = w_n$). Note that for any world $w \in W$, (w) is also a path.

Given two paths $\vec{w} = (w_1, G_1, \ldots, G_{n-1}, w_n)$, $\vec{v} = (v_1, J_1, \ldots, J_{k-1}, v_k)$, we say that $\vec{w}$ is an initial segment of $\vec{v}$ ($\vec{w} \sqsubseteq \vec{v}$) iff $n \leq k$, $w_i = v_i$ for any $1 \leq i \leq n$ and $G_i = J_i$ for any $1 \leq i \leq n-1$. Then, $\vec{v} \setminus \vec{w} = (w_n J_n, \ldots, J_{k-1}, v_k)$. If $\vec{v} = \vec{w}$, then $\vec{v} \setminus \vec{w} = (tail(\vec{v}))$.

Given a path $\vec{w} = (w_1, G_1, \ldots, G_{n-1}, w_n)$ and a group $J \subseteq Ags$, we say that $\vec{w}$ is a J-path iff $J \subseteq \bigcap_{1\leq i\leq n-1} G_i$. Note that paths of the length 1, i.e. paths of the form (w), are vacuously J-path for any $J \subseteq Ags$.

6.3.7. Definition (Unraveling). For any $\mathbf{C_2}$-frame:

$$F = (W, R_\Box, \{R_G\}_{G\subseteq Ags}, \{U_C\}_{C\in Gr}).$$

We define its unraveling as:

$$\mathfrak{F} = (\mathfrak{W}, \mathfrak{R}_\Box, \{\mathfrak{R}_G\}_{G\subseteq Ags}, \{\mathfrak{U}_C\}_{C\in Gr})$$

1. $\mathfrak{W} = \{\vec{w} \mid \vec{w} \text{ is a path on } F\}$;
2. $\vec{w}\mathfrak{R}_\Box\vec{v}$ iff 1) $\vec{w}$ and $\vec{v}$ have a common initial segment, i.e. $\exists\vec{g} : \vec{g} \sqsubseteq \vec{w}$ and $\vec{g} \sqsubseteq \vec{v}$;
3. $\vec{w}\mathfrak{R}_G\vec{v}$ iff 1) $\vec{w}$ and $\vec{v}$ have some common initial segment $\vec{g}$; 2) $\vec{w} \setminus \vec{g}$ and $\vec{v} \setminus \vec{g}$ are G-paths;
4. $\vec{w} \in \mathfrak{U}_C$ iff $tail(\vec{w}) \in U_C$;

6.3.8. Proposition. *Given any* $\mathbf{C_2}$*-frame* $F = (W, R_\Box, \{R_G\}_{G\subseteq Ags}, \{U_C\}_{C\in Gr})$*, its unraveling* $\mathfrak{F} = (\mathfrak{W}, \mathfrak{R}_\Box, \{\mathfrak{R}_G\}_{G\subseteq Ags}, \{\mathfrak{U}_C\}_{C\in Gr})$ *is a* $\mathbf{C_1}$*-frame.*

Proof:
It is a routine to check that $\mathfrak{R}_i$, $\mathfrak{R}_\Box, \mathfrak{R}_G$ are equivalence relations, s.t. $\mathfrak{R}_i \subseteq \mathfrak{R}_\Box$.

$\mathfrak{U}_C$ is $\mathfrak{R}_C$-closed. To see that, assume $\vec{w} \in \mathfrak{U}_C$ and $\vec{w}\mathfrak{R}_C\vec{v}$. Let $tail(\vec{w}) = w$ and $tail(\vec{v}) = v$. Then, $\vec{w}$ and $\vec{v}$ have some common initial segment $\vec{g}$, and $\vec{v} \setminus \vec{g}$ and $\vec{w} \setminus \vec{g}$ are C-paths. Let $tail(\vec{g}) = g$. Then, $g = w_0R_{C_1} \dots R_{C_{n-1}}w_n = w$ and $g = v_0R_{G_1} \dots R_{G_{m-1}}v_m = v$ for some $w_0, \dots, w_n \in W$, $v_0, \dots, v_m \in W$ and $C_1, \dots, C_n, G_1, \dots, G_m \subseteq Ags$, such that $C \subseteq C_1 \cap \dots \cap C_n \cap G_1 \cap \dots \cap G_m$. By (3) of Def. 6.3.3 and transitivity of R_C, gR_Cw and gR_Cv, hence, wR_Cv. Then, as U_C is R_C-closed, $v \in U_C$. Since $v = tail(\vec{v})$, $\vec{v} \in \mathfrak{U}_C$.

Property (5) of Def. 6.3.3 holds for $\mathfrak{F}$: for every world $\vec{w} \in \mathfrak{W}$, $\mathfrak{R}_\Box(\vec{w}) \subseteq \mathfrak{U}_{Ags}$. To see why, take any path $\vec{w} \in \mathfrak{W}$, such that $tail(\vec{w}) = w$. Since $F \in \mathbf{C_2}$, by Definition 6.3.3, $R_\Box(w) \subseteq U_{Ags}$. Take any $\vec{v} \in \mathfrak{R}_\Box(\vec{w})$. Let $tail(\vec{v}) = v$. By definition of $\mathfrak{R}_\Box$, $\vec{w}$ and $\vec{v}$ have some common initial segment $\vec{g}$. Let $tail(\vec{g}) = g$. Then, there exist paths $(g, G_1, \dots, G_n, w)$ and $(g, J_1, \dots, J_k, v)$. Since $R_{G_1}, \dots, R_{G_n}, R_{J_1}, \dots, R_{J_k} \subseteq R_\Box$, $gR_\Box \dots R_\Box w$ and $gR_\Box, \dots, R_\Box v$, and by transitivity of $R_\Box$, $wR_\Box v$. Since, by (5) of Def. 6.3.3 $\forall v \in W(wR_\Box v \to v \in U_{Ags})$, $v \in U_{Ags}$. As $tail(\vec{v}) = v$ and $v \in U_{Ags}$, $\vec{v} \in \mathfrak{U}_{Ags}$.

Property (6) of Def. 6.3.3 holds for $\mathfrak{F}$. Let $\vec{w}\mathfrak{R}_\Box\vec{v}_1, \dots, \vec{w}\mathfrak{R}_\Box\vec{v}_j$, $\vec{v_1} \in \mathfrak{U}_{C_1}, \dots, \vec{v_j} \in \mathfrak{U}_{C_j}$ and $C_1 \cap \dots \cap C_j \neq \emptyset$. Let $tail(\vec{w}) = w$, $tail(\vec{v}_i) = v_i$ for $i \in \{1, \dots, j\}$. Then, analogously to the proof of the previous property, $wR_\Box v_1, \dots, wR_\Box v_j$ and $v_1 \in U_{C_1}, \dots, v_j \in U_{C_j}$. Therefore, there exists $u \in W$, such that $wR_\Box u$ and $u \in U_{C_1} \cap \dots \cap U_{C_j}$. Take a path $\vec{u} = (\vec{w}, \emptyset, u)$. Then, $\vec{w}\mathfrak{R}_\Box\vec{u}$ and $\vec{u} \in \mathfrak{U}_{C_1} \cap \dots \cap \mathfrak{U}_{C_j}$.

So it is left for us to show that $\vec{w}\mathfrak{R}_G\vec{v}$ iff $(\vec{w}, \vec{v}) \in \bigcap_{i\in G} \mathfrak{R}_i$. Left-to-right direction. Assume $\vec{w}\mathfrak{R}_G\vec{v}$. It means that $\vec{v}$, $\vec{w}$ share some initial segment $\vec{g}$ and $\vec{w} \setminus \vec{g}$, $\vec{v} \setminus \vec{g}$ are G-paths. By definition of a G-path, for every $i \in G$, every G-path is a i-path. Hence, $\vec{v}$, $\vec{w}$ share some initial segment $\vec{g}$ and $\vec{w} \setminus \vec{g}$, $\vec{v} \setminus \vec{g}$ are i-paths for every $i \in G$. By definition of $\mathfrak{R}_i$, it means that $(\vec{w}, \vec{v}) \in \bigcap_{i\in G} \mathfrak{R}_i$. Right-to-left

direction. Assume $(\vec{w},\vec{v}) \in \bigcap_{i\in G}\mathfrak{R}_i$. By definition of $\mathfrak{R}_i$, $(\vec{w},\vec{v}) \in \bigcap_{i\in G}\mathfrak{R}_i$ means that for every $i \in G$, $\vec{w}$ and $\vec{v}$ have some initial segment $\vec{g}_i$ and $\vec{w} \setminus \vec{g}_i$, $\vec{v} \setminus \vec{g}_i$ are i-paths. By definition of a path, $\vec{w},\vec{v}$ are finite. Hence, given a set of such common initial segments $\{\vec{g}_i\}_{i\in G}$, there is the longest path $\vec{g} \in \{\vec{g}_i\}_{i\in G}$, i.e. for every $\vec{t} \in \{\vec{g}_i\}_{i\in G}$, $\vec{g} \setminus \vec{t}$ is defined. Moreover, $\vec{w} \setminus \vec{g}$, $\vec{v} \setminus \vec{g}$ are i-paths for every $i \in G$. Hence, $\vec{w}$, $\vec{v}$ have an initial segment $\vec{g}$, such that $\vec{w} \setminus \vec{g}$, $\vec{v} \setminus \vec{g}$ are G-paths. Then, by definition of $\mathfrak{R}_G$, $\vec{w}\mathfrak{R}_G\vec{v}$. Taking it all together, we have shown that $\mathfrak{F} \in \mathbf{C_1}$. □

6.3.9. Lemma. *Given a* $\mathbf{C_2}$*-frame*

$$F = (W, R_\Box, \{R_G\}_{G\subseteq Ags}, \{U_G\}_{G\subseteq Ags})$$

and its unraveling

$$\mathfrak{F} = (\mathfrak{W}, \mathfrak{R}_\Box, \{\mathfrak{R}_G\}_{G\subseteq Ags}, \{\mathfrak{U}_G\}_{G\in Gr}),$$

the function $f : \mathfrak{W} \to W$*, such that* $f(\vec{w}) = tail(\vec{w})$*, is a bounded morphism.*

Proof:
The function f is obviously surjective, since every world $w \in W$ is a tail of at least one path (w). We claim that f is a bounded morphism. That is, for every relation R and $\mathfrak{R}$, we are to prove the following properties. Zig: $\vec{w}\mathfrak{R}\vec{v}$ entails $f(\vec{w})Rf(\vec{v})$; Zag: $f(\vec{w})Rv$ entails that there exists $\vec{v}$, such that $f(\vec{v}) = v$ and $\vec{w}\mathfrak{R}\vec{v}$.

(Zig for $R_\Box$ and $\mathfrak{R}_\Box$) Assume $\vec{w}\mathfrak{R}_\Box\vec{v}$. Then, $\vec{w}$ and $\vec{v}$ share some initial segment $\vec{g}$, which means that there exists a world $tail(\vec{g}) = g$, such that there is a path from g to $f(\vec{w})$ and from g to $f(\vec{v})$, where every step of these paths is some of $\{R_G\}_{G\subseteq Ags}$ relations. Since every R_G-relation is a subset of $R_\Box$, those paths are $R_\Box$-paths. And since $R_\Box$ is transitive, $gR_\Box f(\vec{w})$ and $gR_\Box f(\vec{v})$. Since $R_\Box$ is Euclidean and symmetric, $f(\vec{w})R_\Box f(\vec{v})$.

(Zag for $R_\Box$ and $\mathfrak{R}_\Box$) Assume $f(\vec{w})R_\Box v$. Let $tail(\vec{w}) = w_n$. Then, $w_nR_\Box v$. Then, let $\vec{v}$ be a path that extends $\vec{w}$ with $(w_n, \emptyset, v)$, i.e. $\vec{v} \setminus \vec{w} = (w_n, \emptyset, v)$. $\vec{w}$ and $\vec{v}$ share an initial segment, namely, $\vec{w}$. Hence, $\vec{w}\mathfrak{R}_\Box\vec{v}$. Moreover, since $tail(\vec{v}) = v$, $f(\vec{v}) = v$.

(Zig for R_G and $\mathfrak{R}_G$) Assume $\vec{w}\mathfrak{R}_G\vec{v}$. Then, $\vec{w}$, $\vec{v}$ share some initial segment $\vec{g}$, such that $\vec{w} \setminus \vec{g}$,$\vec{v} \setminus \vec{g}$ are G-paths. Let $tail(\vec{w}) = w$, $tail(\vec{v}) = v$, $tail(\vec{g}) = g$. Given that R_G is transitive and the definition of a path, it means that gR_Gw and gR_Gv. Since R_G is symmetric, wR_Gg and gR_Gv. Using transitivity again, wR_Gv, i.e. $f(\vec{w})R_Gf(\vec{v})$.

(Zag for R_G and $\mathfrak{R}_G$) Assume $f(\vec{w})R_Gv$. Let $tail(\vec{w}) = w$. Then, wR_Gv. There exists a path $\vec{v} = (\vec{w}, G, v)$, i.e. $\vec{v} \setminus \vec{w} = (w, G, v)$. Note that $\vec{w},\vec{v}$ have the initial

segment $\vec{w}$ and $\vec{w} \setminus \vec{v}$ is a G-path ($\vec{w} \setminus \vec{w}$ is vacuously a G-path). Hence, $\vec{w}\mathfrak{R}_G\vec{v}$ and $f(\vec{v}) = v$.

Case of $\mathfrak{U}_C$ and U_C is trivial, since $\vec{w} \in \mathfrak{U}_C$ iff $f(\vec{w}) \in U_C$.

So that, f is a bounded morphism from $\mathfrak{F}$ onto F. Since F was arbitrary, we can conclude that for every $\mathbf{C_2}$-frame there exists a $\mathbf{C_1}$-frame, such that the former is a bounded morphic image of the latter. □

6.3.10. THEOREM ($\mathbf{C_1} \subseteq SF_{Ags}(\mathbf{pSTIT})$). *Every* $\mathbf{C_1}$*-frame is a* U_{Ags}*-based subframe of some* **pSTIT***-frame.*

Proof:
Take any $\mathbf{C_1}$-frame $F = (W, R_\Box, \{R_G\}_{G \subseteq Ags}, \{U_C\}_{C \in Gr})$ for some set of agents $Ags = \{1, \ldots, n\}$. W.l.o.g., we assume that F is point-generated (otherwise, the procedure we are to present should be done for every point-generated subframe of F, after which we should take their disjoint union). Note that in a point-generated frame, $R_\Box$ is a universal relation, since for any other relation R_G, $R_G \subseteq R_\Box$. Then, by (5) of Definition 6.3.3, $W = U_{Ags}$. We are to construct a **pSTIT**-frame $\mathcal{F}$, such that F is its U_{Ags}-generated subframe.

To come up with an atemporal **pSTIT**-frame, we need to construct an atemporal Kripke STIT frame and a R_C-closed set U_C for every $C \in Gr$. We begin with an atemporal Kripke STIT frame. Take the tuple $(W, R_\Box, \{R_i\}_{i \in Ags})$ from F. The only thing that differentiates it from a Kripke STIT frame is that it may not satisfy the independence of agents constraint. We will reformulate the constraint for technical reasons.

Every R_i is an equivalence relation, so R_i yields a partition $[R_i] = \{R_i(w) | w \in W\}$ of W. We can index elements of each partition as $[R_i] = \{X_1^i, \ldots, X_{Card([R_i])}^i\}$. In such a case, the independence of agents constraint is satisfied iff $X_{i_1}^1 \cap \ldots \cap X_{i_n}^n \neq \emptyset$ for every $i_1 \leq Card([R_1]), \ldots, i_n \leq Card([R_n])$.[4]

Since we are not sure if independence of agents holds for F, there might be elements $X_{i_1}^1 \in [R_1], \ldots, X_{i_n}^n \in [R_n]$, such that $X_{i_1}^1 \cap \ldots \cap X_{i_n}^n = \emptyset$. To fix it, we need to make all such intersections non-empty. So we extend W to a larger set of possible worlds – $\mathcal{W}$ – by putting a new world for every such empty intersection. As a result, we will get an extension of the set of possible worlds W to a new set $\mathcal{W}$, and every relation R_i defined on W will be extended to $\mathcal{R}_i$ defined on $\mathcal{W}$.

First, we create a *skeleton* of a **STIT**-frame. For every agent $i \in Ags$, let $[\mathcal{R}_i] = \{E_1^i, \ldots, E_{Card([R_i])}^i\}$ be a collection of sets of possible worlds, which are to serve as $\mathcal{R}_i$-equivalence classes. Our task is to describe elements of all these sets

[4]Note that this reformulation is valid only for point-generated frames, where all worlds are $R_\Box$-related.

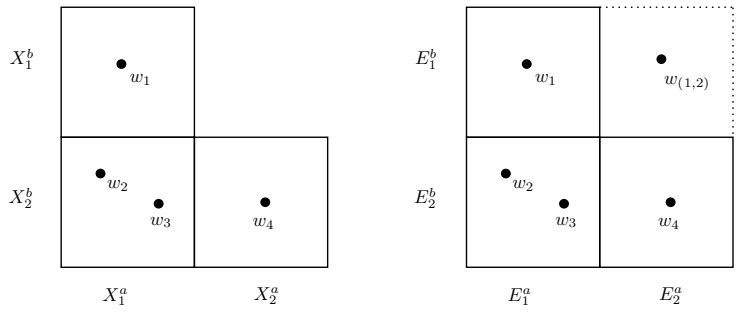

Figure 6.3: Example of constructing an atemporal **STIT**-frame (on the right) from a $\mathbf{C_1}$-frame (on the left) for two agents. In the initial $\mathbf{C_1}$-frame, the independence of agents fails, since $X_2^a \cap X_1^b = \emptyset$. In a constructed **STIT**-frame, $E_2^a \cap E_1^b = w_{(1,2)}$, where $w_{(1,2)}$ is fresh.

in a manner such that for every $E_1 \in [\mathcal{R}_i], \ldots, E_n \in [\mathcal{R}_n]$, $E_1 \cap \ldots \cap E_n \neq \emptyset$. We do it as follows. For every $E_{j_1}^1 \in [\mathcal{R}_1], \ldots, E_{j_n}^n \in [\mathcal{R}_n]$:

$$E_{j_1}^1 \cap \ldots \cap E_{j_n}^n = \begin{cases} X_{j_1}^1 \cap \ldots \cap X_{j_n}^n & \text{if } X_{j_1}^1 \cap \ldots \cap X_{j_n}^n \neq \emptyset \\ \{w_{(j_1,\ldots,j_n)}\} & \text{otherwise} \end{cases}$$

where every $w_{(j_1,\ldots,j_n)}$ is just a unique "fresh" possible world, i.e. $w_{(j_1,\ldots,j_n)} \notin W$.

Knowing the elements of every intersection of the form $E_1 \cap \ldots \cap E_n$, we know the elements of every equivalence class E_i: $w \in E_i$ iff there exists $E_1 \in [\mathcal{R}_1], \ldots, E_{i-1} \in [\mathcal{R}_{i-1}], E_{i+1} \in [\mathcal{R}_{i+1}], \ldots, E_n \in [\mathcal{R}_n]$, such that $w \in E_1 \cap \ldots \cap E_i \cap \ldots \cap E_n$.

So we define a tuple $(\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i \in Ags})$, where:

- $\mathcal{W} = \{w \in E_1 \cap \ldots \cap E_n \mid E_1 \in [\mathcal{R}_1], \ldots, E_n \in [\mathcal{R}_n]\}$, i.e. $\mathcal{W}$ consists of W and all the added "fresh" worlds;
- $\mathcal{R}_\Box = \mathcal{W} \times \mathcal{W}$;
- For any $w, v \in \mathcal{W}$, $w \mathcal{R}_i v$ iff there exists $E \in [\mathcal{R}_i]$ such that $w, v \in E$;

See Figure 6.3 for the example of the construction. $(\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i \in Ags})$ is a an atemporal Kripke STIT frame. Every $[\mathcal{R}_i]$ is a partition of $\mathcal{W}$, hence, $\mathcal{R}_i$ is an equivalence relation on $\mathcal{W}$. Since $\mathcal{R}_\Box$ is a total relation, $\mathcal{R}_i \subseteq \mathcal{R}_\Box$. Since for every $E_1 \in [\mathcal{R}_1], \ldots, E_n \in [\mathcal{R}_n]$: $E_1 \cap \ldots \cap E_n \neq \emptyset$, independence of agents holds. Standardly, $\mathcal{R}_C = \bigcap_{i \in C} R_i$ for every $C \subseteq Ags$. Note that $W \subseteq \mathcal{W}$ and for any $w, v \in W$, $w R_i v$ iff $w \mathcal{R}_i v$.

To finish our construction, we need to define group plans, i.e., a collection of sets $\{\mathcal{U}_C\}_{C \in Gr}$. We do it in the following manner:

- $\mathcal{U}_{Ags} = W$;
- $\mathcal{U}_G = \{w \in \mathcal{W} \mid \exists v \in W : v \in U_C \text{ and } w\mathcal{R}_C v\}$

According to Definition 6.2.1, $\mathcal{F} = \{\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i\in Ags}, \{\mathcal{U}_C\}_{C\in Gr}\}$ is an atemporal **pSTIT**-frame. As we have already shown, $\{\mathcal{W}, \mathcal{R}_\Box, \{\mathcal{R}_i\}_{i\in Ags}\}$ is a Kripke STIT frame. By definition, for every $C \in Gr \setminus \{Ags\}$, $\mathcal{U}_C$ is a $\mathcal{R}_C$-closed set.

$\mathcal{U}_{Ags}$ is $\mathcal{R}_{Ags}$-closed as well. To see it, assume $w \in \mathcal{U}_{Ags}$ and $w\mathcal{R}_{Ags}v$. As $w \in \mathcal{U}_{Ags}$, $w \in W$. As $w\mathcal{R}_{Ags}v$, $w\mathcal{R}_i v$ for any $i \in Ags$. Then, $w \in E_1 \cap \ldots \cap E_n$ iff $v \in E_1 \cap \ldots \cap E_n$ for any $E_1 \in [\mathcal{R}_1], \ldots, E_n \in [\mathcal{R}_n]$. As $w \in W$, it is not "fresh", i.e., it is a member of some $X_1 \in [R_1], \ldots, X_n \in [R_n]$, where $X_1, \ldots, X_n \subseteq W$. Then, $v \in X_1, \ldots, X_n$. Hence, $v \in W$. By definition, $\mathcal{U}_{Ags} = W$, so $v \in \mathcal{U}_{Ags}$. As v was an arbitrary $\mathcal{R}_{Ags}$-successor of an arbitrary $w \in \mathcal{U}_{Ags}$, $\mathcal{U}_{Ags}$ is R_{Ags}-closed.

Finally, (2b) of Def. 6.2.1 holds. To see it, let $C_1 \cap \ldots \cap C_k \neq \emptyset$ for some $C_1, \ldots . C_n \in Gr$. Assume that $\mathcal{U}_{C_1} \neq \emptyset, \ldots, \mathcal{U}_{C_k} \neq \emptyset$. Then, by our construction, $U_{C_1}, \ldots, U_{C_k} \neq \emptyset$. As the frame F satisfies (2b) of Def. 6.2.1, there is a world $v \in W$, such that $w \in U_{C_1} \cap \ldots \cap U_{C_k}$. As $U_{C_1} \subseteq \mathcal{U}_{C_1}, \ldots, U_{C_k} \subseteq \mathcal{U}_{C_k}$, $w \in \mathcal{U}_{C_1} \cap \ldots \cap \mathcal{U}_{C_k}$. Hence, $\mathcal{U}_{C_1} \cap \ldots \cap \mathcal{U}_{C_k} \neq \emptyset$.

Note that $\mathcal{U}_{Ags}$-based subframe of $\mathcal{F}$ is equal to the initial $\mathbf{C_1}$-frame F. $\mathcal{U}_{Ags} = W$; for every relation $\mathcal{R} \in \{\mathcal{R}_\Box, \mathcal{R}_G\}$, $\mathcal{R} \cap (\mathcal{U}_{Ags} \times \mathcal{U}_{Ags}) = R$ and for every $C \in Gr$: $\mathcal{U}_C \cap \mathcal{U}_{Ags} = \mathcal{U}_C \cap W = U_C$.

In other words, F is a $\mathcal{U}_{Ags}$-generated subframe of an atemporal **pSTIT**-frame, $\mathcal{F}$.

Since F was an arbitrarily chosen $\mathbf{C_1}$-frame, we can conclude that every $\mathbf{C_1}$-frame is a U_{Ags}-generated subframe of some **pSTIT**-frame. □

6.3.11. Theorem (Soundness and completeness). *Given any set of formulae* $\Gamma \subseteq \mathcal{L}_{pSTIT}$ *and any formula* $\varphi \in \mathcal{L}_{pSTIT}$*, the following holds:*

$$\Gamma \vdash_{\mathbf{L}^p_G} \varphi \Leftrightarrow \Gamma \models_{SF_{Ags}(\mathbf{pSTIT})} \varphi.$$

Proof:
($\Rightarrow$) By a routine check of the axioms' validity.

($\Leftarrow$) For contradiction, assume $\Gamma \nvdash_{\mathbf{L}^p_G} \varphi$. Then, $\Gamma \cup \{\neg\varphi\}$ is $\mathbf{L}^p_G$-consistent. By Lemmas 6.3.5, 6.3.9, there exists a pointed $\mathbf{C_1}$-model (M, w), such that $M, w \models \Gamma \cup \{\neg\varphi\}$. By Theorem 6.3.10, there is an atemporal **pSTIT**-model $\mathcal{M}$, such that $\mathcal{M}_{Ags}, w \models \Gamma \cup \{\neg\varphi\}$, where $\mathcal{M}_{Ags}$ is the U_{Ags}-based submodel of $\mathcal{M}$. As a consequence, $\Gamma \not\models_{SF_{Ags}(\mathbf{pSTIT})} \varphi$. □

As we have established that any formula is satisfiable on $SF_{Ags}(\mathbf{pSTIT})$-frames iff it is satisfiable on $\mathbf{C_2}$-frames iff it is $\mathbf{L}^p_G$-consistent, we can establish decidability by a standard fmp-via-filtration argument.

6.3.12. THEOREM. *If a formula $\varphi \in \mathcal{L}_{pSTIT}$ is satisfiable, then it is satisfiable on a $\mathbf{C_2}$-model of size at most $2^{3|sub(\varphi)|}$.*

Proof:
Assume $\varphi \in \mathcal{L}_{pSTIT}$ is $SF_{Ags}(\mathbf{pSTIT})$-satisfiable. Then, it is satisfiable on some $\mathbf{C_2}$-model:

$$M = (W, R_\Box, \{R_C\}_{C\subseteq Ags}, \{U_C\}_{C\in Gr}, V)$$

Let $Cl(\varphi) \subseteq \mathcal{L}_{pSTIT}$ be a minimal set of formulae, such that:

1. $Cl(\varphi)$ is subformulae-closed;
2. $\{\varrho_C\}_{C\in Gr} \subseteq Cl(\varphi)$;
3. If $\varrho_C \in Cl(\varphi)$ then $[stit]_C\varrho_C, \Diamond\varrho_C, \Box\varrho_C \in Cl(\varphi)$ for any $C \in Gr$.

If $|sub(\varphi)| = m$ and $|Gr| = n$, then $|Cl(\varphi)| \leq m+4n$: φ has m-many subformulae, $|\{\varrho_C\}_{C\in Gr}| = |Gr| = n$. We add at most three new formulae $[stit]_C\varrho_C, \Diamond\varrho_C, \Box\varrho_C$ to any subformulae $\varrho_C \in sub(\varphi)$.

Let $\sim$ be an equivalence relation on W, such that $w \sim v$ iff $M, w \models \psi$ iff $M, v \models \psi$ for all $\psi \in Cl(\varphi)$. Let $[w] = \{v \in W \mid v \sim w\}$ denote the equivalence class of w w.r.t. $\sim$. Then, we define filtration of M through $Cl(\varphi)$, denoted as M', as follows:

$$M' = (W', R'_\Box, \{R'_C\}_{C\subseteq Ags}, \{U'_C\}_{C\in Gr}, V')$$

1. $W' = \{[w] \mid w \in W\}$. Note that $|W'| \leq 2^{|Cl(\varphi)|} \leq 2^{(|sub(\varphi)|+4|Gr|}$;
2. $[u]R'_\Box[v]$ iff for any $\Box\psi \in Cl(\varphi)$, $M, u \models \Box\psi$ iff $M, v \models \Box\psi$;
3. For any $C \subseteq Ags$, $[u]R'_C[v]$ iff:
   (a) $[u]R'_\Box[v]$;
   (b) $[u]R'_{C'}[v]$ for any $C' \subseteq C$;
   (c) for any $[stit]_C\psi \in Cl(\varphi)$, $M, u \models [stit]_C\psi$ iff $M, v \models [stit]_C\psi$;
4. For any $C \in Gr$, $[u] \in U'_C$ iff $M, u \models \varrho_C$;
5. For any $p \in Var$, $V'(p) = \{[v] \in W' \mid v \in V(p)\}$.

By a standard filtration lemma [BBW06, p.268], $M, w \models \psi$ iff $M', [w] \models \psi$ for any $\psi \in Cl(\varphi)$. To complete the proof, we need to prove that M' is a $\mathbf{C_2}$-model.

It is straightforward that $R'_\Box, \{R'_C\}_{C\subseteq Ags}$ are equivalence relations, such that $R'_{C_1} \subseteq R'_{C_2} \subseteq R'_\Box$ for any $C_2 \subseteq C_1 \subseteq Ags$.

(5) of Def. 6.3.3 holds for M'. Take any world $[w] \in W'$. As M is a $\mathbf{C_2}$-model, $v \in U_{Ags}$ for any $v \in R_\Box(w)$. Since $wR_\Box w$, $w \in U_{Ags}$. Take any $[v] \in W'$, such that $[w]R'_\Box[v]$. Since $w \in U_{Ags}$, $M, w \models \varrho_{Ags}$. Since $R_\Box(w) \subseteq U_{Ags}$, $M, w \models \Box\varrho_{Ags}$ By Def. of $Cl(\varphi)$, $\varrho_{Ags}, \Box\varrho_{Ags} \in Cl(\varphi)$. Take any $[u] \in W'$, such that $[w]R'_\Box[u]$. Then, w and u agree on all $\Box$-formulae in $Cl(\varphi)$, so $M, u \models \Box\varrho_{Ags}$, and consequently, $M, u \models \varrho_{Ags}$. Therefore, $[u] \in U'_{Ags}$. As $[u]$ was an arbitrary $R'_\Box$-successor of $[w]$, $R'_\Box([w]) \subseteq U'_{Ags}$. Since $[w] \in W'$ is an arbitrary world, for every world $[w] \in W'$, $R'_\Box([w]) \subseteq U'_{Ags}$.

It is left for us to show that each U'_C is R'_C-closed, and (6) of Def. 6.3.3 holds.

As for R'_C-closure of every U'_C, assume $[w] \in U'_C$ and $[w]R'_C[v]$. Then, $\varrho_C \in Cl(\varphi)$ and $M, w \models \varrho_C$. Consequently, $M, w \models [stit]_C\varrho_C$. From $\varrho_C \in Cl(\varphi)$ it follows that $[stit]_C\varrho_C \in Cl(\varphi)$. $[w]R'_C[v]$ entails that $M, w \models [stit]_C\varrho_C$ iff $M, v \models [stit]_C\varrho_C$. So $M, v \models [stit]_C\varrho_C$. By reflexivity of R_C, $M, v \models \varrho_C$. Then, $[v] \in U'_C$.

As for (6) of Def. 6.3.3, we are to prove that $R'_\Box([w]) \cap U'_{C_1} \neq \emptyset, \ldots, R'_\Box([w]) \cap U'_{C_j} \neq \emptyset$ entails $R'_\Box([w]) \cap U'_{C_1} \cap \ldots \cap U'_{C_j} \neq \emptyset$ for every $C_1, \ldots, C_j \subseteq Ags$, such that $C_1 \cap \ldots \cap C_j \neq \emptyset$. Assume that $[w]R'_\Box[v_1], \ldots, [w]R'_\Box[v_j]$ and $[v_1] \in U'_{C_1}, \ldots, [v_j] \in U'_{C_j}$. Then, $M, v_1 \models \varrho_{C_1}, \ldots, M, v_j \models \varrho_{C_j}$; therefore, $M, v_1 \models \Diamond\varrho_{C_1}, \ldots, M, v_j \models \Diamond\varrho_{C_j}$. $\varrho_{C_1}, \ldots, \varrho_{C_j} \in Cl(\varphi)$ entails $\Diamond\varrho_{C_1}, \ldots, \Diamond\varrho_{C_j} \in Cl(\varphi)$. As $[w]R'_\Box[v_1], \ldots, [w]R'_\Box[v_j]$, w agrees with $v_1, \ldots, v_j$ on all $\Box$ (hence, $\Diamond$)-formulae in $Cl(\varphi)$. Hence, $M, w \models \Diamond\varrho_{C_1} \wedge \ldots \wedge \Diamond\varrho_{C_j}$. Then, as M satisfies 6 of Def.6.3.3, there should be a world $u \in W$, such that $wR_\Box u$ and $M, u \models \varrho_{C_1} \wedge \ldots \wedge \varrho_{C_j}$. Take $[u] \in W'$. As $wR_\Box u$, w and u agree on all $\Box$-formulae, so $[w]R'_\Box[u]$. Since $M, u \models \varrho_{C_1} \wedge \ldots \wedge \varrho_{C_j}$, $[u] \in U'_{C_1} \cap \ldots \cap U'_{C_j}$. □

6.3.13. THEOREM. *The satisfiability problem of a given formula $\varphi \in \mathcal{L}_{pSTIT}$ in SF_{Ags}(**pSTIT**) is decidable.*

Proof:
Take any formula $\varphi \in \mathcal{L}_{pSTIT}$. By Theorems 6.3.9, 6.3.10 and 6.3.12, if φ is satisfiable in SF_{Ags}(**pSTIT**), it is satisfiable in a $\mathbf{C_2}$-model of size at most $2^{|sub(\varphi)|+4|Gr|}$. So, we can nondeterministically guess the pointed $\mathbf{C_2}$-model (M, w) and check whether $M, w \models \varphi$ in polynomial time [BRV01]. □

6.3.2 Temporal group STIT logic with disjoint groups

In this subsection, we return to the standard group STIT logic and show that a family of its fragments is finitely axiomatisable and decidable. Our results will easily be transferable to the analogous fragments of plan-restricted group STIT logic. First, let us review the results about well-behaved group STIT fragments

we already have in the literature.

Two kinds of fragments are usually studied: temporal and atemporal. The latter received slightly more attention, since dropping the temporal modalities from the language significantly reduces the complexity of the formalism, in both the everyday and computational sense of the word "complexity". In [Sch12], Schwarzentruber studied the fragments of atemporal STIT logic, where group modalities are restricted. Namely, assume we have an arbitrary collection of subsets $\mathcal{G} \subseteq Ags$. Then, we define the corresponding fragment of $\mathcal{L}_{STIT}$ as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid [stit]_C\varphi$$

where $C \in \mathcal{G}$. In other words, we lack modalities for some sets of agents. As Schwarzentruber showed, such fragments may be decidable and axiomatisable, depending on the structure of $\mathcal{G}$. Assume that $Ags = \{1, \ldots, k{\times}m{+}m{-}1\}$, where k and m are integers, s.t. $k \geq 0$ and $m \geq 2$. Then, the corresponding fragment is decidable and axiomatisable, if $\mathcal{G}$ is the subset of the lattice $Lat \subseteq 2^{Ags}$, as depicted on the Haase diagram in Figure 6.4.

To the best of our knowledge, this family of atemporal fragments in the richest we know of so far. It is rather expressive: for instance, we can reason about individual agents and the grand coalition (i.e., $\mathcal{G} = \{\{i\} \mid i \in Ags\} \cup \{Ags\}$) or chains of coalitions (i.e., $\mathcal{G} = \{J_1, \ldots, J_n\}$, such that $J_1 \subseteq \ldots \subseteq J_n$). Yet it suffers from one important limitation: we cannot handle disjoint non-singleton groups. Think of opposing voting coalitions, sports teams, etc. Such groups are pairwise disjoint, and cannot be embedded in Lat, as every non-empty non-singleton group in Lat contains $\{1, \ldots, m-1\}$.

As for temporal fragments, we have the results for next-time STIT logics, often referred to as XSTIT logics. The grammar of the corresponding fragments is defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid [stit]_C X\varphi$$

where $C \subseteq Ags$. Here, we get to reason about time only in scope of $[stit]_C$-modality, and only about the immediate successors. In [Pay14], a finite sound and complete axiom system for this fragment is established, and decidability is proven. Notably, the semantics in this paper is generalised: instead of Kripke STIT models, a broader class (which we will later define as pre-models in Def. 6.3.16) is used. As for axioms for validities in Kripke STIT models, the richest temporal fragment we know of is due to [Lor13], where the only non-singleton group of the language is Ags:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid G\varphi \mid H\varphi \mid [stit]_i\varphi \mid [stit]_{Ags}\varphi$$

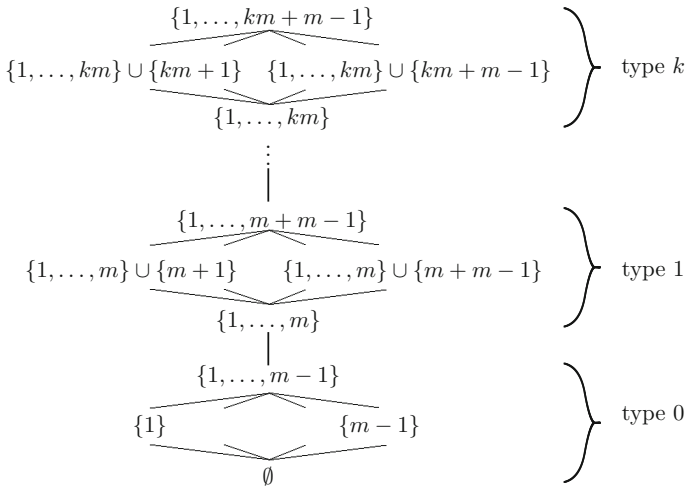

Figure 6.4: The lattice of Ags's subsets. The figure is from [Sch12, Figure 11, p.1016].

where $i \in Ags$. In Lorini's paper, X and Y are lacking, since discreteness of time is not assumed. The axiomatisation has a non-standard, infinite inference rule, which prevents us from establishing decidability of the logic. In what follows, we present the metalogical results for a richer temporal fragment, where we have $[stit]$-modalities for disjoint groups, as well as X and Y.

Given the finite set of agents $Ags = \{1, \ldots, n\}$, let $\mathcal{G} = \{C_1, \ldots, C_k\} \subseteq 2^{Ags}$ be a partition of Ags. Intuitively, elements of $\mathcal{G}$ are proper groups that we assume to be pairwise disjoint. Then, we recursively define the fragment $\mathcal{L}^{t\mathcal{G}}_{STIT}$ of $\mathcal{L}^{t}_{STIT}$ as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid X\varphi \mid Y\varphi \mid G\varphi \mid H\varphi \mid [stit]_i\varphi \mid [stit]_C\varphi$$

where $i \in Ags$, $C \in \mathcal{G}$. Note that for any set of agents $C \in Gr \setminus \mathcal{G}$, $[stit]_C\varphi$ is not a well-formed formula of $\mathcal{L}^{t\mathcal{G}}_{STIT}$. It is handy to denote the set of agents and groups that have their $[stit]$-modalities in $\mathcal{L}^{t\mathcal{G}}_{STIT}$ as $Mod = Ags \cup \mathcal{G}$.

$\mathcal{L}^{t\mathcal{G}}_{STIT}$ is interpreted over Kripke STIT frames (as in Def. 3.2.8) in a standard way. It makes sense to slightly rewrite Def. 3.2.8 for purely technical reasons.

6.3.14. Definition (Kripke STIT frames **F**). A Kripke STIT frame is a tuple of the form:

$$F = (W, R_{\prec}, R_{\Box}, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}})$$

1. $W \neq \emptyset$ is a set of time points;
2. $R_{\Box}, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}}$ are equivalence relations on W, such that:

(a) $R_i \subseteq R_\Box$ for every $i \in Ags$;

(b) *Group actions*: for every $C \in \mathcal{G}$, $R_C = \bigcap_{i \in C} R_i$;

(c) **Independence of agents (IoA):** for any sequence of $R_\Box$-connected time points $(w_1, \ldots, w_n)$ and **any pairwise disjoint indices** $I_1, \ldots, I_n \in Mod$, $\bigcap_{i \in \{1,\ldots,n\}} R_{I_i}(w_i) \neq \emptyset$;

3. $R_\prec$ is a serial relation on W. $R_<$ is the transitive closure of $R_\prec$. $R_\prec$ has the following properties:

(a) *No branching backwards*: if $vR_\prec w$ and $uR_\prec w$, then $u = v$;

(b) *No branching forwards*: if $wR_\prec u$ and $wR_\prec v$, then $u = v$;

(c) **No choice between undivided histories (NCBUH)**: for every $C \in \mathcal{G}$: $R_\prec \circ R_\Box \subseteq R_C \circ R_\prec$;

(d) *Irreflexivity of* $R_<$: If $wR_\Box v$, then $(w, v) \notin R_<$ and $(v, w) \notin R_<$.

Abbreviations: for every $w \in W$, $H_w = \{v \in W \mid wR_< v \text{ or } vR_< w \text{ or } w = v\}$ is the history to which w belongs.

Two definitions of **F**-frames (Def. 3.2.8 and Def. 6.3.14) are equivalent. They differ only in the formulations of the *Independence of agents* (IoA) and *No choice between undivided histories* (NCBUH) conditions. As for the former, in Def. 3.2.8, unlike in Def. 6.3.14, (IoA) is written just in terms of individual agents, not disjoint indices from Mod. Yet, the equivalence of two formulations follows directly from (2b) of Def. 6.3.14. As for the latter, in Def. 6.3.14, we have rewritten (NCBUH) in terms of groups in $\mathcal{G}$ rather than the grand coalition Ags. The two formulations, however, are equivalent.

6.3.15. Proposition. *Given any Kripke STIT frame* $F = (W, R_\prec, R_\Box, \{R_i\}_{i \in Ags}, \{R_C\}_{C \in \mathcal{G}})$*, the following two conditions are equivalent:*

1. $R_\prec \circ R_\Box \subseteq \bigcap R_{i \in Ags} \circ R_\prec$.
2. $\forall C \in \mathcal{G} : R_\prec \circ R_\Box \subseteq R_C \subseteq R_\prec$.

Proof:
($\Rightarrow$) Assume that $R_\prec \circ R_\Box \subseteq \bigcap R_{i \in Ags} \circ R_\prec$. Let $(w, v) \in R_\prec \circ R_\Box$ for some $w, v \in W$. By our assumption, it follows that $(w, v) \in \bigcap_{i \in Ags} R_i \circ R_\prec$. Then, there exists some $u \in W$, such that $(w, u) \in \bigcap_{i \in Ags} R_i$ and $uR_\prec v$. For any $C \in \mathcal{G}$, $\bigcap_{i \in Ags} R_i \subseteq R_C$. Therefore, $wR_C u$ for any $C \in \mathcal{G}$. Consequently, for any $C \in \mathcal{G}$, there exists a world u, such that $wR_C u$ and $uR_\prec v$, i.e., $(w, v) \in R_C \circ R_\prec$.

($\Leftarrow$) Assume that for any $C \in \mathcal{G}$, $R_\prec \circ R_\Box \subseteq R_C \circ R_\prec$. Let $(w, v) \in R_\prec \circ R_\Box$ for some $w, v \in W$. Then, by our assumption, $(w, v) \in R_C \circ R_\prec$ for any $C \in \mathcal{G}$. In other words, there exists worlds $\{u_{C_i}\}_{C_i \in \mathcal{G}} \subseteq W$, such that $wR_{C_i} u_{C_i}$ and $u_{C_i} R_\prec v$

for every $C_i \in \mathcal{G}$. Take any $u_{C_i}, u_{C_j} \in \{u_{C_i}\}_{C_i \in \mathcal{G}}$. As we have shown, $u_{C_i} R_{\prec} v$ and $u_{C_j} R_{\prec} v$. Then, by (3a), $u_{C_i} = u_{C_j}$. Therefore, for every $C_i \in \mathcal{G}$, $u_{C_i} = u$ for some $u \in W$. Since $w R_{C_i} u_{C_i} = u$ for every $C_i \in \mathcal{G}$, $(w, u) \in \bigcap_{C_i \in \mathcal{G}} R_{C_i}$. By (2b) of Def. 6.3.14, $(w, u) \in \bigcap_{i \in C_1} R_i \cap \ldots \cap \bigcap R_{i \in C_k} R_i$. Since $\mathcal{G} = \{C_1, \ldots, C_k\}$ partitions Ags, $(w, u) \in \bigcap_{i \in Ags} R_i$. So there exists a world $u \in W$, such that $(w, u) \in \bigcap_{i \in Ags} R_i$ and $u R_{\prec} v$, i.e., $(w, v) \in \bigcap_{i \in Ags} R_i \circ R_{\prec}$. □

Next, we define a finite axiom system $\mathbf{L}^{t\mathcal{G}}_{STIT}$, as in Table 6.2, and proceed with proving it to be sound and complete w.r.t. the set of validities $\{\varphi \in \mathcal{L}^{t\mathcal{G}}_{STIT} \mid \mathbf{F} \models \varphi\}$. First, let us provide some intuitive explanations for axioms and their connections with semantics of $\mathcal{L}^{t\mathcal{G}}_{STIT}$.

(CPL)	Tautologies of classical propositional logic
(S5)	S5 axioms and rules of inference for $\Box$, $[stit]_i$, $[stit]_C$
(KD_X)	KD axioms and rules for X
(K_Y)	K axiomss and rules for Y, G, H
(Det)	$\neg Q \neg \varphi \to Q\varphi$ for $Q \in \{X, Y\}$
(Reverse1)	$XY\varphi \leftrightarrow \varphi$
(Reverse2)	$YX\varphi \leftrightarrow \varphi$
(Loop)	$X\varphi \to X\neg Y \bot$
(FP_G)	$G\varphi \leftrightarrow (X\varphi \wedge XG\varphi)$
(IN_G)	$G(\varphi \to X\varphi) \to (X\varphi \to G\varphi)$
(FP_H)	$H\varphi \leftrightarrow (Y\varphi \wedge YH\varphi)$
(IN_H)	$H(\varphi \to Y\varphi) \to (Y\varphi \to H\varphi)$
(Incl)	$\Box\varphi \to [stit]_i\varphi$ for every $i \in Ags$
$(i - C)$	$[stit]_i\varphi \to [stit]_C\varphi$ for every $i \in C \in \mathcal{G}$
(Ind)	$(\Diamond[stit]_{I_1}\varphi_1 \wedge \ldots \wedge \Diamond[stit]_{I_n}\varphi_n) \to \Diamond([stit]_{I_1}\varphi_1 \wedge \ldots \wedge [stit]_{I_n}\varphi_n)$ for any disjoint $I_1, \ldots, I_n \in Mod$
(NCBUH_X)	$[stit]_C X\varphi \to X\Box\varphi$ for any $C \in \mathcal{G}$
(NCBUH_Y)	$Y[stit]_C\varphi \to \Box Y\varphi$ for any $C \in \mathcal{G}$
(MP)	If $\vdash \varphi$ and $\vdash \varphi \to \psi$, then $\vdash \psi$

Table 6.2: Logic $\mathbf{L}^{t\mathcal{G}}_{STIT}$.

(S5) axioms guarantee that $R_{\Box}$, R_i and R_C are equivalence relations. (KD_X) and (K_Y) ensure that $R_{\prec}$ is a serial relation, and $R^{-1}_{\prec}$ is well-defined. (Det) ensures that both X and Y are functional: if the world has an immediate successor (which is always true), then there is at most one immediate successor; if the world has an immediate predecessor, then there is at most one immediate predecessor. In other words, (Det) ensures that $R_{\prec}$ branches neither forwards, nor backwards. (Reverse1) and (Reverse2) axiomatically express that $w R_{\prec} v$ and $v R^{-1}_{\prec} u$ entail $w = u$. (Loop) ensures that whenever $w R_{\prec} v$, there is a world u, such that $v R^{-1}_{\prec} u$. (KD_X), (K_Y), (Det), (Reverse1), (Reverse2), and (Loop) together express

that $R_\prec$ and $R_\prec^{-1}$ are functional relations that are reverse w.r.t. one another. (FP_H), (FP_G) express that $R_<$ and $R_<^{-1}$ contain the transitive closures of $R_\prec$ and $R_\prec^{-1}$, respectively. (Nec) axiomatically expresses that $R_i \subseteq R_\Box$ for every $i \in Ags$. $(i - C)$ corresponds to the fact that $R_C \subseteq \bigcap_{i\in C} R_i$ (but not the other way around!). (Ind) ensures the independence of agents. (NCBUH_X) and (NCBUH_Y) express the no choice between undivided histories: $R_\prec \circ R_\Box \subseteq R_C \circ R_\prec$ and $R_\Box \circ R_\prec^{-1} \subseteq R_\prec^{-1} \circ R_C$, respectively.[5] All these correspondences between axioms and properties of Kripke frames may be established via the Sahlqvist-van Benthem algorithm [Sah75; Ben14]. The curious reader may automatically verify them via, e.g., SQEMA [Vak10].

Some properties of Kripke STIT frames are left axiomatically unexpressed. As we will show later, $\mathcal{L}_{STIT}^{t\mathcal{G}}$ is not expressive enough to capture these conditions.

We proceed with showing that $\mathbf{L}_{STIT}^{t\mathcal{G}}$ is sound and weakly complete w.r.t. the class of Kripke STIT frames. Our proof strategy is the following:

1. Define a class of *pre-frames*: the broader class of structures that satisfy $R_C \subseteq \bigcap_{i\in C} R_i$ of of (2b) in Def. 6.3.14, but not necessarily that $\bigcap_{i\in C} R_i \subseteq R_C$;
2. Build a canonical model M^c for $\mathbf{L}_{STIT}^{t\mathcal{G}}$ in a specific way to ensure that M^c is a pre-frame, thereby proving that $\mathbf{L}_{STIT}^{t\mathcal{G}}$ is sound and complete w.r.t. the class of pre-frames;
3. Finally, show that every pre-frame may be transformed to a $\mathcal{L}_{STIT}^{t\mathcal{G}}$-invariant Kripke STIT frame via a variation of the taking isomorphic copies construction.

6.3.16. Definition (Pre-frames $\mathbf{F}_{pre}$). $F = (W, R_\prec, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}}) \in \mathbf{F}_{pre}$ is a pre-frame iff it satisfies all the conditions of Def. 6.3.14, with one exception. (2b) is weakened: for every $C \in \mathcal{G}$, R_C is an equivalence relation, such that $R_C \subseteq \bigcap_{i\in Ags} R_i$.

$\mathbf{L}_{STIT}^{t\mathcal{G}}$ is sound and weakly complete w.r.t. the class of pre-frames. Here is how we prove it. First, we will define a canonical model M^c for $\mathbf{L}_{STIT}^{t\mathcal{G}}$ in a standard way. Then, we will take a filtration of the canonical model to ensure that $R_<$ and $R_<^{-1}$ are equal to the transitive closures of $R_\prec$ and $R_\prec^{-1}$, respectively. Then, we

[5] (NCBUH_Y) is redundant, as its correspondent follows from the correspondent of (NCBUH_X). For the contradiction, assume that $R_\prec \circ R_\Box \subseteq R_C \circ R_\prec$, but $R_\Box \circ R_\prec^{-1} \not\subseteq R_\prec^{-1} \circ R_C$. Then, from the latter, there exist x_2, y_2, y_1, such that $x_2 R_\Box y_2$ and $y_2 R_\prec^{-1} y_1$, but for no x_1 it holds that $x_2 R_\prec^{-1} x_1$ and $x_1 R_C y_1$. Observe that $y_1 R_\prec y_2$ and $y_2 R_\Box x_2$. Then, by $R_\prec \circ R_\Box \subseteq R_C \circ R_\prec$, there should exist x_1, such that $y_1 R_C x_1$ and $x_1 R_\prec x_2$. By the symmetry of R_C and the fact that $R_\prec^{-1}$ is the reverse of $R_\prec$, $x_2 R_\prec^{-1} x_1$ and $x_1 R_C y_1$. Contradiction. Hence, $R_\Box \circ R_\prec^{-1} \subseteq R_\prec^{-1} \circ R_C$. We still add both axioms, since the axiomatic derivation of (NCBUH_Y) from (NCBUH_X) via (Det), (Reverse) and (Loop) is quite convoluted. We will automatically get the syntactic derivability from the semantic entailment by establishing soundness and completeness results.

unravel the resulting filtration in a non-standard way to obtain a pre-model. Some steps of our proof are fairly standard or repeat the technique for the "until-free" fragment of Linear Time Temporal Logic [Bar25]. Such standard or "copy-paste" steps will be omitted, and the references to the lemmas in the original paper will be given instead.[6]

6.3.17. Definition (Canonical model). A canonical model for $\mathbf{L}^{t\mathcal{G}}_{STIT}$ is the following tuple:

$$M^c = (W^c, R^c_{\prec}, R^c_{<}, R^c_{\Box}, \{R^c_i\}_{i\in Ags}, \{R^c_C\}_{C\in\mathcal{G}}, V^c), \text{ where:}$$

1. $W^c = \{\Gamma \subseteq \mathcal{L}^{t\mathcal{G}}_{STIT} \mid \Gamma \text{ is a mcs of } \mathcal{L}^{t\mathcal{G}}_{STIT}\}$;
2. For $R^c \in \{R^c_{\prec}, (R^c_{\prec})^{-1}, R^c_{<}, (R^c_{<})^{-1}, R^c_{\Box}, R^c_i, R^c_C\}$ and $Z \in \{X, Y, G, H, \Box, [stit]_i, [stit]_C\}$, resp.: $\Gamma R^c \Delta$ iff for all $\varphi \in \mathcal{L}^{t\mathcal{G}}_{STIT}$, $Z\varphi \in \Gamma$ entails $\varphi \in \Delta$;
3. $V^c(p) = \{\Gamma \in W^c \mid p \in \Gamma\}$.

Importantly, using (Det), (Reverse1), (Reverse2) and (Loop) axioms, one can show that $\Gamma R^c_{\prec}\Delta$ iff $\Delta(R^c_{\prec})^{-1}\Gamma$ iff $\Delta = \{\varphi \mid X\varphi \in \Gamma\}$ and $\Gamma = \{\varphi \mid Y\varphi \in \Delta\}$. Truth lemma, i.e., $M^c, \Gamma \models \varphi$ iff $\varphi \in \Gamma$, holds, and the proof of this fact is standard [BRV01].

One can show (e.g., by Sahlqvist correspondence theorems) that M^c is almost an $\mathbf{F}_{pre}$-model. The only conditions that lack their canonical axiomatic correspondents and may fail, are:

1. Irreflexivity of $R^c_{<}$ (3d of Def. 6.3.14): There might be $\Gamma, \Delta \in W^c$, such that $\Gamma R^c_{<}\Delta$ and $\Gamma R^c_{\Box}\Delta$;
2. The transitive closure: $(R^c_{\prec})^* \subseteq R^c_{<}$, but the converse does not hold.

To ensure that $R^c_{<} = (R^c_{\prec})^*$, we take a filtration of the canonical model. Take any consistent formula φ. Then, let Σ be the minimal set of formulae that satisfies the following conditions:

$\Sigma 1)$ $\varphi \in \Sigma$;

$\Sigma 2)$ Σ is subformula-closed;

$\Sigma 3)$ If $G\psi \in \Sigma$, then $X\psi \in \Sigma$ and $XG\psi \in \Sigma$;

$\Sigma 4)$ If $H\psi \in \Sigma$, then $Y\psi$ and $YH\psi \in \Sigma$;

$\Sigma 5)$ $X\psi \in \Sigma$ iff $Y\psi$ in Σ;

$\Sigma 6)$ If $X\psi \in \Sigma$, then $X\neg\psi \in \Sigma$;

[6] To the best of our knowledge, the proof originally appeared in [Gab+80], in a relatively sketchy form. [Bar25] provides additional details of the same proof.

$\Sigma 7$) If $\psi \in \Sigma$, then $\Box\psi \in \Sigma$;

$\Sigma 8$) Σ is closed under Boolean connectives;

$\Sigma 9$) If $[stit]_I\psi_1, \ldots, [stit]_I\psi_n \in \Sigma$, then $[stit]_I(\psi_1 \wedge \ldots \wedge \psi_n) \in \Sigma$.

$\Sigma 10$) If $\langle stit\rangle_I\psi \in \Sigma$, then $[stit]_I\langle stit\rangle_I\psi \in \Sigma$ and $Y\langle stit\rangle_I\psi \in \Sigma$;

$\Sigma 11$) If $\Box\psi \in \Sigma$ or $[stit]_i\psi \in \Sigma$, then $[stit]_C\psi \in \Sigma$ for $C \in \mathcal{G}$, s.t. $i \in G$.

We call such Σ a *closure set of* φ. Note that Σ is finite, up to the logical equivalence. In other words, we can define a finite set $\overline{\Sigma} = \{\overline{\psi_1}, \ldots, \overline{\psi_n}\} \subseteq \Sigma$, such that any $\psi \in \Sigma$ has its corresponding formula $\overline{\psi} \in \overline{\Sigma}$, such that $\vdash_{\mathbf{L}^{t\mathcal{G}}_{STIT}} (\psi \leftrightarrow \overline{\psi})$.[7]

Next, we define the filtration of M^c through Σ as follows:

$$M^f = (W^f, R^f_{\prec}, R^f_{\Box}, \{R^f_i\}_{i\in Ags}, \{R^f_C\}_{C\in\mathcal{G}}, V^f)$$

1. $W^f = \{\Gamma \cap \Sigma \mid \Gamma \in W^c\}$. We denote elements of W^c as s, r, t, x, y, z, etc. For every $s \in W^f$, let $[s] = \{\Gamma \in W^c \mid \Gamma \cap \Sigma = s\}$. Note that $|W^f| \leq 2^{|\overline{\Sigma}|}$, so it is finite.
2. $s_1 R^f_{\prec} s_2$ iff $\exists\Gamma \in [s_1]\exists\Delta \in [s_2](\Gamma R^c_{\prec}\Delta)$;
3. $s_1 R^f_{<} s_2$ iff $\exists\Gamma \in [s_1]\exists\Delta \in [s_2](\Gamma R^c_{<}\Delta)$;
4. $s_1 R^f_{\Box} s_2$ iff for any $\Box\psi \in \Sigma$ and any $\Gamma \in [s_1], \Delta \in [s_2]$, $M^c, \Gamma \models \Box\psi$ iff $M^f, \Delta \models \Box\psi$;
5. $s_1 R^f_i s_2$ iff:
   (a) $s_1 R^f_{\Box} s_2$;
   (b) For any $[stit]_i\psi \in \Sigma$ and any $\Gamma \in [s_1], \Delta \in [s_2]$, $M^f, \Gamma \models [stit]_i\psi$ iff $M^f, \Delta \models [stit]_i\psi$
6. $s_1 R^f_C s_2$ iff:
   (a) $s_1 R^f_i s_2$ for any $i \in C$;
   (b) For any $[stit]_C\psi \in \Sigma$, $\Gamma \in [s_1]$, $\Delta \in [s_2]$: $M^c, \Gamma \models [stit]_C\psi$ iff $M^f, \Delta \models [stit]_C\psi$;
7. $V^f(p) = \{s \in W^f \mid \forall\Gamma \in [s] : p \in \Gamma\}$.

[7] ($\Sigma 8$) preserves finiteness up to the logical equivalence, since for any n logically non-equivalent formulae there are at most 2^{2^n}-many logically non-equivalent Boolean combinations of them. ($\Sigma 9$) preserves finiteness up to the logical equivalence, since $\vdash ([stit]_I\psi_1 \wedge \ldots \wedge [stit]_I\psi_n) \leftrightarrow [stit]_I(\psi_1 \wedge \ldots \wedge \psi_n)$. ($\Sigma 7$) preserves finiteness up to the logical equivalence, since by (S5) for $\Box$, $\vdash \Box\Box\varphi \leftrightarrow \Box\varphi$, $\vdash \Diamond\Box\varphi \leftrightarrow \Box\varphi$, $\vdash \Diamond\Diamond\varphi \leftrightarrow \Diamond\varphi$ and $\vdash \Box\Diamond\varphi \leftrightarrow \Diamond\varphi$. Other conditions do not infinitely multiply formulae in Σ.

By the standard filtration lemma, for every $\psi \in \Sigma$, $M^f, s \models \psi$ iff $\forall \Gamma \in [s] : M^c, \Gamma \models \psi$ [BBW06, p.268]. From the Truth lemma, $\forall \psi \in \Sigma$, $M^f, s \models \psi$ iff $\forall \Gamma \in [s] : \psi \in \Gamma$.

Next, we show that (IoA) and (NCBUH) hold for M^f.

6.3.18. Lemma ((IoA) holds for M^f). *For every $R^f_\Box$-connected $s_1, \ldots, s_n \in W^f$ and for every pairwise disjoint $I_1, \ldots, I_n \in Mod$, there exists $t \in W^f$, such that $s_1 R^f_{I_1} t, \ldots, s_n R^f_{I_n} t$.*

Proof:
Assume that $s_1, \ldots, s_n \in W^f$, such that $s_1 R^f_\Box \ldots R^f_\Box s_n$. Let $I_1, \ldots, I_n \in Mod$ be pairwise disjoint. For every s_i, let $\Box s_i = \{\Box\psi, \Diamond\chi \in \Sigma \mid \Box\psi, \Diamond\chi \in s_i\}$, and

$$\gamma_{s_i} := \bigwedge \{\overline{\psi} \in \overline{\Sigma} \mid [stit]_{I_i}\overline{\psi} \in s_i\}$$

$$\nu_{s_i} = \bigwedge \{\langle stit\rangle_{I_i}\overline{\psi} \mid \overline{\psi} \in \overline{\Sigma}, \langle stit\rangle_{I_i}\overline{\psi} \in s_i\}$$

$\Box s_i$ is the $\Box$-theory of s_i; γ_{s_i} is logically equivalent to the set $\{\psi \in \Sigma | [stit]_i\psi \in s_i\}$, while ν_{s_i} is equivalent to $\{\langle stit\rangle_{I_i}\psi \mid \langle stit\rangle_{I_i}\psi \in s_i\}$. The formula $[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i})$ encodes $[stit]_{I_i}$-theory of s_i. Moreover, by (Σ9) and (Σ10), $[stit]_{I_i}(\gamma_i \wedge \nu_i) \in \Sigma$, so $[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i}) \in s_i$ for every s_i. Since $[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i}) \vdash \Diamond[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i})$ and, by (Σ7), $\Diamond[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i}) \in \Sigma$, $\Diamond[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i}) \in s_i$. For any $x \in W^f$, if $[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i}) \in x$, then x and s_i agree on all $[stit]_{I_i}$-formulae in Σ. Since $s_1 R^f_\Box \ldots R^f_\Box s_n$, $\Box s_1 = \ldots = \Box s_n$. Let us denote the $\Box$-theory of all these worlds as $\Box s_i$. Since every $\Diamond[stit]_{I_i}(\gamma_{s_i} \wedge \nu_{s_i}) \in \Sigma$ is a $\Box$-formula of s_i, $\Diamond[stit]_{I_1}\gamma_{s_1}, \ldots, \Diamond[stit]_{I_n}\gamma_{s_n} \in \Box s_i$. By the axiom (Ind), $\Box s_i \vdash \Diamond([stit]_{I_1}(\gamma_{s_1} \wedge \nu_{s_1}) \wedge \ldots \wedge [stit]_{I_n}(\gamma_{s_n} \wedge \nu_{s_n}))$. So, $[stit]_{I_1}(\gamma_{s_1} \wedge \nu_{s_1}) \wedge \ldots \wedge [stit]_{I_n}(\gamma_{s_n} \wedge \nu_{s_n})$ is consistent with $\Box s_i$. Otherwise, if $\Box s_i \vdash \neg([stit]_{I_1}(\gamma_{s_1} \wedge \nu_{s_1}) \wedge \ldots \wedge [stit]_{I_n}(\gamma_{s_n} \wedge \nu_{s_n}))$ by (RN) for $\Box$, $\Box s_i \vdash \Box\neg([stit]_{I_1}(\gamma_{s_1} \wedge \nu_{s_1}) \wedge \ldots \wedge [stit]_{I_n}(\gamma_{s_n} \wedge \nu_{s_n}))$, contradicting $\Box s_i \vdash \Diamond([stit]_{I_1}(\gamma_{s_1} \wedge \nu_{s_1}) \wedge \ldots \wedge [stit]_{I_n}(\gamma_{s_n} \wedge \nu_{s_n}))$. Then, there exists some mcs $\Gamma \in W^c$, such that

$$\Box s_i \cup \{([stit]_{I_1}(\gamma_{s_1} \wedge \nu_{s_1}) \wedge \ldots \wedge [stit]_{I_n}(\gamma_{s_n} \wedge \nu_{s_n}))\} \subseteq \Gamma$$

Take $t = \Gamma \cap \Sigma$. By definition, $s_i R^f_\Box t$, and t agree on $[stit]_{I_i}$-formulae with s_i for every $i \in \{1, \ldots, i\}$. Therefore, $s_1 R^f_{I_1} t, \ldots, s_n R^f_{I_n} t$. □

6.3.19. Lemma ((NCBUH) holds for M^f). *For any $C \in \mathcal{G}$, $x R^f_\prec y$ and $y R^f_\Box z$, then there exists $w \in W^f$, such that $x R^f_C w$ and $w R^f_\prec z$*

Proof:
Let $x R^f_\prec y$ and $y R^f_\Box z$ for some $x, y, z \in W^f$. We define the characteristic formula for z as follows:

$$\psi_z = \bigwedge_{\overline{\psi}\in z\cap\overline{\Sigma}} \overline{\psi} \wedge \bigwedge_{\overline{\phi}\in\overline{\Sigma}\setminus z} \neg\overline{\phi}$$

Note that, by (Σ8), $\psi_z \in \Sigma$, so $\psi_z \in z$. It is straightforward that for every $\Gamma \in W^f$, $\psi_z \in \Gamma$ iff $\Gamma \in [z]$. Since $\psi_z \in z$ and $\psi_z \vdash \Diamond\psi_z$, and, by ($\Sigma$7), $\Diamond\psi_z \in \Sigma$, we conclude that $\Diamond\psi_z \in z$. Since $yR^f_\Box z$, y and z agree on all $\Diamond$-formulae in Σ, so $\Diamond\psi_z \in y$. Therefore, for every $\Delta' \in [y]$, $\Diamond\psi_z \in \Delta'$. Since $xR^f_\prec z$, by definition of $R^f_\prec$, there exists $\Gamma \in [x]$ and $\Delta \in [y]$, such that $\Gamma R^c_\prec \Delta$. Since $\Diamond\psi_z \in \Delta$ (as for any $\Delta' \in [y]$), $X\Diamond\psi_z \in \Gamma$. Therefore, by the axiom (NCBUH$_X$), $\Gamma \vdash \langle stit\rangle_C X\psi_z$. Then, there exists an mcs $\Lambda \in W^c$, such that $\Gamma R^c_C \Lambda$ and $X\psi_z \in \Lambda$. Since $\Gamma R^c_C\Lambda$, $\Gamma R^c_\Box\Lambda$, so Γ and Λ agree on all $\Box$-formulae. As $X\psi_z \in \Lambda$, there should be some $\Theta \in W^f$, such that $\Lambda R^f_\prec \Theta$ and $\psi_z \in \Theta$. Hence, $\Theta \in [z]$. Let $w = \Lambda \cap \Sigma$. Then, $\exists\Lambda \in [w]\exists\Theta \in [z] : \Lambda R^c_\prec\Theta$. Therefore, $wR^f_\prec z$. Since $\Lambda \in [w]$, $\Gamma \in [x]$ and Γ and Λ agree on all $\Box$- and $[stit]_C$-formulae, $xR^f_C w$. So $xR^f_C wR^f_\prec z$. □

Analogously, we can prove that the following property holds:

$$\forall x, y, z \in W^f(xR^f_\Box y \wedge y(R^f_\prec)^{-1}z) \to \exists w \in W^f(x(R^f_\prec)^{-1}w \wedge wR^f_C z))) \qquad (\text{NCBUH}_Y)$$

We just have to use the axiom (NCBUH$_Y$) instead of (NCBUH$_X$). Note that the weaker property, NY, follows from NCBUH$_Y$, as $R^f_C \subseteq R^f_\Box$:

$$\forall x, y, z \in W^f(xR^f_\Box y \wedge y(R^f_\prec)^{-1}z) \to \exists w \in W^f(x(R^f_\prec)^{-1}w \wedge wR^f_\Box z))) \qquad (\text{NY})$$

6.3.20. Corollary. *In M^f, properties NY and NCBUH$_Y$ hold.*

6.3.21. Lemma. *In the filtrated model M^f, $R^f_< = (R^f_\prec)^*$.*

Proof:
See [Bar25, Proposition 4.1]. □

Note that $R^f_\prec$ is not necessarily functional. There might be $\Gamma, \Delta \in W^c$, such that $\Gamma\cap\Sigma = \Delta\cap\Sigma = s_1$, $\Gamma R^f_\prec \Gamma'$, $\Delta R^f_\prec \Delta'$ and $s_2 = \Gamma'\cap\Sigma \neq \Delta'\cap\Sigma = s_3$. Then, $s_1 R^f_\prec s_2$ and $s_1 R^f_\prec s_3$, while $s_2 \neq s_3$. To fix this, we will show how every filtration can be unravelled into a proper pre-model. Our technique is inspired by [Gab+80].

The construction we are to present is quite convoluted, so we will work through a simple example. Assume M^f is the filtration of an one-agent canonical model (as in Fig.6.5), where:

1. $W^f = \{k, s, t, r\}$;

2. $R^f_{\prec} = \{(k,s),(s,t),(s,r),(t,s),(t,s)\}$ (note that it is not functional);
3. $R^f_{\Box} = R^f_i$ are equivalence relations, such that $[R^f_{\Box}] = [R^f_i] = \{\{k\},\{s\},\{r,t\}\}$;
4. $V^f(p) = \emptyset$ for every $p \in Var$.

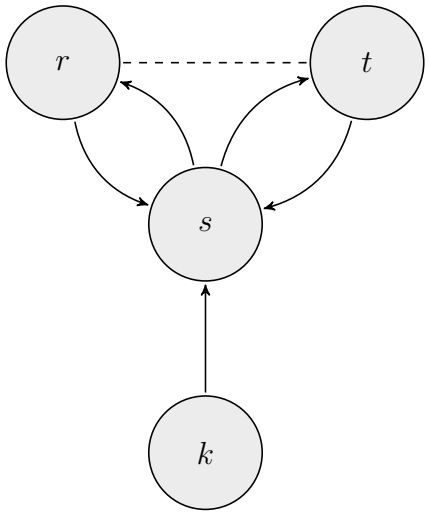

Figure 6.5: Filtration M^f. $R^f_{\prec}$ is represented by solid arrows, $R^f_{\Box}$, R^f_i – by dashed lines. Reflexive arrows for $R^f_{\Box}$ and R^f_i are omitted.

Let $succ(s) = \{x \in W^f \mid sR^f_{\prec}x\}$ and $pre(s) = \{y \in W^f \mid yR^f_{\prec}s\}$ denote the set of immediate successors and predecessors of $s \in W^f$, respectively. Note that $succ(s)$ is finite and $pre(s)$ is either empty or finite for every $s \in W^f$. Let ϵ be a placeholder to denote the absence of immediate predecessors of some worlds. So instead of $pre(s) = \emptyset$, we write $pre(s) = \{\epsilon\}$, meaning the same thing. By convention, $pre(\epsilon) = \{\epsilon\}$ and $succ(\epsilon) = \{\epsilon\} \cup \{s \in W^f \mid pre(s) = \{\epsilon\}\}$.

For each $s \in W^f$, we construct infinite strings $\Lambda_s \subseteq (W^f \cup \{\epsilon\})^{\mathbb{Z}}$, such that for every $(s_i)_{i\in\mathbb{Z}} \in \Lambda_s$:

1. $s_0 = s$;
2. For every non-negative $n \in \mathbb{N}$, $s_{n+1} \in succ(s_n)$;
3. For every non-negative $n \in \mathbb{N}$, if $s_n = t$ occurs infinitely many times in $(s_i)_{i\in\mathbb{N}}$, then so does every $x \in succ(t)$;
4. For every non-positive $i \in \mathbb{Z}$, $s_{i-1} \in pre(s_i)$;
5. For every non-positive $n \in \mathbb{Z}$, if $s_n = t$ occurs infinitely many times in $(s_i)_{i\in\mathbb{Z}\setminus\mathbb{N}^+}$, then so does every $x \in pre(t)$.[8]

[8]By $\mathbb{N}^+$ we denote the set of positive natural numbers, i.e., $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$.

We know that such sequences exist for every world: $\Lambda_s \neq \emptyset$. For example, for every $x \in W^f$, as it has finitely many immediate successors and predecessors, we can define functions $f_x : \mathbb{N} \to succ(x)$ and $g_x : \mathbb{N} \to pre(x)$, such that $f_x^{-1}(y)$ and $g_x^{-1}(z)$ for every $y \in succ(x)$ and $z \in pre(x)$ are infinite. More concrete examples of such functions would be cyclic or connected periodic orderings[9] of $suc(x)$ and $pre(x)$. Then, we construct a sequence $\sigma = (s_i)_{i\in\mathbb{Z}}$ for an arbitrary $s \in W^f$ as follows:

1. $s_0 = s$;
2. For every non-negative $n, i \in \mathbb{N}$, if $s_n = x$ is the ith occurrence of x in σ, then $s_{n+1} = f_x(i)$.
3. For every non-positive $n, i \in \mathbb{Z}$, if $s_n = x$ is the ith occurrence of x in σ, then $s_{n-1} = g_x(-i)$.

Clearly, if any world $t \in W^f$ occurs infinitely many times in the non-negative suffix of σ, then so does every immediate successor of t. Analogously, if t occurs infinitely many times in the non-positive prefix of σ, so does every immediate predecessor of t. Also note that ϵ can occur only in the negative prefix of every sequence, and it is always preceded by $(\epsilon)^\omega$.

Let us construct two sequences for the world $s \in W^f$ from our Example in Fig. 6.5. Note that $succ(s) = \{t, r\}$ and $pre(s) = \{t, k, r\}$. Let $f_s : \mathbb{N} \to succ(s)$ such that $f_s(0) = t$, $f_s(1) = r$, $f_s(2) = t$, $f_s(3) = r$, etc. Let $g_s : \mathbb{N} \to pre(s)$, such that $g_s(0) = t$, $g_s(1) = k$, $g_s(2) = r$, $g_s(3) = t$, etc. As any other world has a unique successor and predecessor, we omit defining analogous functions for them. Then:

$$\sigma_s = (\ldots, \epsilon, \epsilon, k, s, t, s_0 = s, t, s, r, s, t, s, r, \ldots)$$

If we take the same function $g_s : \mathbb{N} \to pre(s)$ and another enumeration $f'_s : \mathbb{N} \to succ(s)$, such that $f'_s(0) = r, f'_s(1) = t, f'_s(2) = r$, etc., we will get the following sequence:

$$\sigma'_s = (\ldots, \epsilon, k, s, t, s'_0 = s, r, s, t, s, r, s, t, \ldots)$$

In both sequences, if a world occurs infinitely many times in the non-negative suffix, so does all of its successors: s, t and r occur infinitely many times in the non-negative suffixes of both sequences, while $succ(s) = \{t, r\}$, $succ(t) =$

[9] By cyclic ordering of a finite set X we intuitively mean arranging elements of X in a circle, and defining the function $f : \mathbb{N} \to X$ clockwise. By connected periodic ordering of X we mean that a function $f : \mathbb{N} \to X$ is such that the sequence $\langle f(i)\rangle_{i\in\mathbb{N}}$ is an infinite string s^ω, where s contains all elements of X. For example, if $X = \{a, b\}$, then $f : \mathbb{N} \to X$, such that $\langle f(i)\rangle_{i\in\mathbb{N}} = (aab)^\omega$ is a connected periodic ordering of X.

$succ(r) = \{s\}$. The only "world" that occurs infinitely many times in the non-positive prefixes of both sequences is ϵ, being the only predecessor of itself.

For every $s \in W^f$, by Λ_s we denote the set of all such sequences σ_s. For every sequence σ_s and any $i \in \mathbb{Z}$, by $\sigma_s(i)$ we denote the world that occupies on the ith position of the sequence.

Finally, we define the pre-model to satisfy φ.

6.3.22. DEFINITION. For every filtration M^f, we define the pre-model M as follows:

$$M = (W, R_\prec, R_\Box, \{R_i\}_{i \in Ags}, \{R_C\}_{C \in \mathcal{G}}, V)$$

1. $W = \{(\sigma_s, i) \mid s \in W^f, \sigma_s \in \Lambda_s, i \in \mathbb{Z}, \sigma_s(i) \neq \epsilon\}$;
2. $(\sigma_s, i) R_\prec (\sigma_t, j)$ iff $\sigma_s = \sigma_t$, $j = i + 1$;
3. $(\sigma_s, i) R_< (\sigma_t, j)$ iff $\sigma_s = \sigma_t$, $j > i$;
4. $(\sigma_s, i) R_\Box (\sigma_t, j)$ iff $i = j$, $\sigma_s(i) R^f_\Box \sigma_t(j)$;
5. $(\sigma_s, i) R_i (\sigma_t, j)$ iff $i = j$ and $\sigma_s(i) R^f_i \sigma_t(j)$;
6. $(\sigma_s, i) R_C (\sigma_t, j)$ iff $i = j$ and $\sigma_s(i) R^f_C \sigma_t(j)$;
7. $V(p) = \{(\sigma_s, i) \mid \sigma_s(i) \in V^f(p)\}$.

For the illustration of pre-model construction for our working example, see Figure 6.6.

We show that M is a pre-frame. Directly from definitions, $R_\Box$, R_i and R_C are equivalence relations, such that $R_i \subseteq R_\Box$, $R_C \subseteq \bigcap_{i \in C} R_i$. It is also straightforward that $R_\prec$ branches neither forwards nor backwards. Obviously, $R_<$ is a transitive closure of $R_\prec$, and if $(\sigma_s, i) R_\Box (\sigma_t, j)$, then neither $(\sigma_s, i) R_< (\sigma_t, j)$ nor $(\sigma_t, j) R_< (\sigma_s, i)$. So it is left for us to prove that the independence of agents and no choice between undivided histories hold. First, we prove one handy lemma.

6.3.23. LEMMA. *For every $(\sigma_s, i) \in W$, $t \in W^f$: if $\sigma_s(i) R^f_\Box t$, then there exists a sequence σ_x, such that $(\sigma_s, i) R_\Box (\sigma_x, i)$ and $\sigma_x(i) = t$.*

Proof:
By induction on $i \in \mathbb{N}$. Base-case: $i = 0$. Then, $\sigma_s(i) = \sigma_s(0) = s$. So $s R^f_\Box t$. Take any sequence $\sigma_t \in \Lambda_t$. By definition, $\sigma_t(0) = t$. Hence, $(\sigma_s, 0) R_\Box (\sigma_x, 0)$.

Induction hypothesis in the positive direction: for $n \in \mathbb{N}$, if $\sigma_s(n) R^f_\Box t$, then there exists a sequence σ_x, such that $(\sigma_s, n) R_\Box (\sigma_x, n)$ and $\sigma_x(n) = t$.

Induction step: Assume $\sigma_s(n+1) R^f_\Box t$. Let $\sigma_s(n+1) = r$ for some $r \in W^f$. Then, $r R^f_\Box t$. By definition of the sequence, there exists some $x = \sigma_s(n)$ and $x R^f_\prec r$. Therefore, $(x, t) \in R^f_\prec \circ R^f_\Box$. Since M^f satisfies (NCBUH), there exists $y \in W^f$, such that $x R^f_C y$ and $y R^f_\prec t$. Since $R^f_C \subseteq R^f_\Box$, $x R^f_\Box y$. Then, $\sigma_s(n) R^f_\Box y$. By (IH),

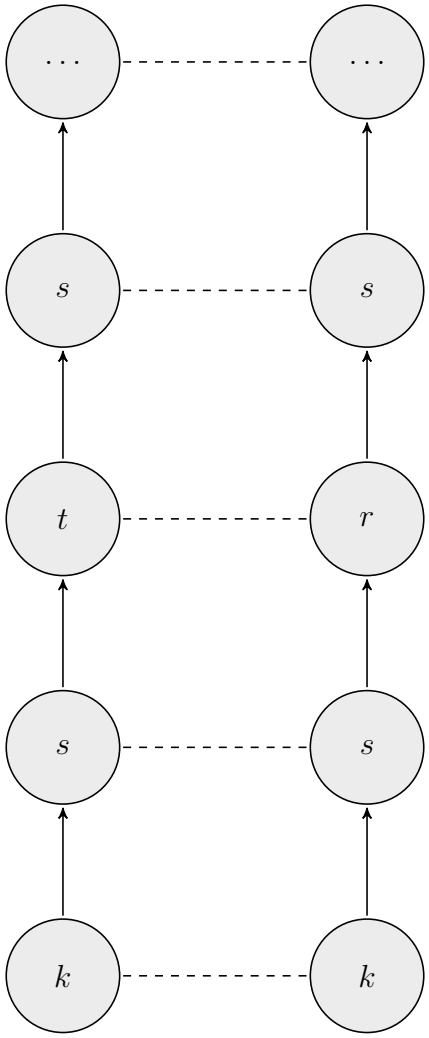

Figure 6.6: Fragment of the pre-model M for M^f in Fig. 6.5. Since the model is infinite, we portray its fragment, with two of the possible sequences $\sigma_k, \sigma'_k \in \Lambda_k$ for k.

there exists a sequence σ_z, such that $\sigma_z(n) = y$. Take any sequence $\sigma'_z \in \Lambda_z$, such that for all $i \leq n$, $\sigma_z(i) = \sigma'_z(i)$, and $\sigma'_z(n+1) = t$. The sequence σ'_z is such that $\sigma'_z(n+1) = t$, and since $rR^f_\Box t$, $(\sigma_s, n+1)R_\Box(\sigma'_z, n+1)$, which witnesses our claim, thereby completing the inductive proof for the positive direction.

Induction hypothesis in the negative direction: for $n \in \mathbb{Z} \setminus \mathbb{N} \cup \{0\}$, if $\sigma_s(n)R^f_\Box t$, then there exists a sequence σ_x, such that $(\sigma_s, n)R_\Box(\sigma_x, n)$ and $\sigma_x(n) = t$.

Induction step: Assume $\sigma_s(n-1)R^f_\Box t$. Let $\sigma_s(n-1) = r$ for some $r \in W^f$. Then, $rR^f_\Box t$. By definition of the sequence, there exists some $x = \sigma_s(n)$ and $rR^f_\prec x$. Therefore, $tR^f_\Box r, r(R^f_\prec)^{-1}x$. Since M^f satisfies NY, there exists $y \in W^f$, such that $tR^f_\Box y$ and $yR^f_\prec x$. Then, $\sigma_s(n)R^f_\Box y$. By (IH), there exists a sequence σ_z, such that $\sigma_z(n) = y$. Take any sequence $\sigma'_z \in \Lambda_z$, such that for all $i \geq n$, $\sigma_z(i) = \sigma'_z(i)$, and $\sigma'_z(n-1) = t$. Then, σ'_z is the sequence, such that $(\sigma'_z, n-1)R_\Box(\sigma_s, n-1)$ and $\sigma'_z(n-1) = t$, which witnesses our claim, thereby completing the inductive proof for the negative direction, and the whole proof in general. □

6.3.24. Lemma (IoA holds for M). *For any $R_\Box$-connected $(\sigma_{t_1}, j), \ldots, (\sigma_{t_n}, j) \in W$ and for every pairwise disjoint $I_1, \ldots, I_n \in Mod$ there exists $(\sigma_z, j) \in W$, such that $(\sigma_{t_1}, j)R_{I_1}(\sigma_z, i), \ldots, (\sigma_{t_n}, j)R_{I_n}(\sigma_z, j)$.*

Proof:
Take any $R_\Box$-connected $(\sigma_{t_1}, j), \ldots, (\sigma_{t_n,j}) \in W$. By definition, there exists $s_1 = \sigma_{t_1}(i), \ldots, s_n = \sigma_{t_n}(i)$, such that $s_1 R^f_\Box \ldots R^f_\Box s_n$. Then, since the filtrated model M^f satisfies independence of agents, there exists $x \in W^f$, such that $s_1 R^f_{I_1} x, \ldots, s_n R^f_{I_n} x$. Since $R^f_{I_1}, \ldots, R^f_{I_n} \subseteq R^f_\Box$, $xR^f_\Box s_1 R^f_\Box \ldots R^f_\Box s_n$. Since $s_1 = \sigma_{t_1}(i)$ and $s_1 R^f_\Box x$, by Lemma 6.3.23, there exists a sequence σ_z, such that $\sigma_z(j) = x$. Then, $(\sigma_z, j) \in W$, while $\sigma_z(j)R^f_{I_i}\sigma_{t_i}(j)$ for every $i \in \{1, \ldots, n\}$. Hence, $(\sigma_z, j) \in \bigcap_{1 \leq i \leq n} R_{I_i}((\sigma_{t_i}, j)) \neq \emptyset$. □

6.3.25. Lemma ((NCBUH) holds for M). *If $(\sigma_s, i)R_\prec(\sigma_s, i+1)$ and $(\sigma_s, i+1)R_\Box(\sigma_t, i+1)$, then for every $C \in \mathcal{G}$ there exists (σ_t, i), such that $(\sigma_s, i)R_C(\sigma_t, i)$ and $(\sigma_t, i)R_\prec(\sigma_t, i+1)$.*

Proof:
Assume that $(\sigma_s, i)R_\prec(\sigma_s, i+1)$ and $(\sigma_s, i+1)R_\Box(\sigma_t, i+1)$. Let $\sigma_s(i) = x$, $\sigma_s(i+1) = y$ and $\sigma_t(i+1) = z$, where $x, y, z \in W^f$. Then, by definition of $R_\prec$ and $R_\Box$, $xR^f_\prec y$ and $yR^f_\Box z$. Since M^f satisfies (NCBUH), there exists w, such that $xR^f_C w$ and $wR^f_\prec z$. From $xR^f_C w$ it follows that $xR^f_\Box w$. Since $x = \sigma_s(i)R^f_\Box w$, by Lemma 6.3.23, there exists some sequence σ'_r, such that $\sigma'_r(i) = w$. Given that $wR^f_\prec z$, we know that there exists a sequence σ_r, such that $\sigma_r(j) = \sigma'_r(j)$ for every $j \leq i$ and $\sigma_r(i+1) = z$. Then, $(\sigma_s, i)R_C(\sigma_r, i)$. If we prove that $(\sigma_r, i)R_C(\sigma_t, i)$, we will demonstrate that $(\sigma_s, i)R_C(\sigma_t, i)$ and $(\sigma_t, i)R_\prec(\sigma_t, i+1)$, thereby terminating the proof.

For contradiction, assume it is not the case that $(\sigma_r, i)R_C(\sigma_t, i)$. Let $\sigma_t(i) = u \in W^f$. Then, it should be the case that $(u, w) \notin R^f_C$. By def. of R^f_C then, either u and w disagree on some $[stit]_C$-formula, or on some $[stit]_i$-formula, where $i \in C$, or on some $\Box$-formula in Σ. We show that all three cases are impossible. First, assume there is a formula $\langle stit\rangle_C\psi$, such that $\langle stit\rangle_C\psi \in w$ and $\langle stit\rangle_C\psi \notin u$, or the other way around. W.l.o.g., assume $\langle stit\rangle_C\psi \in w$ and $\langle stit\rangle_C\psi \notin u$. As $wR^f_\prec z$, by def. of $R^f_\prec$ and (Σ10), $Y\langle stit\rangle_C\psi \in z$. As $\sigma_t(i) = uR^f_\prec z = \sigma_t(i+1)$, $Y\langle stit\rangle_C\psi \in z$ entails that $\langle stit\rangle_C\psi \in u$. Hence, $\langle stit\rangle_C\psi \in w$ entails $\langle stit\rangle_C\psi \in u$ for any $\langle stit\rangle_C\psi \in \Sigma$. The other direction is completely analogous. Hence, $\langle stit\rangle_C\psi \in u$ iff $\langle stit\rangle_C\psi \in w$ for any $\langle stit\rangle_C\psi \in \Sigma$. Then, w and u agree on all $[stit]_C$-formulae in Σ. For the sake of contradiction, assume that w and u disagree on some $\Box$- or $[stit]_i$-formula in Σ. Note that $\Box\psi \vdash [stit]_i\psi$ and $[stit]_i\psi \vdash [stit]_C\psi$ by axioms (Incl) and (i-C), respectively. Then, by (Σ11), there would have been a $[stit]_C$-formula in Σ, on which w and u would disagree, which we have shown to be impossible. So w and u agree on all $\Box$-, $[stit]_i$- (where $i \in C$) and $[stit]_C$-formulae. Hence, $uR^f_C w$, and by definition of R_C, $(\sigma_t, i)R_C(\sigma_r, i)$. By the transitivity of R_C then, $(\sigma_s, i)R_C(\sigma_t, i)$. Consequently, $(\sigma_s, i)R_C \circ R_\prec(\sigma_t, i+1)$. □

6.3.26. Lemma. *For every $\psi \in \Sigma$, $s \in W^f$, $\sigma_x \in \Lambda_s$ and any $i \in \mathbb{N}$, such that $\sigma_x(i) = s$, the following holds:*

$$M, (\sigma_x, i) \models \psi \Leftrightarrow M^f, s \models \psi.$$

Proof:
Cases of propositional variables and Boolean connectives are trivial.

($\Box\psi$). Left-to-right. Assume $\sigma_x(i) = s$ and $M, (\sigma_x, i) \models \Box\psi$ for some $\Box\psi \in \Sigma$. Take any $r \in W^f$, such that $sR^f_\Box r$. By Lemma 6.3.23, we know that there exists a sequence σ_y, such that $\sigma_y(i) = r$, hence, $(\sigma_x, i)R_\Box(\sigma_y, i)$. From $M, (\sigma_x, i) \models \Box\psi$ it follows that $M, (\sigma_y, i) \models \psi$. By (IH), $M^f, r \models \psi$. Since r was an arbitrary $R^f_\Box$-successor of r, all r-successors are ψ-worlds, i.e., $M^f, s \models \Box\psi$. Right-to-left. Assume $M^f, s \models \Box\psi$. Take any $(\sigma_x, i) \in W^f$, such that $\sigma_x(i) = s$. Take any (σ_y, i), such that $(\sigma_x, i)R_\Box(\sigma_y, i)$. From the definition of $R^f_\Box$, $\sigma_s(i) = sR^f_\Box\sigma_y(i)$. Then, $M^f, \sigma_y(i) \models \psi$. By (IH), $M, (\sigma_y, i) \models \psi$. Since (σ_y, i) was an arbitrary $R_\Box$-successor of (σ_x, i), $M, (\sigma_x, i) \models \Box\psi$.

($[stit]_i\psi, [stit]_C\psi$). Analogously to the previous case.

($X\psi$). Left-to-right. Assume $\sigma_x(i) = s$ and $M, (\sigma_x, i) \models X\psi$ for some $X\psi \in \Sigma$. Then, $M, (\sigma_x, i+1) \models \psi$. Let $\sigma_x(i+1) = r$ for some $r \in W^f$. Then, $sR^f_\prec r$. By (IH), $M^f, r \models \psi$. Since $X\psi \in \Sigma$, $\psi \in \Sigma$. Then, for any $\Delta \in [r]$, $\psi \in \Delta$. Since $sR^f_\prec r$, there should exist $\Gamma \in [s]$ and $\Delta \in [r]$, such that $\Gamma R^c_\prec \Delta$. Then, $\psi \in \Delta$ iff $X\psi \in \Gamma$. Since $X\psi \in \Sigma$, $X\psi \in \Gamma'$ for every $\Gamma' \in [s]$. So, by the

standard filtration lemma, $M^f, s \models X\psi$. Right-to-left. Assume $M^f, s \models X\psi$. Then, for any $r \in W^f$, such that $sR^f_{\prec}r$, $M^f, r \models \psi$. Take any (σ_x, i), such that $\sigma_x(i) = s$. Then, $(\sigma_x, i)R_{\prec}(\sigma_x, i{+}1)$ and $s = \sigma_x(i)R^f_{\prec}\sigma_x(i{+}1)$. Since $M^f, s \models X\psi$, $M^f, \sigma_x(i+1) \models \psi$. By (IH), $M, (\sigma_x, i+1) \models \psi$. Hence, $M, (\sigma_x, i) \models X\psi$.

$(Y\psi)$. Analogously to the previous case.

$(G\psi)$. Left-to-right. Assume $\sigma_x(i) = s$ and $M, (\sigma_x, i) \models G\psi$ for some $G\psi \in \Sigma$. Note that if $G\psi \in \Sigma$, then $XG\psi, \psi, \neg\psi, \neg G\psi, X\neg\psi \in \Sigma$. Let $m \in \mathbb{N}$ be the least non-negative number, such that $\sigma_x(m)$ occurs infinitely many times in the positive direction of σ_x. We know that such m exists by definition of σ_x. We fix $\sigma_x(m)$ as $t \in W^f$. Then, we claim that $M^f, t \models G\psi$. For the contradiction, assume that $M^f, t \not\models G\psi$. Then, $M^f, t \models \neg G\psi$. Since $\neg G\psi \in \Sigma$, $\neg G\psi \in \Gamma$ for any $\Gamma \in [t]$. Take any $\Gamma \in [t]$. Since $\neg G\psi \in \Gamma$, by the truth lemma, $M^c, \Gamma \models \neg G\psi$, so there exists $\Delta \in W^c$, such that $\Gamma R^c_{<}\Delta$ and $\neg\psi \in \Delta$. Let $r = \Delta \cap \Sigma$. Since $\neg\psi \in \Delta \cap \Sigma$, $\neg\psi \in \Delta'$ for every $\Delta' \in [r]$. Therefore, $M^f, r \models \neg\psi$. By definition of $R^f_{<}$, since $\Gamma \in [t]$, $\Delta \in [r]$ and $\Gamma R^c_{<}\Delta$, $tR^f_{<}r$. Since $R^f_{<} = (R^f_{\prec})^*$, $tR^f_{\prec}t_1R^f_{\prec}\dots R^f_{\prec}t_n = r$ for some $n \in \mathbb{N}$. By construction of σ_x, since $\sigma_x(m) = t$ occurs infinitely many times in the non-negative direction of σ_x, so does each t_i, including $t_n = r$. Hence, r appears infinitely many times in the positive direction in σ_x. Then, there exists a number $j > i$, such that $\sigma_x(j) = r$. Take the world (σ_x, j). As $M^f, r \models \neg\psi$, by (IH), $M, (\sigma_x, j) \models \neg\psi$, while $j > i$. So $M, (\sigma_x, i) \not\models G\psi$. Contradiction.

So, we have shown that $M^f, \sigma_x(m) \models G\psi$ for the least non-negative m, such that $\sigma_x(m)$ occurs infinitely many times in the positive direction of σ_x. Then, by transitivity of $R_{<}$, we get that $M^f, \sigma_x(k) \models G\psi$ for any $k \geq m$. If $i \geq m$, then $M^f, \sigma_x(i) \models G\psi$, and we are done. Otherwise, $k < m$. Let $h = m - k$. For the contradiction, assume $M^f, \sigma_x(i) \models \neg G\psi$. Then, there exists $\Gamma \in [\sigma_x(i)]$, such that $G\psi \notin \Gamma$. Take any such Γ. If $G\psi \notin \Gamma$, then, by maximal consistency of Γ, $\neg G\psi \in \Gamma$. Then, either $X\neg\psi \in \Gamma$ or $X\neg G\psi \in \Gamma$, as it follows from (FP_G) axiom. Note that $X\neg\psi, X\neg G\psi \in \Sigma$. It cannot be so that $X\neg\psi \in \Gamma$, since then $M^f, \sigma_x(i) \models X\neg\psi$, so by (IH), $M, (\sigma_x, i) \models X\neg\psi$, contradicting $M, (\sigma_x, i) \models G\psi$. Then, $M^f, \sigma_x(i) \models X\neg G\psi$, so by (IH), $M, (\sigma_x, i) \models X\neg G\psi$. From $M, (\sigma_x, i) \models X\neg G\psi$ it follows that $M, (\sigma_x, i+1) \models \neg G\psi$. Then we can repeat the previous argument h times, when we get to $M, (\sigma_x, m) \models \neg G\psi$, reaching the contradiction. So, we conclude that $M^f, \sigma_x(i) \models G\psi$.

Right-to-left. Assume $\sigma_x(i) = s$ and $M^f, s \models G\psi$. Take any (σ_x, j), such that $j > i$. By definition of the pre-model, $\sigma_x(i) = sR^f_{<}\sigma_x(j)$. Let $\sigma_x(j) = t \in W^f$. Then, as by assumption, $M^f, s \models G\psi$, for any $t \in W^f$, if $sR^f_{<}t$, then $M^f, t \models \psi$. Hence, $M^f, \sigma_x(j) \models \psi$. By (IH), $M, (\sigma_x, j) \models \psi$. As (σ_x, j) was an arbitrary $R_{<}$-successor of (σ_x, i), $M, (\sigma_x, i) \models G\psi$.

$(H\psi)$. If (σ_x, i) has infinitely many predecessors, than the proof is analogous to the previous case. We only need to substitute H for G, Y for X and take

the greatest non-positive number $m \in \mathbb{Z}$, such that $\sigma_x(m)$ occurs infinitely many times in the negative direction of σ_x, instead of the least non-negative one, which yields infinitely many occurrences in the positive direction. If (σ_x, i) has no predecessors, then $\sigma_x(i-1) = \epsilon$. By definition of sequences, it is only possible if $\sigma_x(i)$ has no predecessors. Hence, $M, (\sigma_x, i) \models H\psi$ and $M^f, \sigma_x(i) \models H\psi$ holds vacuously for any ψ. So the only case left for us to prove is when (σ_x, i) has finitely many predecessors.

Left-to-right. Assume (σ_x, i) has finitely many predecessors and $M, (\sigma_x, i) \models H\psi$. Take the least number $m < i$, such that $\sigma_x(m) \neq \epsilon$. As $\sigma_x(m-1) = \epsilon$, (σ_x, m) has no predecessors, so $M, (\sigma_x, m) \models H\psi$ holds vacuously. We fix $\sigma_x(m)$ as $t \in W^f$. By definition of sequences, as $\sigma_s(m-1) = \epsilon$, t has no predecessors, so $M^f, t \models H\psi$ vacuously. Let $h = |m| - |i|$. For the contradiction, assume that $M^f, \sigma_x(i) \not\models H\psi$. Then, there exists $\Gamma \in [\sigma_x(i)]$, such that $H\psi \notin \Gamma$, i.e., $\neg H\psi \in \Gamma$. Then, by (FP_H) axiom, either $Y\neg\psi \in \Gamma$ or $Y\neg H\psi \in \Gamma$. Note that $Y\neg\psi, Y\neg H\psi \in \Sigma$. In cannot be so that $Y\neg\psi \in \Gamma$, since then, $M^f, \sigma_x(i) \models Y\neg\psi$, and by (IH), $M, (\sigma_x, i) \models Y\neg\psi$, contradicting $M, (\sigma_x, i) \models H\psi$. Then, $M^f, \sigma_x(i) \models Y\neg H\psi$, i.e., $M^f, \sigma_x(i-1) \models \neg H\psi$. If we repeat this argument h times, we will get to $M, (\sigma_x, m) \models \neg H\psi$, reaching the contradiction. Hence, $M^f, \sigma_x(m) \models H\psi$.

Right-to-left. Assume $\sigma_x(i) = s$ and $M^f, s \models H\psi$. Take any world $(\sigma_x, j) \in W$, such that $j < i$ (if such world does not exist, then (σ_x, i) has no predecessors, so vacuously $M, (\sigma_x, i) \models H\psi$, and we are done). By construction of σ_x, $\sigma_x(j) R^f_< \sigma_x(i) = s$. As $M^f, s \models H\psi$, $M^f, \sigma_x(j) \models \psi$. By (IH), $M, (\sigma_x, j) \models \psi$. As (σ_x, j) was an arbitrary $R_<$-predecessor of (σ_x, i), we conclude that $M, (\sigma_x, i) \models H\psi$.

□

6.3.27. Theorem ($\mathbf{L}^{t\mathcal{G}}_{STIT}$ is weakly complete w.r.t. $\mathbf{F}_{pre}$). *For any $\varphi \in \mathcal{L}^{t\mathcal{G}}_{STIT}$, the following holds:*

$$\models_{\mathbf{F}_{pre}} \varphi \Leftrightarrow \vdash_{\mathbf{L}^{t\mathcal{G}}_{STIT}} \varphi.$$

Proof:
($\Leftarrow$) By the routine validity check.

($\Rightarrow$) By contraposition. Assume that $\nvdash_{\mathbf{L}^{t\mathcal{G}}_{STIT}} \varphi$. Then, $\neg\varphi$ is $\mathbf{L}^{t\mathcal{G}}_{STIT}$-consistent. Build a canonical model M^c as in Def. 6.3.17. There exists a mcs $\Gamma \in W^c$, such that $\neg\varphi \in \Gamma$. Compute the closure set Σ for $\neg\varphi$, build a filtration M^f of M^c through Σ. Let $s \in W^f$ be a world, such that $s = \Gamma \cap \Sigma$. Construct a pre-model M as in Def. 6.3.22. Take any sequence $\sigma_s \in \Lambda_s$. Finally, by Lemma 6.3.26, $M, (\sigma_s, 0) \models \neg\varphi$, witnessing that $\not\models_{\mathbf{F}_{pre}} \varphi$. □

There is the last step to prove that $\mathbf{L}^{t\mathcal{G}}_{STIT}$ is sound and weakly complete w.r.t. Kripke STIT frames: demonstrating that the validity on pre-frames and Kripke STIT frames are the same. Since every Kripke STIT frame by definition is a pre-frame, $\models_{\mathbf{F}_{pre}} \varphi$ entails $\models_{\mathbf{F}} \varphi$. For the converse, we establish that any $\mathbf{F}_{pre}$-frame is a bounded morphic image of some $\mathbf{F}$-frame:

6.3.28. LEMMA. *For every pre-frame $F_1 \in \mathbf{F}_{pre}$, there exists a Kripke STIT frame $F_2 \in \mathbf{F}$, such that there is a bounded morphism $f : F_2 \to F_1$.*

To prove this lemma, we will show how to transform every pre-frame F_1 into a larger F_2-frame of the same signature by taking infinitely many isomorphic copies of every possible world of F_1. This technique was pioneered for a similar proof for arrow logics in [Vak92]. For temporal STIT logic, it was implemented for a fragment with only the grand coalition Ags in [Lor13]. Our proof technique closely resembles the one of [Lor13, Lemma 9].

The construction is again quite tedious, so to make it clearer, we will work through a simple example, depicted on Figure 6.7. Consider a two-agent case, where $Ags = \{1, 2\}$, and $\mathcal{G} = \{Ags\}$. Then, let $F = (W, R_\prec, R_\Box, R_1, R_2, R_{Ags})$ be a pre-frame, such that:

1. $W = \{w_i, v_i \mid i \in \mathbb{N}\}$;
2. $R_\prec = \{(w_i, w_j) \mid j = i + 1\} \cup \{(v_i, v_j) \mid j = i + 1\}$;
3. $R_\Box = R_1 = R_2$ are equivalence relations, such that $[R_\Box] = [R_1] = [R_2] = \{\{w_0, v_0\}\} \cup \{\{w_i\}, \{v_i\} \mid i \geq 1\}$;
4. $R_{Ags} = id_W$, i.e., for every $w, v \in W$, $wR_{Ags}v$ iff $w = v$.

F is a pre-frame, but not a Kripke STIT frame, since $R_{Ags} \subseteq R_1 \cap R_2$, but $R_1 \cap R_2 \not\subseteq R_{Ags}$: $(w_0, v_0) \in R_1 \cap R_2$, but $(w_0, v_0) \notin R_{Ags}$.

Let $F_1 = (W, R_\prec, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}})$ be an arbitrary pre-frame. For every $i \in Ags$, let $\Delta_i = \{R_i(w) \mid w \in W\}$ be a collection of R_i-equivalence classes of F_1. Then, we define the set Δ that corresponds to the collection of $\bigcap_{i\in Ags} R_i$-equivalence classes as follows:

$$\Delta = \{\delta \in \prod_{i\in Ags} \Delta_i \mid \bigcap_{i\in Ags} \delta_i \neq \emptyset\}$$

where for every $\delta \in \prod_{i\in Ags} \Delta_i$, δ_i is the ith projection of δ. For brevity, let $w \in \delta$ mean that $w \in \bigcap_{i\in Ags} \delta_i$. There is a one-to-one correspondence between Δ and the set of $\bigcap_{i\in Ags} R_i$-equivalence classes. By convention, let δ^w denote a unique element $\delta^w \in \Delta$, such that $w \in \delta^w$.

For every $\delta \in \Delta$ and any $C \in \mathcal{G}$, let $\delta|_C$ be the Cth projection of δ. For example, if $\mathcal{G} = \{\{1, \ldots, k\}, \{k+1, \ldots, n\}\}$ and $\delta = (\delta_1, \ldots, \delta_n)$, then $\delta|_{\{1,\ldots,k\}} = (\delta_1, \ldots, \delta_k)$.

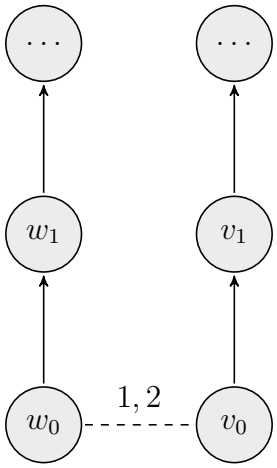

Figure 6.7: Example of a pre-frame that is not a Kripke STIT frame. As usual, bold arrows represent $R_{\prec}$, reflexive arrows for $R_{\Box}$, R_1 and R_2 are omitted.

Informally, every $\delta|_C$ corresponds to some $\bigcap_{i\in C} R_i$-equivalence class of F_1. For brevity, by $w \in \delta|_C$ we mean that $w \in \bigcap_{i\in C} \delta_i$.

For every $\delta \in \Delta$ and $C \in \mathcal{G}$:

$$\Gamma^C_\delta = \{R_C(w) \mid w \in \delta|_C\}$$

I.e., Γ^C_δ is a collection of R_C-equivalence classes that are included into the $\bigcap_{i\in C} R_i$-equivalence class that corresponds to $\delta|_C$. For instance, $\delta = R_1(w_0) \cap R_2(w_0) = \{w_0, v_0\}$ is an $\bigcap_{i\in\{1,2\}} R_i$-equivalence class of our toy example. It contains two $R_{\{1,2\}}$-equivalence classes: $\{w_0\}$ and $\{v_0\}$. Then, $\Gamma^{Ags}_\delta = \{\{w_0\}, \{v_0\}\}$.

W.l.o.g., we assume that $|\Gamma^C_\delta| \leq 2^\omega = |\mathbb{R}|$.[10] Then, for every $\delta, \gamma \in \Delta$ and any $C \in \mathcal{G}$, let k^C_δ be a total surjective function $k^C_\delta : \mathbb{R} \to \Gamma^C_\delta$, such that $\delta|_C = \gamma|_C$ entails $k^C_\delta = k^C_\gamma$.

In example in Figure 6.7, $\Delta_1 = \Delta_2 = \{\{w_0, v_0\}, \{w_1\}, \{v_1\}, \ldots\}$. Consequently, $\Delta = \{(X, Y) \in \Delta_1 \times \Delta_2 \mid X = Y\}$. Let $\delta = \delta^{w_0}$, corresponding to the $R_1 \cap R_2$-equivalence class $\{w_0, v_0\}$. δ contains two R_{Ags}-equivalence classes: $\{w_0\}$ and $\{v_0\}$. Therefore, $\Gamma^{Ags}_\delta = \{\{w_0\}, \{v_0\}\}$. So, we can define $k^{Ags}_\delta : \mathbb{R} \to \Gamma^{Ags}_\delta$, such that for every rational $i \in \mathbb{Q}$, $k^{Ags}_\delta(i) = \{w_0\}$, and for every irrational $j \in \mathbb{R} \setminus \mathbb{Q}$, $k^{Ags}_\delta(j) = \{v_0\}$. For any other $\gamma \in \Delta$ in the example, γ contains only one R_{Ags}-equivalence class, so for any $\gamma \neq \delta$, $k^{Ags}_\gamma(i) = k^{Ags}_\gamma(j)$ for any $i, j \in \mathbb{R}$.

[10] We are safe to make this assumption, since $\mathcal{L}^{tG}_{STIT}$ is a countable language, so there are at most 2^ω-many worlds in a canonical model M^c. As we have shown by the canonical model construction, its filtration and unraveling, every formula that has a pre-model can be satisfied on a pre-model with at most 2^ω-many worlds, hence at most 2^ω-many R_C-equivalence classes for any $C \in \mathcal{G}$.

For every $w \in W$, let

$$TC_w = \{\delta \in \Delta \mid \exists v \in \delta : wR_< v \text{ or } vR_< w \text{ or } w = v\}$$
$$PC_w = \{\delta \in \Delta \mid \exists v \in \delta : vR_< w\}$$
$$FC_w = \{\delta \in \Delta \mid \exists v \in \delta : wR_< v\}$$

Intuitively, TC_w are the elements of Δ that are temporarily connected with w; PC_w and FC_w are δ's that are connected to w via $R_<^{-1}$ and $R_<$, respectively. For instance, $TC_{w_0} = \{\delta^{w_i} \in \Delta \mid i \in \mathbb{N}\}$.

Note that:

$$\text{If } wR_< v \text{ or } vR_< w \text{ or } w = v, \text{ then } TC_w = TC_v \tag{6.1}$$
$$\text{If } wR_\Box v, \text{ then } PC_w = PC_v \tag{6.2}$$

(6.1) holds, since $R_<$ linearly orders H_w. (6.2) holds because of 3c of Def. 6.3.14.

For every world $w \in W$, let Λ_w be a set of all total functions $f : TC_w \to \mathbb{R}^n$ such that, if $f(\delta) = \vec{x}$, then there exists a world $v \in H_w$, such that:

$$v \in k_\delta^C\big(\sum_{i \in C} \vec{x}(i)\big) \text{ for any } C \in \mathcal{G}.$$

Intuitively, for every $\bigcap_{i \in Ags} R_i$-equivalence class $\delta \in TC_w$, f picks the equivalence classes

$$\{X_{C_1}, \ldots, X_{C_k} \mid X_{C_1} \in \Gamma_\delta^{C_1}, \ldots, X_{C_k} \in \Gamma_\delta^{C_k}\}$$

such that for some $v \in H_w$, $v \in X_{C_1} \cap \ldots \cap X_{C_k}$. f does that picking via the tuple of reals $\vec{x} = (r_1, \ldots, r_n)$, choosing a real r_i for every agent $i \in Ags$. The R_C-equivalence class f picks for δ is $k_\delta^C(\sum_{i \in C} \vec{x}(i))$.

For every function $f \in \Lambda_w$, every element $\delta \in TC_w$, every agent $i \in Ags$ and every vector $f(\delta) \in \mathbb{R}^n$, let $f_i(\delta)$ be the ith projection of the corresponding vector. That is, e.g., if $f(\delta) = (r_1, \ldots, r_n)$, then $f_i(\delta) = r_i$ for every $i \in \{1, \ldots, n\}$. Intuitively, $f_i(\delta)$ is the real f picks for i at δ.

For instance, let $f : TC_{w_0} \to \mathbb{R}^2$ be a function, such that $f(\delta) = (0, 0)$ for any $\delta \in TC_{w_0}$. As $0+0 = 0$ and 0 is rational, $w_0 \in \{w_0\} = k_{\delta^{w_0}}^{Ags}(0)$. And for any other $\delta^{w_i} \in TC_{w_0}$, $k_{\delta^{w_i}}^{Ags}(0) = \{w_i\}$, while $w_i \in H_{w_0}$. Therefore, $f \in \Lambda_{w_0}$. Analogously, a constant function $g : TC_{v_0} \to \{(0, \pi)\}$ is in Λ_{v_0}: $\delta^{w_0} = \delta^{v_0}$, so $k_{\delta^{w_0}}^{Ags} = k_{\delta^{v_0}}^{Ags}$. As $0 + \pi = \pi$ is irrational, $k_{\delta^{v_0}}^{Ags}(0 + \pi) = \{v_0\} \ni v_0$.

Next, we prove some useful propositions about such functions.

6.3.29. Proposition. *For any $r \in \mathbb{R}$, if $v \in H_w$ and $v \in k^C_{\delta_w}(r)$, then $v = w$.*

Proof:
If $v \in k^C_{\delta^w}(r)$, then there exists an R_C-equivalence class $X_C \subseteq \delta^w|_C$, such that $v \in X_C$. In other words, there is a world u, such that $(w,u) \in \bigcap_{i\in C} R_i$ and uR_Cv. As $R_C \subseteq \bigcap_{i\in C} R_i$ and $\bigcap_{i\in C} R_i$ is transitive, $(w,v) \in \bigcap_{i\in C} R_i$. Hence, by (2a) of Def.6.3.14, $wR_\Box v$. Then, by (3d), neither $wR_< v$ nor $vR_< v$. Since $v \in H_w$ and neither $wR_< v$ nor $vR_< w$, we obtain that $v = w$. □

6.3.30. Proposition. *For any $r \in \mathbb{R}$ and any $w, v \in W$, such that $wR_\Box v$: there exists $u_1 R_\prec w$, such that $u_1 \in k^C_{\delta^{u_1}}(r)$ iff there exists $u_2 R_\prec v$, such that $u_2 \in k^C_{\delta^{u_2}}(r)$ and $\delta^{u_1} = \delta^{u_2}$.*

Proof:
Assume $wR_\Box v$. Let $u_1 R_\prec w$. Then, $(u_1, v) \in R_\prec \circ R_\Box$. By 3c of Def. 6.3.14, $(u_1, v) \in R_C \circ R_\prec$ for every $C \in \mathcal{G}$. In other words, for every $C \in \mathcal{G}$, there exists u_C, such that $u_1 R_C u_C$ and $u_C R_\prec v$. Analogously to the proof of Proposition 6.3.15, we can show that for all $u_{C_i}, u_{C_j} \in \{u_C\}_{C\in\mathcal{G}}$: $u_{C_i} = u_{C_j} = u_2$ for some $u_2 \in W$. In other words, there exists some world $u_2 \in W$, such that $(u_1, u_2) \in \bigcap_{C\in\mathcal{G}} R_C$ and $u_2 R_\prec v$. Since $\mathcal{G}$ partitions Ags and $R_C \subseteq \bigcap_{i\in C} R_i$ for every $C \in \mathcal{G}$, $(u_1, u_2) \in \bigcap_{i\in Ags} R_i$. Then, by definition, $\delta^{u_1} = \delta^{u_2}$.

So $u_1, u_2 \in \delta^{u_1} = \delta^{u_2}$ and $u_1 R_C u_2$ for every $C \in \mathcal{G}$. Let $R_C(u_1) = k^C_{\delta^{u_1}}(r)$ for some $r \in \mathbb{R}$ (we know that such r exists, since k^C_δ is total and surjective). Then, since $u_2 R_C u_1$, $u_1 \in k^C_{\delta^{u_1}}(r)$ iff $u_2 \in k^C_{\delta^{u_2}}(r)$ (as $\delta^{u_1} = \delta^{u_2}$). □

6.3.31. Corollary. *For any $r \in \mathbb{R}$ and any $w, v \in W$, such that $wR_\Box v$: there exists $u_1 R_< w$, such that $u_1 \in k^C_{\delta^{u_1}}(r)$ iff there exists $u_2 R_< v$, such that $u_2 \in k^C_{\delta^{u_2}}(r)$ and $\delta^{u_1} = \delta^{u_2}$.*

6.3.32. Proposition. *If $f \in \Lambda_w$ and $vR_\Box w$, then there exists $f' \in \Lambda_v$, such that for all $\delta \in PC_w$, $f(\delta) = f'(\delta)$.*

Proof:
Let $f \in \Lambda_w$ and $wR_\Box v$. By (6.2), $PC_w = PC_v$. By Corollary 6.3.31, for every $u \in R^{-1}_<(w)$ there exists $u' \in R^{-1}_<(v)$, such that $\delta^u = \delta^{u'}$ and $uR_C u'$ for any $C \in \mathcal{G}$. Take any $g \in \Lambda_v$. Let $f' : TC_v \to \mathbb{R}^n$ be a function, such that:

$$f'(\delta) = \begin{cases} f(\delta) & \text{if } \delta \in PC_v = PC_w \\ g(\delta) & \text{otherwise} \end{cases}$$

Trivially, $f(\delta) = f'(\delta)$ for every $\delta \in PC_w$. Since $f \in \Lambda_w$, we know that for every $\delta \in PC_w = PC_v$ there exists a world u, such that $uR_< w$ and $u \in k^C_\delta(\sum_{i\in C} f_i(\delta))$ for every $C \in \mathcal{G}$. Then, by Corollary. 6.3.31, for every $u \in R^{-1}_<(w)$, there is

$u' \in R_{<}^{-1}(v)$, such that $uR_C u'$ for every $C \in \mathcal{G}$. Since $f'(\delta) = f(\delta)$ for every $\delta \in PC_w = PC_v$, $k_\delta^C(\sum_{i\in C} f_i(\delta)) = k_\delta^C(\sum_{i\in C} f_i'(\delta))$ for any $C \in \mathcal{G}$ for any $\delta \in PC_w = PC_v$. Then, for every $\delta \in PC_v$, there is a world $u' \in H_v$, such that $u' \in k_\delta^C(\sum_{i\in C} f_i'(\delta))$ for every $C \in \mathcal{G}$. If $\delta \in TC_v \setminus PC_v$, then, since $f(\delta) = g(\delta)$ and $g \in \Lambda_v$, we know that there is a world $u \in \{v\} \cup FC_v$, such that $u \in k_\delta^C(\sum_{i\in C} f_i'(\delta))$ for every $C \in \mathcal{G}$. Hence, $f' \in \Lambda_v$. □

6.3.33. Proposition. *If $v \in H_w$, then $\Lambda_w = \Lambda_v$.*

Proof:
($\subseteq$) Assume $v \in H_w$. That is, either $wR_{<}v$ or $vR_{<}w$ or $w = v$. Then, $H_w = H_v$ and $TC_w = TC_v$. Take any $f \in \Lambda_w$. We are to show that $f \in \Lambda_v$. That is, for every $\delta \in TC_v$, if $f(\delta) = \vec{x}$, then there exists a world $u \in H_v$, such that for any $C \in \mathcal{G}$:

$$u \in k_\delta^C(\sum_{i\in C} \vec{x}(i))$$

We know that $f \in \Lambda_w$. Take any $\delta \in TC_w$. Let $f(\delta) = \vec{x}$. Since $f \in \Lambda_w$, there exists a world $u \in H_w$, such that for for any $C \in \mathcal{G}$: $u \in k_\delta^C(\sum_{i\in C} \vec{x}(i))$. Since $H_w = H_v$, $u \in H_w$ iff $u \in H_v$. Consequently, $f \in \Lambda_w$ iff $f \in \Lambda_v$.

($\supseteq$) Analogously to the other direction. □

Finally, we are all set up to define the construction of our Kripke STIT frame.

6.3.34. Definition. Given any $\mathbf{F}_{pre}$-frame $F = (W, R_\prec, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}})$, we define a frame $\mathfrak{F}$ as follows:

$$\mathfrak{F} = (\mathfrak{W}, \mathfrak{R}_\prec, \mathfrak{R}_<, \mathfrak{R}_\Box, \{\mathfrak{R}_i\}_{i\in Ags}, \{\mathfrak{R}_C\}_{C\in\mathcal{G}})$$

- $\mathfrak{W} = \{(w, f) \mid w \in W, f \in \Lambda_w\}$. We often write simply w_f instead of (w, f), meaning the same thing;
- $w_f \mathfrak{R}_\Box v_g$ iff $wR_\Box v$ and $f(\delta) = g(\delta)$ for all $\delta \in PC_w$;
- for any $i \in Ags$, $w_f \mathfrak{R}_i v_g$ iff $wR_i v$ (or, equivalently, $\delta_i^w = \delta_i^v$), $f_i(\delta^w) = g_i(\delta^v)$ and $f(\delta) = g(\delta)$ for all $\delta \in PC_w$;
- for any $C \in \mathcal{G}$, $w_f \mathfrak{R}_C v_g$ iff $(w, v) \in \bigcap_{i\in C} R_i$ (or, equivalently, $\delta^w|_C = \delta^v|_C$), $f_C(\delta^w) = g_C(\delta^v)$ and $f(\delta) = g(\delta)$ for all $\delta \in PC_w$;
- $w_f \mathfrak{R}_\prec v_g$ iff $wR_\prec v$ and $f = g$;
- $w_f \mathfrak{R}_< v_g$ iff $wR_< v$ and $f = g$;

$\mathfrak{F}$ can be extended to a model $\mathfrak{M} = (\mathfrak{F}, \mathfrak{V})$ by a valuation function $\mathfrak{V} : Var \to 2^{\mathfrak{W}}$, such that for any $p \in Var$, $\mathfrak{V}(p) = \{w_f \in \mathfrak{W} \mid w \in V(p)\}$.

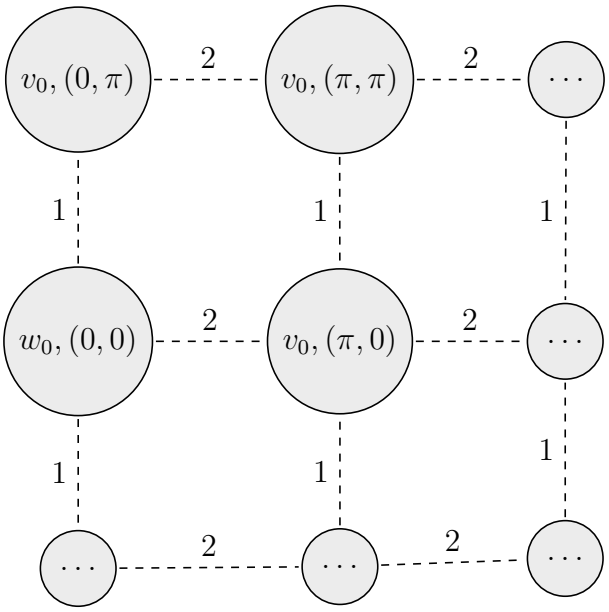

Figure 6.8: Constructed Kripke frame $\mathfrak{F}$ for the pre-frame F in Figure 6.7. As the frame is infinite, we show only the fragment of one $\mathfrak{R}_\Box$-equivalence class of (w_0, f). Let $\delta^{w_0} = \delta$, and $f \in \Lambda_{w_0}$, such that $f(\delta) = (0, 0)$. We portray each world $u_l \in \mathfrak{W}$ as a pair $\langle u, l(\delta)\rangle$. Note that $\mathfrak{R}_{Ags} = \mathfrak{R}_1 \cap \mathfrak{R}_2 = id_{\mathfrak{W}}$.

It is left for us to prove two claims: 1) $\mathfrak{F}$ is a Kripke STIT frame; 2) there exists a bounded morphism $\mathfrak{f} : \mathfrak{F} \to F$.

6.3.35. Lemma. *For every $\mathbf{F}_{pre}$-frame F, the corresponding constructed frame $\mathfrak{F}$ is a Kripke STIT frame in accordance with Definition 3.2.8.*

Proof:
We are going to follow Def. 3.2.8 step-wise, showing that every property holds.

1. $\mathfrak{W} \neq \emptyset$. Note that $W \neq \emptyset$. The functions $\{k^C_\delta\}^{C\in\mathcal{G}}_{\delta\in\Delta}$ are total and surjective. Therefore, $\Lambda_w \neq \emptyset$ for every $w \in W$. Hence, $\mathfrak{W} = \{w_f \mid f \in \Lambda_w\} \neq \emptyset$.
2. $\mathfrak{R}_\Box, \{\mathfrak{R}_i\}_{i\in Ags}$, $\{\mathfrak{R}_C\}_{C\in\mathcal{G}}$ are equivalence relations. It follows directly from their definitions and the fact that $R_\Box, \{R_i\}_{i\in Ags}$, $\{R_C\}_{C\in\mathcal{G}}$, $=$ are equivalence relations.
   (a) For any $i \in Ags$, $\mathfrak{R}_i \subseteq \mathfrak{R}_\Box$. If $w_f \mathfrak{R}_i v_g$, then $wR_i v$, $f_i(\delta^w) = g_i(\delta^v)$ and $f(\delta) = g(\delta)$ for all $\delta \in PC_w$. From $wR_i v$ it follows that $wR_\Box v$ by (2a). Then, $wR_\Box v$ and $f(\delta) = g(\delta)$ for all $\delta \in PC_w$. Hence, by definition, $w_f \mathfrak{R}_\Box v_g$.
   (b) *Group actions:* for any $C \in \mathcal{G}$, $\mathfrak{R}_C = \bigcap_{i\in C} \mathfrak{R}_i$. Follows directly from definitions of $\mathfrak{R}_C$ and $\mathfrak{R}_i$.
   (c) Independence of agents holds. See Lemma 6.3.36.

3. $\mathfrak{R}_\prec$ is serial. Take any $w_f \in \mathfrak{W}$. Since $R_\prec$ is serial, there exists $v \in W$, such that $wR_\prec v$. Then, $v \in H_w$. By Lemma 6.3.33, $\Lambda_w = \Lambda_v$. Hence, $f \in \Lambda_v$, and $v_f \in \mathfrak{W}$. Then, $w_f \mathfrak{R}_\prec v_f$. Since w_f was chosen arbitrarily, every world $w_f \in \mathfrak{W}$ has a $\mathfrak{R}_\prec$-successor, i.e., $\mathfrak{R}_\prec$ is serial.

4. $\mathfrak{R}_\prec$ does not branch backwards. Assume $v_f \mathfrak{R}_\prec w_f$ and $u_f \mathfrak{R}_\prec w_f$. Then, $vR_\prec w$ and $uR_\prec w$. By (3a), $u = v$. Then, $u_f = v_f$. Hence, $\mathfrak{R}_\prec$ does not branch backwards.

5. $\mathfrak{R}_\prec$ does not branch forwards. Analogously to the previous condition.

6. No choice between undivided histories. See Lemma 6.3.37;

7. If $w_f \mathfrak{R}_\Box v_g$, then neither $w_f \mathfrak{R}_< v_g$ nor $v_g \mathfrak{R}_< w_f$. If $w_f \mathfrak{R}_\Box v_g$, then $wR_\Box v$. By (3d) of Def. 6.3.14 (which by definition holds for every pre-frame), neither $wR_< v$ nor $vR_< w$. In other words, for no positive natural $n \in \mathbb{N}$, $wR^n_\prec v$ or $vR^n_\prec w$. Hence, neither $w_f \mathfrak{R}^n_\prec v_g$ nor $v_g \mathfrak{R}^n_\prec w_f$.

□

6.3.36. Lemma ((IoA) for $\mathfrak{F}$). *For every* $(w_1, f_1), \ldots, (w_n, f_n) \in \mathfrak{W}$, *if* $(w_1, f_1)\mathfrak{R}_\Box \ldots \mathfrak{R}_\Box (w_n, f_n)$, *then there exists a world* $(v, g) \in \mathfrak{W}$, *such that* $(w_1, f_1)\mathfrak{R}_1(v, g)$, $\ldots$, $(w_n, f_n)\mathfrak{R}_n(v, g)$.

Proof:
Let $(w_1, f_1), \ldots, (w_n, f_n) \in \mathfrak{W}$, such that $(w_1, f_1)\mathfrak{R}_\Box \ldots \mathfrak{R}_\Box (w_n, f_n)$. By definition of $\mathfrak{R}_\Box$, it follows that $w_1 R_\Box \ldots R_\Box w_n$ and $f_1(\delta) = \ldots = f_n(\delta)$ for every $\delta \in PC_{w_1}$. By (6.2), $PC_{w_1} = \ldots = PC_{w_n}$.

Since $w_1 R_\Box \ldots R_\Box w_n$, by (2c), there exists a world u, such that $w_1 R_1 u$, $\ldots$, $w_n R_n u$. Or, equivalently, $\delta^u_i = \delta^{w_i}_i$ for every $i \in \{1, \ldots, n\}$. Now, let $\vec{x} \in \mathbb{R}^n$ be a tuple of reals, such that $\vec{x}$ agrees on ith coordinate with $f_i(\delta^{w_i})$ for any $i \in \{1, \ldots, n\}$. Then, for every $C \in \mathcal{G}$, since $k^C_{\delta^u}$ is surjective, there exists an R_C-equivalence class $X_C = k^C_{\delta^u}(\sum_{i \in C} \vec{x}(i))$. By (2c), $\bigcap_{C \in \mathcal{G}} X_C \neq \emptyset$. So there exists some world $v \in \bigcap_{C \in \mathcal{G}} X_C$. Fix such a world as v.

Since $v \in k^C_{\delta^u}(\sum_{i \in C} \vec{x}(i))$ for every $C \in \mathcal{G}$, there exists some $w_C \in \delta^u|_C$, such that $w_C R_C v$ for every $C \in \mathcal{G}$. In other words, $(u, w_C) \in \bigcap_{i \in C} R_i$ and $w_C R_C v$. Since $R_C \subseteq \bigcap_{i \in C} R_i$, $(w_C, v) \in \bigcap_{i \in C} R_i$. Since $\bigcap_{i \in C} R_i$ is an equivalence relation (hence, transitive), $(u, v) \in \bigcap_{i \in C} R_i$. Since $\mathcal{G}$ is a partition of Ags, $(u, v) \in \bigcap_{i \in Ags} R_i$. In other words, $\delta^v = \delta^u$. It is also straightforward that $vR_\Box w_i$ for any $i \in Ags$ (since $vR_C u$, $uR_i w_i$ for every $i \in \{1, \ldots, n\}$ and $R_C, R_1, \ldots, R_n \subseteq R_\Box$).

Now, take any function $g \in \Lambda_v$, such that: $g(\delta) = f_1(\delta) = \ldots = f_n(\delta)$ for every $\delta \in PC_{w_1}$; $g(\delta^v) = \vec{x}$. We know that such g exists, since $vR_\Box w_i$, so $PC_v = PC_{w_i}$ for every $i \in \{1, \ldots, n\}$; and $v \in k^C_{\delta^v}(\sum_{i \in C} \vec{x}(i))$ for every $C \in \mathcal{G}$. Finally, it is left for us to show that $(w_i, f_i)\mathfrak{R}_i(v, g)$ for every $i \in \{1, \ldots, n\}$. Since

$\delta^v = \delta^u$, $\delta^v_i = \delta^{w_i}_i$ for every $i \in \{1, \ldots, n\}$. By construction of g, $g_i(\delta^v)$ agrees with $f_i(\delta^{w_i})$ on ith coordinate for every $i \in \{1, \ldots, n\}$. Then, $(w_1, f_1)\mathfrak{R}_1(v, g)$, ..., $(w_n, f_n)\mathfrak{R}_n(v, g)$. Since $(w_1, f_1), \ldots, (w_n, f_n)$ were chosen arbitrarily, we can conclude that independence of agents holds in the whole frame $\mathfrak{F}$. □

6.3.37. Lemma ((NCBUH) for $\mathfrak{F}$). *If $w_f\mathfrak{R}_{\prec}v_f$ and $v_f\mathfrak{R}_{\Box}u_g$, then, for every $C \in \mathcal{G}$, there exists $x_g \in \mathfrak{W}$, such that $w_f\mathfrak{R}_C x_g$ and $x_g\mathfrak{R}_{\prec}u_g$.*

Proof:
Assume that $w_f\mathfrak{R}_{<}v_f$ and $v_f\mathfrak{R}_{\Box}u_g$. Note that, by def. of $\mathfrak{R}_{\Box}$, $f(\delta) = g(\delta)$ for all $\delta \in PC_v = PC_u$, and $\delta^w \in PC_v$.

From $w_f\mathfrak{R}_{<}v_f$ and $v_f\mathfrak{R}_{\Box}u_g$ it follows that $wR_{\prec}v$ and $vR_{\Box}u$. By (3c), for every $C \in \mathcal{G}$ there exists $x_C \in W$, such that $wR_C x_C$ and $x_C R_{<} u$. As we have shown in Prop 6.3.15, $x_{C_1} = x_{C_2}$ for every $C_1, C_2 \in \mathcal{G}$. Then, there exists a single world $x \in W$, such that $\forall C \in \mathcal{G}$: $wR_C x$ and $xR_{\prec}u$. Note that, since $xR_{\prec}u$, $x \in H_u$. And since $u_g \in \mathfrak{W}$ (i.e., $g \in \Lambda_u$), by Prop. 6.3.33, $g \in \Lambda_x$. Then, $x_g \in \mathfrak{W}$. Also note that, since $wR_C x$ for any $C \in \mathcal{G}$, while $R_C \subseteq \bigcap_{i \in C} R_i$ and $\mathcal{G}$ partitions Ags, $(w, x) \in \bigcap_{i \in Ags} R_i$, i.e., $\delta^w = \delta^x$.

It is left for us to show that, for every $C \in \mathcal{G}$, $w_f\mathfrak{R}_C x_g$ and $x_g\mathfrak{R}_{\prec}u_g$. First, note that, as we have shown, $wR_C x$ for every $C \in \mathcal{G}$. In other words, for any $i \in \mathbb{R}$, $w \in k^C_{\delta^w}(i)$ iff $x \in k^C_{\delta^w}(i)$, or, given that $\delta^w = \delta^x$, $w \in k^C_{\delta^w}(i)$ iff $x \in k^C_{\delta^x}(i)$. Since $f(\delta) = g(\delta)$ for every $\delta \in PC_v = PC_u$ and $\delta^w = \delta^x \in PC_v$, $f_C(\delta^w) = g_C(\delta^x)$ for every $C \in \mathcal{G}$. Since $\delta^w = \delta^x$, $\delta^w|_C = \delta^x|_C$ for every $C \in \mathcal{G}$. Since $v_f\mathfrak{R}_{\Box}u_g$, $f(\delta) = g(\delta)$ for every $\delta \in PC_v = PC_u$, while $PC_w \subseteq PC_u$. Hence, $f(\delta) = g(\delta)$ for every $\delta \in PC_w$. To conclude, for every $C \in \mathcal{G}$, $w_f\mathfrak{R}_C x_g$, and, since $xR_{\prec}u$, $x_g\mathfrak{R}_{\prec}u_g$. □

It is left for us to prove that there is a bounded morphism from $\mathfrak{F}$ onto F_1. Let $\mathfrak{f} : \mathfrak{W} \to W$ be a function, such that $\mathfrak{f}(w_f) = w$ for every $w \in W$. Obviously, $\mathfrak{f}$ is surjective. Next, we show that it is a bounded morphism. We do it only for the case of $\mathfrak{R}_C/R_C$. Proofs for other cases are straightforward and identical to [Lor13, p.384].

(Zig) Assume $w_f\mathfrak{R}_C v_g$. By definition, $f_C(\delta^w) = g_C(\delta^v)$ and $\delta^w|_C = \delta^v|_C$. By definition of $k^C_{\delta^w}$ and $k^C_{\delta^v}$, $k^C_{\delta^w} = \delta^C_{\delta^v}$. Then, $w \in k^C_{\delta^w}(\sum_{i \in C} f_i(\delta^w)) = k^C_{\delta^v}(\sum_{i \in C} g_i(\delta^v))$ $\ni v$. Hence, w and v are in the same R_C-equivalence class, i.e., $wR_C v$.

(Zag) Assume $\mathfrak{f}(w_f)R_C v$. Then, $wR_C v$. Since $wR_C v$, $wR_{\Box}v$, hence, $PC_w = PC_v$. Also, since $R_C \subseteq \bigcap_{i \in C} R_i$, from $wR_C v$ it follows that $\delta^w|_C = \delta^v|_C$. Let $g \in \Lambda_v$ be any function, such that $g(\delta) = f(\delta)$ if $\delta \in PC_w$ and $g_C(\delta^w) = f_C(\delta^v)$. We know that such $g \in \Lambda_v$, exists, because of Proposition 6.3.32 and the fact that $v \in k^C_{\delta^w}(\sum_{i \in C} f_i(\delta^w))$. Then, it is straightforward that $w_f\mathfrak{R}_C v_g$ and $\mathfrak{f}(v_g) = v$.

This step completes our proof of Lemma 6.3.28. Now we can establish that logics of pre-frames and Kripke STIT frames coincide.

6.3.38. Lemma. *For any* $\varphi \in \mathcal{L}^{t\mathcal{G}}_{STIT}$, $\models_{\mathbf{F}} \varphi$ *iff* $\models_{\mathbf{F}_{pre}} \varphi$.

Proof:
Left-to-right. By Def. 6.3.14 and 6.3.16, $\mathbf{F} \subseteq \mathbf{F}_{pre}$, i.e., every Kripke STIT frame is a pre-frame. Therefore, $\models_{\mathbf{F}_{pre}} \varphi$ entails $\models_{\mathbf{F}} \varphi$. Right-to-left. As Lemma 6.3.28 shows, $\mathbf{F}_{pre}$ is a bounded morphic image of $\mathbf{F}$. Then, by the standard preservation results [BRV01], $\models_{\mathbf{F}} \varphi$ entails $\models_{\mathbf{F}_{pre}} \varphi$. □

6.3.39. Theorem ($\mathbf{L}^{t\mathcal{G}}_{STIT}$ is sound and complete w.r.t. Kripke STIT frames). *For any* $\varphi \in \mathcal{L}^{t\mathcal{G}}_{STIT}$, *the following holds:*

$$\vdash_{\mathbf{L}^{t\mathcal{G}}_{STIT}} \varphi \Leftrightarrow \models_{\mathbf{F}} \varphi.$$

Proof:
Follows directly from Theorem 6.3.27 and Lemma 6.3.38. □

Interestingly, $\mathbf{L}^{t\mathcal{G}}_{STIT}$, being sound and complete, lacks the non-standard irreflexivity rule that is used for axiomatisation in, e.g., [Lor13]:

$$\text{If } \vdash (\Box\neg p \wedge \Box(Gp \wedge Hp)) \rightarrow \varphi, \text{ then } \vdash \varphi, \text{ given that } p \notin sub(\varphi)$$

This positive trait of our axiom system gives us hope that $\mathbf{L}^{t\mathcal{G}}_{STIT}$ might be decidable. Unfortunately, we cannot establish this via the standard fmp-via-filtration argument, since canonical pre-models (Def.6.3.22) we construct after the filtration are tree-like, hence infinite. Moreover, the sequences we use in the construction of the canonical pre-models are not necessarily periodic, which makes things more complicated. Yet we might be able to reduce the satisfiability of $\mathcal{L}^{t\mathcal{G}}_{STIT}$-formulae to the non-emptiness problem for finite state automata on infinite trees, as in [EJ99]. We leave this investigation for future work.

In the next subsection, we turn to the atemporal fragments of group STIT logic with disjoint groups. As atemporal STIT models are not tree-like, the completeness and decidability are easier to establish.

6.3.3 Atemporal STIT with disjoint groups

The results from the previous subsection are easily translatable to a smaller fragment, atemporal STIT with disjoint groups, whose formal language $\mathcal{L}^{\mathcal{G}}_{STIT}$ is recursively defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid [stit]_i\varphi \mid [stit]_C\varphi$$

For $\mathcal{L}^{\mathcal{G}}_{STIT}$, not only can we finitely axiomatise it, but also prove that the corresponding logic is decidable. First, we prove that the logic $\mathbf{L}^{\mathcal{G}}_{STIT}$ (see Table 6.3) is sound and strongly complete w.r.t. atemporal STIT frames as in Def. 3.2.10. The proof is just a simplified version of Theorem 6.3.39, so we skip most of the details. Then, we show that the satisfiability problem for $\mathbf{L}^{\mathcal{G}}_{STIT}$ is decidable.

(CPL)	Tautologies of classical propositional logic
(S5)	S5 axioms and rules of inference for $\Box$, $[stit]_i$, $[stit]_C$
(Nec)	$\Box\varphi \to [stit]_i\varphi$ for every $i \in Ags$
$(i-C)$	$[stit]_i\varphi \to [stit]_C\varphi$ for every $i \in C \in \mathcal{G}$
(Ind)	$(\Diamond[stit]_{I_1}\varphi_1 \wedge \ldots \wedge \Diamond[stit]_{I_n}\varphi_n) \to \Diamond([stit]_{I_1}\varphi_1 \wedge \ldots \wedge [stit]_{I_n}\varphi_n)$ for any disjoint $I_1, \ldots, I_n \in Mod$
(MP)	If $\vdash \varphi$ and $\vdash \varphi \to \psi$, then $\vdash \psi$

Table 6.3: Logic $\mathbf{L}^{\mathcal{G}}_{STIT}$

As for the completeness proof, we repeat our steps from the previous section:

1. Define a class of atemporal pre-frames $\mathbf{F}^a_{pre}$;
2. Prove that $\mathbf{L}^{\mathcal{G}}_{STIT}$ is sound and strongly complete w.r.t. $\mathbf{F}^a_{pre}$;
3. Prove that $\mathcal{L}^{\mathcal{G}}_{STIT}$ does not distinguish between atemporal pre-frames and atemporal Kripke STIT frames.

Once again, we will first reformulate the definition of atemporal Kripke STIT frames in an equivalent way (this time, only rewriting (IoA)), to simplify our work.

6.3.40. Definition (Atemporal Kripke STIT frames $\mathbf{F}^a$). A tuple:

$$F = (W, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}}) \in \mathbf{F}^a$$

is an atemporal Kripke STIT frame, where:

1. $W \neq \emptyset$;
2. $R_\Box, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}}$ are equivalence relations on W, such that:
   (a) For any $i \in Ags$: $R_i \subseteq R_\Box$;
   (b) Group actions: for any $C \in \mathcal{G}$, $R_C = \bigcap_{i\in C} R_i$;
   (c) Independence of agents: for any disjoint $I_1, \ldots, I_n \in Mod$, if $w_1 R_\Box \ldots R_\Box w_n$, then $R_{I_1}(w_1) \cap \ldots \cap R_{I_n}(w_n)$;

6.3.41. Definition (Atemporal pre-frames $\mathbf{F}^a_{pre}$). A tuple

$$F = (W, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}})$$

is an atemporal pre-frame iff F satisfies all the properties of the atemporal Kripke STIT frame from Def.6.3.40, except of the condition (2b), which is weakened: for every $C \in \mathcal{G}$, $R_C \subseteq \bigcap R_i$.

6.3.42. THEOREM ($\mathbf{L}^{\mathcal{G}}_{STIT}$ is sound and complete w.r.t. $\mathbf{F}^{a}_{pre}$). *For any set of formulae $\Gamma \subseteq \mathcal{L}^{\mathcal{G}}_{STIT}$ and a formula $\varphi \in \mathcal{L}^{\mathcal{G}}_{STIT}$, the following holds:*

$$\Gamma \vdash_{\mathbf{L}^{\mathcal{G}}_{STIT}} \varphi \Leftrightarrow \Gamma \models_{\mathbf{F}^{a}_{pre}} \varphi.$$

Proof:
Each condition in Def. 6.3.41 is the first-order correspondent of the axiom of $\mathbf{L}^{\mathcal{G}}_{STIT}$, which is Sahlqvist formula (see the correspondence in the proof of Theorem 6.3.27). Therefore, soundness and strong completeness automatically follows by Sahlqvist correspondence theorems [Sah75]. □

6.3.43. LEMMA. *For every atemporal pre-frame $F = (W, R_\Box, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in\mathcal{G}})$ there exists an atemporal Kripke STIT frame $\mathfrak{F} = (\mathfrak{W}, \mathfrak{R}_\Box, \{\mathfrak{R}_i\}_{i\in Ags}, \{\mathfrak{R}_C\}_{C\in\mathcal{G}})$, such that F is a bounded morphic image of $\mathfrak{F}$.*

Proof:
Analogously to Lemma 6.3.28. We just simplify the construction of the Kripke STIT frame by postulating Λ_w as the set of functions $f : \{\delta^w\} \to \mathbb{R}^n$, such that if $f(\delta^w) = \vec{x}$, then for any $C \in \mathcal{G}$:

$$w \in k^C_{\delta^w}(\sum_{i\in C} \vec{x}(i))$$

All other steps are just as in Lemma 6.3.28. □

6.3.44. THEOREM ($\mathbf{L}^{\mathcal{G}}_{STIT}$ is sound and complete). *For any set of formulae $\Gamma \subseteq \mathcal{L}^{\mathcal{G}}_{STIT}$ and a formula $\varphi \in \mathcal{L}^{\mathcal{G}}_{STIT}$, the following holdS:*

$$\Gamma \vdash_{\mathbf{L}^{\mathcal{G}}_{STIT}} \varphi \Leftrightarrow \Gamma \models_{\mathbf{F}^{a}} \varphi.$$

Proof:
($\Rightarrow$) By a straightforward validity check.

($\Leftarrow$) By contraposition. Assume $\Gamma \nvdash_{\mathbf{L}^{\mathcal{G}}_{STIT}} \varphi$. Then, $\Gamma \cup \{\neg\varphi\}$ is $\mathbf{L}^{\mathcal{G}}_{STIT}$-consistent. By Theorem]6.3.42, there exists an atemporal pointed pre-model (M, w), such that $M, w \models \Gamma \cup \{\neg\varphi\}$. By Lemma 6.3.43, there exists a pointed atemporal Kripke STIT model $(\mathfrak{M}, w_f)$, such that $\mathfrak{M}, w_f \models \Gamma \cup \{\neg\varphi\}$. Hence, $\Gamma \not\models_{\mathbf{F}^{a}} \varphi$. □

As for satisfiability, having Theorem 6.3.39 at hand, we know that a formula is satisfiable in a class of atemporal Kripke STIT frames iff it is $\mathbf{L}^{\mathcal{G}}_{STIT}$-consistent iff it is satisfiable in $\mathbf{F}^{a}_{pre}$. Next, we show that every formula $\varphi \in \mathcal{L}^{\mathcal{G}}_{STIT}$ is satisfiable in $\mathbf{F}^{a}_{pre}$ iff it is satisfiable on some frame with at most $2^{|sub(\varphi)|}$ worlds.

6.3.45. THEOREM ($\mathcal{L}^{\mathcal{G}}_{STIT}$ has a strong finite model property). *Any $\varphi \in \mathcal{L}^{\mathcal{G}}_{STIT}$ is satisfiable in $\mathbf{F}^{a}_{pre}$ iff it is satisfiable on some pre-frame with at most $2^{|sub(\varphi)|}$ worlds.*

Proof:
Let $M = (W, R_{\Box}, \{R_i\}_{i\in Ags}, \{R_C\}_{C\in Ags}, V)$ be a pre-model, such that, for some $w \in W$, $M, w \models \varphi$. Let

$$M^w = (W^w, R^w_{\Box}, \{R^w_i\}_{i\in Ags}, \{R^w_C\}_{C\in\mathcal{G}}, V^w)$$

be a point-generated submodel of M: $W^w = R_{\Box}(w)$, and for every relation R, $R^w = R \cap (W^w \times W^w)$. By a standard preservation results [BRV01], $M^w, w \models \varphi$.

Let $sub(\varphi)$ be a set of subformulae of φ. Let $\sim$ be an equivalence relation on W^w, such that $w_1 \sim w_2$ iff for any $\psi \in sub(\varphi)$, $M^w, w_1 \models \psi$ iff $M^w, w_2 \models \psi$. For any $u \in W^w$, let $[u] = \{v \in W^w \mid u \sim v\}$. Then, we define a filtration of M^w through $sub(\varphi)$ as follows:

$$M' = (W', R'_{\Box}, \{R'_i\}_{i\in Ags}, \{R'_C\}_{C\in\mathcal{G}}, V')$$

1. $W' = \{[v] \mid v \in W^w\}$. Note that W' is finite, and its cardinality is at most $2^{|sub(\varphi|)}$;
2. $[u]R'_{\Box}[v]$ iff for any $\Box\psi \in sub(\varphi)$, $M^w, u \models \Box\psi$ iff $M^w, v \models \Box\psi$;
3. For any $i \in Ags$, $[u]R'_i[v]$ iff $[u]R'_{\Box}[v]$ and for any $[stit]_i\psi \in sub(\varphi)$, $M^w, u \models [stit]_i\psi$ iff $M^w, v \models [stit]_i\psi$;
4. For any $C \in \mathcal{G}$, $[u]R'_C[v]$ iff $[u]R'_i[v]$ for any $i \in C$ and for any $[stit]_C\psi \in sub(\varphi)$, $M^w, u \models [stit]_C\psi$ iff $M^w, v \models [stit]_C\psi$;
5. For any $p \in Var$, $V'(p) = \{[v] \in W' \mid v \in V^w(p)\}$.

First, we prove that M' is a pre-model. Directly from the definition of M', it follows that $R'_{\Box}, \{R'_i\}_{i\in Ags}, \{R'_C\}_{C\in\mathcal{G}}$ are equivalence relations, such that $\forall i \in Ags : R'_i \subseteq R'_{\Box}$, $\forall C \in \mathcal{G} : R'_C \subseteq \bigcap_{i\in C} R_i$. It is left for us to prove that M' satisfies the independence of agents constraint.

Let $[u_1], \ldots, [u_n] \in W'$, such that $[u_1]R'_{\Box} \ldots R'_{\Box}[u_n]$. Since $W^w = R_{\Box}(w)$, $u_1 R_{\Box} \ldots R_{\Box} u_n$. Since M^w is the pre-model, it satisfies independence of agents. Hence, there exists $v \in W^w$, such that $u_1 R^w_1 v, \ldots, u_n R^w_n v$. Since $R^w_i \subseteq R^w_{\Box}$ for any $i \in Ags$, $v R^w_{\Box} u_1 R^w_{\Box} \ldots R^w_{\Box} u_n$. So, by semantics of $\Box$, $M^w, v \models \Box\psi$ iff $M, u_i \models \Box\psi$

for any $i \in Ags$ and for any formula ψ. Therefore, $[u_1]R'_{\Box}[v], \ldots, [u_n]R'_{\Box}[v]$. Since $u_1R_1^w v, \ldots, u_nR_n^w v$, by semantics of $[stit]_i$, for any $i \in \{1, \ldots, n\}$, $M^w, u_i \models [stit]_i\psi$ iff $M^w, v \models [stit]_i\psi$ for any $[stit]_i\psi \in sub(\varphi)$. Then, $[u_1]R'_1[v]$,...,$[u_n]R'_n[v]$. Hence, $R'_1([u_1]) \cap \ldots \cap R'_n([u_n]) \neq \emptyset$.

It is left for us to show that for any $\psi \in sub(\varphi)$, $M^w, w \models \psi$ iff $M', [w] \models \psi$. We do it by induction on ψ. Cases of atomic propositions and Boolean connectives are trivial. For each of the remaining cases, the proofs are analogous, so we explicitly show it only for $[stit]_i\psi$.

($\Rightarrow$) Assume $M^w, w \models [stit]_i\psi$ for some $[stit]_i\psi \in sub(\varphi)$. Then, by def. of R'_i, $[w]R'_i[v]$ iff $M^w, w \models [stit]_i\gamma \Leftrightarrow M^w, v \models [stit]_i\gamma$ for any $[stit]_i\gamma \in sub(\varphi)$. Since $[stit]_i\psi \in sub(\varphi)$, $M^w, v \models [stit]_i\psi$. By reflexivity of R_i, $M^w, v \models \psi$. By induction hypothesis, $M', [v] \models \psi$. Since $[v]$ was chosen arbitrarily, for every $[v] \in W'$, such that $[w]R'_i[v]$, $M', [v] \models \psi$. Then, $M', [w] \models [stit]_i\psi$.

($\Leftarrow$) Assume $M', [w] \models [stit]_i\psi$ for some $[stit]_i\psi \in sub(\varphi)$. For a contradiction, assume $wR_i^w v$ and $M^w, v \not\models \psi$. By induction hypothesis, $M', [v] \not\models \varphi$. Since $wR_i^w v$, $M^w, w \models [stit]_i\gamma$ iff $M^w, v \models [stit]_i\gamma$ for any $\gamma \in \mathcal{L}^{\mathcal{G}}_{STIT}$. Then, $[w]R'_i[v]$. But then $M', [w] \not\models [stit]_i\psi$. Contradiction. Hence, for every $v \in W^w$, if $wR_i^w v$, then $M^w, v \models \psi$, i.e., $M^w, w \models [stit]_i\psi$. □

6.3.46. THEOREM. *The satisfiability problem of a given formula $\varphi \in \mathcal{L}^{\mathcal{G}}_{STIT}$ is decidable.*

Proof:
Take any formula $\varphi \in \mathcal{L}^{\mathcal{G}}_{STIT}$. By Theorem 6.3.45, if φ is satisfiable, it is satisfiable on a pre-model of the size at most $2^{|sub(\varphi)|}$. So, we can nondeterministically guess the pre-model $M = (W, R_{\Box}, \{R_i\}_{i \in Ags}, \{R_C\}_{C \in \mathcal{G}}, V)$ and check whether $M, w \models \varphi$ for some $w \in W$ in polynomial time [BRV01]. □

To conclude, we have demonstrated that the fragment of standard group STIT logic with disjoint groups has some desirable metalogical properties. As for plan-restricted group STIT logic, our results are easy to transfer to the analogous fragment of $\mathcal{L}^t_{pSTIT}$. All we have to do is add the following axiom for $\{\varrho_C\}_{C \in \mathcal{G}}$:

$$[stit]_C\varrho_C \leftrightarrow \varrho_C$$

If we perceive ϱ_C as a 0-ary modality, the axiom is a Sahlqvist formula, whose first-order correspondent expresses the fact that U_C is R_C-closed:

$$\forall w \in W(w \in U_C \leftrightarrow \forall v \in W(vR_Cw \rightarrow v \in U_C))$$

which is essentially the condition (2a) of Def. 6.2.1. As for (2b) of Def. 6.2.1, it is automatically satisfied for our fragments, given that all groups in $\mathcal{G}$ are pairwise

disjoint. Moreover, the choice of language restriction can be justified in $\mathcal{L}^t_{pSTIT}$ semantically: we can work with the class of **pSTIT**-frames, where $U_C \neq \emptyset$ iff $C \in \mathcal{G}$. In other words, only those sets of agents constitute groups for which we have the corresponding modalities in our language.

6.3.4 Discussion: is our definition of plans too weak?

Strengthening the definition of a group via explicit introduction of joint agreed-upon plans gives us not only conceptual but also some technical advantages. Our representation of a plan is quite general: it is any subset of group members' available actions. Such a representation might appear to be too weak. For instance, it allows *non-interchangeable* plans. Informally, a group's plan is interchangeable iff members of the group can fulfil the plan by acting independently of one another, without any need to coordinate their actions. Consequently, non-interchangeable plans require coordination between group members.

The notion of interchangeability was initially introduced in [Nas50] and adopted in [TD17] in the context of group plans. The interchangeability of a plan closely resembles the independence of agents constraint, since both require independence of action choices. Given a group $G = \{1, \ldots, g\}$, G's plan U_G is interchangeable iff for any worlds $w_1, \ldots, w_g \in W$, such that $w_1 R_\Box \ldots R_\Box w_g$ and $w_1, \ldots, w_g \in U_G$, the following holds:

$$R_1(w_1) \cap \ldots \cap R_g(w_g) \subseteq U_G$$

i.e., any aggregation of members' actions that are consistent with the plan is a group action. If a plan is interchangeable, group members can independently choose any actions of their own that are consistent with the plan and be sure that the plan will be fulfilled, without any need to communicate and coordinate their choice of actions. In [TD17], it is argued that only interchangeable plans are plausible, given that the agents cannot communicate to act conditionally on the choice of others. We allowed non-interchangeable plans, since we reasoned about individual and group actions causally: we have made no assumptions about agents' knowledge and abilities to communicate. Therefore, groups may have non-interchangeable plans if they have sufficient communication sources to coordinate their actions accordingly. In the following section, we explicitly show what communication sources are necessary and sufficient.

6.4 Solving coordination problems via communication

This section is based on [KDB26].

Coordination problems have received significant attention in the fields of theoretical economics and philosophy [Sch58; HS88b; Bac06; Sug03]. Unlike in most of these publications,[11] we will study coordination problems where agents can communicate with one another. Namely, our goal is to identify the necessary and sufficient conditions of communication to avoid coordination failures. Given an arbitrary coordination problem, who must inform whom about their actions to prevent coordination failures? We present an epistemic game-theoretical framework for addressing these questions. In this framework, we semantically prove our main theorem: the coordination problem can be solved via communication if and only if every subset of agents whose actions are *interdependent* can engage in semi-private communication within that set.

The section is organised as follows: in Section 6.4.1, we recall basic definitions and results on coordination games and fix the terminology and notation. We slightly restate the result of [TD17]: if we assume that agents can communicate *publicly*, then any coordination problem can be solved by adopting an optimal, interchangeable plan. In Section 6.4.2, we weaken the assumption regarding who can communicate and show that the necessary and sufficient communication channels depend on the plan: agents must be able to communicate to avoid coordination failures iff their actions are interdependent. After introducing the formal language and semantics in Section 6.4.3, we formally state and semantically prove our main theorem about the necessity and sufficiency of semi-private communication in critical sets (Section 6.4.4). In Section 6.4.4, we discuss whether semi-private communication is indeed necessary, emphasising why fully private communication – some agents sharing information, while outsiders do not know that this happens – is not sufficient. We conclude and outline directions for future work in Section 6.4.5.

6.4.1 Coordination in deontic games: preliminaries

Throughout the section, we will work with a class of *deontic games*. In deontic games, agents share a common interest and must align their actions to serve that interest [TH20].[12] Such common interest may be, e.g., to fulfil their jointly agreed-upon plan, or to maximise their utility. For our purposes, the nature of agents' common interest is inessential. Since multiple joint actions may meet everybody's interest, agents need to coordinate accordingly.

6.4.1. Definition (Deontic games). A deontic game is a tuple:

$$G = (Ags, \{A_i\}, d),$$

[11] A notable exception is a paper of [TD17], where it is assumed that group members are capable of pre-play public communication to establish common knowledge about the adoption of the *plan*.

[12] In deontic games introduced by [TH20], agents have a shared moral code rather than a common interest. This difference in interpretations is not essential for our purposes.

where Ags is a nonempty finite set of agents; for every $i \in Ags$, A_i is a nonempty finite set of actions available for i; $d : \prod A_i \to \{0, 1\}$ is a function that partitions action profiles into deontically ideal ones $d^{-1}(1)$ and deontically non-ideal ones $d^{-1}(0)$. There exists at least one action profile $\alpha \in \prod_{i \in Ags} A_i$, such that $d(\alpha) = 1$. Intuitively, an action profile is deontically ideal iff it promotes the common interest of all agents. We suppose that there is at least one such action profile.

Given any game $G = (Ags, \{A_i\}_{i \in Ags}, d)$, we use the following notational conventions. $Dom(G) := \prod_{i \in Ags} A_i$, the set of all action profiles will be referred to as a domain of the game. If $Ags = \{1, \ldots, n\}$ and $\alpha = \langle a_1, \ldots, a_n \rangle \in Dom(G)$, then for any $i \in Ags$: $\alpha(i) := a_i$. Analogously, for any $C \subseteq Ags$ and any $\alpha \in Dom(G)$, $\alpha(C) \in \prod_{i \in C} A_i$ is a projection of α to C, e.g. $\langle a_1, a_2, a_3 \rangle(\{1, 3\}) := \langle a_1, a_3 \rangle$. For any $\alpha, \beta \in Dom(G)$, $C \subseteq Ags$, $\alpha \equiv_C \beta$ iff $\alpha(C) = \beta(C)$.

6.4.2. Example. Three friends, Anne, Boris and Chloe, live between two train stations: Right and Left. Chloe has severe conflicts with both Anne and Boris. In case she meets either of the two alone, she cannot avoid an argument. Nevertheless, if she meets both Anne and Boris, she will not start an argument, because she knows the third party will interfere.

Every day, Anne, Boris and Chloe have to take a train to the nearest city, which they can do at either of the two train stations. Neither of them prefers either of the two stations to the other. The common interest of Anne, Boris and Chloe is to prevent public arguments, since they damage their reputation as a friend group.

We formalise Example 6.4.2 as a deontic game $G = (Ags, \{A_i\}_{i \in Ags}, d)$, where: $Ags = \{a, b, c\}$ is the set of players, Anne, Boris and Chloe, respectively. For every $i \in Ags$, $A_i = \{r_i, l_i\}$: r_i is i's action to go to Reft, l_i – to go to Light; d is a deontic ideality function, such that $d(\langle r_a, r_b, l_c \rangle) = d(\langle l_a, l_b, r_c \rangle) = d(\langle r_a, r_b, r_c \rangle) = d(\langle l_a, l_a, l_c \rangle) = 1$, and for any other action profile α, $d(\alpha) = 0$. In other words, it is deontically ideal for Ags that Chloe does not meet either Anne or Boris without a third party.

	l_a	r_a
l_b	1	0
r_b	0	1

l_c

	l_a	r_a
l_b	1	0
r_b	0	1

r_c

Figure 6.9: Coordination game for Example 6.4.2

Anne may go to either of the two stations only if Boris goes there as well. Analogously, Boris can go to either of the two stations only if Anne comes along. If Boris and Anne cannot agree upon the station they should go to, there is no way to knowingly avoid the undesirable outcome.

There are two generally acknowledged ways to prevent coordination failures in similar cases, when agents cannot communicate: introducing focal points and revising the solution concept. A *focal point* (or *Schelling point*) is a solution to a coordination problem that agents tend to choose by default. Thomas Schelling showed the existence of focal points experimentally [Sch58]. Schelling asked his students where and when they would go in case they needed to meet a person in New York City without having a chance to agree on time and place in advance. The most common answer was noon at the Grand Central Terminal. There is no clear reason why this time and place is preferred. This choice is grounded in nothing but the custom of New Yorkers to use Grand Central Terminal as a meeting place. When and how exactly the focal points form is still an open question.

Given that the focal points do not necessarily exist in every case, they do not entirely solve our problem. Sometimes, coordination problems can be solved by revised solution concepts. For example, if deontically ideal action profiles somehow differ in terms of payoffs, e.g. one Pareto-dominates others, or one of them is less risky,[13] it is reasonable to say that agents are rationally pressured to take this difference into account as well [HS88a]. Another approach is switching to *team reasoning*: first, agents think about what is best for all of them *as a group*, and then individually deliberate on how they should act in order to fulfil their share of the best group behaviour [Bac06]. Both approaches are really powerful tools that can solve a wide range of coordination problems, e.g., the stag hunt game and the like. [Dui18] amends team reasoning with a more elaborate participatory-reasoning schema, which can be applied to typical collective action problems. Nonetheless, [Dui18] argues that participatory reasoning may still fail to solve cooperation problems because of external or coordination uncertainty.

If we assume that agents have sources of pre-play communication that allow them to adopt or revise their plan and make it commonly known, then they can agree to narrow their choice between the otherwise equally preferred joint actions [TD17]. Technically speaking, plan adoption is modelled by updating the ideality function: in the updated game, group actions that go against the plan are no longer deontically ideal.

6.4.3. Definition (Plan updates). Given a deontic game $G = (Ags, \{A_i\}_{i \in Ags}, d)$, a plan $P \subseteq Dom(G)$ is an arbitrary set of group actions. The update of G with P, denoted as $G \uparrow P$, is the deontic game $G \uparrow P = (Ags, \{A_i\}_{i \in Ags}, d^{\uparrow})$, where $d^{\uparrow}(\alpha) = d(\alpha)$ if $\alpha \in P$; otherwise, $d^{\uparrow}(\alpha) = 0$.

As it was shown in [TD17], agents should adopt *optimal* and *interchangeable*

[13] Informally, given two action profiles α, β, α *risk-dominates* β iff when each player assigns some probability to the possibility that the other players might not chose their joint strategy, choosing α yields a higher expected payoff than choosing β.

plans, i.e., the set of the group's chosen joint actions should contain only deontically ideal joint actions and presuppose independence between agents' individual actions.

6.4.4. Definition. Given a deontic game $G = (Ags, \{A_i\}_{i \in Ags}, d)$, an individual action $a_i \in A_i$ of agent i is called **acceptable** iff it is compatible with some deontically ideal action profile, i.e., for some $\alpha \in Dom(G)$, $d(\alpha) = 1$ and $\alpha(i) = a_i$. Analogously, given any subgroup $C \subseteq Ags$, a subgroup action $a_C \in \prod_{i \in C} A_i$ is acceptable iff there exists an action profile $\alpha \in Dom(G)$, such that $d(\alpha) = 1$ and $\alpha(C) = a_C$.

A plan $P \subseteq Dom(G)$ is **optimal** iff for every $\alpha \in P$, $d(\alpha) = 1$. A plan $P \subseteq Dom(G)$ is **interchangeable** iff for any $\alpha_1, \alpha_2 \in P$, for any disjoint $A, B \subseteq Ags$, there exists an action profile $\alpha_3 \in P$, such that $\alpha_1 \equiv_A \alpha_3$ and $\alpha_2 \equiv_B \alpha_3$. See Figure 6.10 for an example of an interchangeable plan.

	1_a	2_a	3_a
1_b	1	1	0
2_b	0	0	0
3_b	1	1	0

Figure 6.10: Example of a two-player coordination game with an interchangeable plan. $A_a = \{1_a, 2_a, 3_a\}$, $A_b = \{1_b, 2_b, 3_b\}$, $P = \{\langle 1_a, 1_b\rangle, \langle 1_a, 3_b\rangle, \langle 2_a, 1_b\rangle, \langle 2_a, 3_b\rangle\}$, Note that $P = \{1_a, 2_a\} \times \{1_b, 3_b\}$, where $\{1_a, 2_a\}$ are acceptable for a and $\{1_b, 3_b\}$ are acceptable for b.

As it was shown in [TD17], given a set of agents $Ags = \{1, \ldots, n\}$, a deontic game $G = (Ags, \{A_i\}_{i \in Ags}, d)$ and an optimal and interchangeable plan P, individual actions $a_1 \in A_1, \ldots, a_n \in A_n$ are acceptable in $G \uparrow P$ iff $\langle a_1, \ldots, a_n\rangle \in P$. In other words, the optimal and interchangeable plan narrows the set of acceptable actions for every agent in such a way that every aggregation of members' acceptable actions is in the plan and deontically ideal. That way, every agent can choose any of her acceptable actions without knowing what others are doing, and the deontically ideal outcome will be reached.

The set of deontically ideal actions in Example 6.4.2 is not interchangeable: $d(\langle r_a, r_b, r_c\rangle) = d(\langle l_a, l_b, l_c\rangle) = 1$ and $\langle l_a, l_b, l_c\rangle \equiv_{\{b,c\}} \langle r_a, l_b, l_c\rangle \equiv_{\{a\}} \langle r_a, r_b, r_c\rangle$, but $d(\langle r_a, l_b, l_c\rangle) = 0$. If Anne, Boris and Chloe have sufficient sources of pre-play communication to establish common knowledge among them, then they can adopt an interchangeable plan to narrow their choice and avoid coordination failures. For example, they can agree that Anne and Boris should go to Left: $P = (\alpha \in Dom(G) \mid \alpha(\{a, b\}) = \langle l_a, l_b\rangle)$. P is optimal and interchangeable. In that case, both Anne and Boris know they should go to Left, while Chloe knows that she can go to any station, and as long as everyone follows the plan, Chloe will not meet Anne or Bob without a third party.

The solution of [TD17] underlines the key point: if agents can establish common knowledge among them via pre-play communication, any coordination problem can be solved. For every coordination game G, there exists at least one optimal and interchangeable plan P. Take any action profile $\alpha \in Dom(G)$, such that $d(\alpha) = 1$ (by Definition 6.4.1, such an action profile exists). Trivially, the singleton plan $P = \{\alpha\}$ is optimal and interchangeable. The solution of [TD17] requires a very strong assumption about agents' ability to communicate. Namely, [TD17] consider the adoption of the group's plan to be *public*, which results in the plan being commonly known among all members. As they note, "*[t]hese idealizing assumptions [about public adoption of the plan] are similar to those that underlie the logic of public announcements*" [TD17, Footnote 14]. Therefore, for a group of agents to adopt the plan, they should be able to communicate *publicly*: agents' messages should simultaneously reach everyone, everyone gets notified that the message has reached anyone, and so on.

6.4.2 Weakening the announcements

So far, we have explored solving coordination problems under two assumptions about agents' ability for pre-play communication. If we assume that agents cannot communicate at all, the coordination problem can be solved if there exists a focal point [Sch58] or if deontically ideal outcomes differ in terms of pay-offs or riskiness [HS88a; Bac06]. On the contrary, if we assume that agents can communicate publicly, they can solve any coordination problem by adopting an optimal interchangeable plan [TD17]. The first assumption seems too weak: most of the time, group members can share at least some information with at least some of their peers. On the contrary, the second assumption is too strong: rarely can agents publicly announce what strategies they have decided to follow. Often, only some agents have sources of communication, and their messages can reach only some subset of group members. In this subsection, we relax the assumption of public communication. We show in which cases only private communication inside some subgroups is enough to prevent coordination failures.

As it is shown by [TD17], there is no need for agents to communicate iff the set of deontically ideal outcomes is interchangeable, or, in other words, if they can achieve deontically ideal states by acting independently from one another. We show that, given any deontic game, it is necessary and sufficient that only agents whose actions are dependent on one another can sequentially let one another know what they are about to do.

To illustrate our point, we go back to Example 6.4.2. Intuitively, the actions of Anne and Boris are interdependent: to achieve a deontically ideal outcome, whatever station Anne chooses, Boris should choose the same station as well, and vice versa. On the contrary, Chloe's actions are independent of the joint actions of Anne and Boris. Chloe's choice of action does not restrict Anne and Boris in

their choices, and vice versa. Therefore, it suffices, e.g, for Anne to let Boris know which station she will go to. Suppose Anne privately tells Boris she will go to Left. Then, Boris will know that he should choose Left as well, otherwise, either Anne or he will meet Chloe without a third party. Chloe can safely choose any station she prefers, and since both Anne and Boris will be at Left, Chloe will either meet neither or both of them. Then, the plan of three agents is not interchangeable, but coordination failure can be avoided via private communication. The example indicates that public communication may not always be necessary.

We want to formally verify our intuition about the necessity and sufficiency of private communication between interdependent agents. Therefore, we need to reason about not only (in)dependence between all agents and their common knowledge, as in the case of publicly adopted interchangeable plans, but also about (in)dependence and common knowledge inside subgroups as well. Therefore, first, we define a formal language that is rich enough to express these notions. Then, we define a semantical framework – ex-ante epistemic game models – to interpret that formal language. Finally, we prove the theorem about the necessity and sufficiency of private communication between interdependent agents in that language.

6.4.3 Formal language, semantics

We fix a finite set of agents $Ags = \{1, \dots, n\}$ and finite sets of the corresponding action names, e.g. for any $i \in Ags$, $A_i = \{1_i, \dots, j_i\}$ for some $j \in \mathbb{N}$. We recursively define the language $\mathcal{L}_{GEL}$ as follows:

$$\pi := a_i \mid \neg\pi \mid (\pi \vee \pi)$$

$$\mathcal{L}_{GEL} \ni \varphi := \delta \mid \pi \mid \neg\varphi \mid (\varphi \vee \varphi) \mid \Box\varphi \mid K_i\varphi \mid \mid C_X\varphi \mid [\pi]_X\varphi$$

where $a_i \in A_i$, $i \in Ags$, $X \subseteq Ags$. δ stands for "*a deontically ideal outcome is reached*". a_i means that agent i has done the action a_i. $\Box\varphi$ means that it is necessarily true that φ. $K_i\varphi$ reads as "*agent i knows that* φ". $C_X\varphi$ reads as "*it is commonly known in X that* φ" (we will abbreviate C_{Ags} as just C). $[\pi]_X\varphi$ reads as "*if it is semi-privately announced to the members of X that* π*, then* φ *holds*". By semi-private announcement of π to X we mean the following: all members of X establish common knowledge among them that π, while other agents only get to know that members of X have reached common knowledge *whether* π.[14]

[14]To the best of our knowledge, the concept of semi-private (or semi-public) announcement first appears in [BMS23].

Note that $[K_i a_j]_X \varphi$, $[C_X \varphi]_X \psi$, $[[\varphi]_X \psi]_Y \theta$ etc. are not well-formed formulas: we only allow announcements about actions, not about other announcements or agents' knowledge. Notational conventions: we abbreviate $\wedge, \to, \leftrightarrow$ via $\neg, \vee$ in a standard way, as well as modalities' duals: $\widehat{K}_i := \neg K_i \neg$, $\widehat{C}_X := \neg C_X \neg$, $\langle \pi \rangle_X \psi := \neg [\pi]_X \neg \psi$. $\widehat{K}_i \varphi$ reads as "*i considers it possible that* φ", $\widehat{C}_X \varphi$ – "*it is commonly considered possible among* X *that* φ", $\langle \pi \rangle_X \varphi$ – "π *is semi-privately announced to* X*, and after the announcement,* φ *is true*".

In order to interpret this language, we will use *epistemic game models* [Aum76].

6.4.5. Definition (Epistemic game models). Given a deontic game G, an epistemic game model is a tuple $M = (G, W, \{\sim_i\}_{i \in Ags}, \sigma)$, where:

- G is a deontic game;
- W is a non-empty set of possible worlds;
- $\sim_i \subseteq W \times W$ is an equivalence relation on possible worlds, representing agent's knowledge:
- $\sigma : W \to Dom(G)$ is a function that takes a world and returns its corresponding outcome of the game.

For brevity, $\Delta := \{w \in W \mid d(\sigma(w)) = 1\}$, i.e., worlds in Δ correspond to the deontically ideal outcomes.

Semi-private communication on epistemic game models is modeled via *semi-private updates*:

6.4.6. Definition (Semi-private updates). For any epistemic game model $M = (G, W, \{\sim_i\}_{i \in Ags}, \sigma)$, a set of worlds $W' \subseteq W$ and a group $X \subseteq Ags$, $M^X_{W'} = (G, W, \{\sim_i^{(X,W')}\}_{i \in Ags}, \sigma)$ is an updated model, where:

- G, W, σ are defined just like in M;
- $\sim_i^{(X,W')} = \begin{cases} \sim_i \cap (W' \times W') & \text{if } i \in X \\ \sim_i & \text{otherwise} \end{cases}$

$M^X_{W'}$ models the state of affairs after it was semi-privately announced to members of X that the outcome of the game lies in W'.

We will work with a certain subclass of epistemic game models, which we call *ex ante* models. Namely, ex ante models depict the knowledge of agents in games of complete information that they have before any of them has decided upon their action. In other words, every agent in the ex ante model knows neither her own action nor the action of any other player, and it is commonly known which strategies every agent has, what outcomes are deontically ideal, and who knows and does not know what.

6.4.7. Definition (Ex ante models). Given any deontic game G, an ex ante model for this game is an epistemic game model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$, such that:

1. There is a one-to-one correspondence between the domain of the game and the set of possible worlds, i.e. σ is a bijective function;
2. For any agent $i \in Ags$, $\sim_i = W \times W$.

From now on, we denote the class of ex-ante models as $\mathbf{E}$.

6.4.8. Definition (Semantics for $\mathcal{L}_{GEL}$). For any epistemic game model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$, let $\sim_X = \left(\bigcup_{i\in X} \sim_i\right)^*$ be a transitive closure of $\bigcup_{i\in X} \sim_i$. Then, for every model M and a world $w \in W$, we inductively define $\models$ relation as follows:

$$
\begin{array}{lcl}
M, w \models \delta & \Leftrightarrow & w \in \Delta \\
M, w \models a_i & \Leftrightarrow & \sigma(w)(i) = a_i \\
M, w \models \neg\varphi & \Leftrightarrow & M, w \not\models \varphi \\
M, w \models \varphi \vee \psi & \Leftrightarrow & M, w \models \varphi \text{ or } M, w \models \psi \\
M, w \models \Box\varphi & \Leftrightarrow & \forall v \in W :\ M, v \models \varphi \\
M, w \models K_i\varphi & \Leftrightarrow & \forall v \in W\, (w \sim_i v \Rightarrow M, v \models \varphi) \\
M, w \models C_X\varphi & \Leftrightarrow & \forall v \in W\, \big(w \sim_X v \Rightarrow M, v \models \varphi\big) \\
M, w \models [\pi]_X\varphi & \Leftrightarrow & M, w \not\models \pi \text{ or } M^X_{[\![\pi]\!]}, w \models \varphi
\end{array}
$$

where $[\![\pi]\!] = \{w \in W \mid M, w \models \pi\}$. From now on, we denote the updated models as follows: $M^X_\pi := M^X_{[\![\pi]\!]} = (G, W, \{\sim_i^{X,\pi}\}_{i\in Ags}, \sigma)$. $M \models \varphi$ iff $M, w \models \varphi$ for all $w \in W$. $\mathbf{E} \models \varphi$ iff $M \models \varphi$ for any ex-ante model M.

After we have defined the semantics for $\mathcal{L}_{GEL}$, we proceed with expressing some crucial notions that we need to formally define necessary and sufficient conditions for communication channels to avoid coordination failures. Given a set of agents $X \subseteq Ags$, let $\vec{A}_X = \{\bigwedge_{i\in X} a_i \mid \forall i \in X : a_i \in A_i\}$ be a set of conjunctions of X-members' actions names. Informally, every formula $\varphi \in \vec{A}_X$ is a name of some joint action of X. We can express the acceptability of (joint) actions. Namely, given a deontic game G and its ex ante model M, an action a_i is acceptable iff $M \models \Diamond(a_i \wedge \delta)$, i.e., there exists a deontically ideal world where a_i is done. The acceptability of any joint action $\varphi_X \in \vec{A}_X$ is expressed analogously: $M \models \Diamond(\varphi_X \wedge \delta)$.

On ex-ante models, we can express *independence* between actions of some subgroups via acceptability. Given any $A, B \subseteq Ags$, actions of A and B are independent, $A \,||\, B$, iff an aggregation of any acceptable joint actions of A and B

constitute an acceptable joint action of $A \cup B$:

$$A \,||\, B := \bigwedge_{\varphi_B \in \vec{A}_B} \bigwedge_{\varphi_A \in \vec{A}_A} (\Diamond(\varphi_A \wedge \delta) \wedge \Diamond(\varphi_B \wedge \delta)) \to \Diamond(\varphi_A \wedge \varphi_B \wedge \delta) \tag{6.3}$$

Consequently, we can define interchangeability of the set of deontically ideal action profiles as independence of all disjoint subgroups:

$$||_{Ags} := \bigwedge_{A \cap B = \emptyset} A \,||\, B$$

Alternatively, given $Ags = \{1, \ldots, n\}$,

$$||_{Ags} := \bigwedge_{a_1 \in A_1} \ldots \bigwedge_{a_n \in A_n} (\Diamond(a_1 \wedge \delta) \wedge \ldots \wedge \Diamond(a_n \wedge \delta)) \to \Box((a_1 \wedge \ldots \wedge a_n) \to \delta) \tag{6.4}$$

It is a routine to check that both definitions of $||_{Ags}$ are equivalent. Note that on ex-ante models (or any models, where σ is bijective), there is an equivalent definition of $||_{Ags}$:

$$\mathbf{E} \models ||_{Ags} \leftrightarrow \big(\bigwedge_{a_1 \in A_1} \ldots \bigwedge_{a_n \in A_n} (\Diamond(a_1 \wedge \delta) \wedge \ldots \wedge \Diamond(a_n \wedge \delta)) \to \Diamond(a_1 \wedge \ldots \wedge a_n \wedge \delta)\big)$$

since there is a single world, where $a_1 \wedge \ldots \wedge a_n$ is true.

We can express agents' knowledge about the acceptability of their actions. Namely, given a pointed model (M, w), $M, w \models \widehat{K}_i(a_i \wedge \delta)$ iff i considers a_i acceptable. Consequently, $M, w \models K_i(a_i \to \neg\delta)$ means that i knows that her action a_i is unacceptable. Agents' actions may not be independent in general, but if we look only at actions that agents themselves consider acceptable, they might well be independent.

Given a pointed epistemic game model (M, w) of a deontic game G, we say that G is *epistemically interchangeable* at (M, w) iff every agent considers it possible to fulfil the plan and the agglomeration of any actions that agents themselves consider acceptable is acceptable. Formally:

$$||^{\sim}_{Ags} := \bigwedge_{i \in Ags} \widehat{K}_i\delta \wedge \bigwedge_{a_1 \in A_1} \ldots \bigwedge_{a_n \in A_n} (\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_n(a_n \wedge \delta)) \to \Box((a_1 \wedge \ldots \wedge a_n) \to \delta) \tag{6.5}$$

In the case of epistemic interchangeability, a deontically ideal outcome will be reached without additional communication, given that no one will do an action

that she knows to be unacceptable. If the agents have achieved epistemic interchangeability, even if the set of deontically ideal outcomes is not interchangeable, they can use their knowledge to act *as if* that set were interchangeable. They can just avoid doing actions that they know to be unacceptable, and some deontically ideal outcome will inevitably be reached. Note that on ex-ante models, just like on any other models with bijective σ, $||^{\sim}_{Ags}$ has an equivalent definition:

$$\mathbf{E} \models ||^{\sim}_{Ags} \leftrightarrow$$
$$\Big(\bigwedge_{i \in Ags} \widehat{K}_i \delta \wedge \bigwedge_{a_1 \in A_1} \dots \bigwedge_{a_n \in A_n} (\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_n(a_n \wedge \delta)) \rightarrow \Diamond(a_1 \wedge \ldots \wedge a_n \wedge \delta)\Big)$$

since there is a unique world, where $a_1 \wedge \ldots \wedge a_n$ is true.

It is straightforward that from $M, w \not\models A||B$ it follows that $M, w \not\models (A \cup X)||(B \cup Y)$ for any $X, Y \subseteq Ags$. In other words, dependence is *monotonic*: as soon as actions of two subgroups depend on one another, actions of any of their supergroups are dependent as well. If we want to capture the genuine, irreducible dependence between subgroups, we need the notion of a *critical set*.[15] Informally, $X \subseteq Ags$ is a critical set iff there is a dependence that involves all members of X.

6.4.9. Definition (Critical sets in deontic games). Given an ex-ante model of some deontic game M, a set $X \subseteq Ags$ is called *critical* iff there is a *critical partition* of X onto X_1 and X_2, i.e. a partition such that:

1. $M \not\models X_1||X_2$, i.e., X_1 and X_2 are not independent;
2. $M \models (X_1 \cap B)||(X_2 \cap B)$ for any $B \subset X$, i.e., dependence between X_1 and X_2 cannot be reduced to the dependence between their proper subgroups.

Actions of a critical set's members are interdependent: a good choice of strategy of any member is restricted by the choices of all other members.

As for semi-private announcements, we say that an announcement $[\pi]_X$ is *internal* iff π is a formula that is built only from action names $\{A_i\}_{i \in X}$. In other words, $[\pi]_X$ is an internal semi-private announcement iff only information about strategies of some members of X is announced in X. An announcement $[\pi]_X$ is *individual* iff π is built only from action names A_i of some single agent $i \in Ags$ and Boolean connectives, i.e., $[\pi]_X$ is *individual* iff only information about actions of one agent is announced. Assume $i \in X \subseteq Ags$ and π is a formula built from A_i and Boolean connectives. Then, $[\pi]_X$ is an *internal individual semi-private announcement*. Such announcements model semi-private messages of i to X about

[15] [NN13] introduces the notions of critical sets and critical partitions for purely technical reasons, to establish a completeness proof for their logic.

i's choice of strategy: it might be the exact choice of i's strategy or some partial solution.

By semi-private *disclosure of i's action to X* we mean the announcement, after which i's action is commonly known in X, while outsiders only get to know that agents in X get to know the exact action of i, whatever this action is. Such disclosures are also expressible in our language:

6.4.10. Definition (Semi-private disclosure of action). For a deontic game $G = (Ags, \{A_i\}_{i\in Ags}, d)$, its ex ante model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$ and a world $w \in W$, the semi-private disclosure of i's action to X is the semi-private announcement of i's action to X, i.e. $[\sigma(w)(i)]_X$, followed by semi-private announcements of negations of all other i's actions to X, i.e. $[\neg i_j]_X$ for any $i_j \in A_i \setminus \{\sigma(w)(i)\}$. For example, if $A_i = \{i_1, i_2, i_3\}$ and $\sigma(w)(i) = i_1$, i.e. i does i_1 in w, then the semi-private disclosure of i's action to X in w is the following announcement: $\langle i_1\rangle_X\langle\neg i_2\rangle_X\langle\neg i_3\rangle_X$. It is a routine to check that this announcement indeed works as a semi-private disclosure of i's action to X, i.e., that:

$$M, w \models \langle i_1\rangle_X\langle\neg i_2\rangle_X\langle\neg i_3\rangle_X C(\bigwedge_{i_j\in A_i} i_j \to C_X i_j)$$

From now on, we will use the abbreviation $M, w \models [i]_X\varphi$ for a semi-private disclosure of i's action $\sigma(w)(i)$ to X.

Returning to Example 6.4.2, we can now formally demonstrate how pre-play semi-private communication can solve the coordination problem. Let $M = (G, W, \{\sim\}_{i\in Ags}, \sigma)$ be the ex-ante model for the deontic game for Example 6.4.2, presented in Section 6.4.1. Note that the set $\{a, b\}$ is critical: $M \not\models a||b$, while trivially $M \models a||\emptyset$ and $M \models \emptyset||b$.

Without loss of generality, we can assume that Anne goes to Right, while Boris and Chloe go to Left. Could they have avoided this undesirable outcome by pre-play communication?

Let $\sigma(\langle r, l, l\rangle) = w$ be the real world. Suppose Anne had semi-privately disclosed her action to Boris. Then, Anne would still consider going Right acceptable and no longer consider it possible to go left:

$$M, w \models [a]_{\{a,b\}}\big(K_a\neg l_a \wedge \widehat{K}_a(r_a \wedge \delta)\big)$$

Then, Boris will consider the only strategy of his acceptable, to go to the same station as Anne, i.e., Right. Knowing where Anne will go, Boris knows that if he does not go to the same station, either he or Anne will meet Chloe without a third party, which is deontically non-ideal:

$$M, w \models [a]_{\{a,b\}}(K_b(l_b \to \neg\delta) \wedge \widehat{K}_b(r_b \wedge \delta))$$

As a result, after the semi-private disclosure, both Anne and Boris will consider only a single action of each of them acceptable – to go to Right – while Chloe will still consider both of her actions acceptable. Any combination of Anne, Boris and Chloe's actions that they consider acceptable is acceptable, i.e., after the announcement, they will reach the epistemic interchangeability: $M, w \models [a]_{\{a,b\}}||^{\sim}_{Ags}$.

As insiders of communication, Anne and Boris will know that they have reached epistemic interchangeability. As an outsider, Chloe will not know what action Anne has disclosed to Boris, but since the disclosure was semi-private, she will learn that Anne and Boris know enough to choose the same train station. Therefore, Chloe will also get to know that they have reached epistemic interchangeability. Moreover, after the semi-private disclosure of Anne's strategy, it will be commonly known: $M, w \models [a]_{\{a,b\}}C(||^{\sim}_{Ags})$. So it will be commonly known that everyone knows enough to act as if the set of deontically ideal outcomes were interchangeable. In other words, it will be commonly known that agents can avoid coordination failure by choosing strategies they consider acceptable, so no further communication is needed.

6.4.4 When semi-private communication is enough: formal proof

In this subsection, we prove our main theorem: semi-private communication via internal individual announcements in critical sets is a necessary and sufficient condition for solving any coordination problem. We limit communication to *internal individual* semi-private announcements to exclude vicious circles in our explanation. If we had not restricted ourselves to individual announcements and allowed groups to announce their choice of actions, we would need to explain how these groups solve coordination problems to make such truthful announcements. The announcements should be internal to exclude cases of mind-reading: only the agent herself can announce what she has chosen to do. First, we show the necessity of such communication. In other words, we show that the common knowledge of epistemic interchangeability cannot be reached by a sequence of announcements if such announcements do not include at least one member of at least one critical set as an insider.

6.4.11. Theorem (Communication in critical sets is necessary). *Take any game* $G = (Ags, \{A_i\}_{i\in Ags}, d)$ *with its ex ante game model* $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$ *and any possible world* $w \in W$. *Take any sequence of individual internal semi-private announcements* $[\alpha_1]_{X_1} \dots [\alpha_n]_{X_n}$, *such that* $M, w \models \alpha_1 \wedge \dots \wedge \alpha_n$ *and there is a critical set* $X \subseteq Ags$, *s.t.* $X \not\subseteq \bigcup_{1\leq i\leq n} X_i$. *Then,* $M, w \not\models \langle\alpha_1\rangle_{X_1} \dots \langle\alpha_n\rangle_{X_n} C(||^{\sim}_{Ags})$.

Proof:
See Theorem 6.4.19 in Appendix (Sec. 6.4.6). □

Now we will show that communication inside only critical sets is sufficient. To make our formulation precise and justify it, we point out that epistemic interchangeability presupposes that everyone considers it possible to reach a deontically ideal state. Therefore, to reach the common knowledge of epistemic interchangeability, agents should announce they are about to do actions that they consider acceptable, given what they know. Going back to Example 6.4.2: if Anne announces that she is to go Left, Boris cannot announce that he goes to Right to solve the coordination problem: if such announcements are truthful, then Anne and Boris will know they will not reach a deontically ideal state. Having this in mind, we formulate the sufficiency condition as follows:

6.4.12. Theorem (Communication in critical sets is sufficient). *For a deontic game $G = (Ags, \{A_i\}_{i\in Ags}, d)$, take its ex-ante epistemic game model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$. Let $\mathcal{X} = \{X_1, \dots, X_n\}$ be the collection of all critical sets of G. Then, there exists a sequence of individual internal semi-private announcements $[\alpha_1]_{X_1} \dots [\alpha_q]_{X_n}$ such that:*

1. *For any set Y, if $Y \notin \mathcal{X}$, then there is no announcement $[\alpha]_Y$. In other words, communication happens only inside critical sets.*
2. *At every mth step of the sequence $[\alpha_1]_{X_1} \dots [\alpha_m]_{X_j}$, if α_{m+1} is an expression about actions of $i \in Ags$, then $M \models [\alpha_1]_{X_1} \dots [\alpha_m]_{X_j} \widehat{K}_i(\alpha_{m+1} \wedge \delta)$, i.e. at every step the announced action of the agent is considered acceptable by that agent after all the announcements that were already made.*
3. *$M \models [\alpha_1]_{X_1} \dots [\alpha_n]_{X_n} C(||^{\sim}_{Ags})$, i.e., after the sequence of announcements, the common knowledge about epistemic interchangeability is established.*

Proof:
See Theorem 6.4.20 in Appendix (Sec.6.4.6). □

Interestingly, the result of [TD17] follows from Theorems 6.4.19,6.4.20. If the commonly known plan is already interchangeable, then the only critical set is $\emptyset$, therefore, the absence of communication between agents is sufficient to solve the coordination problem and there is no necessity for any agents to communicate.

Discussion: would full privacy be sufficient?

Why is it necessary to achieve common knowledge that everyone has enough information? Why is it not enough to achieve a state where everyone actually has enough information, even if unknown to others? For example, what would be different in the scenario of Example 6.4.2 if Anne had communicated privately to Boris, unbeknownst to Chloe? Chloe still could have done her share – just come to any station – without knowing that others had enough information to independently do their shares.

We can find good arguments in favour of the necessity of common knowledge and hence semi-private communication in the literature on *group identification*. As we have argued in Section 2.2, agents can reason either individualistically, i.e. ask themselves "*what should I do?*" or engage in *team reasoning*, i.e. ask themselves "*what we as a group should do?*" In the second scenario, the individual choices of group members' strategies should follow from the choice of the optimal joint strategy of the group. Nevertheless, it is not always obvious which mode of practical reasoning one should employ. As pointed out by [Sug03], often one should engage in team reasoning only if one is sure that others identify themselves as group members and reason "socially" as well. In case one lacks reasons to believe that other group members are engaged in team reasoning, one may lose the commitment to act towards the common good.

We can make a similar point with respect to (absence of) sufficient communication. Some deontic games require agents to communicate; ergo, if one lacks reasons to believe that others communicate in the way that the game requires them to do, one may lose their commitment to promote common interest and choose some unacceptable action that promotes one's own interest.

6.4.5 Envoi

In this section, we have explored the solutions to coordination problems via preplay communication. Namely, we have specified the necessary and sufficient conditions of communication, via which group members can prevent coordination failures: it is semi-private communication inside the critical sets. We have left some questions untouched, inviting further research.

We have figured out *who* should be able to communicate to solve the coordination problem, but we have not specified *what* should be communicated. So far, we have allowed any individual internal semi-private announcements: agents are free to disclose any truthful information about their own actions. This freedom follows from the idealization that every agent is omniscient with respect to her own actions: every agent can freely decide what she is to do and disclose it to anyone she can communicate with. This idealization should be relaxed. We adopt partial plans exactly because sometimes it is impossible or too costly for us to settle upon the exact action we are to take. Some plans may pressure agents to disclose their exact strategy to avoid coordination failures, while these agents may lack the resources to decide upon their exact strategy. Such restrictions should also be taken into account.

Finally, we have ignored the order of announcements, while the first agent to announce her strategy restricts the acceptable choices of others, which may result in some advantage to her. This obstacle points out that we need to explore *priority orders* among group members that often takes place in such scenarios.

6.4.6 Appendix: proofs for Section 6.4

6.4.13. LEMMA. *For an ex-ante model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$ and two disjoint subsets $A, B \subseteq Ags$, if $M \not\models A||B$, then there is a critical partition C_1, C_2, such that $C_1 \subseteq A$ and $C_2 \subseteq B$.*

Proof:
See proof of Lemma 3 in [NN13]. □

6.4.14. LEMMA. *For any ex-ante model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$, $M \not\models A \,||\, B$ for any non-trivial (but not necessarily critical) partition A, B of a critical set C.*

Proof:
See Lemma 2 in [NN13, p.67]. □

6.4.15. LEMMA. *Given a coordination game $G = (Ags, \{A_i\}_{i\in Ags}, d)$, it ex-ante model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$ and any agent $i \in Ags$, either $M \models i||\bar{i}$ or there exists a critical set C, such that $i \in C$ and $\{i\}, C \setminus \{i\}$ is a critical partition.*

Proof:

Take any agent $i \in Ags$ and assume that $M \not\models i||\bar{i}$. Since Ags is finite, so $\bar{i}$ is. Note that trivially $M \models i||\emptyset$. Hence, there exists a (not necessarily unique) $\subseteq$-minimal non-empty set $B \subseteq \bar{i}$, such that $M \not\models i||B$. Fix such $B \subseteq \bar{i}$. We claim that $B \cup \{i\}$ is a critical set and $B, \{i\}$ is a critical partition. By our assumption, $M \not\models i||B$. Moreover, since B is a $\subseteq$-minimal subset of $\bar{i}$ which is not independent from i, for any proper subset $C \subset B$, $M \models i||C$. Take any subset $D \subset B \cup \{i\}$. If $i \in D$, then $\{i\} \cap D = \{i\}$ and $D \cap B = C$ for some $C \subset B$. In such a case, $M \models \{i\} \cap D||B \cap D$, since it is equivalent to $M \models i||C$, where C is some proper subset of B, and B is a $\subseteq$-minimal set, which is not independent from i. If $i \notin D$, then $M \models \{i\} \cap D||B \cap D$, since it is equivalent to $M \models \emptyset||B \cap D$, which holds trivially. □

6.4.16. LEMMA. *For any $X \subseteq Ags$, let $||^{\sim}_X := \big(\bigwedge_{i\in X}\bigwedge_{a_i\in A_i} \widehat{K}_i(a_i \wedge \delta)\big) \to \Diamond\big(\bigwedge_{i\in X}\bigwedge_{a_i\in A_i} a_i \wedge \delta\big)$. Given any ex ante epistemic game model $M = (G, W, \{\sim_i\}_{i\in Ags}, \sigma)$, any world $w \in W$ and any sequence of individual announcements $[\pi_1]_{X_1}, \ldots, [\pi_n]_{X_n}$, if $M, w \models ||^{\sim}_X$ for some $X \subseteq Ags$, then:*

$$M, w \models [\pi_1]_{X_1} \ldots [\pi_n]_{X_n}(||^{\sim}_X).$$

Proof:
W.l.o.g., let $X = \{1, \ldots, n\}$ for some $n \in \mathbb{N}$. Assume that $M, w \models ||^{\sim}_X$, i.e. for any $a_1 \in A_1, \ldots, a_n \in A_n$, $M, w \models \big(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_n(a_n \wedge \delta)\big) \rightarrow \Diamond(a_1 \wedge \ldots \wedge a_n \wedge \delta)$.

For brevity, let $M_! = (G, W, \{\sim_i^!\}_{i \in Ags}, \sigma)$ be a model that we get after the announcements $[\pi_1]_{X_1}, \ldots, [\pi_n]_{X_n}$ in M. Let $!A_i = \{a_i \in A_i \mid \exists v \in W : w \sim_i^! v$ and $M, v \models a_i \wedge \delta\}$ be a set of actions that i still considers acceptable after the announcements. Note that for every agent $i \in Ags$, $!A_i \subseteq A_i$. Take any $b_1 \in !A_1, \ldots, b_n \in !A_n$. By definition, $M_!, w \models \widehat{K}_1(b_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_n(b_n \wedge \delta)$. Since $b_1 \in !A_1, \ldots, b_n \in !A_n$ and $!A_1 \subseteq A_1, \ldots, !A_n \subseteq A_n$, $b_1 \in A_1, \ldots, b_n \in A_n$. Then, by our initial assumption, $M, w \models \Diamond(b_1 \wedge \ldots \wedge b_n \wedge \delta)$. Then, there exists a world $v \in W$, such that $v \in \Delta$ and $\sigma(v)(i) = b_i$ for every $i \in \{1, \ldots, n\}$. Since the set of possible worlds W and the function σ in the updated model $M_!$ are the same, $M_! \models \Diamond(b_1 \wedge \ldots \wedge b_n \wedge \delta)$. Hence, $M_!, w \models \Diamond(b_1 \wedge \ldots \wedge b_n \wedge \delta)$. Given our assumption that b_i is an arbitrary member of $!A_i$ for every $i \in X$, $M_!, w \models (\widehat{K}_1(b_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_n(b_n \wedge \delta)) \rightarrow \Diamond(b_1 \wedge \ldots b_n \wedge \delta)$ for every $b_1 \in A_1, \ldots, b_n \in A_n$ (if $b_i \notin !A_i$ for some $i \in X$, then the implication is vacuously true). In other words, $M_!, w \models ||^{\sim}_X$ Then, by Def. 6.4.8, $M, w \models [\pi_1]_{X_1} \ldots [\pi_n]_{X_n}(||^{\sim}_X)$. □

6.4.17. Corollary. *If $M, w \not\models [\alpha_1]_{X_1}, \ldots, [\alpha_n]_{X_n}(||^{\sim}_X)$, then $M, w \not\models ||^{\sim}_X$.*

6.4.18. Lemma. *Given any ex ante model $M = (G, W, \{\sim_i\}_{i \in Ags}, \sigma)$, if $M, w \models \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$ for some $i \in Ags$, $\{1, \ldots, j\} \subseteq Ags$, s.t. $i \notin \{1, \ldots, j\}$, then for any internal semi-private announcement $[\pi]_{\{1,\ldots,j\}}$:*

$$M, w \models [\pi]_{\{1,\ldots,j\}} \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$$

Proof:
Assume that $M, w \models \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$. If $M, w \not\models \pi$, then vacuously $M, w \models [\pi]_{\{1,\ldots,j\}} \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$.

Let $M, w \models \pi$. Then, since $M, w \models \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$, there exists $w_1, \ldots, w_n$ and $v_1, \ldots, v_j \in W$, such that $w \sim_i w_1, \ldots, w \sim_i w_j$, $w_1 \sim_1 v_1, \ldots, w_n \sim_j v_j$ and and $M, v_1 \models \delta \wedge a_1, \ldots, M, v_j \models \delta \wedge a_j$. Obviously, $\sim_l^{(X,\pi)}$ is reflexive for any $l \in Ags$. Then, $w \sim_l^{(X,\pi)} w$, $v_1 \sim_l^{(X,\pi)} v_1, \ldots, v_j \sim_l^{(X,\pi)} v_j$ for any $l \in Ags$. Then, $M_\pi^X, v_1 \models \widehat{K}_1(a_1 \wedge \delta), \ldots, M_\pi^X, v_j \models \widehat{K}_j(a_j \wedge \delta)$.

Since $i \notin \{1, \ldots, j\}$, $\sim_i^{(X,\pi)} = \sim_i = W \times W$. Therefore, $w \sim_i^{(X,\pi)} v_1, \ldots, w \sim_i^{(X,\pi)} v_n$. Consequently, $M_\pi^X, w \models \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$, i.e. $M, w \models [\pi]_X \widehat{K}_i(\widehat{K}_1(a_1 \wedge \delta) \wedge \ldots \wedge \widehat{K}_j(a_j \wedge \delta))$. □

6.4.19. Theorem (Communication in critical sets is necessary). *For any ex ante epistemic game model $M = (G, W, \{\sim_i\}_{i \in Ags}, \sigma)$, any possible world $w \in W$ and*

any sequence of individual internal semi-private announcements $[\pi_1]_{X_1} \dots [\pi_n]_{X_n}$, *such that* $M, w \models \pi_1 \wedge \dots \wedge \pi_n$ *and there is a critical set* $X \subseteq Ags$, *s.t.* $X \not\subseteq \bigcup_{1 \leq i \leq n} X_i$: $M, w \not\models \langle \pi_1 \rangle_{X_1} \dots \langle \pi_n \rangle_{X_n} C(||^{\sim}_{Ags})$.

Proof:
Take any deontic game $G = (Ags, \{A_i\}_{i \in Ags}, d)$ and its ex ante game model $M = (G, W, \{\sim_i\}_{i \in Ags}, \sigma)$. Take any world $w \in W$ and any critical set $X \subseteq Ags$. Assume any sequence of individual internal semi-private announcements $[\pi_1]_{X_1} \dots [\pi_n]_{X_n}$, such that $M, w \models \pi_1 \wedge \dots \wedge \pi_n$ and $X \not\subseteq \bigcup_{1 \leq i \leq n} X_i$. In other words, $[\pi_1]_{X_1} \dots [\alpha_n]_{X_n}$ is a sequence of truthful announcements in subgroups that does not cover X, i.e. at least some agents in X are outsiders for all the announcements in the sequence. For brevity, let $X = \{1, \dots, l\}$, $X \cap \bigcup_{1 \leq i \leq n} X_i = Y = \{1, \dots, k\}$ for some $k < l$. Agents $\{i \in Ags | k{+}1 \leq i \leq l\}$ are members of the critical set X that are outsiders for all the announcements that were made. Since outsiders do not learn any new factual information, for every $j \in X \setminus Y$, $M, w \models \widehat{K}_j(a_j \wedge \delta) \leftrightarrow \langle \pi_1 \rangle_{X_1} \dots \langle \pi_n \rangle_{X_n} \widehat{K}_j(a_j \wedge \delta)$. Since X is a critical set, from Lemma 6.4.14 it follows that $M \not\models Y || X \setminus Y$, i.e. there exist some joint strategies $\bigwedge_{i \in Y} a_i \in \vec{A}_Y, \bigwedge_{j \in X \setminus Y} a_j \in \vec{A}_{X \setminus Y}$, such that $M, w \models \Diamond(\bigwedge_{i \in Y} a_i \wedge \delta) \wedge \Diamond(\bigwedge_{j \in X \setminus Y} a_j \wedge \delta) \wedge \Box((\bigwedge_{i \in Y} a_i \wedge \bigwedge_{j \in X \setminus Y} a_j) \rightarrow \neg\delta)$, i.e. there is some acceptable joint action of Y and some acceptable joint action of $X \setminus Y$, such that the aggregation of these two joint actions is unacceptable. Agents outside Y do not know what exactly agents inside Y have learned, and consider it possible that agents inside Y still consider $\bigwedge_{1 \leq i \leq k} a_i$ acceptable. In other words, by Lemma 6.4.18 it follows that $M, w \models \langle \alpha_1 \rangle_{X_1} \dots \langle \alpha_n \rangle_{X_n} \widehat{K}_j(\widehat{K}_1(a_1 \wedge p) \wedge \dots \wedge \widehat{K}_k(a_k \wedge p))$ for every $j \in X \setminus Y$. As they, as outsiders, still consider the same actions of themselves acceptable, every outsider considers $\bigwedge_{k+1 \leq j \leq l} a_j$ acceptable. Therefore, for every $j \in X \setminus Y$, we get the next picture:

$$M, w \models \langle \pi_1 \rangle_{X_1} \dots \langle \pi_n \rangle_{X_n} \widehat{K}_j\big(\widehat{K}_1(a_1 \wedge \delta) \wedge \dots \wedge \widehat{K}_l(a_l \wedge \delta) \wedge \neg\Diamond(a_1 \wedge \dots \wedge a_l \wedge \delta)\big)$$

which means that $M, w \models \langle \pi_1 \rangle_{X_1} \dots \langle \pi_n \rangle_{X_n} \widehat{K}_j(\neg ||^{\sim}_{Ags})$, therefore, $M, w \not\models C(||^{\sim}_{Ags})$. We can conclude that after any sequence of individual internal semi-private announcements that do not cover all critical sets, we will not reach the state where it is commonly known that the set of deontically ideal outcomes is epistemically interchangeable. Hence, communication that covers critical sets is a necessary condition of reaching such a state. □

6.4.20. Theorem (Communication in critical sets is sufficient). *Take any deontic game* G *with its ex ante epistemic game model* M. *Let* $\mathcal{X} = (X_1, \dots, X_n)$ *be a collection of all critical sets of* G. *Then, there exists a sequence of individual internal semi-private announcements* $[\pi_1]_{X_1} \dots [\pi_q]_{X_n}$ *such that:*

1. *For every critical set $X_i \in \mathcal{X}$, there is at least one corresponding announcement $[\pi_i]_{X_i}$, and for any set Y that is not critical there is no announcement $[\pi]_Y$;*
2. *At every mth step of the sequence $[\pi_1]_{X_1} \dots [\pi_m]_{X_j}$, if π_{m+1} is an expression about actions of $i \in Ags$, $M, w \models [\pi_1]_{X_1} \dots [\pi_m]_{X_j} \widehat{K}_i(\pi_{m+1} \wedge \delta)$, i.e. at every step the announced action of the agent is considered acceptable by that agent after all the announcements that were already made.*
3. *For every world $w \in W$, $M, w \models [\pi_1]_{X_1} \dots [\pi_q]_{X_n} C(||\widetilde{}_{Ags})$.*

Proof:

W.l.o.g., assume $\mathcal{X} = \{X_1, \dots, X_n\}$ and $X_i = \{x_i^1, \dots, x_{j_i}^1\} \subseteq Ags$ for every $i \in \{1, \dots, n\}$, and $A_i = \{0, 1\}$ for every $i \in Ags$.[16] Take any deontically ideal world, i.e., $w \in W$, such that $M, w \models \delta$. For brevity, for any $j \in Ags$ let $w_j := \sigma(w)(j)$ be the j's action in w. Then, we construct a sequence, where for every critical set $X_i \in \mathcal{X}$, every member of X_i semi-privately discloses her strategy at w to all members of X_i :

$$[\mathcal{X}] := [w_{x_1^1}]_{X_1} \dots [w_{x_{j_n}^n}]_{X_n}$$

For brevity, let $M_! = (G, W, \{\sim_i^!\}_{i \in Ags}, \sigma)$ be a model that we get after performing the sequence of announcements $[\mathcal{X}]$ in M.

We claim that $[\mathcal{X}]$ satisfies all three requirements. (1) is trivially satisfied, given that $\mathcal{X} = \{X_1, \dots, X_n\}$. (2) is satisfied as well. Take and any world $v \in W$ and the mth step of the sequence $[\mathcal{X}]$ for some arbitrary $m \in \mathbb{N}$ that is less or equal to the length of $[\mathcal{X}]$. Assume that mth step of the sequence is $[w_{x_l^m}]_{X_m}$ and $j+1$th step is $[w_{x_{l+1}^m}]_{X_m}$. If $M, v \not\models w_{x_1^1} \wedge \dots \wedge w_{x_l^m}$, then vacuously $M, v \models [w_{x_1^1}]_{X_1} \dots [w_{x_l^m}]_{X_m} \widehat{K}_{x_{l+1}^m}(w_{x_{l+1}^m} \wedge p)$. Assume that $M, v \models w_{x_1^1} \wedge \dots \wedge w_{x_l^m}$. Then, note that $M, w \models w_{x_1^1} \wedge \dots \wedge w_{x_l^m}$, i.e. in both w and v all the propositions that are announced so far are true. Hence, in the updated model

$$(w, v) \in \sim_j^!$$

for every $j \in Ags$, while $M_!, w \models (w_{x_{l+1}^m} \wedge \delta)$. Hence:

$$M, v \models [w_{x_1^1}]_{X_1} \dots [w_{x_l^m}]_{X_m} \widehat{K}_{x_{l+1}^m}(w_{x_{l+1}^m} \wedge \delta)$$

[16] These assumptions will not lead to loss of generality, since the construction of the sequence of announcements we are about to present can be easily generalised. Our assumption that there are non-empty critical sets is harmless, since if the only critical set is $\emptyset$, then the set of deontically ideal outcomes is already interchangeable, and it is commonly known, hence, no communication is required. Our assumption that every agent has just two strategies is harmless, since, as the reader will see further in the proof, we will construct the needed sequence of announcements from disclosures of agents' strategies. For generalization, we should replace any announcement of an agent's strategy with its semi-private disclosure.

It is left for us to prove that $M \models [\mathcal{X}]C(||^{\sim}_{Ags})$. We do it by proving that $M, v \models [\mathcal{X}](||^{\sim}_{Ags})$ for any $v \in W$. Take any world $v \in W$. If $M, v \not\models w_{x^1_1} \wedge \ldots \wedge w_{x^n_{j_n}}$, then $M, v \models [\mathcal{X}](||^{\sim}_{Ags})$ holds vacuously. For contradiction, assume that $M, v \models w_{x^1_1} \wedge \ldots \wedge w_{x^n_{j_n}}$ and $M_!, v \not\models ||^{\sim}_{Ags}$. Then, by Corollary 6.4.17, $M, v \not\models ||^{\sim}_{Ags}$. Since M is an ex ante model, $\widehat{K}_i\varphi \leftrightarrow \Diamond\varphi$ for any $\varphi \in \mathcal{L}$. Then, from $M, w \not\models ||^{\sim}_{Ags}$ it follows there exists acceptable actions $a_1 \in A_1, \ldots, a_n \in A_n$, such that $(a_1, \ldots, a_n)$ is unacceptable. Then, the set of deontically ideal outcomes in G is not interchangeable. From this, it follows that there exists at least one non-empty critical set $X \subseteq Ags$.

Let $\mathcal{C} = X_1 \cup \ldots \cup X_n$ be a union of all critical sets in G. By construction of $[\mathcal{X}]$, every agent in $\mathcal{C}$ semi-privately discloses her strategy to all members of her critical set, *including herself*. Then, for every $j \in \mathcal{C}$, for any $t \in W$, $v \sim^!_j t$ iff $\sigma(t)(j) = w_j$. In other words, after the sequence of announcements, every member of $\mathcal{C}$ considers a single action of herself possible: the one that she does in w (and given that $M, w \models \delta$, those actions are acceptable). Then, trivially $M_!, v \models (\bigwedge_{j\in\mathcal{C}} \widehat{K}_j(a_j \wedge p)) \rightarrow \Diamond(\bigwedge_{j\in\mathcal{C}} a_j \wedge \delta)$. But since $M_!, v \not\models ||^{\sim}_{Ags}$, there exists some action $b_j \in A_j$ for every $j \in Ags \setminus \mathcal{C}$, such that $M_!, v \models \bigwedge_{j\in Ags\setminus\mathcal{C}} \widehat{K}_j(b_j \wedge \delta) \wedge \neg\Diamond(\bigwedge_{j\in Ags\setminus\mathcal{C}} b_j \wedge \bigwedge_{k\in\mathcal{C}} w_k \wedge \delta)$. In other words, there should exist some actions of members of $Ags \setminus \mathcal{C}$, such that the members consider that actions of themselves acceptable, but the aggregation of these actions with $\sigma(w)(\mathcal{C})$ is unacceptable. Then, there exists an acceptable joint action of $\mathcal{C}$, i.e. $\sigma(w)(\mathcal{C})$, and an acceptable joint action of $Ags \setminus \mathcal{C}$, i.e. $\prod_{j\in Ags\setminus\mathcal{C}}\{b_j\}$, such that the composition of these two joint actions is unacceptable. Then, $M \not\models \mathcal{C}||Ags \setminus \{\mathcal{C}\}$. Then, by Lemma 6.4.13, there should exist a critical partition C_1, C_2, such that $C_1 \subseteq \mathcal{C}$ and $C_2 \subseteq Ags \setminus \mathcal{C}$. Obviously, $C_1 \neq \emptyset$ and $C_2 \neq \emptyset$. But by our assumption, $Ags \setminus \mathcal{C}$ is a set of agents that are not members of any critical set, i.e. for any critical set C, $C \cap (Ags \setminus \mathcal{C}) = \emptyset$. Contradiction. Hence, the only possibility left is that $Ags = \mathcal{C}$. But then every agent should have disclosed her own strategy to herself, i.e. for every agent there is a single strategy that she considers possible: her acceptable strategy at w. Then, given that $M_!, w \models \delta$, trivially $M_!, v \models \bigwedge_{j\in Ags} \widehat{K}_j(w_j \wedge \delta) \rightarrow \Diamond(\bigwedge_{j\in Ags} w_j \wedge \delta)$. But it contradicts our initial assumption that $M_!, v \not\models ||^{\sim}_{Ags}$. Hence, $M_!, v \models ||^{\sim}_{Ags}$. Since $v \in W$ was arbitrarily chosen, $M_! \models ||^{\sim}_{Ags}$. Then, since $\sim^!_{Ags}(w) \subseteq W$ for any $w \in W$, $M_! \models C(||^{\sim}_{Ags})$, i.e. $M \models [\mathcal{X}]C(||^{\sim}_{Ags})$. □

Chapter 7

Conclusion

To briefly summarise what we have done in this thesis. We have contributed to the development of multi-agent modal logics (MAMLs) by making a first step towards refining the formal treatment of groups.

Most of MAMLs, from the well-known and developed [Fag+04; AHK02; Pau02] to the recent ones [Mog13; GJ22; Gor23], treat any set of agents as a group. In Chapter 2, we have presented philosophical arguments against such a view. Our main take-home message from this chapter is as follows: for a set of agents to be a group, they should share some motivational states; further, their actions should be guided by these states. There are different philosophical positions as to what the shared motivational states are and how they should be explained. Methodological individualists argue that the shared motivational states are reducible to the aggregation of members' individual motivational states. Methodological holists, on the contrary, argue that such shared states are irreducible, although they do not presuppose the autonomous existence of groups over and above their members. In Sections 2.1 and 2.2, we have presented the representative theories of both camps. Specifically, in Section 2.1, we have given our own take on Bratman's theory of shared intention, arguing for the need for initiators. A group's shared intention may exist in the form of participatory intentions of the group's members, if at least some of the group members unconditionally intend to do their part of the joint action, even if others will not join. In Section 2.2, we showed how Michael Bacharach's understanding of team reasoning may be seen as a holistic theory of collective intention. Namely, intentions are sensitive to the decision problems in the context of which they are formulated. Collective decision problems cannot be reduced to the aggregation of individual decision problems, as some irreducible distinctions emerge in them.

In Chapter 3, we have looked into multi-agent epistemic logics, as well as multi-agent modal logics of agency, and shown that the weak, aggregative definition of groups brings both conceptual and technical problems. Namely, if we take any set of agents to be a group, we allow ourselves to ascribe common knowledge, joint abilities and actions to them. In the most innocent case, we will get pragmatically strange ascriptions of common knowledge to agents who might not even be aware of one another's existence: e.g., the reader of this text and the king of the Netherlands commonly know that $p \vee \neg p$. In the worst case, we get completely ungrounded collective responsibility attributions: whenever simultaneous actions of agents have caused some outcome, we can take these agents to be collectively responsible for that outcome. For instance, if the killer has murdered someone and the only potential witness happened not to pass by the crime scene, preventing the crime from being witnessed, both of them are collectively responsible for the covered-up murder. There is no difference between them and the organised group of a murderer and a lookout. Moreover, given that any set of agents is a group, whenever we have n agents, there are 2^n-many groups. This obstacle creates redundant metalogical and computational complexity. In case of group STIT logic, it even leads to undecidability and unaxiomatisibility.

In the following two chapters, we have developed formal tools to model shared motivational states, namely, two logics of collective intention. The formalisms differ in their philosophical assumptions: the former is individualistic, while the latter is holistic. We have suspended our philosophical judgment on the individualism versus holism debate. In Chapter 4, we have first developed a single-agent logic of conditional intention, then extended it to a multi-agent case, resulting in a formal language $\mathcal{L}^I_{CL_\alpha}$. In $\mathcal{L}^I_{CL_\alpha}$, we have formalised the view, according to which a shared intention may be reduced to an aggregation of group members' interdependent intentions. We have demonstrated this view to be consistent, as well as normatively, ontologically and conceptually individualistic. To obtain a theory of a rational shared intention, it suffices to postulate individuals who have conditional intentions that obey certain rationality norms. Chapter 5 presents a holistic version of the logic of collective intention. In this formalism, agents' intentions exist in the context of their decision problems. Some of them share collective decision problems that are irreducible to fusions of individual problems. Besides being suitable for formalising the methodologically holistic view on collective intentionality, the formalism we obtained contributes to solving the side-effects problem: we can distinguish between expected side-effects and intended consequences of agents' intentional actions.

In Chapter 6, we have shown how the introduction of shared motivational states brings conceptual and technical advantages to MAMLs. We have done so for the case of group STIT logic, as it is one of the most expressive formalisms to reason about actions and abilities in multi-agent systems. Namely, we have explicitly introduced shared motivational states of agents in the form of joint, agreed-upon

plans. Consequently, we ascribed group agency only to sets of agents who share some plan and act upon it. We demonstrated how this distinction allows for better modelling of collective responsibility attribution. Then, we examined fragments of group STIT logic, as well as semantic generalisations, based on our restriction of what groups are. We have obtained finite axiomatisability and decision procedures for the satisfiability problem for the corresponding fragments/semantic variants. Finally, we have defined necessary and sufficient conditions on communication channels that group members should have to avoid coordination failures when acting upon their plans.

Of course, the work we have done in this thesis is by no means the ultimate and final solution to the problems we have posed. Till the end of this section, we point out some potential directions for future work.

Firstly, we have made strong idealisations about agents' epistemic status. E.g., in Section 5.2, we have assumed that group members commonly know what their collective decision problem is and what they collectively intend. In Section 6.2, a similar assumption was implicitly made about joint, agreed-upon plans: the plans are commonly known among agents who have adopted them. Such assumptions are plausible when we reason about relatively small groups with simple common goals: singing duos, dancing partners, small friend groups, where everyone knows one another. In the case of larger groups, common knowledge may be unattainable. For instance, Dutch citizens form a social group. Obviously, not all Dutch citizens are acquainted with one another, so they cannot establish common knowledge among them. Let us abstract away from this issue and, for the sake of argument, assume that they actually can obtain common knowledge, e.g., via mass media. Is there a commonly known shared intention among all Dutch citizens? One could arguably say that a constitution expresses what the nation collectively intends to achieve and maintain.[1] By the constitution, the Kingdom of the Netherlands is a constitutional monarchy. Does it mean that a Dutch citizen cannot be clueless about the constitution or support republicanism? These questions indicate that our assumptions should be relaxed. We need to explore how we can identify "loose" groups that "lack a codified organisational structure, and where the communication channels between group members are either unreliable or not completely open" [RS19, p.4523]. Consequently, we may need to look into weaker forms of groups' informational states: not only common knowledge and belief, but, e.g., pooled knowledge, virtual belief, majority judgment.

Secondly, we have taken groups and their motivational states to be static. In reality, groups fall apart, and new coalitions are formed. Joint plans can be adopted, revised, and abandoned altogether. This is true not only for collective but also for individual (conditional) intentions. For instance, it seems rational to

[1]Similar arguments were made about, e.g., the European Union, in terms of joint commitment as a shared motivational state, in [Gil13].

revise a conditional intention after getting to know whether the condition holds. Adding dynamic extensions to the formalisms developed throughout the thesis may be an interesting endeavour.

Thirdly, we have gained some positive metalogical results only for fragments of group STIT logic, while there are similar formalisms that are more computationally motivated. So it will be interesting to test how analogous fragments of Strategy Logic [Mog13; CHP10], Alternating-Time Temporal Logic [AHK02] and variations thereof behave in terms of their axiomatisation and decidability/computational complexity.

Finally, as we have briefly mentioned, we are not completely alone in our attempt to introduce a philosophically informed definition of groups into MAMLs. The work on logic of social networks[Bal+19; CH15; SLG13; GPS22; PS17] and intensional groups [BS23; BCR21; HS23] appear to be highly relevant to our concerns. In the spirit of our argument about initiators in Section 2.1, this paragraph may be regarded as a disclosure of author's unconditional intention to do his part in what might become a joint activity to bring these lines of research together.

Bibliography

[ÅA12] Thomas Ågotnes and Natasha Alechina. "Epistemic coalition logic: completeness and complexity". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 2012, pages 1099–1106 (cited on page 57).

[AB22] Aldo Iván Ramírez Abarca and Jan Broersen. "A Stit Logic of Responsibility." In: *AAMAS*. 2022, pages 1717–1719 (cited on pages 53, 61).

[Abr+16] Samson Abramsky, Juha Kontinen, Jouko Vaananen, and Heribert Vollmer. *Dependence logic: theory and applications*. Birkhauser, 2016 (cited on page 155).

[AHK02] Rajeev Alur, Thomas A Henzinger, and Orna Kupferman. "Alternating-time temporal logic". In: *Journal of the ACM (JACM)* 49.5 (2002), pages 672–713 (cited on pages 3, 31, 32, 225, 228).

[ÅHW11] Thomas Ågotnes, Wiebe van der Hoek, and Michael Wooldridge. "On the logic of preference and judgment aggregation". In: *Autonomous Agents and Multi-Agent Systems* 22.1 (2011), pages 4–30 (cited on pages 3, 32).

[Alb02] Luca Alberucci. "The modal μ-calculus and the logic of common knowledge". PhD thesis. PhD thesis, Institut für Informatik und angewandte Mathematik, Universität Bern, 2002 (cited on page 36).

[Alo22] Maria Aloni. "Logic and conversation: the case of free choice". In: *Semantics and Pragmatics* 15 (2022), pages 5–1 (cited on page 73).

[ANB98] Hajnal Andreka, Istvan Nemeti, and Johan van Benthem. "Modal languages and bounded fragments of predicate logic". In: *Journal of philosophical logic* 27.3 (1998), pages 217–274 (cited on page 155).

[Ari13] Aristotle. *Politics*. University of Chicago Press, 2013 (cited on page 7).

[Arr63] Kenneth J. Arrow. *Social Choice and Individual Values*. 2nd edition. Yale University Press, 1963 (cited on pages 9, 130).

[Aum76] Robert J Aumann. "Agreeing to Disagree". In: *The Annals of Statistics* 4.6 (1976), pages 1236–1239 (cited on pages 10, 36, 211).

[Aun66] Bruce Aune. "Intention and foresight". In: *The Journal of Philosophy* 63.20 (1966), pages 652–654 (cited on page 116).

[Bac06] Michael Bacharach. *Beyond Individual Choice: Teams and Frames in Game Theory*. Edited by Natalie Gold and Robert Sugden. Princeton University Press, 2006 (cited on pages 9, 10, 21–25, 27, 28, 205, 207, 209).

[Bal+19] Alexandru Baltag, Zoé Christoff, Rasmus K Rendsvig, and Sonja Smets. "Dynamic epistemic logics of diffusion and prediction in social networks". In: *Studia Logica* 107.3 (2019), pages 489–531 (cited on page 228).

[Bar+10] Nicholas Bardsley, Judith Mehta, Chris Starmer, and Robert Sugden. "Explaining focal points: Cognitive hierarchy theory versus team reasoning". In: *The Economic Journal* 120.543 (2010), pages 40–79 (cited on page 23).

[Bar25] Stefano Baratella. "On the weak completeness of a fragment of linear temporal logic". In: *Journal of Logic and Computation* 35.5 (2025) (cited on pages 179, 182).

[Baz13] Saba Bazargan. "Complicitous liability in war". In: *Philosophical Studies* 165.1 (2013), pages 177–195 (cited on pages 54, 159, 162).

[BBS18] Alexandru Baltag, Rachel Boddy, and Sonja Smets. "Group knowledge in interrogative epistemology". In: *Jaakko Hintikka on knowledge and game-theoretical semantics* (2018), pages 131–164 (cited on pages 114, 118).

[BBW06] Patrick Blackburn, Johan van Benthem, and Frank Wolter. *Handbook of modal logic*. Elsevier, 2006 (cited on pages 123, 172, 181).

[BCR21] Marta Bílková, Zoé Christoff, and Olivier Roy. "Revisiting epistemic logic with names". In: *arXiv preprint arXiv:2106.11493* (2021) (cited on page 228).

[BCS20] Alexandru Baltag, Ilaria Canavotto, and Sonja Smets. "Causal agency and responsibility: a refinement of STIT logic". In: *Logic in high definition: Trends in logical semantics*. Springer, 2020, pages 149–176 (cited on page 53).

[Ben12] Martin Mose Bentzen. "Stit, Iit, and deontic logic for action types". PhD thesis. University of Amsterdam, 2012 (cited on page 160).

[Ben13] Johan van Benthem. "CRS and Guarded Logics: A Fruitful Contact". In: *Cylindric-like Algebras and Algebraic Logic*. Springer, 2013, pages 273–301 (cited on page 155).

[Ben14] Johan van Benthem. "Modal correspondence theory". PhD thesis. University of Amsterdam, 2014 (cited on page 178).

[Ber18] Francesco Berto. "Aboutness in imagination". In: *Philosophical Studies* 175.8 (2018), pages 1871–1886 (cited on page 114).

[Ber22] Francesco Berto. *Topics of thought: The logic of knowledge, belief, imagination*. Oxford University Press, 2022 (cited on pages 115, 116, 123).

[BG23] Bob Beddor and Simon Goldstein. "A question-sensitive theory of intention". In: *The Philosophical Quarterly* 73.2 (2023), pages 346–378 (cited on pages 61, 64, 90, 114, 116, 118, 119, 121).

[BHT06] Jan Broersen, Andreas Herzig, and Nicolas Troquard. "From coalition logic to STIT". In: *Electronic Notes in Theoretical Computer Science* 157.4 (2006), pages 23–35 (cited on pages 40, 47, 151).

[BK88] Jon Barwise and H. Jerome Keisler. "Three Views of Common Knowledge". In: *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*. Edited by Moshe Y. Vardi. 1988, pages 365–379 (cited on page 36).

[BL18] Joseph Boudou and Emiliano Lorini. "Concurrent game structures for temporal STIT logic". In: *17th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*. 2018, pages 381–389 (cited on page 49).

[BM12] Johan van Benthem and Ştefan Minică. "Toward a dynamic logic of questions". In: *Journal of Philosophical Logic* 41 (2012), pages 633–669 (cited on page 118).

[BMS23] Alexandru Baltag, Lawrence S Moss, and Sławomir Solecki. "Logics for epistemic actions: completeness, decidability, expressivity". In: volume 1. 2. MDPI, 2023, pages 97–147 (cited on page 210).

[BPX01] Nuel Belnap, Michael Perloff, and Ming Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, 2001 (cited on pages 31, 32, 46, 49).

[Bra13] Michael E Bratman. *Shared agency: A planning theory of acting together*. Oxford, UK: Oxford University Press, 2013 (cited on pages 9–11, 13, 15, 16, 104, 107).

[Bra87] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Philosophy. Cambridge, MA: Harvard University Press, 1987. ISBN: 0674458184 (cited on pages 61, 64, 65, 68, 89, 90, 114, 157).

[Bra92] Michael E Bratman. "Shared cooperative activity". In: *The Philosophical Review* 101.2 (1992), pages 327–341 (cited on page 152).

[Bra99] Michael Bratman. *Faces of intention: Selected essays on intention and agency*. Cambridge University Press, 1999 (cited on page 17).

[Bro11a] Jan Broersen. "Deontic epistemic stit logic distinguishing modes of mens rea". In: *Journal of Applied Logic* 9.2 (2011), pages 137–152 (cited on page 61).

[Bro11b] Jan Broersen. "Making a start with the stit logic analysis of intentional action". In: *Journal of philosophical logic* 40 (2011), pages 499–530 (cited on page 47).

[BRV01] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, 2001 (cited on pages 31, 34, 75, 139, 140, 146, 173, 179, 199, 202, 203).

[BS06] Alexandru Baltag and Sonja Smets. "Conditional doxastic models: A qualitative approach to dynamic belief revision". In: *Electronic notes in theoretical computer science* 165 (2006), pages 5–21 (cited on page 65).

[BS23] Marta Bílková and Igor Sedlar. "Epistemic logics of structured intensional groups". In: *arXiv preprint arXiv:2307.05056* (2023) (cited on pages 34, 228).

[Cam82] Kenneth Campbell. "Conditional intention". In: *Legal Studies* 2.1 (1982), pages 77–97 (cited on page 61).

[CGM25] Davide Catta, Rustam Galimullin, and Aniello Murano. "First-order coalition logic". In: *arXiv preprint arXiv:2505.06960* (2025) (cited on page 111).

[CH15] Zoé Christoff and Jens Ulrik Hansen. "A logic for diffusion in social networks". In: *Journal of Applied Logic* 13.1 (2015), pages 48–77 (cited on page 228).

[Che75] Brian F Chellas. "Basic conditional logic". In: *Journal of philosophical logic* (1975), pages 133–153 (cited on pages 65, 69, 70).

[Che80] Brian F Chellas. *Modal logic: an introduction*. Cambridge University Press, 1980 (cited on page 34).

[Chi64] Roderick M Chisholm. "Human freedom and the self". In: *Free Will*. Blackwell, 1964 (cited on page 117).

[CHP10] Krishnendu Chatterjee, Thomas A Henzinger, and Nir Piterman. "Strategy logic". In: *Information and Computation* 208.6 (2010), pages 677–693 (cited on page 228).

[CL90a] Philip R Cohen and Hector J Levesque. "Intention is choice with commitment". In: *Artificial intelligence* 42.2-3 (1990), pages 213–261 (cited on pages 60, 61, 64, 80, 81, 87, 92, 114, 118).

[CL90b] Philip R Cohen and Hector J Levesque. "Persistence, intention, and commitment". In: *Reasoning about actions and plans* (1990), pages 297–340 (cited on page 61).

[CL99] Xiaoping Chen and Guiquan Liu. "A logic of intention". In: *Proceedings of the 16th international joint conference on Artifical intelligence-Volume 1*. 1999, pages 172–177 (cited on pages 114, 116, 118).

[CM09] Roberto Ciuni and Rosja Mastop. "Attributing distributed responsibility in stit logic". In: *International Workshop on Logic, Rationality and Interaction*. Springer. 2009, pages 66–75 (cited on page 53).

[Dav01] Donald Davidson. *Essays on Actions and Events: Philosophical Essays Volume 1*. Oxford, GB: Clarendon Press, 2001 (cited on page 8).

[DK25] Hans van Ditmarsch and Roman Kuznets. "Wanted dead or alive: Epistemic logic for impure simplicial complexes". In: *Journal of Logic and Computation* 35.6 (2025), exae055 (cited on page 40).

[DTP21] Hein Duijf, Allard Tamminga, and Frederik van de Putte. "An impossibility result on methodological individualism". In: *Philosophical Studies* 178.12 (2021), pages 4165–4185 (cited on page 9).

[Dui+21] Hein Duijf, Jan Broersen, Alexandra Kuncová, and Aldo Iván Ramírez Abarca. "Doing without action types". In: *The Review of Symbolic Logic* 14.2 (2021), pages 380–410 (cited on page 157).

[Dui18] Hein Duijf. "Responsibility voids and cooperation". In: *Philosophy of the social sciences* 48.4 (2018), pages 434–460 (cited on page 207).

[Dur95] Emile Durkheim. *The Rules of Sociological Method*. New York: The Free Press, 1938[1895] (cited on page 9).

[DV02] Barbara Dunin-Keplicz and Rineke Verbrugge. "Collective intentions". In: *Fundamenta Informaticae* 51.3 (2002), pages 271–295 (cited on page 32).

[EJ99] Ernest Allen Emerson and Charanjit Jutla. "The complexity of tree automata and logics of programs". In: *SIAM Journal on Computing* 29.1 (1999), pages 132–158 (cited on page 199).

[Eps14] Brian Epstein. "What is individualism in social ontology? Ontological individualism vs. anchor individualism". In: *Rethinking the Individualism-Holism Debate: Essays in the Philosophy of Social Science*. Springer, 2014, pages 17–38 (cited on page 8).

[Eps16] Brian Epstein. "Social ontology". In: *The Routledge companion to philosophy of social science*. Routledge, 2016, pages 260–273 (cited on page 7).

[Fag+04] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004 (cited on pages 3, 31, 32, 35, 225).

[Fer09] Luca Ferrero. "Conditional intentions". In: *Noûs* 43.4 (2009), pages 700–741 (cited on pages 17, 61–64).

[Fer15] Luca Ferrero. "Ludwig on conditional intentions". In: *Methode: Analytic Perspectives* 6 (2015) (cited on page 63).

[FH87] Ronald Fagin and Joseph Y Halpern. "Belief, awareness, and limited reasoning". In: *Artificial intelligence* 34.1 (1987), pages 39–76 (cited on pages 40, 90, 138).

[FHV92] Ronald Fagin, Joseph Y Halpern, and Moshe Y Vardi. "What can machines know? On the properties of knowledge in distributed systems". In: *Journal of the ACM (JACM)* 39.2 (1992), pages 328–376 (cited on page 166).

[Fin88] John Finnis. *Intention and side-effects*. Toronto,Canada: Faculty of Law, University of Toronto, 1988 (cited on page 89).

[For84] James William Forrester. "Gentle murder, or the adverbial samaritan". In: *The Journal of Philosophy* 81.4 (1984), pages 193–197 (cited on page 89).

[Fra18] Harry Frankfurt. "Alternate possibilities and moral responsibility". In: *Moral responsibility and alternative possibilities*. Routledge, 2018, pages 17–25 (cited on page 53).

[Gab+80] Dov Gabbay, Amir Pnueli, Saharon Shelah, and Jonathan Stavi. "On the temporal analysis of fairness". In: *Proceedings of the 7th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. 1980, pages 163–173 (cited on pages 179, 182).

[Gau+15] Benoit Gaudou, Andreas Herzig, Dominique Longin, and Emiliano Lorini. "On modal logics of group belief". In: *The Cognitive Foundations of Group Attitudes and Social Interaction*. Edited by Andreas Herzig. Volume 5. Springer, 2015, pages 75–106 (cited on page 34).

[GD06] Valentin Goranko and Govert van Drimmelen. "Complete axiomatization and decidability of alternating-time temporal logic". In: *Theoretical Computer Science* 353.1-3 (2006), pages 93–117 (cited on page 42).

[Gil13] Margaret Gilbert. *Joint Commitment: How We Make the Social World*. New York, NY: Oxford University Press, 2013 (cited on pages 9, 227).

[Gil92] Margaret Gilbert. *On social facts*. Princeton University Press, 1992 (cited on page 152).

[Gil99] Raanan Gillon. "Foreseeing is not necessarily the same as intending". In: *BMJ: British Medical Journal* 318.7196 (1999), page 1431 (cited on page 90).

[Gio19] Alessandro Giordani. "Axiomatizing the Logic of Imagination". In: *Studia Logica* 107.4 (2019), pages 639–657 (cited on pages 128, 138).

[GJ22] Valentin Goranko and Fengkui Ju. "A logic for conditional local strategic reasoning". In: *Journal of Logic, Language and Information* 31.2 (2022), pages 167–188 (cited on pages 95, 225).

[GJT13] Valentin Goranko, Wojciech Jamroga, and Paolo Turrini. "Strategic games and truly playable effectivity functions". In: *Autonomous Agents and Multi-Agent Systems* 26.2 (2013), pages 288–314 (cited on page 42).

[Gor23] Valentin Goranko. "Logics for strategic reasoning of socially interacting rational agents: an overview and perspectives". In: *Logics* 1.1 (2023), pages 4–35 (cited on page 225).

[GPS22] Rustam Galimullin, Mina Young Pedersen, and Marija Slavkovik. "Logic of visibility in social networks". In: *International Workshop on Logic, Language, Information, and Computation*. Springer. 2022, pages 190–206 (cited on page 228).

[Grä99] Erich Grädel. "On the restraining power of guards". In: *The Journal of Symbolic Logic* 64.4 (1999), pages 1719–1742 (cited on page 155).

[Gro+04] Davide Grossi, Frank Dignum, Lamber MM Royakkers, and John-Jules Meyer. "Collective obligations and agents: who gets the blame?" In: *International Workshop on Deontic Logic in Computer Science*. Springer. 2004, pages 129–145 (cited on page 3).

[GS07] Natalie Gold and Robert Sugden. "Collective intentions and team agency". In: *The Journal of Philosophy* 104.3 (2007), pages 109–137 (cited on pages 10, 26, 27).

[Haw18] Peter Hawke. "Theories of aboutness". In: *Australasian Journal of Philosophy* 96.4 (2018), pages 697–723 (cited on page 119).

[HB95] John Horty and Nuel Belnap. "The deliberative stit: A study of action, omission, ability, and obligation". In: *Journal of philosophical logic* 24.6 (1995), pages 583–644 (cited on pages 48, 53).

[HIH22] Sarah Hiller, Jonas Israel, and Jobst Heitzig. "An axiomatic approach to formalized responsibility ascription". In: *International Conference on Principles and Practice of Multi-Agent Systems*. Springer. 2022, pages 435–457 (cited on page 130).

[Hin62] Jaakko Hintikka. "Knowledge and belief: An introduction to the logic of the two notions". In: *Studia Logica* 16 (1962) (cited on pages 31, 35).

[HM90] Joseph Y Halpern and Yoram Moses. "Knowledge and common knowledge in a distributed environment". In: *Journal of the ACM (JACM)* 37.3 (1990), pages 549–587 (cited on page 37).

[HMT71] Leon Henkin, J. Donald Monk, and Alfred Tarski. *Cylindric Algebras*. Volume 1. Amsterdam: North-Holland, 1971 (cited on pages 154, 155).

[HÖB20] Peter Hawke, Aybüke Özgun, and Francesco Berto. "The fundamental problem of logical omniscience". In: *Journal of Philosophical Logic* 49.4 (2020), pages 727–766 (cited on pages 114–116).

[Hoe22] Daniel Hoek. "Questions in Action". In: *Journal of Philosophy* 119.3 (2022), pages 113–143 (cited on pages 114, 118).

[Hor01] John F Horty. *Agency and deontic logic*. Oxford University Press, 2001 (cited on pages 40, 46, 53, 133).

[Hor94] John Horty. "Moral dilemmas and nonmonotonic logic". In: *Journal of philosophical logic* 23 (1994), pages 35–65 (cited on page 67).

[HS08] Andreas Herzig and François Schwarzentruber. "Properties of logics of individual and group agency." In: *Advances in modal logic* 7 (2008), pages 133–149 (cited on pages 47, 58, 155, 156).

[HS23] Merlin Humml and Lutz Schröder. "Common Knowledge of abstract groups". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 37. 5. 2023, pages 6434–6441 (cited on page 228).

[HS88a] J. C. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, 1988 (cited on pages 207, 209).

[HS88b] John C Harsanyi and Reinhard Selten. "A general theory of equilibrium selection in games". In: *MIT Press Books* 1 (1988) (cited on page 205).

[Hum23] Human Rights Watch. *Russia: Supreme Court Bans "LGBT Movement" as "Extremist"*. Human Rights Watch. Nov. 2023. https://www.hrw.org/news/2023/11/30/russia-supreme-court-bans-lgbt-movement-extremist (visited on 02/12/2026) (cited on page 3).

[HZ79] Aanund Hylland and Richard Zeckhauser. "The impossibility of Bayesian group decision making with separate aggregation of beliefs and values". In: *Econometrica: Journal of the Econometric Society* (1979), pages 1321–1336 (cited on page 130).

[KDB26] Daniil Khaitovich, Hein Duijf, and Jan Broersen. *Solving coordination games with minimal communication*. Preprint. 2026. https://dkhaito.github.io/coordination.pdf (cited on pages 151, 204).

[Kelng] Mikayla Kelley. "A Control Theory of Action". In: *Australasian Journal of Philosophy* (forthcoming) (cited on page 117).

[Kha26a] Daniil Khaitovich. *Logic of Conditional Intention*. Under review. 2026. https://dkhaito.github.io/logic_of_conditional_intention.pdf (cited on pages 59, 60).

[Kha26b] Daniil Khaitovich. *Plan-restricted group STIT logic*. Under review. 2026. https://dkhaito.github.io/pSTIT.pdf (cited on pages 151, 156).

[KLM90] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. "Nonmonotonic reasoning, preferential models and cumulative logics". In: *Artificial intelligence* 44.1-2 (1990), pages 167–207 (cited on page 75).

[Kno03] Joshua Knobe. "Intentional action and side effects in ordinary language". In: *Analysis* 63.3 (2003), pages 190–194 (cited on page 89).

[Kno06] Joshua Knobe. "The concept of intentional action: A case study in the uses of folk psychology". In: *Philosophical studies* 130 (2006), pages 203–231 (cited on page 89).

[KÖ25] Daniil Khaitovich and Aybüke Özgun. "Hyperintensional Intention". In: *Proceedings of TARK 2025*. Edited by Adam Bjorndahl. EPTCS. To appear. Waterloo, Australia: Open Publishing Association, 2025, pages 44–60 (cited on pages 89, 90, 113).

[KP93] Kurt Konolige and Martha E Pollack. "A representationalist theory of intention". In: *IJCAI*. Volume 93. 1993, pages 390–395 (cited on pages 64, 114–116, 118).

[Kro24] Wessel Kroon. "Knowledge as Issue-Relevant Information". Available at https://eprints.illc.uva.nl/id/eprint/2316/1/MoL-2024-06.text.pdf. Master's thesis. University of Amsterdam, 2024 (cited on pages 114, 118).

[KT84] Daniel Kahneman and Amos Tversky. "Choices, values, and frames." In: *American psychologist* 39.4 (1984), page 341 (cited on page 117).

[Kun+25] Alexandra Kuncová, Jan Broersen, Hein Duijf, and Aldo Iván Ramírez Abarca. "Ability and knowledge: from epistemic transition systems to labelled stit models". In: *Autonomous Agents and Multi-Agent Systems* 39.1 (2025), page 2 (cited on page 57).

[Kur+03] Agi Kurucz, Frank Wolter, Michael Zakharyaschev, and Dov M Gabbay. *Many-dimensional modal logics: Theory and applications*. Elsevier, 2003 (cited on page 154).

[Kur13] Agi Kurucz. "Representable cylindric algebras and many-dimensional modal logics". In: *Cylindric-like Algebras and Algebraic Logic*. Springer, 2013, pages 185–203 (cited on page 154).

[Kut00] Christopher Kutz. *Complicity: Ethics and Law for a Collective Age*. New York: Cambridge University Press, 2000 (cited on pages 54, 162).

[Lac21] Jennifer Lackey. *The epistemology of groups*. Oxford University Press, 2021 (cited on page 9).

[Leh84] Daniel Lehmann. "Knowledge, common knowledge and related puzzles (extended summary)". In: *Proceedings of the third annual ACM symposium on Principles of distributed computing*. 1984, pages 62–67 (cited on page 35).

[Lew08] David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008 (cited on pages 10, 36).

[Lew88a] David Lewis. "Relevant implication". In: *Theoria* 54.3 (1988), pages 161–174 (cited on page 119).

[Lew88b] David Lewis. "Statements partly about observation". In: *Philosophical papers* 17.1 (1988), pages 1–31 (cited on page 119).

[LH08] Emiliano Lorini and Andreas Herzig. "A logic of intention and attempt". In: *Synthese* 163.1 (2008), pages 45–77 (cited on page 60).

[LHM96] Bernd van Linder, Wiebe van der Hoek, and J -J Ch Meyer. "Formalising motivational attitudes of agents: On preferences, goals and commitments". In: *Intelligent Agents II Agent Theories, Architectures, and Languages: IJCAI'95 Workshop (ATAL) Montréal, Canada, August 19–20, 1995 Proceedings 2*. Springer. 1996, pages 17–32 (cited on page 90).

[LLM14] Emiliano Lorini, Dominique Longin, and Eunate Mayor. "A logical analysis of responsibility attribution: emotions, individuals and collectives". In: *Journal of Logic and Computation* 24.6 (2014), pages 1313–1339 (cited on page 61).

[LM94] Luc Lismont and Philippe Mongin. "On the logic of common belief and common knowledge". In: *Theory and Decision* 37.1 (1994), pages 75–106 (cited on page 36).

[Lor13] Emiliano Lorini. "Temporal logic and its application to normative reasoning". In: *Journal of Applied Non-Classical Logics* 23.4 (2013), pages 372–399 (cited on pages 34, 50, 53, 162, 174, 191, 198, 199).

[LP02] Christian List and Philip Pettit. "Aggregating sets of judgments: An impossibility result". In: *Economics & Philosophy* 18.1 (2002), pages 89–110 (cited on pages 9, 28, 130).

[LP11] Christian List and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Edited by Philip Pettit. New York: Oxford University Press, 2011 (cited on page 7).

[Lud15] Kirk Ludwig. "What are conditional intentions?" In: *Methode: Analytic Perspectives* 4(6) (2015), pages 30–60 (cited on pages 60, 61, 63, 64, 73, 87).

[Man21] Lucas Manfredi. "SEC says GameStop meme stock frenzy fueled by investor sentiment". In: *Fox Business* (Oct. 2021). `https://www.foxbusiness.com/markets/sec-gamestop-investors-meme-stocks` (cited on page 3).

[Mil01] Seumas Miller. *Social action: A teleological account*. Cambridge University Press, 2001 (cited on page 104).

[Min11] Ştefan Minică. "Dynamic logic of questions". PhD thesis. University of Amsterdam, 2011 (cited on pages 114, 118).

[Mog13] Fabio Mogavero. "Reasoning about strategies". In: *Logics in Computer Science: A Study on Extensions of Temporal and Strategic Logics*. Springer, 2013, pages 85–116 (cited on pages 111, 225, 228).

[Mon76] J. Donald Monk. "Cylindric Algebras". In: *Mathematical Logic*. New York, NY: Springer New York, 1976, pages 219–229. ISBN: 978-1-4684-9452-5. `https://doi.org/10.1007/978-1-4684-9452-5_13`. `https://doi.org/10.1007/978-1-4684-9452-5_13` (cited on page 154).

[Nas50] John F Nash Jr. "Equilibrium points in n-person games". In: *Proceedings of the national academy of sciences* 36.1 (1950), pages 48–49 (cited on page 204).

[Nis79] Hirokazu Nishimura. "Is the semantics of branching structures adequate for chronological modal logics?" In: *Journal of Philosophical Logic* (1979), pages 469–475 (cited on pages 49, 50).

[NN13] Pavel Naumov and Britanny Nicholls. "On interchangeability of Nash equilibria in multi-player strategic games". In: *Synthese* 190.Suppl 1 (2013), pages 57–78 (cited on pages 214, 219).

[Nol14] Daniel Nolan. "Hyperintensional metaphysics". In: *Philosophical Studies* 171 (2014), pages 149–160 (cited on page 117).

[ÖB21] Aybüke Özgun and Francesco Berto. "Dynamic hyperintensional belief revision". In: *The Review of Symbolic Logic* 14.3 (2021), pages 766–811 (cited on pages 115, 128).

[Opp20] Karl-Dieter Opp. "Rational Choice Theory and Methodological Individualism". In: *The Cambridge Handbook of Social Theory*. Edited by Peter Kivisto. Cambridge University Press, 2020, pages 1–23 (cited on page 9).

[Owe16] David Owens. "Promises and conflicting obligations". In: *J. Ethics & Soc. Phil.* 11 (2016), page 1 (cited on page 67).

[Özg17] Aybüke Özgun. "Evidence in epistemic logic: a topological perspective". PhD thesis. University of Amsterdam, 2017 (cited on page 34).

[Pac17] Eric Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017 (cited on page 34).

[Pau02] Marc Pauly. "A modal logic for coalitional power in games". In: *Journal of logic and computation* 12.1 (2002), pages 149–166 (cited on pages 31, 32, 40, 42, 43, 225).

[Pay14] Gillman Payette. "Decidability of an xstit logic". In: *Studia Logica* 102.3 (2014), pages 577–607 (cited on page 174).

[Pea18] Wendy Pearlman. "Moral identity and protest cascades in Syria". In: *British Journal of Political Science* 48.4 (2018), pages 877–901 (cited on page 20).

[Pet95] F. V. Hasle Peter Øhrstrøm. "Indeterministic tense logic". In: *Temporal Logic: From Ancient Ideas to Artificial Intelligence*. Dordrecht: Springer Netherlands, 1995, pages 257–269. ISBN: 978-0-585-37463-5. https://doi.org/10.1007/978-0-585-37463-5_24 (cited on page 49).

[Pla07] Jan Plaza. "Logics of public communications". In: *Synthese* 158.2 (2007), pages 165–179 (cited on page 15).

[PMS07] Rohit Parikh, Lawrence S Moss, and Chris Steinsvold. "Topology and epistemic logic". In: *Handbook of spatial logics*. Springer, 2007, pages 299–341 (cited on page 34).

[Pop02] Karl Popper. *The Poverty of Historicism*. Routledge, 2002 (cited on page 9).

[Pri67] Arthur N. Prior. *Past, Present and Future*. Oxford: Clarendon Press, 1967 (cited on page 60).

[PS17] Truls Pedersen and Marija Slavkovik. "Formal models of conflicting social influence". In: *International Conference on Principles and Practice of Multi-Agent Systems*. Springer. 2017, pages 349–365 (cited on page 228).

[PS21] Matteo Plebani and Giuseppe Spolaore. "Subject Matter: A Modest Proposal". In: *Philosophical Quarterly* 71.3 (2021), pages 605–622 (cited on page 119).

[PS96] Henry Prakken and Marek Sergot. "Contrary-to-duty obligations". In: *Studia Logica* 57 (1996), pages 91–115 (cited on page 67).

[Qui75] Anthony Quinton. "The presidential address: Social objects". In: *Proceedings of the Aristotelian society*. Volume 76. JSTOR. 1975, pages 1–viii (cited on page 8).

[Rac24] Matthew Rachar. "Conditional intentions and shared agency". In: *Noûs* 58.1 (2024), pages 271–288 (cited on pages 17, 18).

[Ram23] Aldo Iván Ramírez Abarca. "From ATL to Stit". In: *arXiv preprint arXiv:2302.07332* (2023) (cited on page 151).

[Ree97] C.D.C. (trans.) Reeve. *Plato, Cratylus: translated with introduction and notes*. Indianapolis and Cambridge: Hackett, 1997 (cited on page 7).

[Rey02] Mark Reynolds. "Axioms for branching time". In: *Journal of Logic and Computation* 12.4 (2002), pages 679–697 (cited on pages 47, 49, 50, 152).

[RG97] Anand S Rao and Michael P Georgeff. "Modeling rational agents within a BDI-architecture". In: *Readings in agents* (1997), pages 317–328 (cited on pages 64, 90).

[Roy08] Olivier Roy. *Thinking before acting: intentions, logic, rational choice*. University of Amsterdam, 2008 (cited on pages 114, 116, 118).

[RS19] Olivier Roy and Anne Schwenkenbecher. "Shared Intentions, Loose Groups and Pooled Knowledge". In: *Synthese* 5 (2019), pages 4523–4541. `https://doi.org/10.1007/s11229-019-02355-x` (cited on pages 130, 227).

[Sah75] Henrik Sahlqvist. "Completeness and correspondence in the first and second order semantics for modal logic". In: *Studies in Logic and the Foundations of Mathematics*. Volume 82. Elsevier, 1975, pages 110–143 (cited on pages 166, 178, 201).

[Sal95] David Sally. "Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992". In: *Rationality and society* 7.1 (1995), pages 58–92 (cited on page 22).

[Say32] Francis Bowes Sayre. "Mens rea". In: *Harvard Law Review* 45.6 (1932), pages 974–1026 (cited on page 61).

[Sch11] Benjamin Schnieder. "A logic for 'because'". In: *The Review of Symbolic Logic* 4.3 (2011), pages 445–465 (cited on page 117).

[Sch12] François Schwarzentruber. "Complexity results of STIT fragments". In: *Studia logica* 100.5 (2012), pages 1001–1045 (cited on pages 34, 58, 162, 174, 175).

[Sch58] Thomas Schelling. "The strategy of conflict. Prospectus for a reorientation of game theory". In: *Journal of Conflict Resolution* 2.3 (1958), pages 203–264 (cited on pages 205, 207, 209).

[Sea95] John R. Searle. *The Construction of Social Reality*. New York: Free Press, 1995, page 241. ISBN: 9780029280454 (cited on pages 9, 152).

[Ser21] Marek Sergot. "Some forms of collectively bringing about or 'seeing to it that'". In: *Journal of Philosophical Logic* 50.2 (2021), pages 249–283 (cited on pages 53, 56, 159, 160).

[SG00] Jason Stanley and Zoltán Gendler Szabó. "On quantifier domain restriction". In: *Mind & Language* 15.2-3 (2000), pages 219–261 (cited on page 155).

[She03] Paul Sheehy. "Social groups, explanation and ontological holism". In: *Philosophical Papers* 32.2 (2003), pages 193–224 (cited on page 8).

[Sie21] Maximilian Siemers. "Hyperintensional Logics for Evidence, Knowledge and Belief". Master's thesis. ILLC, University of Amsterdam, 2021 (cited on page 138).

[SLG13] Jeremy Seligman, Fenrong Liu, and Patrick Girard. "Facebook and the epistemic logic of friendship". In: *arXiv preprint arXiv:1310.6440* (2013) (cited on page 228).

[SMK09] Krister Segerberg, John-Jules Meyer, and Marcus Kracht. "The logic of action". In: *Stanford Encyclopedia of Philosophy* (2009) (cited on page 46).

[Sta02] Robert Stalnaker. "Common ground". In: *Linguistics and philosophy* 25.5/6 (2002), pages 701–721 (cited on page 15).

[Sta84] Robert Stalnaker. *Inquiry*. MIT Press, 1984 (cited on pages 116, 118).

[Sta91] Robert Stalnaker. "The problem of logical omniscience, I". In: *Synthese* (1991), pages 425–440 (cited on page 114).

[Sug00] Robert Sugden. "Team preferences". In: *Economics & Philosophy* 16.2 (2000), pages 175–204 (cited on page 27).

[Sug03] Robert Sugden. "The Logic of Team Reasoning". In: *Philosophical Explorations* 6.3 (2003), pages 165–181 (cited on pages 27, 205, 218).

[Sve96] Steven Sverdlik. "Consistency among intentions and the 'Simple View'". In: *Canadian Journal of Philosophy* 26.4 (1996), pages 515–522 (cited on page 65).

[Tar41] Alfred Tarski. "On the calculus of relations". In: *The journal of symbolic logic* 6.3 (1941), pages 73–89 (cited on page 32).

[TD17] Allard Tamminga and Hein Duijf. "Collective obligations, group plans and individual actions". In: *Economics & Philosophy* 33.2 (2017), pages 187–214 (cited on pages 204, 205, 207–209, 217).

[TH20] Allard Tamminga and Frank Hindriks. "The irreducibility of collective obligations". In: *Philosophical Studies* 177 (2020), pages 1085–1109 (cited on pages 130, 205).

[Tho84] Richmond H Thomason. "Combinations of tense and modality". In: *Handbook of Philosophical Logic: Volume II: Extensions of Classical Logic*. Springer, 1984, pages 135–165 (cited on page 50).

[Tol02] Deborah Perron Tollefsen. "Collective intentionality and the social sciences". In: *Philosophy of the social sciences* 32.1 (2002), pages 25–50 (cited on pages 8, 9).

[Tor94] Leendert WN van der Torre. "Violated obligations in a defeasible deontic logic". In: *ECAI*. Volume 94. Citeseer. 1994, pages 371–375 (cited on page 67).

[Tuo05] Raimo Tuomela. "Two basic kinds of cooperation". In: *Logic, Thought and Action*. Springer, 2005, pages 79–107 (cited on page 152).

[Tuo13] Raimo Tuomela. *Social Ontology: Collective Intentionality and Group Agents*. New York, US: Oup Usa, 2013 (cited on page 9).

[Vak10] Dimiter Vakarelov. "Algorithmic Definability and Completeness in Modal Logic". In: *International Symposium on Foundations of Information and Knowledge Systems*. Springer. 2010, pages 6–8 (cited on page 178).

[Vak92] Dimiter Vakarelov. "A modal theory of arrows. Arrow logics I". In: *European Workshop on Logics in Artificial Intelligence*. Springer. 1992, pages 1–24 (cited on page 191).

[Var86] Moshe Y Vardi. "On epistemic logic and logical omniscience". In: *Theoretical aspects of reasoning about knowledge*. Elsevier. 1986, pages 293–305 (cited on pages 114, 116).

[Vel97] J David Velleman. "How to share an intention". In: *Philosophy and Phenomenological Research: A Quarterly Journal* (1997), pages 29–50 (cited on pages 12, 17, 18).

[Ven95] Yde Venema. "Cylindric modal logic". In: *The Journal of Symbolic Logic* 60.2 (1995), pages 591–623 (cited on pages 154, 155).

[WÅ13a] Yi Nicholas Wang and Thomas Ågotnes. "Multi-Agent Subset Space Logic." In: *IJCAI*. 2013, pages 1155–1161 (cited on page 34).

[WÅ13b] Yı N Wáng and Thomas Ågotnes. "Public announcement logic with distributed knowledge: expressivity, completeness and complexity". In: *Synthese* 190.Suppl 1 (2013), pages 135–162 (cited on page 32).

[WÅ20] Yı N Wáng and Thomas Ågotnes. "Simpler completeness proofs for modal logics with intersection". In: *International Workshop on Dynamic Logic*. Springer. 2020, pages 259–276 (cited on page 166).

[Wat52] J. W. N. Watkins. "The Principle of Methodological Individualism". In: *British Journal for the Philosophy of Science* 3.10 (1952), pages 186–189. https://doi.org/10.1093/bjps/iii.10.186 (cited on page 9).

[Web13] Max Weber. *Economy and Society*. Harvard University Press, 2013 (cited on page 9).

[Wri13] Richard W Wright. "The NESS account of natural causation: A response to criticisms". In: *Critical essays on "causation and responsibility"* (2013), pages 13–66 (cited on page 54).

[Yab14] Stephen Yablo. *Aboutness*. Princeton, NJ: Princeton University Press, 2014 (cited on page 130).

[Yaf04] Gideon Yaffe. "Conditional intent and mens rea". In: *Legal Theory* 10.4 (2004), pages 273–310 (cited on page 61).

[Yal18] Seth Yalcin. "Belief as question-sensitive". In: *Philosophy and Phenomenological Research* 97.1 (2018), pages 23–47 (cited on pages 114, 118).

[Zhu10] Jing Zhu. "On the principle of intention agglomeration". In: *Synthese* 175 (2010), pages 89–99 (cited on pages 65, 118).

Abstract

From arbitrary sets to social groups: group agency in multi-agent modal logic

We differentiate between arbitrary sets of agents and organised groups. The difference is especially consequential in our strategic and moral reasoning. We expect group members to act towards their common interest, and we tend to hold individuals morally (and even legally) accountable for the joint actions of the groups they are members of. The distinction between groups and arbitrary sets of agents is mostly neglected by modal logicians who formalise reasoning about multi-agent systems. In logics of several group notions – common knowledge, coalitional powers in games, and collective seeing-to-it-that – groups are idealised as nothing but sets of agents. Such an idealisation is conceptually questionable, and it adds unnecessary technical complexity to the formalisms. This thesis is dedicated to filling this lacuna by introducing a philosophically informed definition of group agency into multi-agent modal logics.

Chapter 2 is dedicated to the philosophical background. The major theories of group agency are presented, divided into two camps: methodological individualism versus holism. Representatives of each of the two camps – Michael Bratman's theory of shared agency and Michael Bacharach's theory of team reasoning – are addressed in detail. Chapter 3 contains the logical background. Some fundamental notions of multi-modal logics, their standard syntax and semantics, are introduced. It is also demonstrated how the idealised definition of groups as arbitrary sets of agents accounts for conceptual flaws in modal logics of common

knowledge (Section 3.1), coalitional powers in games (Section 3.2.1) and seeing-to-it-that (Section 3.2.2).

Chapters 4 and 5 contain two versions of the logic of collective intention, the central notion for the refined definition of group agency. Neither of the two theories is chosen to maintain ontological neutrality. In Chapter 4, Bratman's methodologically individualistic theory is formalised. Namely, it is demonstrated how agents can share the collective intention by having interdependent conditional intentions. For that purpose, a new logic of conditional intention is developed (Section 4.1), as well as its simplified multi-agent version (Section 4.2). In Chapter 5, the hyperintensional logic of collective intention is presented, where intentions are sensitive to the decision problems in the context of which they are formulated. Reflecting the conceptual grounds of the theory of team reasoning, a decision problem of a group (and hence, its collective intention) cannot be reduced to an aggregation of decision problems and intentions of the group's members.

Chapter 6 introduces the philosophically refined definition of group agency to the logic of group seeing-to-it-that (group STIT logic). The focus is on group STIT logic, as, on the one hand, this formalism is more expressive, and on the other hand, it suffers the most from the weak definition of groups, both conceptually and technically. The refinement of group STIT logic is presented, where uncoordinated actions of independent agents are distinguished from joint actions of groups that act upon agreed plans. It is demonstrated how such a distinction improves collective causal responsibility attributions and allows for formalising the subtler notion of complicity (Section 6.2). Further, it is demonstrated how generalised semantics (Section 6.3.1) and language fragments (Sections 6.3.2 and 6.3.3) based on the introduced restriction allow for obtaining finite axiomatisability and decidability of group STIT logic that is otherwise neither decidable nor axiomatisable. The problem of coordinating on an agreed plan via restricted communication is addressed (Section 6.4), which grounds our definition of a group plan as an arbitrary subset of members' possible joint actions.

Samenvatting

Van verzamelingen naar sociale groepen: groepsagentschap in multi-agent modale logica

Wij maken een onderscheid tussen willekeurige verzamelingen van agenten en sociale groepen. Dit verschil is met name van belang voor ons strategisch en moreel redeneren. Wij verwachten dat groepsleden handelen in het gemeenschappelijk belang, en wij zijn geneigd individuen moreel (en zelfs juridisch) verantwoordelijk te houden voor de gezamenlijke handelingen van de groepen waarvan zij lid zijn. Het onderscheid tussen groepen en willekeurige verzamelingen van agenten wordt grotendeels veronachtzaamd door modale logici die redeneren over het formaliseren van multi-agentsystemen. In logica's van verschillende groepsnoties – gemeenschappelijke kennis, coalitiemacht in spelen, en collectief 'seeing-to-it-that' – worden groepen geïdealiseerd als niet meer dan verzamelingen van agenten. Zo'n idealisering is conceptueel twijfelachtig en voegt onnodige technische complexiteit toe aan de formalismen. Dit proefschrift is gewijd aan het opvullen van deze lacune door een filosofisch onderbouwde definitie van groepsagentschap (group agency) te introduceren in modale logica's voor multi-agentsystemen.

Titles in the ILLC Dissertation Series:

ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction

ILLC DS-2021-14: **Álvaro Piedrafita**
On Span Programs and Quantum Algorithms

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions

ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes

ILLC DS-2022-05: **Lwenn Bussière-Carae**
No means No! Speech Acts in Conflict

ILLC DS-2022-06: **Emma Mojet**
Observing Disciplines: Data Practices In and Between Disciplines in the 19th and Early 20th Centuries

ILLC DS-2022-07: **Freek Gerrit Witteveen**
Quantum information theory and many-body physics

ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity

ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms

ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories

ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

ILLC DS-2023-05: **Dean McHugh**
Causation and Modality: Models and Meanings

ILLC DS-2023-06: **Jialiang Yan**
Monotonicity in Intensional Contexts: Weakening and: Pragmatic Effects under Modals and Attitudes

ILLC DS-2023-07: **Yiyan Wang**
Collective Agency: From Philosophical and Logical Perspectives

ILLC DS-2023-08: **Lei Li**
Games, Boards and Play: A Logical Perspective

ILLC DS-2023-09: **Simon Rey**
Variations on Participatory Budgeting

ILLC DS-2023-10: **Mario Giulianelli**
Neural Models of Language Use: Studies of Language Comprehension and Production in Context

ILLC DS-2023-11: **Guillermo Menéndez Turata**
Cyclic Proof Systems for Modal Fixpoint Logics

ILLC DS-2023-12: **Ned J.H. Wontner**
Views From a Peak: Generalisations and Descriptive Set Theory

ILLC DS-2024-01: **Jan Rooduijn**
Fragments and Frame Classes: Towards a Uniform Proof Theory for Modal Fixed Point Logics

ILLC DS-2024-02: **Bas Cornelissen**
Measuring musics: Notes on modes, motifs, and melodies

ILLC DS-2024-03: **Nicola De Cao**
Entity Centric Neural Models for Natural Language Processing

ILLC DS-2024-04: **Ece Takmaz**
Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective

ILLC DS-2024-05: **Fatemeh Seifan**
Coalgebraic fixpoint logic Expressivity and completeness result

ILLC DS-2024-06: **Jana Sotáková**
Isogenies and Cryptography

ILLC DS-2024-07: **Marco Degano**
Indefinites and their values

ILLC DS-2024-08: **Philip Verduyn Lunel**
Quantum Position Verification: Loss-tolerant Protocols and Fundamental Limits

ILLC DS-2024-09: **Rene Allerstorfer**
Position-based Quantum Cryptography: From Theory towards Practice

ILLC DS-2024-10: **Willem Feijen**
Fast, Right, or Best? Algorithms for Practical Optimization Problems

ILLC DS-2024-11: **Daira Pinto Prieto**
Combining Uncertain Evidence: Logic and Complexity

ILLC DS-2024-12: **Yanlin Chen**
On Quantum Algorithms and Limitations for Convex Optimization and Lattice Problems

ILLC DS-2024-13: **Jaap Jumelet**
Finding Structure in Language Models

ILLC DS-2025-01: **Julian Chingoma**
On Proportionality in Complex Domains

ILLC DS-2025-02: **Dmitry Grinko**
Mixed Schur-Weyl duality in quantum information

ILLC DS-2025-03: **Rochelle Choenni**
Multilinguality and Multiculturalism: Towards more Effective and Inclusive Neural Language Models

ILLC DS-2025-04: **Aleksi Anttila**
Not Nothing: Nonemptiness in Team Semantics

ILLC DS-2025-05: **Niels M. P. Neumann**
Adaptive Quantum Computers: decoding and state preparation

ILLC DS-2025-06: **Alina Leidinger**
Towards Language Models that benefit us all: Studies on stereotypes, robustness, and values

ILLC DS-2025-07: **Zhi Zhang**
Advancing Vision and Language Models through Commonsense Knowledge, Efficient Adaptation and Transparency

ILLC DS-2025-08: **Sophie Klumper**
The Gap and the Gain: Improving the Approximate Mechanism Design Frontier in Constrained Environments

ILLC DS-2025-09: **Jordi Rudo Weggemans**
Quantum versus classical resources in computational complexity

ILLC DS-2026-01: **Bryan Eikema**
A Sampling-Based Exploration of Neural Text Generation Models

ILLC DS-2026-02: **Marten Folkertsma**
Empowering Quantum Computation with: Measurements, Catalysts, and Guiding States

ILLC DS-2026-03: **Valentin Richard**
Presuppositional and Dynamic Aspects of Questions

ILLC DS-2026-04: **Puyu Yang**
Bringing Science to the Public: The Role of Wikipedia in Scientific Communication

ILLC DS-2026-05: **Johannes Kloibhofer**
Cycles with Annotations: Non-Wellfounded Proof Theory of Modal Fixpoint Logics

ILLC DS-2026-06: **Danish Kashaev**
Approximation via duality in offline, online and strategic settings

ILLC DS-2026-07: **Oskar van der Wal**
Taking a Step Back: Measuring and Mitigating Bias in Language Models

ILLC DS-2026-08: **Lynn Engelberts**
Classical and Quantum Cryptanalysis of Lattices and Codes

ILLC DS-2026-09: **Xiaoyu Tong**
Dissecting Incongruity: Metaphor and Humor Understanding of Large Language Models

ILLC DS-2026-10: **Qian Chen**
Lattices of Tense Logics: Tabularity, Completeness and Decidability