170.00 mm

170.00 mm

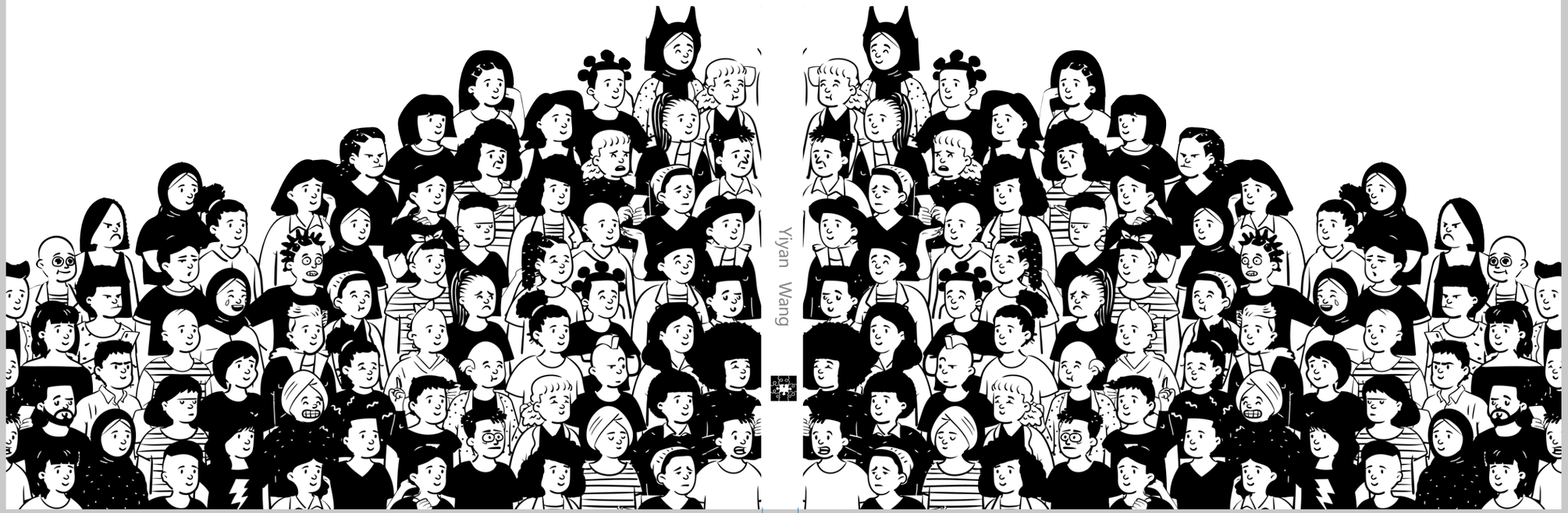
240.00 mm

COLLECTIVE AGENCY · FROM PHILOSOPHICAL AND LOGICAL PERSPECTIVES

COLLECTIVE AGENCY

FROM PHILOSOPHICAL AND LOGICAL PERSPECTIVES

Yiyan Wang



Yiyan Wang

8.00 mm

集体能动性：
从哲学与逻辑的视角看

申请清华大学-阿姆斯特丹大学联合授予

博士学位论文

王奕岩

二〇二三年五月

Collective Agency:
From Philosophical and Logical Perspectives

Dissertation Submitted to

Tsinghua University and University of Amsterdam

in partial fulfillment of the requirement

for a joint doctorate degree

by

Yiyan Wang

May, 2023

Collective Agency:

From Philosophical and Logical Perspectives

Yiyan Wang

Collective Agency:

From Philosophical and Logical Perspectives

ILLC Dissertation Series DS-2023-07

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/

We acknowledge the generous support of a 1-year Chinese Scholarship Council (CSC) scholarship.

Cover design by Youbang.
Printed and bound by Proefschriften.

ISBN: 978-90-833330-9-0

Collective Agency: From Philosophical and Logical Perspectives

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 12 september 2023, te 10.00 uur

door Yiyan Wang
geboren te Shanxi

Promotiecommissie

Promotores:	prof. dr. S.J.L. Smets	Universiteit van Amsterdam
	prof. dr. F. Liu	Tsinghua University
Copromotores:	prof. dr. M.J.B. Stokhof	Universiteit van Amsterdam
Overige leden:	prof. dr. ing. R.A.M. van Rooij	Universiteit van Amsterdam
	prof. dr. H.O. Dijstelbloem	Universiteit van Amsterdam
	dr. F. Russo	Universiteit van Amsterdam
	prof. dr. J.F.A.K. van Benthem	Universiteit van Amsterdam
	prof. dr. H. Tang	Tsinghua University
	prof. dr. D. Zhu	Tsinghua University
	dr. A. Jiang	Tsinghua University

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Dit proefschrift is tot stand gekomen binnen een samenwerkingsverband tussen de Universiteit van Amsterdam en Tsinghua University met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid in de Faculteit der Natuurwetenschappen, Wiskunde en Informatica van de Universiteit van Amsterdam en in de School of Humanities van Tsinghua University.

This thesis was prepared within the partnership between the University of Amsterdam and Tsinghua University with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Science at the University of Amsterdam and in the School of Humanities at Tsinghua University.

to collective agency

we have forged in our quest

摘 要

如今，人们生活在一个庞大而复杂的社会网络中。除了个人的决策和行动，人们每天还要面对各种各样的群体性决策与行动。相比于个体，这些群体性决策与行动显得扑朔迷离。作为某些群体的成员，我们为集体决策做出贡献，但这份贡献与最终决策并非总是一致。同时我们也被排除在某些群体之外，被动地受到他们的影响，但却找不出这些影响的根源。我们习惯于身处重叠的群体关系中，并于其间来回切换身份，支持或反对特定群体的主张，却很少停下来思考一个根本性的问题：当我们谈论集体及其决策时，我们在谈论什么？

本文的核心是关于集体能动性的问题，即在何种意义上我们可以将一个群体视为有能力采取行动的理性主体。文章采用哲学与逻辑两种研究视角。哲学视角主要探讨与集体能动性相关的本体论和认识论问题，梳理了相关哲学史思想，并提出了新的解释：集体能动性的关系观与集体意向性的倾向观。逻辑视角则与关于群体的形式化理论建立联系，忽略掉哲学视角所涉及的心理内容，以建立一个足够形式和客观的逻辑系统，从而将集体的主要性质公理化。

首先讨论的话题是集体能动性的本体论，即关于集体能动性究竟是什么的问题。集体能动性的哲学讨论主要集中在“集体”这一概念的化简问题上。个体主义和笛卡尔式的内在主义长期影响着本领域的正统理论，使其面临不可还原的“集体”概念与本体还原论之间的选择。功能主义和解释主义等非正统理论则重新解释了“主体”概念，并接受其在集体层面上的实现。为充分解释以关系为本质的社会现象，我们提出一种关于集体能动性关系的、整体论的解释，并认为功能主义和解释主义可被整合其中。

在承认“集体”概念不可约性的同时，我们发现“集体”概念与作为精神标志的“意向性”概念极度不相容。为了解释集体意向性是如何可能的，以及为什么我们倾向于将它与“个体意向性”概念进行类比，我们提出了关于意向性的倾向性解释，使得“意向性”概念能够合理地同时施用于个体和集体层面。具体而言，我们将倾向性解释细分为行为的，纯粹心理的和认知的三个方面。在此之上，通过分析不同形式的意向性归因判断，以及引入不可或缺的集体责任视角，我们论证了集体意向性的真实存在。

本文还分析了关于集体主体的哲学理论与关于集体决策的形式理论 (如博弈论) 的联系。这两个领域虽都以集体为研究对象，但却存在明显差异。例如，博弈论具有明显的反心理主义倾向，因为其研究目的在于提供形式且客观的分析。而

从本文提出的关系与倾向的解释视角来看，作为研究对象的个体意向性和集体意向性都不可避免地涉及到心理内容。为解释这种差异并明确不同领域的研究边界，本文分析了关于集体研究中的三个基本概念：意向性、偏好和依赖之间的基本关系，从而提供了关于集体研究的跨越哲学和形式理论的统一图景。

在铺设哲学视角与形式视角间的联系后，逻辑视角成为本文讨论的主题。为了形式表达博弈论中的核心概念，并揭示其与本文哲学观点的联系，我们提出了新的逻辑系统：偏好和函数依赖的逻辑及其混合扩展，并提供了可靠且完备的公理化系统。本文还证明了偏好和函数依赖逻辑的可判定性，并探讨了其在策略形式的非合作博弈和合作博弈中的应用。基于这一形式框架，本文提供了对纳什均衡、帕累托最优和核心等概念的统一解释，并最终明确这些博弈论概念与集体能动性哲学的相关性。

最后，本文总结并澄清了所提出的理论在更广泛研究领域中所处的位置，同时也指出了许多新的问题和富于前景的研究方向，包括哲学和逻辑视角的多个开放性问题。

关键词：集体主体性；集体意向性；整体论；其他条件不变下的偏好；函数依赖；联合力

SAMENVATTING

Mensen leven tegenwoordig in een uitgebreid en ingewikkeld sociaal netwerk. Naast onze eigen beslissingen en handelingen worden we dagelijks geconfronteerd met die van verschillende groepen. Collectieve beslissingen en handelingen zijn complexer en verwarrend vergeleken met die van individuen. Als leden van een collectief dragen we bij aan de beslissingen ervan, maar onze bijdragen komen niet altijd overeen met het resultaat. We kunnen ook buitengesloten worden van bepaalde groepen en passief onderworpen worden aan hun invloeden zonder ons bewust te zijn van de bron. We zijn gewend deel uit te maken van overlappende groepen en kunnen van identiteit veranderen, waarbij we de aanspraken van bepaalde groepen steunen of ons ertegen verzetten. Maar zelden staan we erbij stil: Waar hebben we het over als we het hebben over groepen en hun beslissingen?

Centraal in dit proefschrift staat de kwestie van collectief handelen, d.w.z. de vraag in welke zin we een groep kunnen behandelen als een rationele agent die in staat is tot handelen. We hanteren twee perspectieven: een filosofisch en een logisch perspectief. Het filosofische perspectief bespreekt vooral de ontologische en epistemologische vraagstukken in verband met collectief handelen, zet de relevante filosofische geschiedenis op een rij, en betoogt dat de combinatie van een relationele kijk op collectief handelen en een dispositionele kijk op collectieve intentionaliteit een rationele en realistische beschrijving oplevert. Het logische perspectief is geassocieerd met formele theorieën over groepen. Het gaat voorbij aan de psychologische inhoud van het filosofische perspectief, construeert een logisch systeem dat voldoende formeel en objectief is, en axiomatiseert wat een collectief is.

Het eerste onderwerp dat wordt behandeld is de ontologie van collectief handelen, d.w.z. de vraag wat collectief handelen precies is. De filosofische discussie daarover draait om het reductieprobleem van het begrip collectief. Individualisme en Cartesiaans internalisme hebben orthodoxe theorieën lang beïnvloed en hen voor de keuze gesteld tussen een onherleidbaar concept van een collectief en ontologisch reductionisme. Heterodoxe theorieën zoals het functionalisme en het interpretationisme herinterpreteren het begrip handelen en aanvaarden het als ook gerealiseerd op het niveau van een collectief. Om sociale fenomenen die in essentie relationeel van aard zijn adequaat te verklaren, stellen wij een relationele, holistische beschrijving van collectief handelen voor en beargumenteren

wij dat functionalisme en interpretationisme in een dergelijke beschrijving kunnen worden geïntegreerd.

Hoewel wij de onherleidbaarheid van het begrip collectief erkennen, stellen wij vast dat er een diepe onverenigbaarheid bestaat tussen het begrip collectief en het begrip intentionaliteit als kenmerk van het mentale. Om te verklaren hoe collectieve intentionaliteit toch mogelijk is en waarom wij geneigd zijn het te gebruiken naar analogie van de wijze waarop wij het begrip individuele intentionaliteit gebruiken, onderzoeken wij een dispositionele analyse van intentionaliteit die ons in staat stelt het begrip intentionaliteit zowel op individueel als op collectief niveau te verantwoorden. Daarbij onderscheiden we drie aspecten aan de dispositionele analyse: het gedragsmatige, het mentale, en het cognitieve aspect. Vervolgens betogen wij dat collectieve intentionaliteit reëel is door verschillende vormen van attributieve oordelen over intentionaliteit te analyseren en door het perspectief van onmisbare collectieve verantwoordelijkheid te introduceren.

We analyseren ook hoe filosofische theorieën over collectief handelen zich verhouden tot centrale kenmerken van formele theorieën over collectieve beslissingen, zoals de speltheorie. Hoewel beide gebieden zich bezighouden met collectieven, zijn er ook verschillen tussen beide die moeten worden aangepakt. Zo is de speltheorie duidelijk antipsychologistisch omdat zij een formele en objectieve analyse beoogt te geven. Vanuit het relationele en dispositionele perspectief hebben intentionaliteit op individueel niveau en collectieve intentionaliteit zoals wij die analyseren echter onvermijdelijk een mentale inhoud. Om dit verschil te verklaren en vast te stellen waar de grens ligt, analyseren wij de relaties tussen de drie betrokken basisbegrippen, namelijk intentionaliteit, voorkeur en afhankelijkheid, zodat een uniform beeld ontstaat van de collectieve theorie dat van toepassing is op zowel filosofische als formele theorieën.

Na het leggen van de verbinding tussen filosofische en formele perspectieven, wordt het logische perspectief het thema van onze discussie. Om speltheoretische concepten te kunnen uitdrukken en ze te verbinden met ons filosofisch perspectief, presenteren we een logica van voorkeur en functionele afhankelijkheid en de hybride uitbreiding ervan en geven we een axiomatisering die deugdelijk en sterk compleet is. De beslisbaarheid van deze logica wordt ook bewezen. De toepassing ervan op het modelleren van niet-coöperatieve en coöperatieve spellen in strategische vorm wordt onderzocht. Het resulterende raamwerk biedt een eenduidige kijk op Nash-evenwicht, Pareto optimaliteit en de kern. De filosofische relevantie van deze speltheoretische noties voor discussies over

collectief handelen wordt expliciet gemaakt.

Ten slotte concluderen wij en verduidelijken wij de positie van onze theorie in het bredere onderzoeksveld over de in het proefschrift behandelde onderwerpen. Ook wijzen wij op vele nieuwe vragen en richtingen die door onze analyse worden gesuggereerd, waaronder filosofische en logische open problemen.

Trefwoorden: collectief handelen; collectieve intentionaliteit; holisme; ceteris paribus voorkeur; functionele afhankelijkheid; coalitionele macht

ABSTRACT

People inhabit a vast and intricate social network nowadays. In addition to our own decisions and actions, we confront those of various groups every day. Collective decisions and actions are more complex and bewildering compared to those made by individuals. As members of a collective, we contribute to its decisions, but our contributions may not always align with the outcome. We may also find ourselves excluded from certain groups and passively subjected to their influences without being aware of the source. We are used to being in overlapping groups and may switch identities, supporting or opposing the claims of particular groups. But rarely do we pause to think: What do we talk about when we talk about groups and their decisions?

At the heart of this dissertation is the question of collective agency, i.e., in what sense can we treat a group as a rational agent capable of its action. There are two perspectives we take: a philosophical and logical one. The philosophical perspective mainly discusses the ontological and epistemological issues related to collective agency, sorts out the relevant philosophical history, and argues that the combination of a relational view of collective agency and a dispositional view of collective intentionality provides a rational and realistic account. The logical perspective is associated with formal theories of groups, it disregards the psychological content involved in the philosophical perspective, establishes a logical system that is sufficiently formal and objective, and axiomatizes the nature of a collective.

The first topic that is addressed is the ontology of collective agency, i.e., the question what exactly is collective agency. The philosophical discussion of collective agency centres around the reduction problem of the concept of a collective. Individualism and Cartesian internalism have long influenced orthodox theories and made them face the choice between an irreducible concept of a collective and ontological reductionism. Heterodox theories such as functionalism and interpretationism reinterpret the concept of agency and accept it as also realized on the level of a collective. To adequately explain social phenomena that are essentially relational in nature, we propose a relational, holistic account of collective agency and argue that functionalism and interpretationism can be integrated into such an account.

While acknowledging the irreducibility of the concept of a collective, we find that there is a deep incompatibility between the concept of a collective and the concept of

intentionality as the mark of the mental. To explain how collective intentionality nevertheless is possible and why we tend to use it analogously to how we use the concept of individual intentionality, we explore a dispositional account of intentionality which enables us to give an account of the concept of intentionality at both the individual and collective level. Specifically, we subdivide the dispositional account into three aspects: behavioral, purely mental, and cognitive. We then argue that collective intentionality is real by analyzing different forms of attributive judgments of intentionality and by introducing the perspective of indispensable collective responsibility.

We also analyze how philosophical theories about collective agency relate to central features of formal theories about collective decisions, such as game theory. Although the two fields are both concerned with collectives, there are also differences that need to be addressed. For example, game theory is clearly anti-psychologistic since its aim is a formal and objective analysis. However, from the relational and dispositional perspective, intentionality at the individual level and collective intentionality as we analyze it, inevitably involve mental content. In order to explain this difference and identify where the boundary is, we analyze the relationships between the three basic concepts involved, namely intentionality, preference, and dependency, so as to provide a unified picture of collective theory across philosophical and formal theories.

After paving the nexus between philosophical and formal perspectives, the logical perspective becomes the theme of our discussion. To be able to express game theoretical concepts and to connect them to our philosophical perspective, We present a logic of preference and functional dependence and its hybrid extension, and provide an axiomatization which is sound and strongly complete. The decidability of this logic is also proved. Its application to modeling non-cooperative and cooperative games in strategic form is explored. The resulting framework provides a unified view of Nash equilibrium, Pareto optimality, and the core. The philosophical relevance of these game-theoretical notions to discussions of collective agency is made explicit.

Finally, we conclude and clarify the position of our theory in the broader field of research on the topics addressed in the thesis. Also, we point out many new questions and directions suggested by our analysis, including philosophical and logical open problems.

Keywords: collective agency; collective intentionality; holism; ceteris paribus preference; functional dependence; coalitional power

TABLE OF CONTENTS

SOURCES OF THE CHAPTERS

- Chapter 2 is based on the following paper:

  Wang, Yiyan, and Stokhof, Martin (2022). "A Relational Perspective on Collective Agency." *Philosophies* 7 (3):63.

  Authors contributions: Yiyan Wang and Martin Stokhof initiated the project and discussed the central arguments together. Yiyan Wang developed the original draft. Martin Stokhof reviewed and edited it to the final edition.

- Chapter 3 is based on the following paper:

  Wang, Yiyan (2022). "Towards a Dispositional Account of Intentionality." Under review.

  which develops ideas from the following paper:

  Wang, Yiyan (2020) "Intentionality as Disposition." In: Liao Beishui, Wáng Yi (eds) *Context, Conflict and Reasoning. Logic in Asia: Studia Logica Library.* Springer, Singapore.

- Chapter 5 is based on the following paper:

  Chen, Qian, Shi, Chenwei, and Wang, Yiyan (2022). "Reasoning about Dependence, Preference, and Coalitional Power." Under review.

  which develops ideas from the following paper:

  Shi, Chenwei, and Wang, Yiyan (2021). "Pareto Optimality, Functional Dependence and Collective Agency." Manuscript. Tsinghua - Amsterdam Joint Research Centre for Logic, Department of Philosophy, Tsinghua University.

  Authors contributions: Chenwei Shi and Yiyan Wang initiated the project. The three authors contributed equally to the latter paper.

CHAPTER 1 INTRODUCTION

1.1 Motivations

As early as the beginning of the formation of human civilization, in addition to the natural blood ties, primitive production activities have required competition and cooperation between people, and simple groups bonds are maintained within a tribe. With the development of productivity, such group bonds become more complicated, individuals can play corresponding roles in different social groups simultaneously, and a social network with multiple superimposed relationships gradually emerges. With the development of the level of trade and the birth of the concept of the legal person, some fixed group relationship models become institutional facts, often confined in law. In modern society, everyone lives in complex social interactions since birth and belongs to different collectives. A prominent feature of the information age is that the Internet continuously presents us with distant information. As long as we care, we are connected to cultural, economic, and political events anywhere on the earth. We may ignore things happening in our community but are extremely concerned about events in the other hemisphere. Similar cultural transmissions generate new social relationships at any time, and new circles and new groups are generated one after another. We are fixed in some groups for reasons of work and life. We can also join some groups because of interests and leave because of changing interests. Constantly in the midst of group change, naturally, we want to examine those social concepts that shape our lives, and explore their essence to understand what it is that we are actually involved in. Therefore, we have a central question that runs through the thesis: what are we talking about when we talk about a collective?

This dissertation will take a philosophical and logical perspective on this question, which means that our analysis is more focused on general, abstract conceptual analysis than on specific situations, which will only appear in examples. We shall use the term "group" to refer generally to a set of multiple individuals and the term "collective" to refer further specifically to groups that have formed a stable structure. There have been philosophical and logical studies on groups before. Philosophical research on groups focuses on questions concerning the existence of the concept of collective and its essence, agency, reduction problems, etc. These discussions originate from the philosophy of mind and the philosophy of action, and thus relate to the core concept discussing agent and ac-

tion: intentionality. Logical research in this area has created many logics involving group issues. Formal methods are usually used to define group-related concepts, such as the discussion of common knowledge in epistemic logic and the discussion of common obligation in deontic logic. In addition, coalition logic directly describes group behavior in the setting of cooperative game theory, which manifests the close relationship between logic and other formal frameworks. Our research will take philosophy as well as logic into account. On the one hand, we will be concerned with philosophical discussions of the agency and intentionality of a collective. On the other hand, we will also pay attention to how a collective is generated in social interaction and hope to promote interdisciplinary understanding of game theory and its corresponding logical characterization. In order to clarify the various directions involved, we will introduce some relevant topics in the following.

Philosophies of collective agency

In philosophy, an agent is usually defined as an entity that has the capacity to act, and agency refers to the implementation or manifestation of such capacity. At the individual level, agency is usually explained by the concepts of intentional action and causal links between mental states and action events. The idea of explaining agency through intentional action can be traced back to the work of Hume and even Aristotle, while the direct source of the resurgence of this trend in modern philosophy is the work of Anscombe (1957) and Davidson (1963). Although there are differences between their analyses, an analytical model of agency and agential behavior is established through the concept of intentionality.

The primary reason for using intentional action, or intentionality, to explain agency comes from an argument that intentionality is more fundamental than actions, which is briefly described below. Agency is related to the capability of action of an entity, and thus how to interpret action is the key to grasping agency. Objectively speaking, an action is an event, including time, place, actor, action object, and other elements. There are infinitely many descriptions of the same act-event, but only some of them can be said to be intentional. For example, suppose a construction worker is renovating a room in a residential building. He only intentionally completes the work and earns wages, but he does not intentionally make noise and disturb the people living in that building, even if the two descriptions refer to the same event. The concept of action is inherently active, volitional, and contains a necessary connection between the agent and its action. But the

description of unintentional behavior does not include such a necessary connection. In other words, for any behavioral event, only if there is a description of it as an intentional action, can we say there is a necessary connection between the behavior and the actor, and judge it as an action. Therefore, intentional action and intentionality become the key to analyzing the concept of agency.

However, there are questions concerning the standard theory from the philosophy of action, which mainly stem from two considerations. One is based on the concern that the conventional definition of the agent is not enough to distinguish human beings from other agents, and that therefore aims to find a definition that singles out human agents (cf. Mele (1995), Bratman (2007)). The other is based on concerns about the broader concept of agency. Here it is argued that the traditional method is limited to the analysis of the human individual, ignoring that individuals are socially situated. For instance, simple biological agents that have not developed into creatures that have psychological content but that can conduct actions, are free of intentionality and the link between internal states and behavior. Similar concerns arise with artificial intelligence, which does not have a biological brain but can judge the environment and perform actions by itself. These considerations also seem to apply to case of collectives, for example, the government as a collective agent that does not have a collective biological brain but still functions in social life. Standard explanations fall short of explaining how these properties of collectives can be explained on the basis of analogous properties of individuals; thus, there are attempts to relax the requirements from different perspectives, such as functionalism (cf. Pettit (2009a)) and interpretationism (cf. Tollefsen (2002)). Since we are concerned with collective agency, we focus more on the second approach, and we will come back and further review these theories in Chapter 2.

Philosophies of collective intentionality

As discussed in the previous section, the interpretation of agency is closely related to the concept of intentionality. In modern philosophy, intentionality was first introduced by Brentano (1874), and it basically means the aboutness of human consciousness. For example, “I plan to go to the van Gogh Museum,” “I prefer to have herring for dinner,” “I believe it will not rain today,” and so on. Each manifestation of intentionality contains a particular intentional type, namely intention, preference, belief, etc., as a psychological state, and an intentional object, namely “going to the van Gogh Museum”, “having herring for dinner”, “it not raining today”, etc., as a propositional content. Except for aboutness,

that human consciousness is about something, intentionality expressions do not provide any existential guarantees about the intentional objects ('the feature of non-existence'), nor do they determine various specific properties of the intended objects ('the feature of indeterminacy'), and intentionality is also used as an exclusive mark of psychological phenomenon ('the feature of the mark of mental'). We will discuss these properties of intentionality further in Chapter 3.

It is not difficult to find that the examples mentioned above of individual intentionality also apply to groups, such as "We are going to the van Gogh Museum," "We prefer to have herring for dinner," and "We believe it will not rain today." In these sentences, the subject of the intention is a group "we," i.e., a "we-subject," which performs the intentional action. However, the consequent problem is that if intentionality is interpreted as the directional relationship between mental states and intentional objects, we need to admit a certain collective mental state. And then we can, in fact need, to ask is this mental state the superposition of each member's intention? Or a separate irreducible subject?

At first sight, this question may seem inappropriate because the concepts of intentionality and intentional actions were originally used to describe individuals and individual actions, and they were even regarded as the exclusive mark of psychological phenomena. Thus, one might argue, it is not necessary for us to continue to use these concepts in the collective situation. Nevertheless, at least in everyday conversation, collective intentionality is often present in our everyday discourse, and each of us can fully accept and use it without any ambiguity. On the other hand, there is a profound connection between the collective and the individual, and we cannot completely separate the group situation from the discussion of individual intentionality: collective intentionality always affects individual behavior. Therefore, since the 1990s, philosophical discussions on collective intentionality have gradually emerged. Approaches that have been influential in this field are Bratman (2014); Gilbert (2006); Searle (2010); Tuomela (2013); List et al. (2011). These works discuss questions like whether collective intentionality exists, what its nature is, and whether it can be reduced to individual intentionality from different perspectives. We will continue to discuss these themes in Chapters 2 and 3.

Formal theories of social groups

Generally speaking, formal theories refer to the type of system that establishes a class of assumptions and rules in mathematical terms to study and understand various scenarios and agents' behaviour in reality. There are many formal theory approaches to the study of

collectives in the social sciences; here, we only briefly review two main strings of them, namely social choice theory and game theory.

Rational choice theory studies the decision-making process of individuals based on their personal preferences. Social choice theory focuses on how to bundle the preferences of individual members into a group preference. Social choice theory takes individual opinions, preferences, interests, etc., as inputs and picks an appropriate aggregation function to generate corresponding opinions, preferences, interests, etc., of a group as the output, exploring the effects of variant aggregation functions (cf. List (2022)). An intuitive example is voting, where each individual gives his or her own opinion, and majority rules are used to make the alternative with the highest number of votes the group opinion.

Social choice theory can be traced back to Condorcet's paradox① (cf. de Condorcet (1785)). Modern social choice theory is based on Arrow's impossibility theorem (cf. Arrow (1951)), which basically means that under certain assumptions (universal domain, ordering of social preference, weak Pareto principle, independence of irrelevant alternatives, non-dictatorship), no function can aggregate a group preference from the individual preferences given by two or more agents over three or more alternatives. Therefore, Arrow argues that the utilities based on individual preferences cannot be compared interpersonally because their roots are internalized in human psychology. Scholars of contemporary social choice theory tend to hold the opposite view. Researchers like Sen (1970) and Harsanyi (1990) believe that people with common backgrounds and experiences can compare their utilities. They follow Arrow's formal method and turn to study how to relax some assumptions to obtain effective and practical aggregation functions (cf. May (1952), Thomson (2001), List et al. (2011)). From a philosophical point of view, social choice theory insists on individualism in the sense of methodology; that is, it only presupposes the input value of preferences and some aggregation rules at the individual level, and then generates the concept of preference at the group level. In other words, those concepts of the collective are reducible.

In comparison, game theory focusses more on the interaction of individual preferences and choices and attempts to analyze how this interaction affects the overall decision-making (cf. Ross (2021)). Modern game theory stems from the work of von Neumann

① Generally speaking, Condorcet's paradox refers to the failure of majority rule in aggregating individual preferences to social ones, characterized by the non-transitivity of aggregate preferences. A straightforward example is to assume that there is a group G with three voters, 1, 2 and 3, who rank three alternatives x, y, and z, respectively. And the results are $x >_1 y >_1 z$, $y >_2 z >_2 x$, and $z >_3 x >_3 y$. By the majority rule, we can conclude that the aggregate preference for the group is satisfied that $x >_G y$, $y >_G z$, and $z >_G x$, clearly unsatisfied transitivity. This paradox is also known as the cyclic majority or paradox of voting.

et al. (1947), which established a general mathematical formalization of different types of games and concepts such as expected utility. Since the 1950s, the development of game theory has been prosperous, and various landmark concepts have been proposed, such as Nash equilibrium (cf. Nash Jr (1950)), the core (cf. Gillies (1959)), etc.

In a game, each player faces two or more strategies and has to choose among them. Players know how to make choices to respond to others' (anticipated) decisions, and their optimal strategy is the one that maximises the outcome for a player, or for a group of players. Whether a coalition can be set in advance depends on the nature of the game, viz., on whether it is a cooperative game or a non-cooperative game. Non-cooperative games are based on the absence of coalitions, with each player making individual decisions based on the individual decisions of other players, focusing solely on the procedural level, that is, the interaction between agents' strategies. Cooperative games, on the contrary, allow various coalitions and their interrelationships to exist, abstract away from detailed individual interactions and focus on how the coalitions that agents can form affect the decisions of their members.

We can use cooperative games to study situations where coalitions have been formed. But, game theorists have also studied the process of group formation in the framework of non-cooperative game, and this work has another kind of theory, that of team reasoning. Research on team reasoning proves that if we can change the thinking mode of all players or make each of them subject to additional agreements, we can achieve the optimal collective strategy that cannot be endogenously achieved in the original game. In particular, Bacharach (1999) points out that it is the team reasoning pattern of the so-called 'we-perspective' (members as team reasoner) rather than caring for the team from a personal perspective (members as team benefactor) that is indispensable in coordinate games. Similar to the philosophical divergence, different game theorists have offered different views on the reason behind team reasoning (cf.Bacharach et al. (2006); Sugden (2003); Hurley (2005); Karpus et al. (2017)).

Exploring the problem of collective agency in the context of both formal theories is promising work. The most striking feature of these theories is that they all use individual preference as a basic concept, rather than the more basic concept of desire, or the key concept in philosophical discussion: intentionality. So it will be part of our work to analyze the connection, the boundaries, and the scope of application of these basic concepts, which will be discussed in depth in Chapter 4.

Logics of multiple agents

Modern logic can provide a powerful formal analysis of multi-agent scenarios: a great many aspects of the interaction in a multi-agent system can be described and their logical properties analysed. Firstly, a multi-agent system involves the dynamic interaction between the agents' individual cognitive states, deontic norms and specific actions, while contemporary logic, especially modal logic, provides a detailed description and interpretation of the cognitive, temporal, moral and actional dynamics. Secondly, multi-agent interaction involves a large number of reasoning decisions, and a relatively stable coalition will be formed through the interactive iteration, in which decision-making reasoning will focus on the gains and losses of collective interests. These features can also be captured by modern logic. Finally, modern logic can provide an effective semantics for various concepts involved in multi-agent interaction and give clear explanations for concepts such as knowledge, belief, goal, intention, and coalition.

The description of an agent cannot be given only in extensional terms (such as those of propositional logic and first-order logic), but requires an intensional description of the agents' knowledge and belief (thus requiring the use of an intensional logic, such as modal logic). The scenario of multiple agents brings various interpersonal interactions, such as the nested relationship between the agent's beliefs. Thus higher-order information flow enters the discussion, and we can further define concepts like "everyone knows", "public knowledge", and "distributed knowledge" through the cognitive interaction between agents. In addition, multiple agents make coalitions possible, which leaves logic to consider complex issues such as how to determine coalitions and how to make them effective. In addition to coalitions, the multi-agent scenario also introduces social norms into the discussion. The concepts of permission, prohibition, and responsibility bring up new and complex topics. And all of these problems also have a dynamic dimension. The facts in the world, the cognitive content in the minds, the coalitions formed and the norms adhered to are constantly changing. Change involves the description of tense and the changed state at the next moment, which are thus additional problems that multi-agent logic needs to face.

Generally speaking, the development of multi-agent logic has two strands: the modelling of rational agents' cognitive content and action, and the characterization of the strategic structure interactions. The main instruments of the former include epistemic logic, intention logic, dynamic epistemic logic, deontic logic (in terms of social attitudes towards norms), etc. The main instruments of the latter include coalition logic,

alternating-time temporal logic (ATL), and judgment aggregation logic, etc.

The epistemic strand

Epistemic logic analyzes the behaviour of rational agents through concepts of knowledge, belief, desire, intention, etc. Concepts like knowledge and belief mainly involve one's passive representation of the external world. While concepts of desire and intention mainly involve one's tendency to take actions to actively transform the external world to conform to one's preferences. In basic epistemic logic (cf. Fagin et al. (1995); Meyer et al. (1995)), let variable i range over a set of agents $Ag = \{1., , , .n\}$, $K_i\phi$ means that agent i knows ϕ, and the following axioms are given:

Axioms	Inference Rules
(Kn1) ϕ is propositional tautology.	(MP) $\vdash \phi, \vdash \phi \rightarrow \psi \Rightarrow \vdash \psi$
(Kn2) $K_i(\phi \rightarrow \psi) \rightarrow (K_i\phi \rightarrow K_i\psi)$	(Nec) $\vdash \phi \Rightarrow \vdash K_i\phi$
(Kn3) $K_i\phi \rightarrow \phi$	
(Kn4) $K_i\phi \rightarrow K_iK_i\phi$	
(Kn5) $\neg K_i\phi \rightarrow K_i\neg K_i\phi$	

This particular system is often referred to as S5. Kn2 means an agent can make deductions based on what he/she knows. Kn4 and Kn5 refer to positive and negative introspection, respectively; that is, an agent knows what it knows and knows what it does not know. Kn3 means that what an agent knows is true. Note that when we replace Kn3 with the weaker constraint $\neg K_i \perp$, we get the logic KD45 for beliefs, where an agent i believes ϕ is represented as $B_i\phi$. Possible world semantics is usually used to interpret the formulas in epistemic logic, where the accessibility relation is interpreted as the cognitive indistinguishability relation between an agent's epistemic states. With the language of basic epistemic logic, we can formalize some concepts regarding group knowledge. For instance, let $\{1, 2, ..., n\} = G \subseteq Ag$, aforementioned "everyone knows" is defined as $E_G\phi = K_1\phi \wedge K_2\phi \wedge ... \wedge K_n\phi$ (Everyone in the group G knows ϕ.). "Distributed knowledge" means the knowledge derived after summarizing group members' knowledge. Suppose that group $G = \{1, 2\}$, $K_1(\phi \rightarrow \psi)$ and $K_2\phi$, we have the distributed knowledge of G as $(K_1(\phi \rightarrow \psi) \wedge K_2\phi) \rightarrow D_G\psi$. "Common knowledge" $C_G\phi$ is defined as the infinite conjunction: $E_G\phi \wedge E_GE_G\phi \wedge E_GE_GE_G\phi \wedge ...$

Knowledge is not always static; in fact, our daily knowledge is constantly changing and updating. Dynamic epistemic logic (as first developed in van Benthem (1989) and Plaza (1989), for recent overview, see van Ditmarsch et al. (2007)) studies the influence

of actions on one's knowledge. A typical example would be a public announcement $[\phi]\psi$ expressing that after ϕ is announced, ψ will be the case. (cf. Plaza (1989)). More refined research is the semantics of action models (cf. Baltag et al. (2004)), which generalize changes in the cognitive states of multiple agents after performing particular actions. Regarding multi-agent interaction, the prominent research is logics of social networks (cf. Liu (2009); Seligman et al. (2013); Liu et al. (2014)), which highlight the mutual influence of knowledge and belief among multiple agents under different social network structures, as well as the final group knowledge or belief state under such influence.

In addition to developing systems that can be use to represent knowledge and belief about the external world, logicians also focus on the pro attitudes that an agent would act by to change the world. Prominent work in this strand is intention logic (cf. Cohen et al. (1990); Galliers (1989); Levesque et al. (1990)). Such a study mainly follows the work of Bratman (1987), which interprets the concept of intention as the complex of one's beliefs, desires, goals and other mental states. An individual agent acting according to his/her intentions will adopt the intentions he/she believes in and abolish the ones he/she does not believe. Once the intention is formed, the agent will commit to the intentional action, and such a commitment will be abandoned when the intention is satisfied, or the related belief is changed. In a long-term plan, the fulfilment of one intention is often accompanied by engaging in another, subsequent one. In order to express the above meaning, Cohen and Levesque defined four basic modalities for intention logic, which are two epistemic modalities (Bel i ϕ) ("agent i believes in ϕ") and (Goal i ϕ) ("agent i has a goal of ϕ"), and two actional modalities (Happens α) ("action α happens at the next moment") and (Done α) ("action α just happened").[①] Based on these basic modalities, intention logic defines the concepts of *Persistent Goal* and *Intention*, which are represented as (P-Goal i ϕ) and (Intend i α), respectively. Another important work on pro attitudes is BDI logic (cf. Bratman et al. (1988); Rao et al. (1998); Wooldridge (2003)), in which modalities like Bel, Des and $intend$ are added to the branching time logic to express beliefs, desires and intentions. Another research on intention and action also based on branching time logic is STIT logic (beginning with Belnap et al. (1990), recent development see cf. Broersen (2011); Lorini et al. (2016); etc.), which is designed to portray what is going to happen ("see to it that ϕ ").

① These definitions, including the previously mentioned public announcement operator, involve logics about time and about action, which are not reviewed in detail here due to limitations of space. Interested readers may refer to Plaza (1989) and Cohen et al. (1990).

In addition to passive and active mental states, logicians also consider people's social attitudes toward norms. Research in this direction follows the tradition of deontic logic (cf. von Wright (1951)) and connects it with the multi-agent setting, for instance, Shoham et al. (1992a); Shoham et al. (1992b) study norms in multi-agent systems.

The strategic strand

This strand focuses on the state in which an agent (individual or collective) can exert its ability to achieve particular goals under certain conditions. Its primary purpose is to analyze the strategic ability of the agent in a specific environment through modelling. The need for formal analysis of strategic ability is widespread in artificial intelligence, game theory, philosophy of action and other disciplines. There are numerous studies in this field, to name but a few essential works, such as coalition logic (cf. Pauly (2001)) and ATL (cf. Alur et al. (2002)).

As mentioned above, game theory is a prominent formal method to study group interaction. Coalition logic studies abstract mathematical structures contained in game theory using modal logic as a tool. In coalition logic, let C be the set of agents, then its fundamental operator $[C]\phi$ means that coalition C can cooperate to guarantee that ϕ will be the case. The semantics of this operator is given by the effectivity function $E : S \to (\mathcal{P}(Ag) \to \mathcal{P}(\mathcal{P}(S)))$, which means that coalition C can guarantee ϕ if it can choose a particular set of states S which precisely guarantees ϕ. The basic axioms of coalition logic (cf. Pauly (2002)) are as follows:

Axioms	Inference Rules
($\bot$) $\neg[C]\bot$	(MP) $\vdash \phi, \vdash \phi \to \psi \Rightarrow \vdash \psi$
(N) $\neg[\varnothing]\neg\phi \to [Ag]\phi$	(Nec) $\vdash \phi \Rightarrow \vdash [C]\phi$
(M) $[C](\phi \wedge \psi) \to [C]\psi$	
(S) $([C_1]\phi \wedge [C_2]\psi) \to [C_1 \cup C_2](\phi \wedge \psi)$ where $C_1 \cap C_2 = \varnothing$	

Here $[\varnothing]$ is the empty coalition, that is, no one participates, and $[Ag]$ is the grand coalition composed of all the agents in Ag. (N) indicates that if the empty coalition cannot guarantee $\neg\phi$, then the grand coalition ensures ϕ. (M) represents the feature of monotonicity inherited from the effectivity function. (S) indicates that non-overlapping coalitions can combine and achieve more, a property also known as superadditivity. Coalition logic, as well as the effectivity function, focuses on the outcome of a game rather than the intermediate process of how to reach that outcome. Per contrast, studies that center on changes in decisions made at different moments, Alur et al. (2002) further extended the branching

time logic to construct ATL. Similar to coalition logic, ATL can also express sentences such as "coalition C can cooperate to guarantee ϕ", denoted as $\langle\langle C\rangle\rangle\phi$. In addition to game theory, logicians are also interested in the logical characterizon of social choice theory (cf. Endriss (2011); Ågotnes et al. (2011)). Using logic to describe game theory and social choice theory is usually related to preference logic (cf. van Benthem et al. (2009); Liu (2011)) and expected utility (cf. Jamroga (2008)).

Request for a comprehensive study

Combining the above-mentioned directions of philosophy, game theory and logic for the study of a collective, the need for a comprehensive theory emerges. From the perspective of philosophy and game theory, philosophy sticks to the tradition of intentionality to explain collective agency, while game theory uses the quantifiable concept of preference to establish an objective analysis of social facts. In this thesis, we describe both the connections and boundaries between the concepts of intentionality and preference, which will provide a more unified picture and bring some insights into studying social agency.

From the perspective of game theory and logic, logic can provide a transparent description of the mathematical structure behind games, while game theory provides a lot of directions and formal preliminary work for logic. In this thesis, we will characterize several essential collective features in game theory through logic and illustrate the relationship between some of the significant game theory concepts.

In philosophical discussions, we pay more attention to the process of the generation of collective agency from complex interactions between different subgroups in a collective. We hope that these concerns can be expressed and discussed through logical methods. Along this way of thinking, we extend the logic of functional dependence as our basic logical language, which does not assume axioms like superadditivity and thus allows us to talk more details within a group. With the logic of preference and functional dependence and its hybrid extension, we could appropriately express concepts in game theory and connect them to our philosophical perspective.

We will start from the philosophy of collective agency and collective intentionality and discuss questions about ontology and characterization primarily from a philosophical perspective. Next, we will discuss the connection between philosophical and formal theories, mainly through the concept of desire as a bridge between intentionality and preference. Meanwhile, we also show their commonality through the concept of dependency. Finally, we describe philosophical views and game theory concepts through logic, and

take dependency as the core concept to describe the relationships among individuals and those between individuals and collectives. The technical basis of the logic of functional dependence will be briefly introduced in Section 1.4.

1.2 Outline of the Thesis

Here we will briefly list the main arguments and results presented in this dissertation.

Chapter 2 discusses the question "what is collective agency?" from the philosophical, especially ontological perspective. Firstly, we review the philosophical debate on collective agency in recent decades, and point out that they all explicitly or implicitly insist on individualism, and thus have to face the dilemma of inferential circularity. To dig deeper into the reasons behind this phenomenon, we review the root causes of the popularity of individualism, summarize the common paradigms for the argument of physicalistic individualism, and give a refutation. On this basis, we begin to look for an alternative to the individualistic theory. We review existing theories of functionalism and interpretationism and combine them into a unified point of view which we call "relationalism". In addition, we further argue for a thoroughly relationalist position that allows an explanation of not only collective agency but also individual agency. In the rest of this chapter, we explain the complex relationship between collective agency and individual members' actions in the relational framework, and point out that the existence of collective agency does not contradict the display of individual agency.

Chapter 3 follows the threads of the relationalist account. Relationalism and the debate on the irreducibility of the concept of a collective provides an explanation for that fact that collective agency exist, but can not fully explain how collectives come to be. The philosophical analysis of collective agency is an extended application of the philosophy of action, which follows the path of using intentionality as the core concept. By reviewing the definition of intentionality in the literature, we find that the common conception of intentionality is limited by its characterization in terms of individual activities and does not match the collective situation. Meanwhile, we also note that there are attempts to explain intentionality and disposition in terms of each other. Along this line of thought, we put forward the dispositional account of intentionality. In such a framework, individual intentionality and collective intentionality are well coupled and consistent with the relationalist account of Chapter 2. In addition, we construct arguments supporting the dispositional account from epistemological and deontic perspectives, respectively.

From Chapter 4 onwards, formal theories of groups are involved in our topic. We want to formalize the relational framework obtained in Chapters 2 and 3 and construct a logical system to express its core views. Nevertheless, before starting the technical part, bridging philosophical discussion with formal theories is indispensable, and Chapter 4 serves that purpose. We review preference as a foundational concept in formal theories and compare its connection to and difference with intentionality as a foundational concept in philosophy. On this basis, we connect the two through desire as a pivot, construct a more unified explanatory framework, and clarify the scope of application of the philosophical and formal theories. In addition, we discuss another important fundamental concept in social interaction: dependency, and reinterpret it through the relationalist account. In the remainder of this chapter, we represent our philosophical views in the context of game theory.

In Chapter 5, we develop the logic of functional dependence into the logic of preference and functional dependence (LPFD). We give the syntax and semantics of LPFD, and use it to express many game theoretical concepts, such as Pareto optimality and Nash equilibrium, and formalize our basic philosophical views. We also give the axiomatization of LPFD and prove its strong completeness and decidability. Then, we go a step further and explore the situation in cooperative games. In order to improve the expressivity of LPFD, we carry out its hybrid extension (HLPFD) and prove its completeness. In HLPFD, we can express an essential concept in the cooperative game: the core, and through it, we can represent our philosophical views in the context of cooperative games.

Finally, Chapter 6 concludes the main results of this dissertation, and points out a series of promising research directions, both from philosophical and logical perspectives.

1.3 Philosophical and Technical Preliminaries

Here, we will briefly introduce several basic technical terms in philosophy and technical preliminaries in logic in order to serve readers who lack a background in one or the other to understand the dissertation better. Only the most common and basic technical terms and their general meaning will be mentioned in this part; other terms will be introduced in the corresponding sections, and the technical part will not be covered again until Chapter 5.

1.3.1 Technical terms in philosophy

Ontology, epistemology, and methodology. Generally speaking, ontology refers to the philosophical study of being, as well as a series of related concepts, such as existence, entity, reality, becoming, fact, state, disposition, etc. Ontology divides all beings into different categories and studies their commonalities and differences. The main topic of the dissertation is collective agency, which belongs to social existence from the perspective of category, so our research is mainly related to social ontology. Social ontology studies the existence of various objects in social interaction, such as collective, social classes, institutions, corporations, laws, currency, race, gender, languages, artworks, etc., discussing their natures and properties, and questions like whether they exist, how to understand them, how they come into existence, etc. When the term "ontology" or "ontological" is used in the following chapters, it primarily means that the discussion refers to the existential level of the object, i.e., in the most abstract and universal sense, what kind of existence or universal nature is attributed to object under discussion.

In the general sense, epistemology is associated with the philosophical branch concern about knowledge, along with the study of the origins, scope, nature of knowledge and belief, epistemic justification, rationality in beliefs and preferences, and the interactive effect between intentions, beliefs, preferences, etc. Note that all these research subjects are within the cognitive realm of human agents. When we use terms such as "epistemology" or "epistemological" in the dissertation, we mean that the discussion is conducted only on the cognitive level and does not involve any ontological elements, i.e., the existence or reference of some concepts is only out of cognitive necessity or convenience.

Methodology, as its name implies, basically refers to the study of research methods. When we use the term "methodology" or "methodological," we simply mean that the object of discussion is only for the purpose of research presupposition or presentation, but irrelevant to any of its existence or nature in the ontological or epistemological sense.

Individualism and collectivism. Note that individualism and collectivism may have a broader range of application, but here we focus only on the domain of social objects. The individualistic view of social objects holds that only individuals exist, and that the existence of groups, companies, institutions, etc., is to be explained entirely in terms of the individuals from which they are composed. Collectivism, on the other hand, holds that these social objects have existence beyond that of the individuals that make them up. Of course, both of the above viewpoints in themselves are worked out in ways that displays

all kinds of ontological and methodological differences. And analogously, the collectivist view on the existence of social objects comes in a variety of different forms. Modest ontological collectivism, for example, only insists that social objects cannot be fully explained alone by the individuals who compose them. However, extreme ontological collectivism holds that social objects are more real than their individual members. Such a radical theory mainly originates from Hegelian idealism, which holds that the essential attribute of man is sociality. A rough argument is as follows: If the difference between man and animals is that man has rationality, self-consciousness, and moral sense, and these distinguished attributes are essentially derived from society, then the essential nature of man is sociality. As for rationality, its indispensable vehicle of language is a product of society. As for self-consciousness, it needs awareness of others as a precondition, which presupposes the existence of sociality. And as for morality, it always needs to be judged by other people.

Nowadays, some of these arguments have been refuted. For example, anthropological research shows that chimpanzees are also rational and self-conscious to certain degrees; Ethics distinguishes attributive responsibility and accountable responsibility, and the former is evaluated regardless of the social context. Today, the essential idea that modernity brings to the world, especially the West, is individualism. Nevertheless, we cannot claim that collectivism has been defeated, at least not its moderate forms. The following chapters will investigate how to choose between these positions and how to understand social objects.

Holism and reductionism. In popular terms, holism refers to the view that we can understand various systems only as a whole instead of as a mere collection of their component parts: in other words, a system is more than the sum of its constituent parts. Reductionism, on the other hand, claims that a whole is exactly its components' structure and nature, and we can understand it by analyzing its components. Many branches sprout from holism and reductionism; there also have the distinctions of ontological, methodological, etc. When we use the term "holism" or "holistic", it implies that we take a global view of systems and accept that objects in a system other than its constituent parts exist. When we use the term "reductionism" or "reducible", it means the object can exist or be analyzed simply by its structures plus parts.

1.3.2 Logic of functional dependence

As briefly mentioned above, there are differences between philosophical and formal studies of collective agency, and we attempt to capture their commonalities through the concept of dependence. Dependency is everywhere, from the abstract connections between data sets, to the laws of celestial bodies in the natural world, to the linkage between complex interactions in social life and the mental capacity of human agents. In the field of philosophy, the concept of dependence involves both ontological (cf. Tahko et al. (2020)) and causal discussions (cf. Menzies et al. (2020)). In the field of logic, there have been variant studies on dependence relations (cf. Galliani (2021)). Nevertheless, for the need to explain interactions between agents, we only focus on the most abstract, general, minimalistic sense of dependency. Suppose an arbitrary object Q depends on another arbitrary object P, it just means when P is fixed, Q is the same. In the physical world, we usually observe it as object P determines object Q. In the cognitive world, we usually perceive it as the cognitive content about P implies the cognitive content about Q.

To further illustrate such a basic concept of dependence, here we briefly introduce the syntax and semantics of the logic of functional dependence (LFD, Baltag et al. (2021)) in its original narrative. In section 5.2, we will reintroduce it in a game-context narrative for convenience of understanding.[①] From now on, the dependence relationships will be represented in the form of 'Q depends on P' or 'P determines Q'; they mean the same thing. Here is an example.

Example 1 *A simple information structure is given as follows, which includes four variables: p, q, r and s, and four values represent by different rows.*

p	q	r	s
0	2	0	0
1	2	0	1
2	1	1	0
3	0	1	1

We say a variable v globally depends on a variable u if, when the values of u are the same in two rows, then the values of v are the same too. From the table, we can see q depends on p, and r depends on q, and not the other way around. Also we have s does not depend on r, while $\{r, s\}$ determines p, $\{r, s\}$ determines q, $\{r, s\}$ determines $\{p, q\}$, etc.

① Recently, dynamic aspects of LFD have been studied in detail (cf. Baltag et al. (2022)). However, we will not review that since our concern does not cover the dynamic analysis yet, although combining the logic of dynamic functional dependence with repeated games is a promising direction.

The above-mentioned dependencies are global ones since they take all the values into consideration. However, we are also concerned about dependencies at the current values. Such a concern brings the concept of local dependencies: a variable v locally depends on a variable u if, every row matching the current value of u also matches the current value of v. For example, p locally depends on q at (2,1,1,0) but not at (0,2,0,0), q locally depends on r at (0,2,0,0) but not at (2,1,1,0), etc.

LFD is established for further capturing the properties of dependencies. It starts with a set of variables V and a relational vocabulary $(\mathsf{Pred}, \mathsf{ar})$, where Pred is a set of predicate symbols and $\mathsf{ar} : \mathsf{Pred} \to \mathbb{N}$ is an arity map, associating to each predicate $P \in \mathsf{Pred}$ a natural number $\mathsf{ar}(P)$.

Definition 1 (Dependence models, local dependences) *A FOL model is a pair $M = (O, I)$, where O is a non-empty set of objects and I is an interpretation map that assigns to each predicate $P \in \mathsf{Pred}$ a subset of $O^{\mathsf{ar}(P)}$. A dependence model $\mathbf{M}$ is a pair $\mathbf{M} = (M, A)$, where $M = (O, I)$ is a FOL model and $A \subseteq O^{\mathsf{V}}$ is a set of admissible assignments of objects to variables.*

For each $X \subseteq_{\aleph_0} \mathsf{V}$, we define a binary relation $=_X \subseteq A \times A$ such that $a =_X a'$ if and only if $a{\restriction}X = a'{\restriction}X$, i.e., the values of $x \in X$ in a is the same as its values in a'.

To capture functional dependence, LFD uses two operators $\mathbb{D}$ and D in its language.

Definition 2 *The language $\mathcal{L}$ of* LFD *is given by*

$$\varphi ::= P\vec{x} \mid D_X y \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathbb{D}_X \varphi$$

where $P \in \mathsf{Pred}$, $\vec{x} = (x_1, \ldots, x_n)$ is a finite string of variables of length $n = \mathsf{ar}(P)$, $X \subseteq_{\aleph_0} \mathsf{V}$ is a finite set of variables and $y \in \mathsf{V}$ is a variable.

$\mathbb{D}_X \varphi$ says that whenever the variables in X take their current values, φ is the case; $D_X y$ says that whenever the variables in X take their current values, y also takes its current value.

Definition 3 *Truth of a formula $\varphi \in \mathcal{L}$ in a dependence model $\mathbf{M} = (M, A)$ at an assignment $a \in A$ is defined as follows: ($\mathsf{set}(\vec{x})$ for the set $\{x_i : i \in I\}$)*

$$\begin{array}{lll}
\mathbf{M}, a \vDash P\vec{x} & \textit{iff} & a(\vec{x}) \in I^M(P) \\
\mathbf{M}, a \vDash D_X y & \textit{iff} & a(y) = a'(y) \textit{ for all } a' \in A \textit{ with } a =_X a' \\
\mathbf{M}, a \vDash \neg\varphi & \textit{iff} & \mathbf{M}, a \nvDash \varphi \\
\mathbf{M}, a \vDash \varphi \wedge \psi & \textit{iff} & \mathbf{M}, a \vDash \varphi \textit{ and } \mathbf{M}, a \vDash \psi \\
\mathbf{M}, a \vDash \mathbb{D}_X \varphi & \textit{iff} & \mathbf{M}, a' \vDash \varphi \textit{ for all } a' \in A \textit{ with } a =_X a'
\end{array}$$

Note that $=_X$ is an equivalence relation on A and $a =_\varnothing a'$ holds for all $a, a' \in A$. So $\mathbb{D}_\varnothing$ is a universal operator and we define $\mathsf{A}\varphi := \mathbb{D}_\varnothing\varphi$ and $\mathsf{E}\varphi := \neg\mathsf{A}\neg\varphi$. With such a language, we can express some results in Example 1, for instance, we have $(2,1,1,0) \models D_q p$, $(2,1,1,0) \not\models D_r q$, $(0,2,2,0) \models D_r q$ and $(0,2,2,0) \not\models D_q p$, etc.

We will return to the logical work in Chapter 5, where we expand this logic to express several essential concepts in game theory.

CHAPTER 2 A RELATIONAL PERSPECTIVE ON COLLECTIVE AGENCY

The discussion of collective agency involves the reduction problem of the concept of a collective. Individualism and Cartesian internalism have long restricted orthodox theories and made them face the tension between an irreducible concept of a collective and ontological reductionism. Heterodox theories as functionalism and interpretationism reinterpret the concept of agency and accept it as realized on the level of a collective. In order to adequately explain social phenomena that have relations as their essence, in this chapter we propose a relational, holistic account of collective agency and argue that functionalism and interpretationism can be integrated into such an account.

2.1 Introduction

In the broadest sense, the term "agent" refers to those who have the ability to act. Human agency further requires the capacity of rationality and its manifestation. As the name suggests, collective agency means realizing this capacity at the collective level. The interpretation of agency involves a core concept in philosophy of action: intentionality.① Similarly, at the collective level, the interpretation of collective agency involves a parallel concept: collective intentionality.

Research on collective intentionality has flourished in recent decades. Compared with the traditional social science that insists on individualism (see Popper (1962)), research on collective intentionality reflects a strong tendency towards non-individualism. The work of Michael Bratman, Margaret Gilbert, Raimo Tuomela, John Searle and others has considerably broadened our horizons in social philosophy and made people generally accept that intentionality can be shared in social actions. However, in these approaches, the influence of individualism still lingers and that leads them to regard individuals as the cornerstone of the entire theoretical edifice and bearers of collective intentions. And with that, they encounter difficulties facing the tension between an irreducible concept of collective intentionality and ontological individualism.

Recently, some alters have modified the definition of agency and relaxed certain re-

① Interpreting agency using the concept of intentionality is the mainstream tradition in philosophy of action since G. E. M. Anscombe and Donald Davidson.

strictions, for example that any intentional subject should meet the same physical or psychological standards as individuals do, thereby making the existence of collective agency possible. Two notable forms of realism with respect to collective agency are interpretationism (cf. Rovane (1997); Tollefsen (2002)) and functionalism (cf. List et al. (2011)). From the perspective of interpretationism, as long as a system behaves in a way that can be interpreted as rational by others, it can be regarded as an agent. Interpretationism focuses on organizational structures of collectives and believes that these structures guarantee the possibility of rational perspectives on collectives. On the other hand, functionalism believes that agency is embodied in the agent's ability to perform actions that meet certain functional conditions, such as representation, motivation, and rational capacity. Analogously, functionalism characterizes collective agency in terms of its action results and explains collective agency in terms of reliable aggregation of converging opinions of the members of a collective to ensure that a collective's opinions are rational.① In this chapter, it is argued that these two different perspectives on realism of collective agency can be integrated into a relational account.

Various authors (cf. Baier (1997); Meijers (2003); Schmid (2003)) have claimed that we need a relational or holistic explanation of collective intentionality. In line with their arguments, we will demonstrate that collective agency essentially is a combination of relational eventualities② of a collective consisting of members of a collective, the targets of collective actions, structures and functions of a collective. Moreover, for each collective agent, these relational eventualities constitute a relational identity; not a physical existence, such as a table, but a non-physical existence such as friendship, policy or a nation-state. Basically, our claim is that this relational identity cannot be reduced to a simple aggregation of its members, plus the targets of collective actions, plus structures, plus functions.

Before starting the discussion, two premises need to be stated: 1. we only focus on collectives containing specific institutional structures or functions; in other words, a simple collection of individuals does not make a collective; 2. our discussion does not involve spontaneous behaviors, which, in our opinion, constitute an aspect of individual intention-

① The "rational" here indicates logical consistency of a collective's decisions; that is, it is irrational for a collective agent to agree with A and B but to deny $A \wedge B$ at the same time. More detail on interpretationism and functionalism can be found in Section 2.4.

② Following Bach (1986), we use the term 'eventuality' to refer neutrally to events and states. Note that the reference to events or states instead of facts emphasizes the importance of the temporal dimension as a constitutive element. As a set of relational eventualities, collective agents can originate, develop and vanish. In other words, collective agents are not constants but relatively stable states of relations.

ality, which concerns the accessibility and interrelation of the mental and material worlds. However, in this chapter, we focus on the issue of collective intentionality that is on the relation between individual minds and social phenomena.

The structure of the chapter is as follows. In the next section, we will review orthodox theories of collective intentionality and explain why they cannot avoid what is called the "circularity problem". Furthermore, we will discuss why most philosophers are resistant to the existence of collectives in Section 2.3. Specifically, we will start with an objection to the basic argument of physicalistic individualism by showing that the very same argument does not exclude the existence of collective. In Section 2.4, we will return to the existing forms of realism with respect to collective agency and integrate them into a relational account. In Section 2.5, we will respond to the objection that the existence of collective agency will harm individual agency and discuss how collective agency affects individual actions practically in a relational view. A summary and possible future directions are included in the conclusion.

2.2 Irreducible Concepts of Collectives

As Petersson (2007), Schweikard et al. (2021) have pointed out, orthodox theories of Gilbert, Tuomela, Searle and Bratman are trapped in the circularity problem because they persist on individualism in varying degrees. At least two types of individualism need to be distinguished. Ontological individualism holds that a collective is nothing more than a complex of individuals and their actions. Methodological individualism insists that social phenomena can only be explained by showing how they arise from individual actions and the intentional states that underlie and motivate those actions.

Generally speaking, all these theories adhere to ontological individualism. Among these, Gilbert, Tuomela and Searle admit the existence of an irreducible concept of a collective in a methodological sense to some extent. In Gilbert, it is "joint commitment"; in Tuomela, it is "we-intention".[①] Searle further binds the concept of "we-intention" to individuals' physical brains or even brains in vats, manifesting an apparent tendency of internalism. Bratman goes even further by insisting that we can explain collective intentionality without any irreducible concept of a collective; that is, any collective intentionality can be explained by its members' individual intentionality and interdepen-

① It is generally accepted that Wilfrid Sellars' thought of we-intention (cf. Sellars (1963)) was the source of Tuomela's original analysis of we-intention (cf. Tuomela et al. (1988)).

dent relationships. Using these characteristics, we will elucidate how they fall prey to the dilemma of circularity.

2.2.1 Ontological individualism

Margaret Gilbert holds that activities of a collective require the collective to take on a "joint commitment", which in her works means "a kind of commitment of the will. In this case, the wills of two or more people create it, and two or more people are committed by it" (Gilbert (2006), p.134). Such a joint commitment implies a normative relationship among members that each has an obligation to act accordingly and demands others' conforming actions. She further defines "plural subject" as people "are jointly committed to doing something as a body" (Gilbert (2006), p.145). Although Gilbert claims we need to transcend individualism (cf. Gilbert (2000)), she restricts the conceptual basis of "joint commitment" to include only conditional personal commitments (cf. Gilbert (2002)) and explicitly bases her analysis on "the concept of an individual person with his own goals, and so on, does not require for its analysis a concept of a collectivity" (Gilbert (1992), p.435).

Among various criticisms of Gilbert's theory, there is a particular criticism that claims that Gilbert faces the danger of circularity. It asks how a joint commitment to jointly perform an action can be formed without presupposing the concept of a joint action that is meant to be brought about by the plural subject; for, if we say individual agents A and B jointly commit to have a meeting, the concept of the joint action of their meeting is presupposed. It seems that joint commitment presupposes rather than explains the concept of a joint action it is intended to analyze. A possible reply could be that before A and B have the joint commitment and act as a body, each of them is already willing to do so. However, such a reply is not enough because such circularity reappears at the level of willingness. Nevertheless, her theory of creating joint commitments is a normative description rather than a conceptual analysis (cf. Petersson (2007)). Thus, it would not be proper to criticize her theory with circularity problems.

Regarding Tuomela's theory, it distinguishes the form of individual attitudes in "we-mode" versus "I-mode", as "notions of having an attitude and acting as a group member versus as a private person" (Tuomela (2007), p.46). Moreover, he defines "we-intention" as follows: an individual agent has the we-intention to do X if and only if they have a respective individual intention to do X and have certain beliefs about the actions and beliefs of other participants, and there is a mutual belief among the participants about this.

"When there is shared we-intention in this sense we can speak of at least a weak joint intention, one in the I-mode. In the we-mode case we can say that the participant intends to participate in X at least in part because the group intends to perform X" (Tuomela (2006), pp. 41–42). Tuomela's analysis of (we-mode) "we-intention" can be briefly characterized as follows: A member A of a group G has we-intention to do action X if and only if (1) A intends to do their part of X;[①] (2) A believes that others in G will do their parts; (3) A believes that there is a mutual belief among the participating members of G; (4) (1) in part because of (2) and (3) (Tuomela (2006), pp.42–43; Tuomela (2005), pp.340-341).

Observe that although Tuomela clarifies that "my analysis is not meant to be reductive but is rather meant to elucidate the irreducible notion of we-intention in a functionally informative way" (Tuomela (2005), p.358), he also explicitly mentions that "the present approach is ontologically individualistic in the sense that we-mode states and properties are attributed to individuals, severally or jointly, when they function as group members" (Tuomela (2013), p.93). In other words, he believes that although we have a special kind of attitude—we-mode—and a special kind of intention—we-intention—this does not presuppose any existence of collective entities as its bearer.

Tuomela's theory also faces a problem of circularity. By his analysis, if an agent A intends to do their part of X, and has we-intentions that satisfy the conditions (1)–(4) above, it still presupposes a concept of a collective that intended to be constituted by we-intention. So, "we-mode" and "we-intention" do not constitute the concept of a collective, but rather presuppose its existence. In (Tuomela (2005), pp.355-361), he replies that his account does not contain any vicious circularity since the analysis means that a collective is consisted in, instead of constituted by, the participants' having we-intention. However, this is not satisfactory since it is just a verbal way out; he still needs to explain what "consisted in" is.

2.2.2 Ontological individualism with internalism

Searle supports the idea of "we-intention" and rejects Tuomela's analysis of it; he says, "this account is typical in that it attempts to reduce collective intentions to individual intentions plus beliefs. I, on the contrary, am proposing that no such reduction will work, that 'we-intentions' are primitive" (Searle (1990), p.404). However, meanwhile Searle insists that "all intentionality, whether collective or individual, has to exist inside

① "This is a we-mode personal intention essentially of the kind 'I, as a group member, intend to participate"' (Tuomela (2006), p.44).

individuals' head" (Searle (2010), p.44). To Searle, the primitive, irreducible "we"-form is the core feature of cooperation, yet this concept of a collective can only exist in the participants' heads and motivate their contributions. Although insisting on an irreducible concept of a collective, he claims that his theory adheres to methodological individualism (cf. Searle (1990)). In addition, the above conception of individuals as just brains has a strong flavor of internalism.

The combination of internalism and ontological individualism makes his theory more vulnerable to criticism. In addition to the circularity problem, Searle also has to accomplish a seemingly impossible thing: to explain collective actions, which are characterized as interpersonal, in an internalistic way. Moreover, it remains an open question how an irreducible, primitive "we"-form emerges in the individual brain.

2.2.3 Ontological individualism with methodological individualism

Bratman executes a thorough reductionistic analysis in which a collective intention can ultimately be reduced to interrelated individual intentions (cf. Bratman (2006)). He suggests that each participant agent can intend not only their respective contributions but also a joint activity J as such. His analysis runs as follows: "We intend to J if and only if 1. (a) I intend that we J and (b) you intend that we J; 2. I intend that we J in accordance with and because of 1a, 1b, and meshing subplans of 1a and 1b; you intend that we J in accordance with and because of 1a, 1b, and meshing subplans of 1a and 1b. 3. 1 and 2 are common knowledge between us" (Bratman (1999), p.121). Although Bratman claims his theory can radically reduce collective intentionality to the individual level, many situations with a concept of a collective do not seem to yield to his reduction.① In addition, the collective concept "we J " is not fully reduced in his analysis: it reappears at the individual level.

The criticism of circularity is brought against Bratman's work and takes the form of a question: how individuals can refer to joint activity in the absence of "jointness"? How can an agent assume that others will intend as they do without assuming there is a collective intention? It seems individual intentions, as he conceives of them, may not be a sound ground for shared intention but rather a superfluous expression of existing collective intentionality. Bratman replies that in order to achieve "jointness", members' intentions also need to satisfy the following conditions: "each is appropriately responsive to the other;

① For instance, the example of business school (Searle (2010), pp.47–48), we will go back with more details to this in section 3.2.

each is set not to thwart the other and is at least minimally disposed to help the other if needed; there is (feasibility-based) persistence interdependence in relevant intention; there is, as a result of all this, rational pressure in the direction of social rationality; each expects all of this to issue in the joint action each intends; and all this is out in the open" (Bratman (2014), p.95). These conditions are, in his terminology, a form of interconnected functioning between the intentions and actions of the participants. However, again, this interconnected process of intention formation presupposes the concept of a collective it is meant to constitute.

2.2.4 The Gordian knot

As mentioned above, centering on 'joint activity', 'we-mode', or 'joint commitment' while retaining the 'I'-form of the constitute individual intentions will face the danger of circularity. However, this is not the crucial difficulty of individualism, in our opinion.

Humberstone (1997) distinguishes analytical circularity from inferential circularity and claims that the latter is generally vicious, but the former is not. Suppose the general form of an account of a concept K runs as follows: the concept K applies if and only if certain conditions $C_1, C_2, \ldots, C_n$ are obtained. Such an account can be thought of as a putative analysis of K. Viewed as such, an account is analytical circular if the concept K is employed in specifying the conditions $C_1, C_2, \ldots C_n$. Alternatively, we can also think of such an account from the perspective of the bi-directional claim "if and only if." The direction of "only if" just provides necessary conditions for applying the concept K. When it comes to the direction of "if", the situation is different in a crucial sense. As Humberstone formulates it: "the kind of circularity pertinent here will not be like the circularity of a putative analysis or definition, but rather, like the circularity of an argument" (Humberstone (1997), p.250). That conditions $C_1, C_2, \ldots C_n$ obtain are premises of this argument, and that concept K applies is its conclusion. Thus "an account of the application some concept is inferentially circular when any argument or inference from premises claiming the various conditions provided by the account to obtain, to the conclusion that the concept applies, is itself circular in whatever sense an inference or argument can be circular" (Humberstone (1997), p.250). A more usual way to explain the flaw of circularity is this: if a concept is being explained, the explanation should not be understood only by those already possessing the concept. In this sense, the analytical circularity is deficient in the fact that it impedes the transition from an understanding of the terminology in an account to an understanding of concepts. In contrast, the flaw of the inferential circularity

is not only at the level of understanding but also at the level of knowledge: people cannot obtain any new knowledge from a definition that is inferentially circular.

Under such a distinction, Petersson (2007) mentions that the above analyses of acting with collective intentions face analytical circularity. For example, if we say one of the necessary and sufficient conditions for "I push the car intentionally" to be true is that I know I push the car intentionally; that is the case of inferential circularity. Because if I know I push the car intentionally, I already push the car intentionally. Such an analysis may not contribute to our understanding of the term "intentional". However, if we say one of the necessary and sufficient conditions of "we push the car collectively" to be true is that our pushing the car is explained by "joint commitment", "we-intention" or "we-$\boldsymbol{J}$" and further involves each individual's intending on pushing the car collectively, although the concept analyzed is implicitly present in components of the analysis, we do obtain further knowledge compared to the single statement of "we push the car collectively".

In line with Humberstone and Petersson's arguments, we also think the so-called circularity problem these orthodox theories encounter is not severe, since they (at most) presuppose the existence of a concept of a collective instead of the actual existence of a collective. After all, concepts can exist without any assumption that there is something ontological that corresponds to them. Thus the object being explained is different from the one being assumed conceptually, and we do obtain new knowledge.

Furthermore, we believe that the true Gordian knot lies in the attempt to explain collective agency or intentionality solely in terms of individual agency or intentionality. In such an attempt, an irreducible concept of collective agency remains at the individual level, and then the question of where that concept comes from arises. If there is, in fact, no such thing, as ontological individualism claims, the concept of a collective that individuals have should be considered as nothing but an illusion. However, we still need to explain why that illusion exists. The crucial point is why this illusion of a collective plays an essential role in collective actions. The three available answers to the question of the essential role of the concept of a collective are non-satisfactory; they are (1) claim that they are mere illusions; (2) acknowledge that the concept of a collective that cannot be reduced; (3) reduce the concept of a collective to the individual level to dismiss it. The "non-satisfactory" here means that they may not contribute to our understanding of the essential role of the concept of a collective. If we admit that answer (1) is irrelevant to our understanding and that the reduction of approach (3) cannot be completed (as in

Bratman's theory, "we J" reappears at the individual level), only (2) is left to be elaborated. Suppose we put "Acknowledge the concept of a collective that cannot be reduced" as one of the necessary and sufficient conditions for, "The concept of a collective plays an essential role in collective actions". Then it involves inferential circularity rather than merely analytical circularity. This is because, by referring to concepts with irreducibly collective content, we already presuppose that some kind of conception of collective action plays an essential role without providing further analysis of it. In other words, we cannot obtain any new knowledge from answer (2). The whole individualistic explanation fails to provide an understanding of the concept of a collective and its source. This is the crucial challenge for individualistic accounts, especially for those that admit an irreducible concept of a collective. The relevant circularity is not an analytical one but an inferential one. To cut this Gordian knot, we need to explore its source related to our ontological assumptions, which will be discussed in the next section.

2.3 Why Can't We Speak of Collective Entities?

Before presenting our point of view in Section 2.4, we first address the question of why individualism has such a broad audience. Baier (1997) points out that a Cartesian specter has haunted the field of collective agency for a long time and has led to the existing individualistic theories. Meijers (2003) and Schmid (2003) continue to develop this view and refute any individualistic tendencies in the discussion of collective agency. We believe that resisting this Cartesian influence will shed new light on our breaking away from the shackles of the physical bearer to make it possible to acknowledge the existence of collectives. We will illuminate this by reformulating the basic argument of physicalistic individualism and showing that it cannot exclude the existence of collectives.

2.3.1 The basic argument of physicalistic individualism

People always regard intentionality as a mental state possessed by an individual. Correspondingly, we intuitively find it is challenging to accept collective intentionality if this indicates a certain spiritual body of a collective: a Hegelian spirit of a collective possesses such collective intentionality, just as an individual possesses their own intentionality. The Cartesian specter plays a decisive role in this. In Meditations, Descartes' argument of "Cogito ergo sum" constructs the concept of an independent individual mind, which is isolated from the social and natural world. In his narrative, intentionality is not a

property of interpersonal and multi-subjective types of entities. As formulated by Schmid, this tradition brings about a kind of "formal individualism", which restricts intentionality to an 'I'-form and rejects any plural form of intentionality.

Bratman's reductionist theory presents firm ontological individualism plus methodological individualism. Searle's strategy of looking for brains in vats as bearers is obviously internalistic. Gilbert, Tuomela and even Sugden, Gold, etc., (in the economics tradition of team reasoning) adopt a modest approach: they accept concepts of collectives but still insist on ontological individualism. The key observation is that these accounts avoid attributing any concept of a collective (whether reducible or not) to a non-individual collective entity. The reason behind this is that all the authors accept that nothing at the collective level can serve as a vehicle for concepts of a collective. Per contrast, the individual in a physical sense is a perfect vehicle, so in addition to individual intentionality, they ascribe part or all of collective intentionality to the individual, or more directly, to the individual brain. Note that there is a strong physicalist or materialist streak behind each of these theories, even for those who take a modest approach; without loss of generality, we may name such a type of theory more properly as "physicalistic individualism".

Both Cartesian internalism and contemporary physicalism lurk behind individualistic accounts. For narrative convenience, here we briefly reconstruct the basic argument of contemporary physicalistic individualism,[①] which consists of an internalistic argument of others' having intentionality and a reinforced physicalistic condition. The former contains three premises (P1, P2, P3) and a conclusion (C1):

(P1) Internalism in principle: the justifications of our cognitive contents are independent of the outside world.

(P2) Self-awareness: everyone is familiar with themselves and can be aware of their physical body and their cognitive contents, and understand the correlations and differences between them.

(P3) Awareness of others: we can see and touch other people (in a physical sense), although we cannot see and touch cognitive contents of other people, we can be aware of such things. And we can also have the concept of similarity and difference with other humans.

(C1) With (P2) and (P3) as basic intuitive facts, each individual "I" knows themselves well and is aware of others. This "I" would see others as

① For the purposes of this chapter, we will simplify the Cartesian claim by focusing only on internalism and leave out solipsism, i.e., we will take the existence of other individuals for granted.

familiar kinds of stuff and attribute similar cognitive contents to them. According to (P1), "I" do not need any external source to justify knowledge (internalism). Therefore, every individual has the concept of other people having intentionality, independent from any actual connection with such intentionality.

Based on the facts of (P2) and (P3), Cartesian internalism acknowledges the awareness of others having intentionality.

The conclusion (C1) easily fits with physicalism and further strengthens it to a claim of physicalistic individualism. The reinforced physicalistic condition is as follows (P4):

(P4) Physicalistic function of brains: all intentionality requires a physical brain, and in the end is a physical existence.

(C2) With (C1) as the basic conclusion, according to (P4), such intentionality depends only on physical brains, which can only own by individuals. In other words, intentionality only exists on the individual level (me and other persons, even intelligent animals), which is a typical individualistic claim.

There are at least three aspects of this argument of physicalistic individualism that can be thought of differently. Of course, the first is directed at the condition (P1) and is adopted by the authors that oppose the Cartesian tradition mentioned at the beginning of this section. In this way of thinking, theories of collective agency do not succumb to those individualistic constraints and are not limited to dealing with the tension between irreducible concepts of collectives and ontological reductionism. It allows us to embrace a more general relational, holistic account that can explain collectives through primitive relational concepts. The second alternative view is connected with (P4) and argues that a system could have intentionality or agency without any physical bearers and that collective agency can exist in the sense of interpretationism or functionalism.[①] We will come back to these accounts in detail in Section 2.4.

Nevertheless, we think these two alternative ways of thinking are not sufficient to turn our attention to the relational account because they are alternatives to, instead of

① Broadly speaking, in contrast to physicalism and reductionism, emergentism recognizes the existence of collective agency and asserts that such an entity can not be reduced to its members and their interactions. In this sense, interpretationism and functionalism can be regarded as specific ways of explaining emergentism by relaxing several restrictions on agency from their particular perspectives. However, in detail, their concerns are different from those of emergentism; they pay more attention to the way we interpret conceptual objects rather than whether they are reducible to other entities.

refutations of, physicalistic individualism itself. Even if we claim that individualism will face intractable theoretical difficulties and appeal to relationalism or holism to explain collectiveness, individualists can still find a way to escape, such as relaxing restrictions of methodological individualism (as in the work of Gilbert and Tuomela) and claiming that the relational account is nothing but another narrative.

2.3.2 Objection to familiarity

In our opinion, the real problem with physicalistic individualism is that the logical pattern behind it is invalid. That pattern is: (1) "I" am familiar with myself as an agent; (2) as an agent "I" have intentionality; (3) in the same way as with myself, "I" am familiar with others as agents; (4) thus other agents also have intentionality; (5) and those agents that "I" am not familiar with do not have intentionality.

The problem here is that familiarity requires a premise that the objects being compared are of the same type. At first glance, it certainly seems that "I" and other people belong to the same type of human being. However, knowledge of these two is of different types. Our knowledge of ourselves is intuitively distinct from our knowledge of external objects.① Instead of the actual existences of "I" and others, what the internalist agent compares are two different types of knowledge of those existences. Thus, if this distinction holds, the extension from first-person to third-person having intentionality is problematic;② in other words, familiarity does not apply here.

In the argument of physicalistic individualism, the sentence in (C1) "This 'I' would see others as familiar kinds of stuff and attribute similar cognitive contents to them" follows the logical pattern of familiarity. The result of mistakenly applying familiarity to two different types of knowledge is to accept a strict definition of an agent, i.e., all agents must have the exact attributes and organizational structures as the canonical agent such as myself, including necessary physical bearers. Hence, physical brains become a necessary condition for collective agency. Nevertheless, this is absurd. Our awareness of collectives and other people's intentional states is of the same invisible and untouchable

① Here, we merely point out that there is an intuitive difference between knowledge of oneself and knowledge of the external world and that this difference affects the application of familiarity. It is beyond the scope of this chapter to discuss in what sense this distinction holds. For interested readers, see Gertler (2021) for various accounts of the distinction between the two types of knowledge.

② In order for internalism to not collapse into solipsism, it needs to satisfy two conditions: 1. the subject regards perceptual inputs as evidence; 2. the subject has the criteria to distinguish between knowledge of oneself and knowledge of others. Therefore, the distinction between the two types of knowledge is still meaningful in the context of internalism, which affects the application of familiarity.

objects. Yet, just because "I" am familiar with a particular relationship between my physical brain and my cognitive contents, "I" takes for granted that the physical brains of others also have their cognitive contents. However, since a collective does not have a physical brain, although "I" can be aware of a so-called collective, "I" firmly believes it does not exist and cannot have intentionality because "I" am not familiar enough with it compared to a canonical agent like myself.

Based on the distinction of two types of knowledge, we do not need to stick to the definition of an agent in the strict sense. This means our understanding of an agent can be loosened to a level that an agent is only assumed to have sufficiently similar properties and functions as canonical agents; then, we do not need to seek any physical brains for collective agents.

Now the severe problem is that if we abandon the logical pattern of familiarity, we can declare collective existence through a similar argument without disposing of the premise of Cartesian internalism:

(P1) Internalism in principle: the justifications of our cognitive contents are independent of the outside world.

(P2) Self-awareness: everyone is familiar with themselves and can be aware of their physical body and their cognitive contents, and understand the correlation and difference between them.

(P3*) Awareness of collectives: although we cannot see collectives (physically), any person-in-society does perceive the existence of collectives, at least in the conceptual sense. Moreover, a person-in-society can be aware of a different kind of intentionality than that in a collective sense.①

(C*) With (P2) and (P3*) as basic intuitive facts, each individual agent "I" knows myself well and is aware of collectives. Since the definition of an agent no longer requires physical grounds, "I" have no reason to reject the existence of collectives, at least in the conceptual sense. In this

① Many experimental studies in cognitive science and anthropology confirm that infants about 12 to 15 months of age are capable of having collaborative engagements. They can exercise their individual intentionality and recognize the intentionality of others, and more importantly, they can realize the cooperative behavior and the role of others in collaborative engagements and can take in and complete the tasks of other roles. (See Bakeman et al. (1984); Ross et al. (1987); etc.) With the observations that gorillas cannot have collective intentionality and that autistic children cannot fully have collective intentionality, Tomasello et al. (2005) further suggests that the most important distinction between humans and other species is that we understand collective intentionality and can cooperate, rather than merely being rational and intentional.

> way, "I" will realize that such concepts arise from the manifestation of some counterpart of the physical world different from individuals, to which "I" would naturally attribute collective intentionality. According to (P1), justifications of my knowledge are independent of the outside. Therefore, collectives may exist and have their intentionality, and such intentionality depends on something different from individuals; in other words, intentionality can be realized at the level of a collective.

It is an apparent problem that internalism will make our cognition completely independent of any relation that refers to the external. In fact, concerning existence in the ontological sense, internalists cannot obtain any practical knowledge except of themselves. Therefore, the argument for others having intentionality also applies to a collective. In addition, physicalists think that they can find physical grounds for others' intentionality. Although we cannot see or touch collectives and thus cannot find a solid counterpart of collective intentionality, by internalism, our awareness of the existence of concepts of collectives would suggest that we ascribe collective intentionality to the ontological source of such concepts. Suppose there is a so-called dependency between the cognitive contents and the physical body of each individual (and such dependency is essentially an assumption of analogy to myself and may not be confirmed at all). In that case, collective intentionality can also depend on the source of concepts of collectives instead of physical brains.

If the above argument is valid, we may conclude that physicalistic individualism does not bring any substantive constraint. With the premise of internalism and a correct understanding of familiarity, we can conclude that collective intentionality can be attributed to the non-individual source of irreducible concepts of collectives. And that conflicts with individualism. It needs to be pointed out that such an objection against familiarity does not necessarily lead to collective existence. Philosophers who are worried about the collective specter can rest assured. It can only adequately show the absurdity of physicalistic individualism. Adhering to it can neither solve the problem of circularity nor naturally explain collectiveness. This suggests that it may be worthwhile to investigate whether a non-individualistic, relational and holistic account that is more suitable for explaining social phenomena should be adopted. Under such a relational account, we believe that collective intentionality depends on the collection of all the relational eventualities that

generate it.

2.4 The Relational Account of Collective Agency

Orthodox theories that adhere to individualism are based on an intuitive argument about intentionality: whether a subject has an intention is based on whether it has a mind. However, two heterodox theories that acknowledge the existence of collective agency and collective intentionality emerged: functionalism and interpretationism.

2.4.1 Functionalism and interpretationism

The classic functionalist approach of agency believes that a system manifesting agency only requires its behaviors to instantiate the agential model, which is represented as its behaviors meeting some necessary functional conditions. According to contemporary functionalist theory, the mind is an entity that can operate through its functional organization. In a holistic view, each part of the entity is defined by its role in the entire system, and these roles can be multiply realized.① Specifically, Pettit and List believe that a functionalist account of agency should abandon physical constraints: agents must be composed of certain material substance that is the same as that of canonical agents such as humans or animals. Additionally, they believe we should also abandon psychological mode constraints: the process by which the agent forms its inclination to act must be realized through a certain psychological model, such as consciousness. Moreover, a functionalist account should turn to the only condition of ratiocinative capacity: as long as the subject's actions reflect its capacity of rationality, it can be identified as an agent. "The system has to be able to identify some of the demands imposed by the pattern② as regulative or normative requirements, and to let the identification of those demands reinforce conformity and underpin the recognition of non-conformity as a failure" (Pettit (2009b)). Therefore, their theory relaxes the restrictions on agency and maintains that as long as collective actions reflect the collective's rational ability, the collective can have intentional

① A common example is pain. Functionalists may interpret pain as a physical injury, a belief in physical discomfort, a desire to express such discomfort, moaning, restlessness, etc., and assume that anything that satisfies these conditions is capable of being pain. However, in this account, it is not just humans or animals that have internal mental states that are capable of pain, but also those that meet these conditions, even silicon-based creatures, aliens, and other completely different types, can be capable of pain. This is the so-called multiple realization: "pain can be realized by different types of physical states in different kinds of creatures" (Levin (2021)).

② The "pattern" refers to the agential pattern in functionalism, which is the basic condition for behaviors of a system as an agent needs to satisfy. "There are purposes and representations that it is independently plausible to ascribe to the system... and the behavior of the system generally promotes those purposes according to those representations" (Pettit (2009b)).

states and agency. On this basis, their theory examines the issue of how to achieve rationality in collective decision making in voting and claims that if multiple decisions of a collective are logically consistent,① they reflect rationality and, by their definition, can be treated as a collective agent (cf. Pettit (2007); List et al. (2011)).

On the other side, interpretationism maintains that if we can interpret an agent, then the agent is intentional,② and it asks: what assumptions do we need to explain agential behaviors? If we assume that certain premises and successfully explain agential behaviors in terms of them, then our assumptions about agents can be justified, thereby determining the nature of agents. The common assumption is that an agent should obey rational norms, i.e., is rational. Such an assumption is basic since we cannot understand and interpret any agent until we make it. Contemporary interpretationism follows a view called Neo-Lockeism: personhood means to keep a rational first-person point of view. They believe that this first-person perspective is not related to the unified consciousness that Locke originally insisted on, nor is it related to souls or brains, but is only a perspective that can be applied and understood by others. This rational perspective does not need to be bound to any individual soul or animal (cf. Rovane (1997)). Based on the assumption of a rational perspective, interpretationism attributes beliefs, desires, and intentions to agents according to their environment and functions and can further predict agents' behaviors based on those attributed intentions. Furthermore, "these attributions are not made in isolation but holistically" (Tollefsen (2002)). Because there is no link between physical and spiritual aspect, contemporary interpretationism recognizes the existence of collective agency, in the sense of realizing the rational first-person point of view at the collective level. In its analysis of collective intentions, the structure of a collective is regarded as an essential component. "The structure of the organization provides a way of synthesizing the disparate perspectives of individuals into a unified perspective from which goals and subgoals can be set and achieved" (Tollefsen (2002)). Individuals can think and act intentionally in the sense of "collectively" by adopting the collective perspective. Interpretationism believes that a collective is rational, if and only if, it has a certain structure that enables it to achieve its goals effectively. When this hypothesis is established, interpretationism "attributes beliefs, intentions, and desires to organizations the same way we

① For example, there are three propositions (a), (b) and (c) with interrelations: the conjunction of (a) and (b) implies (c), namely "(a) Take preventive measures against COVID-19", "(b) Fighting the epidemic should have the highest priority, whatever it takes", and "(c) Accepting financial loss." According to functionalism, a collective that manifests agency can not simultaneously adopt (a) and (b) but reject (c) because it will be logically inconsistent and indicates the collective cannot exercise its ratiocinative capacity.

② Donald Davidson and Daniel Dennett are prominent in this tradition.

do to individuals" (Tollefsen (2002)).

Interpretationism is a mild realism of agency. It uses a hypothesis-inference approach to judge what assumptions are necessary in a roundabout way. In interpretationism, we can directly describe the attribution of beliefs, desires, intentions to collective agents. On the other hand, functionalism is more a kind of behaviorist judgment about agency. It judges whether a collective has agency by determining whether its actions meet necessary functional conditions. Because functionalism relies only on observable actions, it can be well integrated with specific approaches in the social sciences such as game theory and social choice theory.

2.4.2 Turn to a unified account

In this section we propose a new perspective that starts from the observation that interpretationism and functionalism are different but compatible and can be integrated into a unified account. In our opinion, the most significant difference between them is in their research perspective. In addition to relaxing the physical and psychological constraints on agency, contemporary theories of interpretationism and functionalism loosen the constraints on agency from two aspects: interpretationism relaxes the constraints on the conditions that behaviors of a collective should satisfy; functionalism relaxes the constraints on the internal structure of a collective. If our understanding is correct, it is possible to integrate them into one unified account in which the essence of collective agency is directly revealed.

Our account is based on the observation that we all have no problem accepting and using concepts of collectives in our daily lives, for instance, companies, institutes, nation-states, etc., and also that we have the ability to attribute collective beliefs, desires, etc., to collective agents, as well as to predict and expect their behaviors. For example, "The soccer team has the determination to win"; "The Dutch government believes herd immunity is more effective in fighting COVID-19"; etc.

We want to argue that the ontological sources of these facts are relational eventualities. Regarding collective agency, functionalism and interpretationism deal with two aspects of the relations across a collective, namely the internal relations of a collective (collective structure, that is, the interaction between its members) and the external relations of a collective (collective function, that is, the interaction between the collective and the outside world).

We believe that a collective agent is essentially a collection of relational eventualities.

Specifically, this collection of relational eventualities includes relata (members of a collective and the targets of collective behaviors) and relations (a collective's internal structures and external functions). These relata and relations constitute a recognizable identity of a collective agent, which can then play a role in our decision-making. The irreducible collective intentionality is assigned to its members through internal relations, and an individual can choose to take part in the intentionality of different collectives depending on what net of relations they finds themselves in. For example, an agent could be in multiple collectives simultaneously, such as their company, family, fishing club, sports team, etc. Each collective assigns collective intentionality to its members through its unique structure, and the individual-in-relation has to consider these collective intentions because they are involved in these collective relations.

This is a purely relational explanation. On the one hand, we do not presuppose any existence beyond what is commonly acceptable for our analysis, i.e., only substances and relations,① without any mysterious element, such as a spiritual body. On the other hand, we reject physicalistic individualism and regard the existence of collective agency as a recognizable collection of relational eventualities. This view is consistent with the argument of Schmid (2009) and Schweikard et al. (2021), "It is, rather, to treat collective intentionality as irreducible with respect to its content and mode, and relational with respect to the structures that ground individuals' reference to plural contents and their self-conception as subject-in-relations" (Schweikard et al. (2021)).②

2.4.3 A thorough relational account

Admittedly, some kinds of collective agents are not stable compared to individual agents, for example, a reading group. However, some kinds of collective agents look more stable than individual agents, for example, a university or a government. Either kind maintains its diachronic identity, the difference lies in the source of this maintenance. A collective with institutional structures or functions will not change by variation in members but only by variations in its structures or functions. We will not consider a basketball team a new team just because of its members' changes. Nor will it be consid-

① Although some authors advocate ontological reductionism about relations, most philosophers acknowledge relations as ontological commitments. For detail, see MacBride (2020).

② One of the anonymous reviewers pointed out to us that a combination of organizational structures and functional capacity has already been made in the area of complex adaptive systems. In addition to convergence of opinions, in evolutionary economic theory emergence of generic rule structures plays an essential role. As a result, collective agents have internal processes that allow them to manifest rationality, independent of being observed as doing so. Interested readers can further check Page et al. (2009).

ered a new company just because the shareholders of a company are renewed. However, we will be intuitively aware of changes in collective identity due to changes in the internal and external relations of the collective. For instance, the internal structure has changed, such as the merger of a company with another one into a new enterprise, or the realization of the external functions can no longer meet the necessary conditions, for example, when it is revealed that the actual purpose of a so-called charity organization was to make a profit.

It seems that the relational patterns instead of the relata play a fundamental role in collective agency. The examples above show that a relata's changing does not necessarily lead to a variation of our recognition of a collective, but a relation's changing does the opposite. Can we speak of the relational pattern itself without relata? Of course! For instance, in modal logic, we can either use a concrete symbol Rab to express the relationship between states a and b, or we can use formulas such as T: $\Box\phi \rightarrow \phi$, 4: $\Box\phi \rightarrow \Box\Box\phi$, 5: $\Diamond\phi \rightarrow \Box\Diamond\phi$ to express more abstract relational properties. Similarly, our cognition of collective agency is actually more direct to its abstract relational patterns, and those relata only enter into relational eventualities by being referred to such relations.[①]

The relations in the collection of relational eventualities thus play a critical role in forming the fixed pattern of collective identity. Only when the pattern undergoes a fundamental transformation will we think that the collective has changed. It should be noticed that this does not mean we advocate a theory in which individual and collective agents are defined differently: a collective agent is determined by relations, whereas an individual agent is not. On the contrary, we hold that the concept of agency as such, not just that of collective agency, is essentially a collection of related eventualities that guarantee the realization of a diachronic rational perspective that can be applied and understood by others, and its actions satisfy necessary functional conditions. For an individual, this rational perspective is guaranteed by the stable and rational diachronic self-consciousness of the individual, while for a collective, this rational perspective is guaranteed by the relatively stable diachronic structure of the collective. Concerning functional conditions, we do not agree that the consistency of the system's multiple decisions, as functionalism insists, is a necessary condition for the realization of rationality, but rather, more generally, that functional conditions merely ensure that actions embody the intentionality of the agent.

① Admittedly, our understanding of modal operators and principles depends on a previous understanding of the meta-language definitions in terms of explicit relations and relata. The point here is that even when all the relata have been abstracted, we can still directly talk about the relational properties without referring to any concrete relata. Similar things happen when we talk about the essence of a collective.

In other words, even if multiple decisions of an agent are not consistent with each other, as long as they are the result of the agent's intentionality, we still consider them to satisfy the functional conditions of agency.

We can re-examine the relationship between intention and agency from a relational perspective. In philosophy of action, intentionality is crucial because it is necessary for an agent to carry out their actions. From a relational point of view, intentionality represents the necessary relations that constitute agency; intentions and intentional actions directly bear the accessibility between cognitive contents and the external world. Prima facia, agency seems to be a property of an individual, but in essence, it is a relational property determined by the individual's various relational eventualities. The reason to say that agency is a relational property is that, in the broadest sense, agency expresses the ability to act as well as the capacity to manifest rationality, that presupposes and refers to the interactive relations between the agent and the external world (action), and indicates the nature of these relations to be manifested (rationality).

Thus, from a relational perspective, both individuals and collectives can be in relatively stable relational eventualities and thus can have the possibility to display agency. In terms of external functions, they have various interactions with the external world, and these interactions meet the same conditions, such as representation, motivation, rationality, etc. In terms of internal structures, individuals are fundamentally different from collectives. There are no inter-individual relations within individuals. Per contrast, a collective must have inter-individual relations as a collective structure to ensure its stability. On a relational view, these inter-individual relations are, from the perspective of an individual agent, between such an agent and the outside world since, for each individual, others categorically belong to the outside world too. However, from the perspective of collective agents, inter-individual relations constitute the stable internal relations of a collective. One may think of the internal relations of a collective as a subset of the union of various relations its members as individual agents are involved in. However, it should be noted that the properties of the same inter-individual relation from the individual perspective and from the collective perspective are different. The former is an accessible relation between cognitive content and the outside world, and the latter is part of the internal structure of a collective. Once collective intentionality has been formed, we can only regard the latter kind of inter-individual relations as part of internal structures, and not of relations of accessibility. This is the ontological reason at the micro-level why we have

an irreducible concept of a collective in the context of collective intentionality. Moreover, besides those inter-individual relations that serve as the basic constitutive level of a collective's identity, a collective also contains higher-order relations, which take these inter-individual relations as relata. From a holistic perspective, together with these higher-order relations, all internal relations of a collective form a relational system that manifests the collective's identity. Thus, on such a holistic account, a collective identity grounded on its relational system cannot be adequately explained by its constituent parts. This is the ontological reason at the macro-level why we cannot reduce the concept of a collective. The question of how this collective structure comes into existence is beyond the scope of this chapter. We assume that collectives by themselves have structures, and the discussion in this chapter is within this boundary. Getting rid of this assumption will lead to more complex problems, a further relational account of agency left to future work.

Adopting a holistic perspective has some advantages: First, this is an entirely non-individualist account. We can fundamentally avoid the circularity problems faced by the orthodox theories and use the primitive concept of relation in explaining social phenomena. Second, holism can integrate interpretationism and functionalism into a unified theory and properly explain the irreducibility of the concept of a collective. Third, although this holistic theory recognizes a realism of collective agency, we have not introduced any Hegelian mysterious spiritual entities or any totalitarian specters that Popper has criticized. On the contrary, we only point out the individualist's excessive ontological demands and reaffirm the importance of relations as bona fide, acceptable ontological entities. Moreover, since we have stated that the collective discussed here is structural and does not involve any spontaneous behavior, our account has a solid and defensible basis.

In conclusion, we say a collective has agency if and only if it has a relatively stable relational pattern to realize its rationality. Such a relational pattern includes at least one of the following relations:

(1) A stable interrelationship between its members that constitutes a stable structure;
(2) A stable relationship between its members' choices and the collective action that constitutes a stable procedure.

Any collective with agency that satisfies this condition will be counted as a collective agent. The relational pattern is critical because groups with no relatively stable relational pattern cannot manifest collective agency.

In the sense that the concept of a collective is presupposed in the conditions of an

account, the relational account also involves analytical circularity. Nevertheless, the problem here is not severe because those relations in our conditions exist factually rather than conceptually, which means that the individual-to-collective transition that hampers individualist accounts presents no problem for ours. In the relational account, relations as sufficient and necessary conditions must exist in the ontological sense, which provides the source of the concept of a collective. Furthermore, by acknowledging the essential role of the relational pattern and its content, the relational account provides new knowledge about our understanding of collective agency, thus avoiding inferential circularity. The relational account combines interpretationism and functionalism and thus inherits their tradition of removing superfluous constraints on the concept of an agent and agency. In other words, in the relational account, as long as a group satisfies at least one type of the relatively stable relational pattern, we regard it as a collective agent possessing collective agency. This means that physicalism does not apply in the relational account. Furthermore, as mentioned above, acknowledging ontological relations leads the relational account away from individualism. Therefore, the problem of physicalistic individualism discussed in Section 3 is circumvented in the relational account.

2.5 Relations between Collective Agency and Individual Actions

There has always been a powerful objection to acknowledging the existence of collective agency: it is that collective agency will cause its members to lose their individual agency and become nothing but mechanical or instrumental units, which will constitute a major problem for rational theories that focus on individuals. However, this worry appears unfounded in a holistic account.

This objection reverses the order of one's feasible strategies and one's agency. In a holistic account, the agency of any system is manifested in its ability to evaluate, select and implement all its feasible strategies, and at the same time, reflect on its own rationality. This is not to say that an individual maintains their agency first, then participates in some collective agency and is restricted, i.e., collective agency rests on individual agency. This is the perspective that most analyses work from. Instead, in a relational network, individuals-in-relations show their agency by choosing between the optimal individual strategies and the optimal collective strategies.[①] That is to say, not only does collective

① In some cases, these strategies are inconsistent, such as the prisoner's dilemma.

agency rest on individual agency, but individual agency also depends on its relation to relevant collective agency. There is a back and forth relationship between them, i.e., individual and collective agency are mutually dependent.

This does not mean that the choice of an individual agent can only be 0 or 1, i.e., to either manifest collective agency or their own agency. Note that in our definition of collective agency, a collective is a collection of relational eventualities. This introduces a temporal dimension. In a diachronic process, an individual agent can order the various relations they are involved in and gradually accomplish feasible strategies from different relations. For instance, someone finishes their work as a company's employee, then accompanies their child as a parent, then has dinner with their friend as part of maintaining a friendship, and then may go out cycling for their own pleasure. A variety of collective agencies emerges in the diachronic actions of an individual agent, and individual agency is not compromised. Nevertheless, at any given time, if an individual chooses to implement the collective's optimal strategy, we believe that the collective achieves its agency. If an individual refuses to act in accordance with its status and functions in collective relations, we say the collective does not achieve its agency.

Another reason why the aforementioned objection is not valid is that participation in specific collective relations is always the result of an individual's intentionality. In daily life, individuals may function in multiple collectives. Different social relations bring optimal strategies in different dimensions; however, all things the agent has done are based on their own intentionality. They could choose to engage or not. Individual agency is not static. An agent is constantly updating their preferences and re-evaluating the relations they are in, and in that way they manifest diachronic agency. This is similar to the view of Rovane (2014), "the existence of any agent is always a product of effort and will, and is never a metaphysical given—which is just to say, there are no natural persons." Our behaviors and choices regulate what kind of rational agent, what kind of person we really are.

The relational account also explains the phenomenon of sacrifice. If individual members only play an instrumental role in collective agency, then no one will be motivated to sacrifice individual gains. We would not say that those heroic figures who sacrifice for the collective benefit have no individual agency. On the contrary, this sacrifice manifests a kind of thinking that, after facing a variety of behavioral strategies and knowing all the consequences, the agent chooses to dedicate themselves to others in the collective. It is

precisely because the individual agent evaluates all relevant relational eventualities and chooses to unify their own agency with their role in collective agency that other members feel respect for self-sacrifice and other forms of heroic behaviors.

On a relational account, the influence that collective agency will have on the behavior of its members is mainly divided into two aspects. On the one hand, members' behaviors will be affected in accordance with the particular structure a collective may have. If the collective has a flat structure, that is, all members have equal status in the collective, then the realization of collective agency requires each member to evaluate and decide among the feasible strategies involved in collective relations and ensure that the optimal strategy of the collective is chosen and implemented. If the collective contains a particular hierarchical structure, the hierarchy is reflected in its members' evaluations and decision-making. The subgroups that occupy the most central position should select strategies first regardless of other members, and then the next-level subgroups can choose based on the decision-making of the most central subgroup, and so on. In this sequential mode, the collective's optimal strategy selection is realized at different levels. Although, of course, collective structures in real society are more complicated in detail, it appears that these are the two most straightforward and common structures: either each level achieves collective optimality equally (small work team), or the choices of subgroups in different levels depending on the higher level together achieve the optimal collective strategy (large company). Changes in collective structures will correspondingly affect the decision-making sequence of its members. On the other hand, collective functions also affect the targets of collective actions realized by its members-in-relation. If relations between collective members and behavioral targets change, it will change the criteria of agency for collective actions. For example, changes in the industry a company is engaged in, such as shifting from furniture to grocery supply, will directly affect the regular operation of the whole enterprise in the short term and the specific work of its employees. In summary, changes in structures or functions of a collective substantially affect its members' actions.

2.6 Summary and Ideas for a Future Investigation

In this chapter, we have reviewed several mainstream accounts of collective agency and collective intentionality and we have pointed out that they all manifest the tendency of physicalistic individualism resulting from long-term Cartesian influence. We have shown that the foundation of such physicalistic individualism is fragile and that theories that sub-

scribe to it cannot adequately explain the source of irreducible concepts of a collective. We then have developed a relational account, which regards the subject of collective intentionality as the relational identity composed of all the relative relational eventualities. An individual identifies and acts in a relational identity of a collective through their social relations, thus manifesting collective agency. Collective agency has both internal relations (collective structures) and external relations (collective functions). This confirms that interpretationism and functionalism about agency can be integrated into a relational, holistic account.

As mentioned in the introduction, all the discussions in this chapter are based on two assumptions: a collective has structural characteristics, and we do not discuss spontaneous behaviors. Our relational account is successful with regard to structural collectives.① However, what if we broaden our horizons and discuss more situations, such as non-structural groups? For example, people who suddenly run together due to an alarm; who happen to pass by and offer assistance; and who walk together for no reason; etc. Various social phenomena need to be explained. For groups with opaque structures that do have constituted identities of collectives, it may be appropriate to talk about particular groups' pre-structural or quasi-structural characteristics on a looser basis.②

The relational, holistic account is still in its infancy, but its advantages are apparent enough. Physicalistic individualistic theories will be turned upside down from this perspective, forming a brand-new consistent interpretation of the three-body: individual–collective–society. As Karl Marx puts it, "but the human essence is no abstraction inherent in each single individual. In its reality it is the ensemble of the social relations" (Marx (1970)). Combined with the relational account of collective agency, we will have a refreshed understanding of this sentence.

① Structures or functions of a collective guarantee minimum relational eventualities that constitute the relational identity of a collective.

② The possibility of integrating our relationalist approach into the broader framework of critical realism, as suggested by one of the anonymous reviewers, is also a promising avenue to be explored.

CHAPTER 3 A DISPOSITIONAL ACCOUNT OF INTENTIONALITY

Based on our relational account of (collective) agency, in this chapter, we further analyze the irreducibility of collective intentionality. While acknowledging such irreducibility, we find a deep incompatibility between the concept of a collective and the concept of intentionality as the mark of mental. In order to explain how collective intentionality nevertheless is possible and why we tend to use it analogously to how we use the concept of individual intentionality, we explore a dispositional account of intentionality through which we can give an account of the concept of intentionality at both individual and collective levels. Specifically, we subdivide the dispositional account into three aspects: behavioral, purely mental, and cognitive. We then argue that collective intentionality exists by analyzing different forms of attributive judgments and by introducing the perspective of indispensable collective responsibility.

3.1 Introduction

Intentionality as a philosophical concept usually refer to the aboutness of human consciousness. For instance, "I believe that it is raining;" "I intend to go to the movies;" "I prefer beer to wine." In each case, "I" am in an intentional state, i.e., one that is about something, refers to something. Intentionality also applies to collective. For example, "we intend to go to a bar for a drink;" "We believe that the study of philosophy is beneficial to mankind;" "We prefer argumentation to fantasy." In such cases, intentionality shows itself in the first-person plural form and concerns a collective's states that are about or refer to something. It appears to be common sense that it is the individual that is the subject of intentionality. However, a core issue follows: how do we explain collective intentionality? Can it be structurally reduced to every individual in the group having such intentionality?

In most existing theories, an intentional state consists of two components: the type of a state and its contents, i.e., psychological modes and propositional contents.[①] Also, intentionality has a subject, namely an entity who possesses the intention. Based on these

① For instance, "I" can believe that it is raining, fear that it is raining, or desire that it is raining. In each of these cases, "my" state has the same propositional content, viz., that it is raining. However, "my" states are of different intentional types, that is, different psychological modes: belief, fear, desire. etc.

conceptual features of intentionality, there are three primary types of explanations for collective intentionality: content, mode, and subject accounts.[①] At the same time, we can also classify them as reducible and irreducible accounts according to whether a theory supports the reducible *we- concept*. Approaches that have been influential in this field are Bratman (2014); Gilbert (2006); Searle (2010); Tuomela (2013); List et al. (2011).

However, there is a crucial but often overlooked issue: current analyses of intentionality presuppose ontological homogeneity of a particular kind, e.g., individualism, and then characterize collective intentionality in terms of it. That is, collective intentionality itself may be a separate concept, but intentionality and the ability to intend are individual and can only belong to individuals. Thus, in describing intentionality, there is no need to refer to anything beyond the individual itself. As we will show in the following, one of the sources of individualism is that the concept of intentionality is analysed as referring to a mental capacity, which hampers a unified theory for both individual and collective intentionality.

In this chapter, we want to argue that the traditional criterion of intentionlaity cannot sufficiently explain the concept of collective intentionality, and that what we need is an explanatory model that can satisfy the following conditions: (1) Contrast to individualism but is consistent with ontological relationalism[②], which claims that a collective agent is essentially a collection of relational eventualities, in which includes relata (members of a collective and the targets of collective behaviors) and relations (a collective's internal structures and external functions); (2) Explain collective intentionality naturally without losing any traditional features of the concept of intentionality, that is, directedness, in-existence, and intdeterminacy; (3) It does not contain any Hegelian spiritual entity of collective, which means we admit collective intentionality, but there is no mysterious collective entity that 'own' such ability as an individual 'own' her own intentionality.

Our answer to such a model is a dispositional account of intentionality[③]. Based on an observation of the similarity between the concepts of intentionality and disposition, we divide dispositional relations into three categories: physical dispositions, biological dispositions, and volitional dispositions. We will argue that on the one hand, volitional dispositions distinguish individual intentionality from collective intentionality, while on the other hand, the characteristic difference between individual intentionality and collec-

① For more details, see Schweikard et al. (2021).

② Such a view is proposed in Wang et al. (2022) and a brief review of it will be provided in section 3.5.

③ Thus, what we call the dispositional account here is different from the dispositional account of phenomenal intentionality, e.g., Bourget (2010), Kriegel (2011), etc.

tive intentionality is strongly correlated with physical dispositions.

In order to make good on this claim, we need to complete a few steps. We will review the most prominent theories in the field, evaluate their solution to the core issue in section 3.2, and explain that the mainstream definition of intentionality is insufficient to explain collective intentionality in section 3.3. Then we begin to introduce our dispositional account. First, we compare the traditional dispositional explanation of intentionality with the traditional intentional explanation of disposition and analyze the relationship between the two concepts in section 3.4. Second, in section 3.5, we shortly review our ontological standpoint of relationalism, which is fundamental for understanding the dispositional account. Finally, the dispositional account is provided in section 3.6; To facilitate the analysis, we subdivide the dispositional account into three aspects: behavioral, purely mental, and cognitive[①]; Then, we argue that collective intentionality can exist from the perspective of different manners in attributive judgments and the perspective of indispensable collective responsibility. Let us start with the evaluation of existing theories.

3.2 An Evaluation of Different Approaches

In what follows, to identify the problem we evaluate the representative views by prominent philosophers for both the reducible and irreducible accounts; A basic classification of the different viewpoints has been given in Section 2.2. Furthermore, we give a counterexample to demonstrate that neither reductionist nor irreductionist claims can avoid the persistent problem we have seen above.

For the irreducible-account, we introduce John Searle. He takes the core of collective intentionality as "*we-intention*", insists that this notion is irreducible, and gives an argumentation with the well-known business school example:

BUSINESS SCHOOL CASE 1

Imagine a group of Business School graduates who were taught and come to believe Adam Smith's theory. After graduation day, each goes out in the world to try to benefit humanity by being as selfish as each of them possibly can and by trying to become as individually rich as they can. Each does this in the

① These three aspects are different from the three categories of the concept of disposition mentioned above. The aspect distinction is based on the contents of dispositions. Both individual and collective intentionality display these three aspects of dispositional content, but the three categories of disposition (physical, biological, volitional) do not all have both individual and collective instances. In particular, as we will argue, there are no volitional collective dispositions.

> *mutual knowledge that the others are doing it. Thus there is a goal that each has, and each knows that all the others know that each has it and that they know that each knows that each has it. All the same, there is no cooperation. There is even an ideology that there should be no cooperation. This is a case where the people have an end, and people have common knowledge that other people have that end.*
>
> *BUSINESS SCHOOL CASE 2*
>
> *All the conditions are the same as in CASE 1, except they together make a solemn pact that they will each go out and try to help humanity by becoming as rich as they can and by acting as selfishly as they can. All of this will be done in order to help humanity.*(Searle (2010), p.47)

Searle argues that CASE 2 is a case of collective intentionality while CASE 1 is not. The reason is that there is an obligation assumed by each individual member in CASE 2, but no such pact or promise exists in CASE 1. Only if such shared obligation exists, can we regard it as a case of collective intentionality. But this kind of cooperation is not *implied* by common knowledge or belief together with individual intentions.

On the other hand, Searle puts this "*we-intention*" only in the mind of an individual. He said explicitly, "The only intentionality that can exist is in the heads of individuals. There is no collective intentionality beyond what is in the head of each member of the collective." (Searle (2010), p.55). In other words, Searle insists on conceptual irreducibility while rejecting ontological irreducible-account. Based on his irreducible "*we-intention*," Searle develops his social ontology, in which people with a certain social status can create social entities through declarative behavior. Obviously, we can see that by being engaged in such kind of declarative behavior, social entities like groups can be created. An everyday example is an enterprise as a legal entity. Thus, Searle's theory clearly has a considerable gap: it allows for collective agency of social entities created by people, but such collective agency cannot instantiate collective intentionality. This diagnosis is also endorsed by Baier (1997); Stoutland (1997); Meijers (2003).

We agree with Searle's argument that *we-intentions* are irreducible, and we also agree that declarative behavior can produce social entities. However, we disagree with Searle's assertion that the irreducible *we-intention* can only exist in the individuals. This is a clear instance of Searle's commitment to individualism. Nevertheless, Searle's theory does make important contributions. It provides a strong argument for the irreducibility

of "*we-intention*," and it outlines a basic frame for a social ontology. However, although the irreducible-account represented by Searle supports conceptually irreducible collective terms, it still presupposes an individualist view on their origins. Let us look at the other side.

In favour of a reducible-account, Michael Bratman argues that there is no need to introduce novel conceptual, metaphysical, or normative notions to explain the relation between joint actions and an individual's single action. He limits his attention to "modest sociality", namely, small-scale cases of shared agency by groups of adults that remain constant over time that do not have authority relations to one another. He claims that his theory of shared agency can reduce Searle's "*we-intention*" by providing sufficient reductive conditions:

> A. *Intention condition: We each have intentions that we J; and we each intend that we J by way of each of our intentions that we J and by way of relevant mutual responsiveness in sub-plan and action, and so by way of sub-plans that mesh.*
> B. *Belief condition: We each believe that if the intentions of each in favor of our J-ing persist, we will J by way of those intentions and relevant mutual responsiveness in sub-plan and action; and we each believe that there is interdependence in persistence of those intentions of each in favor of our J-ing.*
> C. *Interdependence condition: There is interdependence in persistence of the intentions of each in favor of our J-ing.*
> D. *Common knowledge condition: It is common knowledge that A-D.*
> E. *Mutual responsiveness condition: Our shared intention to J leads to our J-ing by way of public mutual responsiveness in sub-intention and action* ...(Bratman (2014), p.103)

We can see that Bratman's approach gives a structural account of reducibility, one that is clearly individualistic.We agree with Schmid et al. (2008); Petersson (2007), who argue that Bratman's explanation faces a circularity: an individual cannot refer to a joint activity without that joint activity existing, which means that such references cannot cause that joint activity to exist. Beyond that, we should also take a look at the problems in Bratman's reductive conditions. Although Bratman claims that he has given sufficient conditions to characterise collective intentionality, his theory is not enough to get rid of

Searle's irreducible "*we-intention*." We will show it by means of a counterexample. Let us go back to Searle's business school example, but modify it a bit so that it can qualify for Bratman's small-scale condition. Instead of leaving school, all students go to the same room after graduation and make online investments to achieve their default goals. In this case, the scale is small enough in Bratman's sense, where each individual is equal, and students do not form companies or corporate organizations. At the same time, this example still preserves Searle's distinction, viz., there is an obligation assumed by each individual member in modified CASE 2, but not in modified CASE 1.

Furthermore, by meeting Bratman's condition, and by the construction of the business school with the modified same room restriction, we can immediately see that the students do have the intention that everyone is striving toward the same goal, and that they do have mutual knowledge that everyone else is striving toward the same goal. The students have a mutual sense of what each other's subplans are. (They are now in the same room). Each student really believes that everyone is pursuing the same goal and that all students are responsive to each other's plans and behavior, and they all believe that their intentions to stick to their goals are interdependent. Moreover, there is indeed an interdependence between their intentions to stick to their goals, and all of the above is common knowledge among them. Indeed, students in the same room react to each other's sub-intentions, and the purpose of the shared action can be traced from individual intentions (Both the group and each individual, in the long run, can achieve the desired goal).

This shows that the modified CASE 1 satisfies Bratman's conditions A-E; the only thing we should look into is the object of intention: Bratman's *we-J*. The question is, what is the meaning of *we-J*? Can we get it from the relationship between individuals?

Bratman uses the form "I intend that *we-J*", where *J* refers to a joint activity in which the intending member participates. We can check an earlier version of Bratman's concept of intentionality:

> *We intend to J if and only if:*
>
> 1. *(a) I intend that we J and (b) you intend that we J.*
> 2. *I intend that we-J in accordance with and because of 1(a), 1(b), and meshing subplans of 1(a) and 1(b); you intend that we J in accordance with and because of 1(a), 1(b), and meshing subplans of 1(a) and 1(b).*
> 3. *1 and 2 are common knowledge between us.*(Bratman (1999), p.121)

Thus, we intend *J* only if each of us separately intends *we-J* and it is common knowl-

edge that each of us separately intends *we-J*. What changed here is just the subject of intentionality; the object of intentionality is still the same, *we-J*. Can this joint activity *we-J* be reduced? Whether the answer is affirmative or negative, the fact remains that Bratman's conditions are not adequate:

- If Bratman holds a radical individualistic view and gives a thorough structural explanation, then he would answer that *we-J* can be simplified. Then, the only unclear concept in our analysis of the example, *we-J*, is also reduced into the individual level; This *we-J* is nothing more than the intention of each student regarding their goal. Consequently, Bratman's condition fails in the test of our modified example, because we still fail to generate a group intention from individual intentions (in Searle's sense).
- If Bratman answers that we-J is irreducible, then it seems that the modified example does produce collective intentionality. However, if we jump out of the example to the theory itself, we see that Bratman has acknowledged an irreducible concept, and his standpoints become indistinguishable from that of Searle. This interpretation just translates Searle's irreducible "we-intention" into irreducible "we-J."

Therefore, we believe that Bratman's analysis is insufficient to refute Searle's argument that there is an irreducible "*we-intention*." That is to say, irreducibility does exist in this field, and with it, an irreconcilable tension between individualism and irreducibility arises. In summary, we claim that Bratman's theory has a strong tendency towards individualism, and that his conditions for collective intention are insufficient. Nevertheless, Bratman's theory has made outstanding contributions, especially regarding the power of the mutual connection among individual intentions.

To sum up, we can see that individualism profoundly influences this field. Both reducible and irreducible accounts insist on an individualistic view to explain collective intentionality, and both face the intractable tension between the irreducible concept and the reducible ontology. So, why would one insist on individualism? Does it bring any beneficial results to the explanation of collective intentionality? In view of the discussion above, there does not seem to be any. Instead, it is precise because of the presupposition of individualism that all explanations have to face the difficulty of circularity. In order to locate the source of that individualism, let us spend some time looking into the original

definition of intentionality.

3.3 Does the Brentano-Anscombe Criterion Fit Collective Intentionality?

At the beginning of this chapter, we have given an intuitive meaning of intentionality through examples. This section will review the philosophical discussion on the definition of intentionality and attempt to summarize some of the criteria that are relatively widely accepted and less controversial. It should be noted that the topic of intentionality is a vast field, and a detailed introduction to it is far beyond the scope of this chapter. Even so, some prominent views are worthy of attention in the service of our discussion of collective intentionality.①

Intentionality as a philosophical terminology was first introduced by Brentano (1874), and most of the subsequent discussions are based on his framework. His famous words are as follows:

> *Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on.*
>
> *This intentional in-existence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.* (Brentano (1874), p.68)

① In order to focus on collective intentionality, we omit a large area of philosophical discussion of intensionality-with-s. The reason is that although many analytical philosophers and linguists believe that our locutions that express intentionality-with-t are those that involve intensionality-with-s (c.f. Chisholm (1957)), such a connection itself has been questioned (c.f. Kneale (1968)). From one direction, should the extension of the concept of intentionality include cognitive abilities (e.g., knowing) as well as sensory abilities (e.g., seeing or hearing)? If yes, locutions that express such intentional states refer to the external. From the other direction, the locutions expressing intensionality-with-s are not limited to expressing intentionality-with-t. Locutions about necessity and causality also involve intensionality-with-s. We, therefore, treat the philosophical discussion of intensionality-with-s as a separate topic and shall not review it here.

These two paragraphs cover Brentano's essential assertions about intentionality, from which we can derive three characteristic features of intentionality:

- (Directedness) The primary characteristic of intentionality is directedness or aboutness. Intending, desiring, believing, loving, hating, etc. always refer to some object different from themselves;
- (In-existence) The existence of that object is being referred to is independent of the intentional act itself. This means that in the relation from a personal mind to its terminus, the latter does not need to exist in reality at all①;
- (Mark of mental) Intentionality is the exclusive mark of mental phenomena. Only mental phenomena exhibits it.

Among these features, "Directedness" is relatively uncontroversial and is accepted by more or less everybody. Although there is an extended period during which philosophers discuss intentionality by intensionality-with-s in a linguistic context and focus on the reference problem, the importance of directedness in intentional states has been re-emphasized by Searle. In the opening of Searle (1983), he claims:

> *As a preliminary formulation we might say: Intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world.* (Searle (1983), p.1)

In contrast to "Directedness", "In-existence" is controversial. In Anscombe (1965)②, Anscombe reviews two different meanings of the word "object": one is about things as reality, and the other is just about contents that belong purely to the idea. Then she argues that the existential paradox regarding "intentional object" is because people confuse these two meanings. Thus, Anscombe illustrates that "an intentional object" does not mean "an intentional entity" and claims that there is no sense in questioning the kind of existence of intentional objects as such. In light of this, we can say Anscombe consents to the feature "In-existence" of intentionality.③ In the same paper, she also provides a different version

① Latter, Brentano calls this "quasi-relational" since it is different from the paradigmatic "relational" like similarity and difference or cause and effect where all the relata must exist.

② We discuss the theory of Anscombe as a representative example here. For historical details, see Jacob (2019).

③ Admittedly, the conclusion drawn here seems to be a bit of a leap. The reason is, in that paper, Anscombe's discussion is not limited to the distinction between the two meanings of 'object'; she also spent a significant portion discussing the similarity and distinction between direct/indirect/intentional/material objects. She concludes that direct/indirect/intentional objects are neither the phrase nor what the phrase stands for, and claims that an intentional object does not need to result in a material object, which shows her sympathy to the 'In-existence' feature. Because such discussions go beyond the concerns of this section, we omit the reviews of these details. Interested readers can refer directly to Anscombe (1965).

of the features of intentionality:

> *There are three salient things about intention which are relevant for my subject. First, not any true description of what you do describes it as the action you intended: only under certain of its descriptions will it be intentional. ("Do you mean to be using that pen?"—"Why, what about this pen?"—"It's Smith's pen."—"Oh Lord, no!") Second, the descriptions under which you intend what you do can be vague, indeterminate. (You mean to put the book down on the table all right, and you do so, but you do not mean to put it down anywhere in particular on the table—though you do put it down somewhere in particular.) Third, descriptions under which you intend to do what you do may not come true, as when you make a slip of the tongue or pen. You act, but your intended act does not happen.*
>
> *Intentionality, whose name is taken from intention and expresses these characteristics of the concept intention, is found also in connection with many other concepts.* (Anscombe (1965), p.159)

The general linguistic approach that Anscombe adopts goes well beyond the quest for a general criterion for intentionality. For reasons mentioned in footnote ① on page 52, we refrain from discussing Anscombe's first point but just want to note that it has a deep relationship with the theories of Frege, Quine, and Chisholm on the reference problem. If we set aside the linguistic aspects, the third point from Anscombe is exact the same as Brentano's "In-existence", i.e., the intended object need not exist. Nevertheless, the example given by Anscombe here seems to be very limited; cases where one makes "a slip of the tongue" are quite specific, and moreover, they do imply that something exists, viz., (the expression of) the unintended meaning. However, as a general criterion of intentionality, "In-existence" does not require that there be an unintended result. For example, the mere desire expressed by "I want to do my doctoral defense" already reflects the in-existence of its intended object (the doctoral defense does not happen yet), and there are no unintended outcomes (nothing happens yet, so also nothing unintended).

A unique contribution from Anscombe is her second point, indeterminacy, a non-linguistic version of which can be formulated as follows:

- (Indeterminacy) The specific nature and mode of occurrence of the object to which the intentional state refers may be ambiguous and indeterminate.

Clearly, "Indeterminacy" is different from Berantano's features of "Directedness" and "Mark of mental." As for "In-existence," there is a subtle distinction between that and indeterminacy. "In-existence" focuses on the existence of the object that an intentional state refers to, while "Indeterminacy" focuses on the specific properties that the intended object is supposed to have. It could be its height, width, volume, or even the way it will be brought about. For example, if a group of people says "we intend to have a good time tonight," then, on the one hand, we can question whether they will indeed be having a good time tonight: that is in-existence; on the other hand, we can ask in which way exactly they will have a good time: drinking together, going to see a movie together, etc.? And if drinking, which type and which bar? If a movie, which one and which cinema? And so on. That is indeterminacy.

Different from "Directedness," "In-existence" and "Indeterminacy," in the theory of Brentano, the last feature "Mark of mental" is not a defining characteristic, but rather a feature that defines the scope of intentionality. Although some philosophers accept this (cf. Crane (1998)), there are many different views, some of which suggest that consciousness, instead of intentionality, should be the mark of the mental (cf. Strawson (1994); Searle (1992)), while others argue that intentionality should be the mark of disposition rather than of the mental (cf. Place (1996))①.

Let us tentatively set aside the controversial feature "Mark of mental" and consider it as an ontological or methodological constraint in the background. By merging the rest of features into one, we have a general criterion for the concept of intentionality. Call it Brentano-Anscombe (BA) criterion:

> (BA) A state is intentional only if directed towards an object that needs not exist and is indeterminate.

Now we can finally go back to the main issue of this chapter: collective intentionaltiy. The BA-criterion seemingly suits the collective case; for example, if the proposition "the students intend to hold a reading group this afternoon" is true, then the collective consists of students constituting a specific state (if any) directed towards the event of the reading group. Moreover, such an event does not yet exist in the present and is indeterminate. Since the ways of its happening, e.g., which chapters or paper will be read, which room it will be in, which students will come and which students will speak, and so on, are indeterminate. However, the question lies in whether there is such a state of a collective

① We will come back to elaborate this dispositional point in the next section.

that satisfies the BA-criterion, and if there is, does it also satisfy the background constraint of "Mark of mental"?

As discussed in Section 3.2, if we accept the irreducibility of the concept of a collective, then we must consider how collective intentionality can be possible. Suppose we consider the collective from a relational perspective①, then the collective is an aggregate consisting of stable structural and functional relations between individual members. In that case, we do not need to assume any collective existence as a Hegelian spiritual entity. But note also that there is no mental content presupposed for a collective. If collective intentionality is discussed in this 'physical' sense, it seems the BA-criterion is still satisfied. However, the constraint "Mark of mental" is not applicable because a collective lacks a mind or a brain that can serve as the primary vehicle for mental phenomena.

The crux of this argument is whether it is justified to ascribe mental phenomena to a collective. The traditional scope of application of the BA-criterion in combination with the "mark of the mental" constraint is individual intentionality. So, if we hold on to the mark of the mental constraint on intentionality, we are forced to conclude that collective intentionality is not possible. However, to jump to that conclusion runs counter to the basic observation in our daily lives, that we do use intentional statements to talk about collectives, such as "This team intends to win," "This company is committed to making the deal," or "The nation is confident that the economic crisis will be averted." When we use such statements, we feel that we are effectively talking about the collective and its intentional object and their relationship. And this suggests that the BA-criterion as such is neutral, and that it is the mental constraint that enforces individualism. Thus we may ask what precisely the "mental" applies to? It seems the mental constraint wants to show some unique feature of individual intentionality, but its reference is left vague. Is it just the opposite of physical? Or could we find the explicit unique feature of individual intentionality? To adequately explain this phenomenon, we need to explore a concept that is very close to intentionality: disposition.

3.4 Intentionality and Disposition

At first, we should make a distinction between two kinds of disposition. One is the typical biological disposition, where the manifestations of the disposition are controlled by a negative feedback mechanism that detects deviations from an end state or optimum

① Such a perspective will be elaborated on in section 3.5.

condition and supplies the necessary correction. The other is a typical physical disposition that only shows a capacity or possibility of the object, such as fragility or solubleness, where the form of the ultimate manifestation is determined only by the interaction of 'blind' forces like gravitational or magnetic attractions (cf.Place (1999)).

Philosophical discussions of disposition concentrate on the typical physical one (cf. Choi et al. (2021)). Following this tradition, in what follows, 'disposition' without qualification will refer to the physical sense. A more general description of disposition is as follows:

> *A disposition is a state whereby the entity (substance), whose dispositional property it is, is orientated towards the coming about of a possible future state which does not now exist and may never do so, but which, if it does exist and thus becomes determinate, will constitute a manifestation of that disposition.*(Place (1996), p.105)

It is not difficult to see that dispositions thus conceived share characteristics with intentionality as characterized by the BA-criterion, except for the 'Mark of mental' feature. Each aspect of intentionality has a counterpart in the manifestation of the disposition. This observation has already appeared in previous literature and has led to two prominent interpretations of this similarity relationship. One is that of behavioral dispositional accounts, which assume some form of materialism or philosophical behaviorism, and consider mental states suspected and acceptable only if they can be shown ultimately reducible to the visible or material or externally verifiable (cf. Ryle (1949); Searle (1983); Marcus (1990); Audi (1994)). These accounts have generally aimed to show how all descriptions of mental states can be transformed or reduced to descriptions of other, less objectionable things, which are specifiable in terms of dispositions or functions. However, some anti-behaviorist objections (cf. Putnam (1963); Strawson (1994); Chisholm (1957)) exploit the loose connection between mental states and behavior and argue that in many cases, a mental state of a certain behavioral entity does not necessarily involve a disposition to display that behavior.

Another way of interpreting this is the physical intentionality approach (cf. Place (1996); Molnar (2003); Heil (2003); Martin (2007); Bauer (2016)). Here the claim is that intentionality that satisfies the BA-criterion is not the mark of the mental but of the dispositional. In addition, it is claimed that non-psychological dispositional states are also intentional and have all the essential marks of intentionality. Of course, there are objec-

tions to the physical intentionality thesis (cf. Mumford (1999); Bird (2007); Mumford et al. (2011)). These objections question whether the BA-criterion conditions of intentionality also apply to dispositions and whether intentionality can be regarded as physical.

The proponents of these two views are locked in a long debate. On the one hand, simply reducing intentionality to dispositions is challenged by the fact that dispositions appear to lack executive force in the mental sphere. On the other hand, explaining dispositions in terms of physical intentionality appears to be inconsistent with the distinction between physical and mental directedness. Two positions are behind the two views, physicalist monists are inclined to equate these two concepts, while philosophers who adhere to the distinction between mental and physical reject such a claim, worried that the consequences of such an equation can, at best confuse categorical distinctions, and at worst lead to panpsychism and animism since it brings intentionality as mental properties to the physical world.

The vision of reducing mental states to dispositions for objective analysis is tempting, but this reduction will inevitably lose some of the attributes of intentional states that are more than purely physical. This kind of non-physical property of intentionality seems to underlie the "mark of mental" feature insisted on by Brentano. However, the "mark of mental" feature does not clarify what that mental property of intentionality exactly is. We sympathize with the arguments given by Putnam, Strawson, and Chisholm, but perhaps the discussion here can go one step further. The opposition to the behavioral dispositional account led by Putnam mainly attacked the loose connection between mental state and its associated behavior, as exemplified in his conception of a society made up of "super Spartans" that can feel pain but do not perform any action related to it (cf. Putnam (1963)). In fact, if one considers mental states that have intentionality in the broadest sense, there are many examples of mental states that do not lead to any behavioral disposition, such as a kid who merely desires to go to a good college but makes no effort at all to do so. This does not mean, however, that the dispositional account is entirely useless. Examples like this just point out that it is unnecessary to adhere to behaviorism, i.e., to postulate an associated behavior for any state of mind. Rejecting behaviorism we can still argue that each mental state is minimally connected with a disposition to have mental phenomena and cognitive content. Nevertheless, a question remains: if the disposition is understood in a physical or basic biological sense, how does it refer to non-physical objects such as mental or cognitive content? And so does another one: if a disposition is merely a

tendency or propensity, what guarantees that such a tendency necessarily comes to be realized, at least in the sense of resulting in mental or cognitive content? It appears that the dispositional account lacks at least two key components: a type of disposition capable of targeting mental or cognitive content, and an immanent force or willpower as the cause of the disposition and the drive that ensures its realization.

The physical intentionality approach interprets intentionality as a mark of the disposition itself rather than of the mental, and it applies intentionality broadly to the physical world. This approach rests on the general applicability of the three features of directedness, in-existence, and indeterminacy to both intentionality and disposition. However, as Mumford points out, "There is little reason to think that a material object without a mind is capable of having aims and strivings for events of a certain kind, because to do so would be for it to act, and attributions of action we reserve for things with minds. (Mumford (1999), p.221)" If intentionality can be applied universally to all situations involved in dispositions, then all physical and biological dispositions can be explained by intentionality. If intentionality is only defined by the BA-criterion, there seems nothing wrong with doing so. However, the crux of the question is why we still feel uncomfortable. As Mumford states, the application of intentionality in the physical world adds something that would otherwise be inconsistent with physical and biological dispositions. In other words, if the application of intentionality to these situations does indeed add a mental element to the physical world, then it shows that the concept of intentionality by itself contains some mental property, in line with what Brentano claims to be its "mark of mental" feature. Acknowledging that intentionality contains certain mental characteristics does not mean that its relationship with dispositions is suspended. But it does indicate that beyond their existing nexus, dispositions must satisfy some more mental conditions to achieve intentionality, and intentionality cannot be applied arbitrarily to the dispositional world.

On the basis of the discussion of the above two aspects, we think it is necessary to discuss, at least in the case of individual intentionality, whether, there are some psychological properties different from the BA condition that make intentionality different from a mere tendency. If such mental properties can be identified in the context of individual intentionality, then we can propose a view that can accommodate the advantages of both views: on the one hand, we can make use of the analyzability of dispositions to explain collective intentionality; on the other hand, we can leave room for the mental and guarantee the psychological dimension of the intentionality of the individual members of a

collective. One crucial property of intentionality that has been overlooked so far is the dimension of volition. As Anscombe mentioned:

> *Every intentional action is also voluntary, though again ... intentional actions can also be described as involuntary from another point of view, as when one regrets 'having' to do them. But 'reluctant' would be the more commonly used word.* (Anscombe (1957), §49)

We want to show that voluntariness, which Anscombe claims is a feature of individual intentional actions, is in fact a general feature of individual intentionality as such. In table 3.1, we give an overview of how voluntariness relates to major kinds of intentional states. Some notes for the table: First of all, not all philosophers agree that every mental state is also an intentional state. Some mental states, such as experiencing pain, are not about anything at all and thus are non-intentional. Here we only discuss the intentional mental states, which means there is always a directed object of the intentional states. Second, intentional states can be viewed concretely and abstractly. The intentional states discussed here are in the concrete sense, i.e., they are the intentional states of a subject. We do not want to include intentional states in the abstract sense, i.e., the mere concept of belief, the mere concept of perception, etc. Rather, we try to find a general characteristic for phrases like "I want P", "I believe P", "I feel P", etc. Third, we aim to discuss only conscious intentional states here and claim that any long-term intentional state has corresponding conscious intentional states when it is happening.

As the table shows, an individual cannot have a belief, desire, or intention without volition[①]. If volition is absent, the state cannot be properly called a "belief", or a "desire", or an "intention". Similar views can be found in Audi (2019); Anscombe (1957). For other intentional states the relation between being in that state and volition depends on context. In the above table, "+" indicates when some kind of state has a volition aspect, and "−" marks when that is lacking.

Two general observations on the table need to be mentioned. First, all the intentional states with "+" concern intentional states that are actively formed or possessed by the subject, while the intentional states with "−" are related to the reactive ones. Second, for those intentional states with "−", not only is it improper to say that the subject's possession of

① The case of desire is much more complicated than the example we give here. One example: many philosophers and psychologists also distinguish a concept of "desire" in the passive or reactive sense. For the purpose of discussing conscious intentional states, we will use the word "desire" only in the active sense. For details of various accounts of desire, see Heckhausen (2007); Wall (2009); Tooming (2019); Clark (2020).

Intentional states	Examples
Believe	If agent *A* believes that the earth revolves around the sun, then *A* is volitional to have a belief as such.
Desire	If agent *A* desires to be a philosopher, then *A* is volitional to have a desire as such.
Intention	If agent *A* chooses to start a Ph.D. study intentionally, it is strange to say that *A* is involuntary to her choice.
Knowing^{+}	If agent *A* knows that *P* is true, then the way she knows *P* is true, or retrieves her knowledge of *P* is true, is volitional.
Knowing^{-}	If it is truly depicted that agent *A* knows how to play the piano, then it seems we hardly attribute voluntary or involuntary to the way *A* has such knowledge-how.
Thoughts and mental images^{+}	If agent *A* is thinking about *P*, then *A* is volitional to her thought. If agent *A* has a mental image of Joe, then *A* is volitional to her mental image.
Thoughts and mental images^{-}	Some thoughts or mental images may recurrent in one's mind unwillingly; for example, memes keep popping up in one's head when she is working. It is inappropriate to say that having thoughts or mental images of a meme is voluntary.
Emotions and moods^{+}	If agent *A* hopes to get her paper published, then *A* is volitional to hope as such. If agent *A* has a grudge against someone, then *A* is volitional to have that grudge.
Emotions and moods^{-}	If agent *A* goes camping and encounters a bear, her sudden fear of the bear is involuntary. If agent *A* is frustrated with her failed exam, such a bad mood is involuntary too.
Sensations and perceptions^{+}	If agent *A* claims she feels others' emotions and shows her sympathy, then she is volitional to have sympathy as such. If agent *A* is looking at a cat, she is volitional to have perceptual experience as such.
Sensations and perceptions^{-}	When a musician goes on stage and feels the pressure of a large audience, she becomes nervous about it, but not voluntarily. When one wakes up and sees the bedroom ceiling, she is not volitional to have perceptual experience as such.

Table 3.1 Intentional states and their examples

these mental states is voluntary, but it is also improper to say that the subject's possession of these mental states is intentional. It is markedly weird to claim that an individual 'intentionally' has any reactive mental states, let alone that they manifest any intentionality. So it might be approporiate to remove these mental states with "−" from the category of intentional states altogether. A deeper reason is that the concept of intentionality, defined by the BA-condition, fails as the main characteristic of the concept of agency. The concept of agency, which expresses that an agent has the capacity to act, by itself relates to the concept of volition, which expresses the active relationship between the agent and the directed object. Since none of the examples that fail to include active relations adequately reflect our intended use of 'intentional', this indicates that individual intentionality has a volitional dimension.

Although there are numerous intentional states, after analyzing the most common ones, we can find that with them, volition is an essential and common characteristic of individual intentionality. This is enough to show that in the individual case such directedness involves some inner volitional force. However, neither of the two types of disposition mentioned at the beginning of this section is sufficient for exercising intentionality in the volitional sense, since physical disposition lacks even the primary agency, and biological disposition reflects only the viability of elementary life forms instead of any capacity of volition. Therefore, we may be able to form a definition of intentionality by combining these two aspects: volition and disposition.

Thus, our basic view is as follows. An agent A intends P iff agent A has a volitional disposition to bring about P. Volitional disposition is different from the disposition as possibility or capacity in the physical sense. Rather, it is a biological disposition plus a volition process that determines it[①]. In our view, the difference between physical, biological, and mental directedness is that the last one has an additional dimension of volition. In other words, when a mental state is directed at something, it is not simply a dispositional relation in the 'blind' physical sense or a disposition to survive in the 'naive' biological sense but in an 'intellectual' sense that constitutes an actively, voluntarily, and willingly directing of the agent. Thus, a stone that falls because of gravity or a paramecium's response to external stimuli cannot be interpreted as intentional, nor can "I intend to watch a movie" be explained by non-volitional dispositions alone.

① We use 'determines' in order to avoid going into the extensive discussion of mental causation. We only want to show here that the existence of such a volition-disposition nexus is the most significant characteristic of mental directedness. And we believe we can do so while remaining neutral in the debate on mental causation.

The advantage of this approach is that we can distinguish the physical and psychological parts of intentionality, put all the purely psychological content in the volitional part and effectively analyze the physical content in terms of dispositions on the factual level. This conceptual analysis is necessary for our further analysis of collective intentionality.

3.5 Preliminary: Ontological Relationalism

The dispositional account of collective intentionality is based on the ontological relationalism account of collective agency. This section will briefly review such a fundamental standpoint. This section is devoted to a brief review. For further details the reader is referred to Chapter 2.

Ontological relationalism claims that the ontological sources of social facts of collective agency are relational eventualities, a term that refers neutrally to events and states that can originate, develop, and vanish. It is a fusion of functionalism (cf. List et al. (2011)) and interpretationism (cf. Rovane (1997); Tollefsen (2002)), each of which addresses a different aspect of the relations across a collective. The former deals with collective functions, that is, the interaction between the collective and the outside world, which form the external relations of a collective; The latter deals with collective structures, that is, the interactions between its members, which make up the internal relations of a collective.

Within ontological relationalism, a collective agent is interpreted as a collection of relational eventualities, including relata and relations. The relata of a collective agent contain members of a collective and the targets of collective behaviors, and the relation of a collective agent contain the collective's internal structures and external functions. The recognizable identity of a collective agent is constituted by these relata and relations. Through its internal relations, members partake in collective intentionality, and an individual can choose to participate in the intentionality of different collectives depending on the net of relations she finds herself in.

Furthermore, ontological relationalism claims that the relational patterns rather than the relata play a fundamental role in collective agency. Whereas in most cases, a collective's relata's changing does not necessarily lead to a variation of our recognition of a collective, but a change in its relation does. Furthermore, ontological relationalism is not confined to collective, actually, the concept of agency as such is relational.

By holding a fundamental relational perspective on the concept of agency, ontological relationalism claims both individuals and collectives could maintain relatively stable

relational eventualities, by our definition, namely manifest agency. Regarding external functions, individuals and collectives may have different interactions with the external world but both need to satisfy the same conditions of representation, motivation, rationality, etc., as proposed in List et al. (2011). Regarding internal structure, individuals and collectives are fundamentally different. There are no inter-individual relations within an individual. However, a collective must have such inter-individual relations to ensure its stability and constitute its basic identity. In addition to that, the collective also contains higher-order relations, which take these inter-individual relations as relata. For example, the discipline and supervision departments within enterprises, armies, or political parties have functions and powers targeted at inter-member relations. From a holistic perspective, all the inter-individual relations within a collective, together with these higher-order relations, constitute a relational system that embodies the collective's identity. The collective in the context of ontological relationalism is assumed as structural by itself; thus, it guarantees that there are minimum relational eventualities that constitute the relational identity of a collective and, in such a way, makes our claim defendable.

It should be noted that, from the relational perspective, a collective's internal and external relations act as a truthmaker when we judge whether there is a collective in the ontological sense. In the following section, we will discuss several dispositional aspects for judging collective intentionality, all based on ontological relationalism.

3.6 A Dispositional Account of Collective Intentionality

As described in section 3.4, we explain individual intentionality in terms of volitional dispositions while acknowledging the close similarity between different types of dispositions. Here we will show the close relationship between individual and collective intentionality by explaining individual intentionality in terms of volitional disposition and collective intentionality in terms of physical dispositions. This interpretation is sufficient for those three conditions we mentioned in the introduction: (1) Jump out of the individualistic frame and follow ontological relationalism; (2) Explain collective intentionality naturally without losing any original features of the concept of intentionality; (3) The account does not contain any mysterious part like a Hegelian spiritual entity of collective. And can naturally explain the phenomenon that we are accustomed to using the concept of collective intentionality in the same sense as individual intentionality in everyday language.

We will make good on such an argument in three steps: by working out dispositional analysis of individual and collective intentionality; by using inner and outer manners of attributive judgment; and by invoking the indispensability of collective responsibility. The dispositional analysis will account for the relation between individual and collective intentionality without losing the original nature of the concept of intentionality. Different manners ('inner' and 'outer') of attributive judgment explain that psychological phenomena and cognitive contents that we attribute to the collective do not introduce any mysterious elements. Based on the similarities between different dispositions and different manners of attributive judgment, these two steps will provide a natural account for our everyday tendency to talk about collective intentionality as if in the same way as individual intentionality. The discussion about collective responsibility will strengthen the dispositional demonstration of the existence of collective intentionality from the moral perspective. Furthermore, the attempt to justify collective responsibility constitutes one of the reasons why we tend to use collective intentionality along the lines of individual intentionality. All these steps take us out of the individualistic framework and are consistent with ontological relationalism.

3.6.1 Three dispositional aspects of individual and collective intentionality

We will first give the dispositional account of individual intentionality and then apply it to collective intentionality. Recall that our basic definition of individual intentionality is: An agent A intends P iff agent A has a volitional disposition to bring about P. Following the dispositional explanation in Schwitzgebel (2002), we divide 'volitional disposition' into three aspects[①].

Individual Case

Volitional disposition to do actions: This aspect is mainly related to intention. In the idealized context, intention is intrinsically related to its corresponding intentional action, although this intentional action can be canceled, and the reason would arise from external (world blocks the intended action) or internal (the agent decides to give up her

① The distinct aspects we make here are merely intended to facilitate a systematic analysis of the concept of intentionality, which means we remain neutral on whether intentionality can be reduced to other concepts. In fact, there is a wide range of discussions on such a topic, for instance, reductive phenomenal intentionality theory, tracking theory, conceptual role theory, primitivism, representationalism, separatism, etc. (for details, see Bourget et al. (2019)). Since we focus on all mental states that meet the BA-criterion of intentionality, our attitude towards the reduction issue does not matter.

intention). The behavioral dispositions here refer to verbal and nonverbal behavior, such as the disposition to say that "I will cook stewed beef brisket with tomato and potato for dinner" and go to the kitchen to start to cut the brisket. These dispositions are transcendent compared to barely mental phenomena. They refer directly to the objects or actions in the outside world.

Volitional disposition to have purely mental phenomena: Purely mental phenomena mainly refer to sensory experience and emotional experience. Typical examples would be feeling pain and mourning. In the process of exercising intentionality, intentional states are also accompanied by the volitional disposition to the associated mental phenomena. This is consistent with Brentano's view that every mental phenomenon is of an object in the intentional sense, and for every mental act, there is an intentional consciousness of it. The phenomenal dispositions here refer to those dispositions associated with undergoing active mental experiences.[①] Such as saying silently to oneself, "I have got brisket, potato, and tomato already," and feeling surprised when one finds that there lacks a kind of spices. Phenomenal dispositions are not transcendent; these dispositions only refer to the occurrent mental experience. Usually, intentions come with behavioral as well as phenomenal dispositions, and there would also be intentions with behavioral dispositions but not phenomenal ones, such as sub-intentional actions as a master pianist playing some specific notes in performing a piece. In contrast, other intentional states such as beliefs and desires relate only to phenomenal rather than behavioral dispositions.

Volitional disposition to have cognitive contents: Cognitive states are different from the purely phenomenal dispositions we mentioned above; for instance, when a non-Dutch speaker and a Dutch speaker watch a Dutch television program, they share the same experiential phenomenon but get very different cognitive content. We will name these kinds of states associated with intentionality as cognitive dispositions, which refer to those associated with the agent's cognitive ability, such as "I" know that "I" have to cut the brisket in order to complete the disposition to cook the meal. Cognitive dispositions play an essential role in breaking down the main goals into smaller ones and assembling them to derive the final purpose; this unique function has a close connection with the planning theory in Bratman (2018). Almost all intentional states have associated cognitive dispositions.

① Active mental experience is the mental experience that is induced by an action. It includes direct sensory experience (seeing, hearing) as well as emotional experience. These experiences all require an active stance of the subject. Reactive mental experience comes about without such an active stance. Examples are pain sensations caused by external stimuli, or experiencing anxiety or depression without specific concrete causes. Important to note is that only dispositions that are associated with active mental experience have a volitional dimension.

Although we have set apart intentionality into three aspects of volitional dispositions, this does not mean that each aspect is independent. On the contrary, different dispositions are related to each other. For example, behavioral dispositions are usually based on associated purely mental phenomena and cognitive contents. Mental phenomena are, in most cases, produced accompanied by behavior and cognition, and the information update in cognition is always through mental phenomena and behavioral results. In short, the division into three aspects is only for explanatory clarity, and in fact, they are closely related to each other as a whole.

Collective Case

Things get complicated when it comes to the collective context. For a collective, we can consider it from the perspective where only the collective as such is taken into account (hereafter "collective's perspective"). And we can consider the same collective also from another perspective, where the focus is on the members (hereafter "members' perspective"). Moreover, perhaps there may be a third perspective that both the collective and the members constituting the collective are taking into consideration. For example, when we think of Koninklijke Luchtvaart Maatschappij (Royal Dutch Airlines), we can consider it a company in an abstract sense. Meanwhile, we can also consider it a set of its boards, employees, etc. Furthermore, these two perspectives are not conflicting; we can consider both simultaneously.

The focus of members' perspective is still on each individual as a member of a collective. Regarding an individual as an individual is different from regarding an individual as a member of a collective: to be a member means that the individual is in stable interrelations with other members; in other words, a member also has the volitional dispositions that refer to actions, mental phenomena, and cognitive contents, but these dispositions are interrelated with those of other members of the collective. We do not want to repeat the details here, but it needs to be pointed out that the members' perspective is the perspective that Bratman describes. As discussed in section 3.2, such a superposition of individuals and interrelations is not enough to achieve collective intentionality; at most, it is a shared intentional state, but the owner of which is still each independent individual.

Considering the pure collective's perspective, the most apparent difference from the members' perspective is that we cannot find any vehicle for the existence of collective intentionality; that is, no physical or mental entity can 'own' those collective actions, mental

phenomena, or cognitive contents in the same sense as an individual owns them.[①] Thus, a fundamental distinction between individual intentionality and collective intentionality is the characteristic of the dispositions it has. By 'owning' intentional actions, mental phenomena, and cognitive contents, individual intentionality manifested a prominent volitional dimension since it is the inner power that drives the agency. However, lacking such ownership makes collective intentionality only connect to its behavior, mental phenomena, and cognitive content in the sense of physical dispositions. And that means we can still find dispositional relations in our everyday talk about collective about their actions, mental phenomena, and cognitive content, but those are dispositions without volition. For example, the International Red Cross organizes relief operations for a natural disaster. We would think of the IRC as a collective agent whose actions in organizing relief operations are intentional. Also, the IRC has dispositions to do such intentional action, sympathize with the refugees that are affected by the disaster, and that the IRC acts on the basis of (universal) values. Most IRC members and volunteers have corresponding volitional dispositions at the individual level. However, in terms of the IRC as a collective agent, none of its organizational structure, rules, regulations, or people's shared faith in its existence and development is able to serve as a vehicle for its volition. So when we say that the IRC intends to organize relief operations, it does involve relevant dispositions, but none of them have any volitional dimension. We can conclude that a collective G intends P iff the collective G has dispositions to bring about P, noting that the dispositions here are physical rather than volitional ones.

Disposition to do collective actions: A collective G has dispositions to verbal (collective announcement) and non-verbal behavior (doing things together, like a couple cooking dinner as a collective) in order to bring about P. From a relational perspective, the behavioral dispositions of a collective are actually identical to its functions, that is, the external relations between the collective and its external behavioral objects. This is because the behavioral object of a collective does not exist before the occurrence of the corresponding behavior, and once it occurs, constitutes a manifestation of the collective function. For example, the functions of the Cultural Heritage Agency in the Netherlands are "generating and disseminating knowledge, implementing policy and legislation, administering guarantees and subsidies, searching for innovations in heritage care, and providing practical

① Admittedly, a collective can still have some 'physical' features, such as a company registered in the Cayman Islands for tax purposes, being a legal entity defined as a collective that has existed in the Cayman Islands for a period of time. Nevertheless, these kinds of quasi-physical properties are insufficient to be used as a concrete vehicle of intentionality and cannot manifest the kind of ownership relations in individual intentionality.

advice."(Ministerie van Onderwijs, Cultuur en Wetenschap (2019)) However, before there is a specific collective behavior, such as protecting a particular cultural relic, practical protection does not exist. Only when such collective behavior truly happens does it constitute a manifestation of the function of the Cultural Heritage Agency. Thus collective function satisfies the basic definition of disposition and can be explained in such terms.

Disposition to have collective mental phenomena: The mental phenomena of the members of a collective belong to each individual independently, and in ontological relationalism there is no room for a collective 'owning' an aggregation of such mental phenomena. Even so, expressions like "The government is confident that it is able to prevent the outbreak" still occur in our everyday language. More extreme cases also exist, such as when no individual in a collective has confidence in a particular collective action, but strong confidence arises from being part of a collective, such as people in strikes. A collective mental phenomenon such as this cannot be the superposition of mental phenomena of its members, nor is it easy to find an ontological vehicle in a collective context. However, even so, at least in the epistemological sense, this collective mental phenomenon exists; we can still make collective intentionality judgments through our judgments on the relevant dispositions to have collective mental phenomena.

Disposition to have collective cognitive contents: There are two ways in which we attribute collective cognitive contents. Although the cognitive contents of members belong to each individual independently, through language those cognitive contents can be externalized and can form higher-order knowledge, such as common knowledge. These higher-order cognitive contents can be stored both in each member's head and in public language. From the collective's perspective, the relations connecting such cognitive contents/states among each other are not psychological but purely physical structures. This bottom-up mode of common knowledge is just one way we understand collective cognition content. Usually we rely more on collective actions to judge whether a collective has the corresponding cognitive content, such as observing an enterprise avoid risk in its operation effectively, and judging as follows: "This company knows how to ward off risks." In such a judgment, how to avoid risks is not the common knowledge of the enterprise because we cannot confirm that every employee of the enterprise has perfect knowledge of how to avoid risks. Thus, such a judgment cannot be reduced to the members' perspective. However, like mental phenomena, they exist at least in the epistemological sense, and we can still determine whether a collective has intentionality by judging its relevant

dispositions to have cognitive contents.

Through the above discussion on the classification of individual and collective dispositions, it is not difficult to find that one of the reasons we insist on collective intentionality is based on the similarity between the definition of intentionality and disposition, and on the observation that in the dispositional sense, a collective satisfies a similar relation of intentionality. Furthermore, because of this similarity, we tend to misconstrue the physical nature of the collective in daily life and add mental or cognitive content that should belong to the individual to the collective to use the concept of collective intentionality more freely. Nevertheless, in this part, we only give a dispositional explanation of collective intentionality based on our daily language usage. But a more critical question is still worth considering: how exactly do we make each kind of dispositional judgment?

3.6.2 Inner and outer manners of attributive judgment

For the individual, all three volitional dispositions rest on the individual as a physical being; Once the physical existence of the individual vanishes, no one can think of the existence of individual dispositions for behavior, mental phenomena, or cognitive content. Still, there is a distinction in how such attributions are made. Generally speaking, this distinction exists between the first-person perspective, i.e., the attribution of the three dispositions from the subjective perspective of the agent, and the third-person perspective, i.e., the attribution of the three dispositions in the objective sense. The two perspectives lead to two manners we make attributive judgments: inner, by accessing what happens within our body vs. outer, by observing what happens from the outside. In terms of a manner's directness, the former is direct while the latter is indirect.①

It is worth observing that behavioral dispositions are different from the other two types of disposition since the behavior it points to exists from both perspectives. As Anscombe's dictum "I do what happens" has it: what the agent disposed to do from the first-person perspective is identical to the actions observed from the third-person perspective. Although we may say that a certain action belongs to an individual, we do not mean that the action must exist according to the individual, but as an objective eventuality. As long as the action exists, we make attributive judgments, no matter in which manner, according to its

① This is not to say that all judgments made in an inner manner are intuitive rather than deductive, and all judgments made in an outer manner are deductive rather than intuitive. Nor does it mean that the inner manner of judgment is based on ontological dependence, while the outer way of judgment is based only on epistemological observation and deliberation. We only point out the existence of the difference between two basic judgment manners to help our analysis of collective intentionality, a detailed analysis of the difference is out of the scope of this chapter. For readers interested in the difference between the two manners of judgment, see Gertler (2021).

relations to the agent. Although the exact process of making the judgment is different due to two different perspectives, the object of the judgment is the same.

However, different manners bring different attribution reasons regarding purely mental phenomena and cognitive content. From the first-person perspective, it is evident that both purely mental phenomena and cognitive content rest on awareness of the subject of its existence. However, from the third-person perspective, the reason why purely mental phenomena or cognitive content belong to a specific agent only lies in the reverse inference we can make from the observation of that agent's intentional action. For example, observing that an agent causes a traffic accident and actively makes subsequent compensation, we will judge that the agent has a sense of guilt and knowledge of what is owed. There is a distinction between purely mental phenomena attributed through inner and outer manners; because observers cannot directly observe what the agent's purely mental phenomena exactly are, we can at most have a similar sensory experience to simulate those feelings of the agent.

On the other hand, the cognitive content of the agent is also different from the cognitive content that is attributed to the agent by others. From a holistic perspective, the knowledge that an agent has is part of a complex network and is not (always) fully captured by the content of an externalisation by means of language, which is what is publicly accessible. For example, native speakers attach meaning to some specific words in their language, which may be very different from the sense that non-native speakers attach to them. For native speakers, their language is part of their identity architecture, the foundation of their thoughts, the tools by means of which they know the world, etc.; while for those non-native speakers, these words are to be carried out by unique intentional behavior of learning and these words are associated with him in a completely different way. However, any external observer can still abstract from the behaviors of the native speakers and non-native speakers that they have the same knowledge about a word object, although those distinctions in their minds are abstracted away in such progress.① Furthermore, such differences also apply to the context between two native speakers. Note that not all of the connections between mental (cognitive and emotional) contents that speakers are concerned with need to be shared, even if they grow up with the same language. Because such connections are only partially "grammaticalized" in terms of meaning re-

① Note that the meanings that a native and a non-native speaker attach to a word cannot be so different that they fail to communicate with them: there must be a shared core. That shared core must determine the reference (give or take a few exceptions), and the differences will be situated in the connections with other words and concepts, attitudes and emotions, knowledge of history, and so on.

lations, many of them are personal and depend on subjective experiences and personal history, and more general aspects such as age, gender, etc.

Based on these two intuitive observations, we believe there is a difference between the inner and outer manner in attributing dispositions of purely mental phenomena and cognitive content. In these two cases, not only the specific process of making the judgment is different but also the object of the judgment is different. Nevertheless, even if these differences exist, they are not significant at the individual level. The conclusions of attributive judgments from the first-person and third-person perspectives at the individual level are the same. However, it becomes more significant at the collective level.

According to our ontological view, a collective is a collection of relational eventualities, which means that a collective is categorically distinct from a human brain in physical or biological sense. So at the collective level, we do not have a reliable vehicle to prop up the collective's first-person perspective. Instead, our concept of a collective's identity stems from our knowledge of the collection of relational eventualities, which means that this identity does not precede our knowledge. Therefore, at least in the ontological sense, there is no constant collective identity that can 'own' collective intentionality and be its vehicle. Nevertheless, can we, therefore, claim that all our everyday talk of collective intentionality is just an illusion? This obviously goes against the intuitions of our everyday usage of language. In order to show that (and how) these phenomena can be justified, it is necessary to explore the differences between collective intentionality and individual intentionality in the three dispositional aspects.

Specifically, behavioral dispositions of a collective are similar to that of an individual, the action a collective is disposed to do is the same as the action attributed to that collective by others. Whether individual or collective, we all believe that such intentional action is an objective eventuality, and the subordination is the connection between such eventuality and the agent. In line with Wittgenstein's private language argument and Anscombe's argument of 'one object,' we believe that the objects described by propositions such as "we are talking" are both objectively observable events and cognitively explainable intentional actions, which are identical. Therefore, the attributive judgments we make in both manners, viz., inner and outer, are the same. To some extent, it can be said that the attribution of collective behavioral disposition is not only ontological (based on the external relations of the collective) but also epistemological (based on observation of such behaviors).

The real difficulty lies in explaining purely mental phenomena and cognitive con-

tents in connection with collectives. Ontologically speaking, a collective does not contain anything mental or something that could be the necessary vehicle of mental things. The mental and cognitive elements that we associate with a collective in everyday language exist definitely not in an ontological sense. However, it is essential to point out that there is still an observer's sense of attribution. Besides collective intentions, we also ascribe purely mental phenomena to collectives, including sensory and emotional experiences, such as "the community suffers pain together" and "the army is furious now". A collective has no brain, body, or mind, it lacks the most fundamental physical basis for feeling, let alone emotion, but even so, our expressions are not meaningless. Because we do not think of the existence of a collective first and then attach these feelings or emotions to it, per contrast, we make attributive judgments based on specific behaviors of a collective. For example, according to the "scar literature" produced by some Chinese writers, we ascribe feelings of hurt and emotions of mourning to that collective. This means that although there is no inner manner when it comes to collectives, we still make attributive judgments of a collective regarding mental phenomena through the outer manner. To some extent, it can be said that we judge a collective's dispositions to have mental phenomena only in an epistemological sense, but not in an ontological sense.

On the other hand, as mentioned above, cognitive contents can be attributed to a collective in two ways: integrating the cognitive contents of its members or making abstract judgments based on its collective behavior. Take common knowledge as an example. For the former, an individual's cognitive contents can be externalized using language and combined into common knowledge. Such common knowledge can be defined in terms of the individual level of knowledge (in the objective sense, such as the way of definitions in epistemic logic), although such reduction does not arrive in individual knowledge as inner states. For the latter, our judgment of collective knowledge in daily life mostly comes from attributive judgments based on collective behavior. For instance, if a company sues another company, we will ascribe to the plaintiff company collective knowledge, viz., the knowledge that its interests were harmed. In any case, such collective knowledge is not owned by a collective entity in the ontological sense. It is attributed through the observer's judgments on collective behaviors or through a construction from individuals' externalized knowledge, but these attributive relations only exist in the epistemological sense.

To sum up our views. At the individual level, an agent has a body, a brain, a mind, and an identity by herself, all of which are ontologically prior to her intentionality. So

the attribution of the three aspects of individual intentionality established in both inner and outer manners are the same; to some extent, we can say they exist both in the ontological and epistemological senses. A collective has no body, brain, mind, or identity by itself but only exists as a collection of relational eventualities. Therefore, we need to discuss the manners of attribution of collective dispositions in different aspects. As regards behavioral dispositions, an action is attributed to a collective in both manners; to some extent, such behavioral dispositions exist ontologically as well as epistemologically. But for phenomenal and cognitive dispositions, sensory experiences, emotional experiences, and cognitive contents are attributed to a collective only in the outer manner, i.e., we can only make these dispositional judgments in the epistemological sense.

According to the above analysis, we can still find a way to attribute purely mental phenomena, cognitive contents, and behavior to a collective through dispositional relations. By the dispositional definition of intentionality, we can make the attribution of intentionality at the collective level, i.e., collective intentionality is possible. Our epistemological judgment of collective intentionality does not come out of thin air but stems from a well-based ontological source, i.e., the collective viewed as a collection of relational eventualities, a collective ontologically related to its actions. Moreover, because of differences in attribution, the concept of collective intentionality cannot be reduced to merely individual intentionality.

3.6.3 The indispensability of collective responsibility

Two kinds of moral responsibilities can be distinguished, which are moral responsibility as attributability and as accountability. The former concerns moral evaluative judgments that concern the individual as such. The latter places the individual in a social setting and deals with blame (cf. Strawson (1962); Darwall (1996)); Watson (1996).

Attributive responsibility refers to those responsibilities we attribute to an agent by self-disclosing, which means in some cases where conduct is controlled by an agent's "deepest" values and ends, the related responsibility is defined in terms of the relationship between those values and his/her conduct. Take Watson's example:

> *If I dance clumsily, it is inescapably true of me that I was (on that occasion) a clumsy dancer. But if what I do flows from my values and ends, there is a stronger sense in which my activities are inescapably my own: I am committed to them. As declarations of my adopted ends, they express what I'm about,*

my identity as an agent. (Watson (1996))

Thus the notion of attributive responsibility is related to an agent's practical identity and his/her values and ends. So, could a collective have its own values and ends? The answer from the previous philosophical discussion is yes, at least in some cases. A collective's ends are included in the concept of a collective agent: it lies in the nature of a collective agent that it is capable of goal-oriented behavior. In line with Huebner et al. (2012) and Gilbert (2005), we believe collective values are "in principle enough to unify people into a social group as opposed to a mere aggregate"(Gilbert (2005)), and are analytically constitutive of the collective itself, or, minimally, necessary for the continuation of a collective endeavor. Values written in collective rules are a good example. For instance, the ILLC Diversity Committee is dedicated to providing a welcome and supportive environment for all, regardless of background or identity. The Committee is committed to promoting this value, which in practice will influence the decision-making and specific behavior of the institution or even the university. Now suppose someone disagrees and wants to question the result of such collective behavior. For example, he thinks that the implementation of diversity is not thorough and he is still be discriminated against or thinks that the implementation of diversity causes his application not to be treated fairly on the basis of quality. In such a situation, the ILLC Diversity Committee should accept such attributive responsibilities as a collective agent. In actual life, we do judge in this way; complainants, for instance, usually begin by consulting with the Committee. Therefore, for those collectives who have their values and ends, if such values and ends influence particular collective behaviors, we will say the collective bears the related attributive responsibility as an agent.

For accountable responsibility, in everyday life, we often blame a collective for immoral behavior, such as blaming a country for invading another country, blaming a company for tax evasion, or blaming a political party for corruption. In each of these cases, whom we blame is the collective. Suppose we admit a primary condition that only a moral agent can be held responsible. Then it follows that if our daily blames are justified and not illusory, then such collectives as listed above should exist as moral agents.

Moral individualists may object, usually for the following two reasons: First, unlike individuals, a collective cannot form intentions and therefore cannot be understood to act or cause harm as a collective; Second, As distinct from its individual members, a collective cannot be understood as morally blameworthy; it cannot bear moral responsibility

(cf. Weber (1978); Lewis (1948); Sverdlik (1987)). These reasons arise from a deep normative assumption that actions originate from intention and that responsibility requires corresponding blameworthiness, that blameworthiness arises from deviant behavior, and that deviant behavior arises from (at least morally) bad intention. Thus, collective intention is the premise of attributing collective responsibility. If collective intentionality does not exist as claimed by individualists, there is no foundation for collective responsibility.

However, in this subsection, we try to show that if we jump out of the individualistic stance we can find another way to show that a collective can be a moral agent and have responsibility. Then under the same normative assumption, we may have an alternative way to argue for the existence of collective intentionality.

As Cooper (1968) points out, not only do we ascribe responsibility to a collective in everyday practice but those collective responsibilities cannot be analyzed in terms of individual responsibility, "because the existence of a collective is compatible with varying membership. No determinate set of individuals is necessary for the existence of the collective"(Cooper (1968), p.260). The kind of collectives that Cooper is referring to, viz., those that do not vary with the individuals that make them up is consistent with the examples we have given above, and they all refer to the object of our analysis, namely the collective with a stable structure.①

Most philosophers (cf. French (1984); Copp (2007); List et al. (2011); Hindriks (2018); Collins (2019)) believed that a collective with sufficient organizational structures, such as a nation, a university, a political party, or a company, are sufficiently independent of their members to count a moral agents. Because there exists a sound structure to ensure the collective decision-making conforms to the criteria of the general definition of agent: representation of the external environment, goal-oriented motivation, and manifestation of essential rationality. From a functional perspective, such a collective can formulate and understand moral causes and act strictly in accordance with them, just as individuals do. Under ontological relationalism, the process of collective decision-making is restricted and guaranteed by the structure and function of the collective. Therefore, if we put aside the standpoint of individualism and consider the moral agent purely from the perspective of functionalism of interpretationism, the structured collective can fully meet its requirements.

However, even if we admit collectives as moral agents, there is still a question of

① There are also deontological analyses of unstable collectives, such as crowds and other accidentally formed groups, but that topic is beyond the scope of this chapter. For interested readers, see de Haan (2021).

whether the responsibility of a collective behavior should be assigned to or be accounted for by the collective. Let us imagine the following example:

> Company A suffered losses in the first quarter. According to the contract, company B should be compensated 5 million euros. Due to the large amount and company A's current deficient performance, the two companies agreed to pay the compensation in installments for a period of five years. During the five years, company A carried out a thorough reform from top to bottom and changed its business direction according to the market orientation, so that in the fourth year of the compensation, the last employee who was working there when they signed the compensation agreement left the company. Furthermore, all the owners or shareholders of the company have changed too. One year later, Company A is still in compliance and has paid all arrears at the end of the five-year period.

This example is quite possible. If we insist that there is no collective responsibility, the so-called collective responsibility is just the superposition of each member's responsibility. In such a case, the collective responsibility can be distributed entirely over all its members, i.e., with no redundant responsibilities. In other words, each owner or shareholder, and those higher or lower employees involved in a concrete irregular operation that causes losses are responsible for their part of the compensation of 5 million euros.① Thus, to say that Company A is responsible for 5 million euros refers to the aggregation of all its member's responsibilities. However, the problem occurs since the collective responsibility still exists when all the original employees and shareholders leave. As a moral agent, an individual cannot unconditionally transfer her responsibility unless it can be shown that her intentional action is actually affected by external forces; then, her responsibility can be transferred to the one who imposes such forces. In the current case, either we can say that the superposition of each individual responsibility cannot eliminate the collective responsibility, and ultimately (a part of) the whole responsibility can only be attributed to the company as a collective. Or we can say that the so-called individual responsibility

① The connection between the responsibility of a company and those of its owners and employees is more complex and diverse, depending as it does on different legal frameworks. Generally speaking, a small business owner may be solely responsible for all the debts of his company. Those employees who were directly involved in possible irregularities that caused the losses may bear their part of the responsibilities through salary cuts. But this is not part of the legal responsibility. In a large company, besides the shareholders, the debt liability also refers to higher employees (CEO, CFO, etc.) and those employees at the lower level who are involved in irregularities that caused the losses.

is caused by his collective identity, so these responsibilities can only be transferred to the collective when the individual leaves the company. In that situation, too, we can conclude the existence of collective accountable responsibility.

In line with Copp (2007) and Pettit (2009a), we believe that in order to avoid the emergence of irreducible responsibility and the loss of moral accounting resulting from individualistic interpretation, we should admit the existence of collective responsibility. In other words, a collective with sufficient organizational structure can act as a moral agent and assume moral responsibility in the normative sense. Moreover, depending on the different nature of the collective in particular instances, its moral responsibility can be distributed (partially) over its members.

The emphasis on the indispensability of collective responsibility is due to our purpose to strengthen our argument for collective intentionality. The previous discussion on collective responsibility focused on the idea that responsibility needs action and action needs intention, and thus the core of the discussion was the question whether collective intention exists. However, if we can conclude from purely normative considerations as such that the structural collective can act as a moral agent and bear its own moral responsibility, we can use this notion backward to argue that the collective can be intentional. Recall the normative assumption that responsibility requires corresponding blameworthiness and that is based on deviant behavior, and deviant behavior springs from (at least morally) bad intentions. So through collective action and corresponding collective responsibility, we can conclude that the collective has, at least morally wrong, intentions. Reviewing our discussion on the dispositional classification of collective intentionality, the bad intention here belongs to the disposition to do collective behavior, and the attributive judgment of such disposition is established under both manners. Therefore, from the perspective of collective responsibility, we can also draw the conclusion that collectives have intentionality.

Given the above explanation, we can ask the question why some people nevertheless do tend to think of collective intentionality along the lines of individual intentionality. The answer comes in three aspects: First, it is because of the similarity between different types of dispositions: volitional and physical; Second, it is because of the similarity between different manners of attributions; Third, it is because of our attempt to justify collective responsibility. The way in which we talk about these things in everyday language tends to blur these distinctions. However, when we reflect on these situations, we nevertheless can

discern the underlying distinctions and become confused about our incorrect usage. By clarifying collective intentionality in a dispositional way, our interpretation can make such a concept possible without presupposing any Hegelian mental entity. Which addresses the tension we pointed out in section 3.2.

3.7 Summary and Ideas for a Future Investigation

This chapter starts from the debate in the mainstream theory of collective agency and accepts the irreducibility of the concept of collective. However, such irreducibility has a fundamental incompatibility with the concept of intentionality. This incompatibility forces us to review the standard criterion of intentionality and gradually screen out which one is incompatible with the collective. We found that it is ambiguous to insist on a mental constraint. In this way, we study the relationship between intentionality and disposition, and bring both individual and collective intentionality into a unified disposition-based account, which can explain their similarity naturally. With this account, we demonstrate the existence of collective intentionality by analyzing different forms of attributive judgments and by introducing the indispensable collective responsibility.

Positioning our dispositional account in philosophical tradition may clarify it further. First of all, our way of interpreting is dispositionalist, but our definition of dispositions contains more than it does usually. Beyond dispositions confined to physical and biological cases, we add volitional dispositions to capture the unique feature of individual intentionality. Meanwhile, through the analysis of volition, we can naturally explain both individual and collective intentionality in terms of disposition. Secondly, our account is not dispositional behaviorism since our dispositions are also directed to mental and cognitive content and do not necessarily require dispositions to the associated behavior. Finally, our dispositional account adheres to irreductionism at both the collective and individual levels. Although we can have a unified understanding of individual and collective intentionality through dispositions, collective intentionality and individual intentionality are indispensable concepts from each other, they are typically different type of disposition.

On the basis of ontological relationalism of agency and dispositional account of intentionality, we further establish a unified foundation for discussing individual and collective. Upon which we are able to explore a broader range of topics; for example, linking philosophy of action or social ontology with general social sciences such as game theory or social choice theory, or with general social philosophy or phenomenological explanations

of intentionality, and comparing the similarities and differences between their underlying concepts of concern. In a unified framework, these comparisons can become more explicit and facilitate a macro understanding of the various parallel studies on collective.

Through the dispositional account, we can view the relationship between an individual and the world and the one between people in a more unified way. In this light, perhaps some of the literature we are familiar with will take on new meaning, like Jane Austen's famous words:"It is not time or opportunity that is to determine intimacy;—it is disposition alone. Seven years would be insufficient to make some people acquainted with each other, and seven days are more than enough for others.(Austen (2004), Chapter12)"

CHAPTER 4 INTENTIONALITY, PREFERENCE AND DEPENDENCY

In this chapter, we analyze how philosophical theories about collective agency relate to ideas from formal theories about collective decisions, such as game theory and social choice theory. Although the two fields are both concerned with collectives, differences in their relationship still need to be addressed. For example, both game theory and social choice theory are clearly anti-psychologistic since their aim is a formal and objective analysis. However, from the relational and dispositional perspective, intentionality at the individual level and collective intentionality in the epistemic sense inevitably involve mental content. In order to explain this difference and make clear where the boundaries are, we analyze the primary relationship between the involved three basic concepts, namely intentionality, preference, and dependency, so as to investigate how and why they differ and what they have in common.

4.1 Introduction

In previous chapters, we have explored the concepts of collective agency and collective intentionality from a philosophical perspective and have formulated two views, consistent with on another, namely, relationalism about agency (Chapter 2) and dispositionalism about intentionality (Chapter 3). The following two lines of research are worth exploring. One is in a more concrete direction, compared to philosophy, which takes all kinds of specific social situations as its objects, involving sociology, anthropology, social psychology, etc., and which can be expected to produce empirically validated results. The other is a more abstract direction, which takes the abstract relations among groups, individuals, and their behavioral objects as its objects, involving formal theories and logic, revealing the connection between them and which can be anticipated to produce various formal results. Of course, these two directions are not always in conflict, and some approaches, like game theory, focus both on abstract relationships and concrete situations.

The application direction is important, and scholars with a social science background may carry out this kind of research and achieve fruitful results. Nevertheless, since our experience and interest are more in the formal direction, we will only focus on the relationship between our philosophical viewpoint and formal theories. The ultimate goal is to

provide a new logical system in which we can axiomatize our ideas (at least the core part) and illuminate their relation to other formal theories of collectives. This will be done in chapter 5. In order to achieve this goal, some conceptual groundwork is necessary. That is what the present chapter is devoted to: we will compare the underlying ideas of existing formal theories with those of our philosophical approach, so as to clarify the differences and establish a common basis for the new logic.

Specifically, this chapter aims to link our philosophical interpretation with existing formal theories as a way towards the development of a unified picture of collective agency. Three core concepts underpin this unifying picture: intentionality, preference, and dependency. Starting from the relationship between these three concepts, we will elaborate on a series of issues, such as which concept is more fundamental, how dependency affects individual choice, and why psychological content can be abstracted at the collective level. After a detailed exploration of these basic concepts, we will formulate our philosophical views in the context of game theory.

In order to make good on this purpose, we need to complete a few steps. In Section 4.2, we will give a brief review of formal theories and elucidate the main issues. Then we will use the concept of desire as the hub to link the concepts of preference and intentionality in Section 4.3. In Section 4.4, we discuss interdependency relations, mainly from the perspective of ontological relationalism, to show its position in the philosophical framework. After these fundamental discussions, we turn to the context of game theory in Section 4.5 and express the specific requirements our philosophical views place on formal analysis through some examples. The summary and future directions are in conclusion.

4.2 Collective Agency in Formal Theories

There are many formal approaches to the study of collectives in the social sciences, such as social choice theory, game theory, etc. The most striking feature of these theories is that they all use individual preference as a primitive concept, rather than the more basic concept of desire, or the key concept in philosophical discussion: intentionality. Generally speaking, social choice theory deals with the problem of how to aggregate a given set of individual inputs (usually choices or preferences) into a collective decision, exploring the different effects of variant aggregation functions (cf. List (2022)). In comparison, game theory focuses more on the interaction of individual preferences and choices and attempts to analyze how this interaction affects the overall decision-making (cf. Ross (2021)). Ex-

ploring the problem of collective agency within both formal theories is promising work. However, we will only touch upon the question of how a philosophical account and a formal account compare within the context of game theory, and the reasons for that are two-fold. First, social choice theory already presupposes the range of a collective from which collective decisions can be derived through certain aggregate functions. For example, experts in a committee voting on a project, or a jury voting on whether a criminal is guilty, the so-called groups in all such scenarios are given and characterized by a pre-set aggregation function. In contrast, game theory allows us to talk about more general cases, in which it does not presuppose any collective in the beginning but instead discusses how the concept of a collective arises naturally from specific settings. Since we are also interested in how collective agency comes about and how collective and individual agency relate, we better illustrate the problem by means of the more general point of view of game theory. Second, the practical reason is that our logical analysis in the next chapter will be based on game theory, including narrative, definitions, examples, etc. So this chapter use game theory as an example and can also serve as an informal introduction to our formal work. Besides, there is already excellent work on the relationship between social choice theory and philosophical functionalism, such as List et al. (2011).

In game theory, the term "agent" that is used in philosophy is usually replaced as "player". Each player faces two or more strategies and has to choose among them. The term "strategy" means the way to play the game in which players know how to make choices to respond to the (anticipated) decisions made by other players, and the best option of a player or a group of players is their optimal strategy. This so-called "optimal" in the context of game theory means that the strategy has the largest payoff, and payoff means assignments obtained from utility functions, which map individual preferences onto a set of real number, with larger numbers representing higher preferences. Since this thesis focuses on collective problems, it is natural for us to consider optimality with respect to groups and explore how the group optimum is defined in game theory. Is it something that can be set in advance, or does it have to be generated from the concept of individual optimality? The answer introduces a fundamental distinction in game theory: that between cooperative and non-cooperative games.

Non-cooperative games are based on the absence of coalitions, with each player making individual decisions based on the individual decisions of other players. Cooperative games, on the contrary, allow various coalitions and their interrelationships to exist, i.e.,

they allow players to pre-commit to a collective strategy to realise a collective desideratum. However, the processes through which such collective strategies and desiderata come about, are assumed as a given, and are not analysed within the theory of cooperative games itself. Non-cooperative games focus on the procedural level, that is, the interaction between agents' strategies, and focus on the utility maximization of individual choices in a given situation. Cooperative games abstract away from these detailed interactions and focus on the coalitional level, that is, how the coalitions that agents can form affect the decisions of their members. Note that 'cooperative' and 'non-cooperative' are technical terms and do not represent whether the game players cooperate or compete. In other words, cooperative games can model extreme competition, and non-cooperative games can model intimate cooperation.① In the remainder of this chapter, we will only discuss non-cooperative games. But in the next chapter, we will use logical tools to show the close connection between some basic concepts of cooperative and non-cooperative games.

It should be stated that a presupposition of non-cooperative games implies a lurking idealization of individualism. Many philosophers (cf. Jones (2005); Godfrey-Smith (2009); Stokhof et al. (2011)) pointed out that there are two different ways for a formal theory to deal with parameterized properties of natural objects: abstraction and idealization. Abstraction assigns a specific value to some quantitative parameter of its object, and such parameters are present in the resulting model. Its primary motivation is methodological and practical. Idealization, on the other hand, simply ignores a qualitative feature of its object and does not present such features in the generative model. Its motivation is mainly ideological and theoretical. In non-cooperative games, each player making individual decisions based on the individual decisions of other players is a typical example of abstraction. Possible agreement between individuals, or the possible collective relationship formed by them, is completely ignored and eliminated from the game model. Combined with our criticism of individualism in previous chapters, it seems that the non-cooperative game-theoretic approach faces the same dilemma. Nevertheless, we will show that this underlying individualistic presupposition does not prevent us from expressing the concept of a collective in non-cooperative games, since dependencies still exist (cf. below on dependencies vs. relations).

The choices a player makes in a game constantly interact with the choices of other players, and such interaction is mainly reflected by the interdependency of the agents'

① See Chatain (2016) for more details regarding cooperative and non-cooperative games.

choices. That is, given the choices of other players, an agent chooses their most optimal strategy. In other words, each individual's choice depends on the choices of other players in the game, and this dependence relationship is mutual. Such dependencies can be understood through some everyday examples. Consider soldiers on the front line, who can choose their relative positions in the unit, such as going first or bringing up the rear (in extreme cases, deserting). The commander is deciding between attacking and retreating. In conventional positional warfare (regardless of other factors such as shelling), when the commander chooses to attack, the soldier's optimal choice is to move to the rear of the unit, because the most likely casualties will be in the front. When the commander chooses to retreat, the soldier's optimal choice should be to move to the head of the unit, since the tail is vulnerable to pursuit. Of course, for each soldier, the optimal solution is to flee to save his own life, because no matter whether he wins or loses one battle, there is still a high chance of death in the next one. Using game-theoretical thinking, commanders can anticipate this behavior and use backward inference to set anti-retreat troops to shoot deserters, eliminating the option for soldiers to avoid battle to force them to have the determination to fight to the death. Many similar examples can be found; For instance, in Chinese history, Xiang Yu[①] ordered his soldiers to break the cooking caldrons and sink the boats after crossing. By ruining the soldiers' alternatives, the only option they have is to fight to the death, thus strengthening their morale and forming a stable collective decision.

The existence of this kind of interdependency is the first premise of game theory, leading to an essential concept: Nash equilibrium. A Nash equilibrium (hereafter, equilibrium) is a stable state that results from intrinsic mutual restrictions. It is a state in which no player can make better choices if the other players stay put. As long as those mutual restrictions exist, in the technical sense, any combination of individual strategies that does not qualify as an equilibrium does not have the possibility of realization. Of course, in real life, as people are not perfect reasoners, any outcome could be the result of actual game play. We will come back to this later in section 4.4 where we will compare this kind of mutual restriction in non-cooperative games with relations that constitute collectives, and investigate how the latter kind of dependencies relate to our relational account

① Xiang Yu was a legendary general at the end of the Qin Dynasty, was in charge of the core force to overthrow the Qin regime, and was later defeated by Liu Bang, the first emperor of the Han Dynasty. The story quoted here takes place in 208 BC, when Xiang Yu went to rescue the king of Zhao, who was besieged by Qin troops at Julu county. After crossing the Zhanghe River, he ordered his troops to eat a full meal and carry enough dry food for three days, and then sank the ferry and broke the cooking pot, showing that he would not withdraw or turn back, and finally won the victory.

of collective agency. For now, we temporarily stick to introducing basic concepts in game theory.

In any non-cooperative game, due to the absence of external agreement, agents risk potentially huge losses if they arbitrarily choose a nonequilibrium collective optimal strategy, such as in the well-known "Prisoners' Dilemma" as follows:

		Prisoner2	
		cooperate	confess
Prisoner1	cooperate	2,2	0,3
	confess	3,0	1,1

Table 4.1 Prisoners' Dilemma I

The two rows are Prisoner 1's possible strategies, and the two columns are Prisoner 2's possible strategies. In each resulting state, the row's payoff is always listed first, followed by the column's payoff. When more players get involved, their payoffs will be listed following the two. If both players cooperate rather than confess, neither will be sentenced, so the payoff in the state (cooperate, cooperate) is (2,2). If one of them chooses to snitch and blame the crime on the other, the snitcher will be rewarded instead of sentenced, and the wretch will face a heavy sentence, so the corresponding states are (3,0) and (0,3). Moreover, if both of them choose to confess, then they all get sentenced, but not more severely than if one person confesses to the crime; thus, the corresponding state is (1,1).

It is important to note that the Prisoners' Dilemma is essentially a mathematical matrix structure with two rows and two columns, in which only exists relations between numbers. We can make up any kind of concrete story to explain it, such as that of the commander and the soldiers mentioned before. We just need to fill attack and retreat in the commander's optional strategies and move to rear or front in the soldier's options. ①

An intuitive and quick observation would be that (2,2) is the optimal solution to this game because they have the largest sum and seem to benefit the most. However, this is where the notion of equilibrium comes into play. Suppose the resulting state is (2,2). Prisoners 1 and 2 will find that it is more profitable to confess if the other chooses to cooperate, thus changing the state to (3,0) or (0,3). The two states will eventually change to (1,1) because the other player finds the cooperation fruitless, thinking along the same lines. Another way this might come about is that both prisoners simply note the sum payoff

① It is in everyone's best interest to attack and win, but in a situation of poor morale, the commander may send his soldiers to their deaths and flee, or the soldiers may flee and leave behind the commander. In the worst case, both the commander and the soldiers give up.

of two strategies and chose the higher payoff strategy, so the outcome state is (1,1).

Some studies (cf. Bacharach (1999); Bacharach et al. (2006); Sugden (2003); etc.) point out that if we can change the thinking mode of all players or make each of them subject to additional agreements, we can achieve the optimal collective strategy that cannot be endogenously achieved in the original game, like (2,2) in table 4.1. This line of thinking is similar to our philosophical discussion on collective agency, where we also highlighted additional conditions that a collective agent needs to satisfy to maintain its ability to realize rational behavior constantly. In particular, Bacharach (1999) points out that it is the team reasoning pattern of the we-perspective (members as team reasoners) rather than that in which players care for the team from a purely personal perspective (members as team benefactors) as indispensable in coordinate games. Namely, the considerations of the members of a group should shift from the question of "*what should I do for us?*" to "*what should we do?*" to achieve an optimal outcome for their group. Similar to the philosophical divergence on the source of collective intentionality, several game theorists have offered different views on the reason behind team reasoning. For example, Bacharach et al. (2006) speculates that reasoning patterns depend on the mental framework of the decision-makers. Sugden (2003) argues that decision-makers can endorse a particular reasoning scheme, but such choice may lie outside the individual rationality. This means team reasoning needs guarantee from each member, which is, according to him, agreement or joint commitment. Hurley (2005) suggests that group reasoning may be consciously chosen within rational deliberation. Furthermore, Karpus et al. (2017) proposed a formal measurement of mutual advantage and suggested that team reasoning is a switch from personal best-response to mutually advantageous when the former cannot resolve the current decision problems definitively.

In contrast, the opposing voice holds that any condition change outside the game will eventually be reflected inside the game, thus fundamentally changing the current situation we are facing. In other words, the solution realized in this way corresponds to other games instead of the current game (cf. Binmore (1994); Stirling (2012); Ross (2014)). For example, the original game in table 1 will be changed by team reasoning into the game in table 4.2. Because if the agent adopts team reasoning now, then his preference order should have been changed according to it, i.e., cooperative options are sequentially preferred over non-cooperative options. Furthermore, if the utility function is defined to assign values to the individual preference, such a change should be reflected in the game

model, where the payoff of cooperation is at least as large as the payoff of confession. As mentioned above, the game model is essentially a purely mathematical structure. The influence of (cooperate, cooperate)'s payoff changes from (2,2) to (4,4) is not only on the value of the current state, but also on the relationship between the current state and other possible states; in a word, it changes the entire game. In fact, just by definition, we can find that (4,4) itself is one of the equilibrium of the new game, without adopting any kind of team reasoning.

		Prisoner2	
		cooperate	confess
Prisoner1	cooperate	4,4	0,3
	confess	3,0	1,1

Table 4.2 Prisoners' Dilemma II

The above doubts are reasonable, and adopting team reasoning does change the game model, but that does not mean team reasoning is meaningless. At best, it just means that in some environments, we simply cannot have any collective thinking at all. In other words, the existence of teams depends on the current environment; they cannot take root without suitable soil. On the other hand, note that after imposing team reasoning in the Prisoner's Dilemma, the new model (in table 4.2) actually has two equilibria (4,4) and (1,1).① Players cannot move from (1,1) to (4,4) just by following equilibrium and individual optimality, which means we need some notion of collective optimality to achieve this change. Several studies of team reasoning (cf. Hodgson (1967); Bacharach et al. (2006)) argue that a reasonable understanding of collective optimality should imply Pareto optimality. Pareto optimality roughly means that the group reaches a state where it cannot make any members better without making other members worse. But do note, though, that Pareto optimality does not ensure that there is no chance for any player to exploit the current collective state by deviating from their current strategies. For instance, in the standard Prisoners' Dilemma in table 4.1, (2,2) is Pareto optimal, but players still have reasons to deviate from it. In other words, Pareto optimality may be unstable. Recall that stability in game theory is usually associated with the notion of equilibrium. From the perspective of stability, it seems that equilibrium should also be taken as a necessary condition for achieving a stable collective state.

① By definition of equilibrium, if the current state is (1,1), there is no reason for either player to change their strategy since unilateral changes would result in worse outcomes.

We summarize the main points of the above review and briefly explain how it relates to our philosophy of collective agency as follows:

1. In game theory, there is also an attempt to pursue collective agency: team reasoning. In the context of game theory, a group demonstrates collective agency if it ends up with one of its collectively optimal states among those feasible for it.
2. Not all situations can produce a state similar to collective agency, and the game setting can avoid the generation of any collective factors, such as the Prisoners' Dilemma. In other words, collective agency is contextual, i.e., a state that can be generated and vanish depending on changes in its environment.
3. Pursuing transient collective agency is meaningless since agency requires at least a consistent rationality of action choices for a certain period. In the context of game theory, this means that a collective does not change its choice easily in a given setting; that is, collective choice is stable. Stability is guaranteed by the concept of equilibrium in game theory, so equilibrium should be one of the necessary conditions for collective agency.
4. Equilibria alone are not sufficient. In games with multiple equilibria, the collective has no incentive to switch from the non-optimal equilibrium to the optimal one. Therefore, to realize such an incentive, we need to admit that concepts like Pareto optimality are one of the necessary conditions of collective agency.

To sum up, collective agency can be discussed in the context of game theory as long as one is willing to make accept several fundamental presuppositions; namely, collective agents are rational and contextual dependent and their actions aim for Pareto optimality and Nash equilibrium.

It can be seen that our philosophical outlook is consistent with these fundamental presuppositions, such as the properties of collective rationality, the contextual dependence of the collective, the stability of the collective, and the prioritization of the collective interest, etc. In particular, formal theories are mainly based on two core concepts: individual preferences and mutual dependence, which are considered sufficient to generate a collective preference. In other words, no concepts such as collective preferences are presupposed; instead, collective preferences come from the interactions between individual members' preferences. A similar view can be found in our claim of ontological relationalism, in which individual intentionality and stable interactions among them (represented as collective structures and functions) endogenously form collective agency. This means that

a combination of philosophy and formal theory may be feasible—still, there is a gap between the two fields. The most pressing theoretical difficulty is that the formal theory of the collective is based on preference, while the philosophy of the collective is based on intentionality.

When we comprehensively explain some specific social situations or further explain the relationship between individual psychological content and collective behavior, we must explore beyond the scope of formal theories. This implies that the philosophy and formal theory of collectives are basically consistent; their difference in basic concepts of intentionality and preference is due to the division of labor when faced with specific problems. Philosophy inclines to seek unified answers for general cases, and intentionality suits the such purpose. While theories tend to achieve analytical results from models in which only certain characteristics of the object are idealized into objective, comparable value; thus, they pick the preference. As we have mentioned, psychological content occupies a place in the philosophical interpretation of relationalism and dispositionalism, but apparently, these contents are not included in the kind of formal analysis that game theory exemplifies. However, a formal discussion of collective agency has to involve many of the best results from game theory, which makes it necessary for us to clarify the relationship between the concepts of intentionality and desire in philosophy and the concept of preference in formal theories.

4.3 Desire as the Hub for Preference and Intentionality

4.3.1 Preference in game theory

Game theory defines an economic agent as an entity with preferences and compares the object of preferences in terms of utility. According to the traditional interpretation, utility refers to the ranking of subjective welfare of various possible options by economic agents. What drives the agents to evaluate the utility of different options can be income, natural resources, moral pursuit, or any other desideratum. This interpretation implies that individual preference is directly related to some psychological process of the agent; in other words, it is the external expression of the individual's subjective likes and dislikes. Such an interpretation is heavily influenced by Bentham's utilitarianism. With the development of behaviorism and empiricism in the 1930s, economists pursued objectivity and refused to hold unobserved and unmeasurable psychological processes as the basis of utility. As a result, the revealed preference theory proposed by Samuelson (1938) re-

defines utility as a purely technical concept and completely drives out the psychological content. Under the new explanation, an economic agent maximizing the utility of their action simply means they consistently take those actions to improve the possibility of its realization. This new definition sounds like a tautology, and it actually is. It fixes the explanation of preferences at the level of behavior. From then on, the economic agent is the entity that takes action to maximize utility, action is the process of maximizing utility, and utility is the object that the economic agent tries to maximize. These concepts interlock at the basic level and are not explained further, thus forming a starting point for the whole theory. Questions concerning individual preference are excluded from the system. From a philosophical point of view, the adoption of this particular starting point of the theory can be questioned, because philosophy always favors universal and uniform answers and is not content to stop at artificial boundaries. However, from the point of view of formal theory, this presupposition is consistent and acceptable. Recall the afore-mentioned distinction between idealization and abstraction; what game theory does here is not to deny or ignore various fundamental features of the agent, either ontologically or ideologically, but to abstract away overly psychological details for methodological and practical considerations. Through such abstraction, game theory can provide quantitative parameters for preferences and their derived utilities, thus serving the objectivity of formal analysis.

We have no intention to attack the prespecified part within a formal system, but it should be pointed out that when our attempts to explain collectives and their properties go beyond the scope of the formal system and pursue a unified picture, whether preference can be served as the background color needs to be questioned. There are at least two ways in which we need more systematic answers than the interlock of preferences: when we are not content with tautological presuppositions and attempt to ask further questions about the sources of these preferences, and when formal theory is applied to the analysis of concrete situations. As a normative system, game theory only abstractly explains the minimum and most necessary part of rationality in the situation, but an abundance of additional phenomena emerges in the application process, including mental contents. This line of thought leads to a concept we usually associate with preferences: desire.

4.3.2 Desire in philosophy

Although the concept of desire is well known to everyone, it is not easy to come up with a comprehensive and reasonable theory of it. This is because our daily understanding of desire is confined to specific examples such as the want to drink, the preference for

tea, or the desire for family love. However, among these situations there is no unified understanding of their nature, causes, and manifestations. A philosophical interpretation of the concept of desire needs to include not only an explanation of the concept itself but also some familiarity with the existing varieties of the concept.

A relatively intuitive, simple, and conservative explanation of desire regards dispositions to act as the most significant nature of desire. For example, if I desire to have coffee in the morning, this means I am disposed to get some coffee, and disposed to feel good about coffee, and maybe disposed to think coffee is good for my work, etc. This is the most widely-held theory: action-based theories of desire. As Anscombe mentioned, "The primitive sign of wanting is trying to get."(Anscombe (1957)) When we say someone desires something, we mean that she is disposed to act to bring it about. Whether such an act will succeed and such a desired object will become true does not hamper her desire for it. Along this way, a naive action-based definition of desire is

> A creature desires p is the creature to be disposed to act so as to bring about p.①

However, finding a counterexample to this naive definition is easy. Suppose a gymnast is prone to perform a challenging routine and is more likely to make a mistake. Then it follows from the definition that the athlete desires to miss because she is disposed to act so as to miss. Given this, a more sophisticated version of the action-based definition of desire is

> A creature desires p is the creature to be disposed to take whatever actions it believes are likely to bring about p. (cf. Smith (1987); Smith (1994))

However, this version still has counterexamples, such as a musician choosing to play something technically tricky and making a mistake. The musician knows the difficulty and the possibility of mistakes, but she still chooses to do it. Even so, it is hard to say that the musician is eager to make mistakes. And a more general and direct criticism is whether desire should be defined by dispositions to act. Is desire driving our actions, or is it something else independent?

Another widely-held way to interpret desire highlights the concept of pleasure. A problem with the action-based definition is that it cannot distinguish desire and the judgment of goodness. What distinguishes these two, in the view of many philosophers (cf.

① The reader may find that this definition is precisely the behaviorist aspect of dispositionalism as mentioned in Chapter 3.

McDowell et al. (1978); Scanlon (2000)), is the satisfaction of pleasure. People's judgments of goodness were related to their dispositions to act but irrelevant to whether such acts bring pleasure. On the contrary, the concept of desire is naturally associated with pleasure, even more so than dispositions to action. Desire motivates us to act contingently, but it is associated with some feelings necessarily. A simple definition along this way is

> A creature desires p is the creature to be disposed to take pleasure from situations seeming to contain p and displeasure from situations seeming to contain $\neg p$. (cf. Strawson (1994))

Two reasons suggested by Galen Strawson defend such a definition: first, exhibiting desire requires consciousness, and in terms of consciousness, pleasure and displeasure are most closely linked to desire. Second, it is possible to imagine living things with no disposition to act, but that do have dispositions to feelings of pleasure and displeasure, and subsequently to desire objects that can produce pleasure for them. (cf. Strawson (1994)). Other philosophers questioned whether pleasure could be identified with desire, since it is more likely that the satisfaction or dissatisfaction of desire causes pleasure or displeasure. They are like causes and effects, which are ontologically distinct (cf. Davis (1982); Schroeder (2004)).

Other represented definitions of desire are good-based theories of desire (cf. Price (1989); Byrne et al. (1997); Broome (1991)), attention-based theories of desire (cf. Scanlon (2000)), and learning-based theories of desire (cf. Dretske (1991); Schroeder (2004)). We do not want to go into the detail of every theory of desire that has come up in the history of philosophy, but it is essential to point out that each theory faces counterexamples and fails to describe every concrete situation of desire. If we take the risk of combining these single-feature theories of desire, the resulting definition is not well-defined because it will cause a contradiction in the interpretation and value of desire. According to the unified single-feature definitions, we arrive at a pluralistic definition of desire, in which each feature (the disposition to act, the pleasure from satisfaction, the pursuit of goodness, etc.) is a universal characteristic of desire, and the fulfilment of any one of them counts as desire. However, an implicit problem is how to deal with cases that do not fit into some features of the unified definition. For instance, if a person is disposed to work, his purpose is only to earn money to sustain his life, and there is no other feature of desire, such as

pleasure and good.① At this point, according to the unified single-feature definition of desire, one observer might conclude that the observed subject has a desire to work based on the fact that the subject works hard every day. Per contrast, another observer might conclude that the subject does not want to work because the subject does not obtain any pleasure from it. Furthermore, even if we identify an agent with a desire to work and a desire to watch a football game, we still lack a comprehensive measure of how strong or weak the two desires are. There is no direct conversion relationship between the strength of behavioural disposition and the strength of pleasure, so the theory needs to add more content to reconcile the differences between various features of desire, and the resulting theory would be far beyond our needs. ② The core reason for this phenomenon lies in the fact that every feature in the unified definition claims to be the universal feature of desire, so we have to examine every aspect of the case and judge whether it is desire and what is its strength, thus providing a hotbed for contradictions.

In contrast, a more feasible approach is to acknowledge that the concept of desire cannot be defined by grasping only one particular aspect. On the contrary, it is a concept of Wittgensteinian family resemblance, which satisfies various characteristics in different contexts of language use. In other words, the concept of desire is composed of a series of overlapping similarities, none of which is common enough to cover all its extensions. Following this line of thinking, we can adopt a holistic account of desire, which comes in two primary forms: functionalism and interpretationism. In the functionalist interpretation, a desire is an internal state-type that plays sufficiently many of the causal roles suggested by its various similarities, such as dispositions to act, pleasure, judgments of goodness, sustained attention, reward-based learning, etc. (e.g., Lewis (1972)). In the interpretationist form, desires are treated as states of the creature so that it displays acceptable behaviors that fit these similarities that can be rationally interpreted as having desires (e.g., Davidson (2001)). When judging desire, we can adopt these two perspectives at the same time. As long as an agent's ability can satisfy the conditions of possessing desire, or we intuitively adopt desire as the explanation when rationalizing an agent's performance, the mental

① Defenders of the pluralistic definition of desire might argue that this example is too extreme because people usually do not have only one purpose. Nevertheless, it is essential to note that our identification of purpose or desire depends heavily on the contextuality of the description. Different descriptions of the same action lead to different intentionality conclusions. Not all desires in all contexts can be considered relevant, and there may be contexts where only one desire is relevant. Thus, such a defense is actually further evidence of interpretationism, that we always identify an agent's desire from an observer's perspective by rationalising his/her action.

② Admittedly game theory adopts a strategy of "monetising" the desire and compares the utility outcomes quantitively. However, such a strategy already involves a conversion process that maps the power of agents' mental states into some comparable parameters.

states of the agent in these situations can be judged as desire. Since interpretationism and functionalism begin by acknowledging that no single feature can universally describe desire, we are exempt from examining every aspect of the phenomenon in the first place. As long as the subject we observe has the disposition to work, or the observer can use the concept of desire to rationalize the behaviour of the observed, we can judge that the observed desires to work. This kind of judgment is more relaxed, and there is no need to examine every aspect of the case, thus eliminating the possibility of contradiction. Therefore, by adopting a holistic account of desire, we avoid counterexamples, meanwhile maintaining a well-defined concept of desire that present a unified interpretation and value.

4.3.3 Desire or preference, which is fundamental?

In the holistic account, we regard all mental states that fit the family resemblance similarities of desire as desires, and these desires can be compared in intensity. Making these comparisons we do not need to go back to the similarity level, because we can get the strongness and weakness of different desires just from the strength of the agent's feelings toward them. Such comparisons, of course, involve different situations. For desires with purely external constraints, agents have to make the comparison due to external reasons. For instance, when an agent is faced with a life-threatening situation, it is natural to activate the survival instinct, and her desire to survive is the indisputably strongest. For desires partially conditioned by external circumstances, an agent will make her decision by combining her own feeling and external constraints, such as not having enough money in her pocket to pay for food and drink, so comparisons must be made between desires for each of them. The last one refers to a purely internal comparison of agents, such as a person who, without external constraints[①], wants to decide whether to spend the next four years studying for a doctorate or making films.

The comparison between various desires implies that desires can be measured in terms of specific parameters. We can have three ways to grasp these parameters in the holistic account, one is from the agent's perspective, and the two others are from the observer's perspective. The agent's perspective is more psychological, in which those parameters

① Admittedly, such a condition is idealized. There is almost no comparison of desire without external constraints in real life. After all, a desire by itself is a kind of intentionality that is in any case always related to an object (existing or imagined). Even those comparisons of desires on purely mental phenomena or cognitive contents, such as one's desire for different dreams, still depend to varying degrees on one's knowledge of the outside world and thus inherit more or less some of the basic external constraints. We do not intend to discuss such matters too much, but just point out that, at least in the theoretical sense, there are arguably comparisons of desires that are entirely free from external constraints.

could only be known by the agent since only she directly knows her mental states that play causal roles in action. The second method is the interpretationist way with verifying (the idealized interpretationist way), in which the observer makes the best possible backward induction based on the phenomenon, and then asks the observed to verify whether the induced mental state is valid. This approach reflects reality as much as possible and is typical in daily life. Nevertheless, such a way is excluded from game theory since it requires too many communication steps. The more conditions a theory requires, the narrower its scope of application. Conditions as strong as effective communication would significantly limit the scope of game theory, ruling out a large number of non-cooperative games without communication. The third and last one is the interpretationist way without verifying. By rationalizing without communication, although what the observers have is different from the original desire, these analyses derive directly from the agent's action fact and thus do not rely on verification, unlike the previous cases that involve subjective elements. However, from a philosophical point of view, interpretationism cannot wholly avoid psychological factors because the mental state obtained by rationalizing the agent's action still belongs to some speculative psychological process. Even so, compared with the first path of pure internal content and the second path of lack of efficiency, interpretationism without verifying by the agent is more in line with the needs of formal theories. Often in an interpretationist way without verifying, the strength of desire is measured in terms of how much control an agent has over subsequent actions. Suppose other things being equal, and each desire requires a distinct action that cannot be satisfied simultaneously. One desire is stronger than the other means the agent is disposed to do the actions required by the stronger desire.

Most philosophical theories that analyze desire pay little attention to the strength of desire. Nevertheless, one exception is decision theory, in which the primary concept, that of preference, is directly based on the premise of comparing different desires and steadily choosing the stronger desire (e.g., von Neumann et al. (1947)). Another exception is Stalknaker's reductionist interpretation of "want/desire", where he argues "wanting something is preferring it to certain relevant alternatives, the relevant alternatives being those possibilities that the agent believes will be realized if he does not get what he wants.(Stalnaker (1984))" Which implies a preorder between possible worlds that include or do not include the desideratum. However, from the perspective of philosophical explanation of desire, it is doubtful whether a stable preference can be formed. The reason is that desire changes

all the time. It seems that there is no constant preference in our daily life, i.e., one in which the desire for all objects forms a fixed order. For example, if I prefer coffee to wake me up this morning, I do not have a preference to prioritize coffee, milk, juice, tea, etc. Instead, I just think of coffee and decide to drink coffee, regardless of any other unnecessary objects. In addition, daily desire often changes based on the environment. For example, when a person is in a good mood, she wants to travel more, and when she is in a bad mood, her desire to travel will decrease. However, it seems that a person's preference for travel is not directly related to whether she is in a good mood or not. We can say that a person likes to travel, but at the moment, she is busy and does not want to travel. The first example deals with how preferences actually arise, and the second deals with how we understand preferences and desires. The latter is relatively simple because if we use holism to explain desire, it allows the existence of desires without current dispositions to act or to feel pleasure. Meanwhile, it is intuitive in such situations to define preference as habitual choices between different desires, independent of incidental ones.

However, the question of how we understand preference, is more complicated. Decision theorists take preference as their key concept and see human beings as making choices between options, and such choice-action shows their preference and agency. As mentioned above, the expected utility of an action is the expected pleasure or relief from suffering from the object, according to the older approach, or it is the degree to which the agent would be inclined to choose it, according to the modern approach (cf. Skyrms (1990)). Whether we take utility as pleasure or choice worthiness, it is compatible with the holistic definition of desire. Nevertheless, the question remains which one is more fundamental.

If we choose desires to be the more fundamental concept, as long as desire has the dimension of strength, we can easily form an agent's preference, at least partially. However, if preferences are basic, we have to answer how a consistent, constant, total, and prior preference comes out naturally, which seems like a conundrum in a realistic framework. As Pollock (2006) argues, just to encode the preference that could be generated from just three hundred basic facts about desires, the number of objects of such a total preference is an order of billion billions at the very least. This is highly problematic because the human brain is limited at the physical level. So, from a realistic perspective, it is difficult for the human brain to understand and calculate pairwise preferences as basic concepts.Therefore it appears to be much more realistic to assume that basic desires are fundamental concepts.

So it seems that using preference as a theoretical basis abstracts desires away. Is this abstraction worth it? Depending on what we want to talk about, there are different conclusions. If we tend to avoid any psychological factors, with their inherent instability, when discussing formal systems, a simple and definite starting point is more attractive. When we talk about preference in formal systems, we know that philosophically it derives from desire, but for the sake of formal certainty, this psychological element has been abstracted away, and no explanation is forthcoming for those given preferences. Per contrast, philosophical theories of desire can serve as a background of the formal theory, and desire enters into the discussion only when we try to question further the source of given preferences and when we try to apply formal theories to concrete situations. As von Neumann and Morgenstern put it, "every measurement — or rather every claim of measurability — must ultimately be based on some immediate sensation, which possibly cannot and certainly need not be analyzed any further. In the case of utility the immediate sensation of preference — of one object or aggregate of objects as against another — provides this basis" (von Neumann et al. (1947), p.16).

4.3.4 General discussion: intentionality rather than desire

As discussed in the analysis of intentional states in Chapter 3, desire is a state that by itself intentional. In all respects, it conforms to our criteria of individual intentionality (from Chapter 3), such as directedness (it is always directed towards an object or a state of affairs), in-existence, indeterminacy (the directed object is in-existent and indeterminate), and volitionalness (the agent is willing to desire so).

Desire is one of the main psychological states that reflect intentionality because of its purpose and directivity. Can all the preferences and choices of an individual agent be reduced to desires? Some counterexamples are worth considering. An extreme one is addiction. A person may smoke one cigarette after another while writing because of addiction, but he does not have a clear desire to do so, and may only have some accompanying unconscious thoughts. More generally, a person may make a particular choice only because of his belief, knowledge, mental image, perception, etc. Instead of just desire, each of these states by itself could be reason enough for the agent to make a choice.

Without loss of generality, an agent's preference or choice comes from their long-term or current state of mind, and this mental part involves volition. This fit into our dispositionalist interpretation of intentionality, according to which agent A intends P iff agent A has a volitional disposition to bring about P. We believe that the agent's disposition,

instead of just desire, towards behavior, pure psychological phenomena, and cognitive content in individual scenarios is sufficient to provide fundamental philosophical support for asking the question as to the source of preference and applying formal theories.

On the other hand, we also note that there are efforts in game theory to relate the interpretation of the rationality of economic agents to the intentional stance. As Ross (2021) puts it, game theory supposes that economic agent has the following rationalities: First, the ability to assess each outcome in terms of their contribution to the agent's welfare. Second, to calculate practical actions matching each outcome and assess which path is the most realistic. Third, make decisions between alternatives and select the most preferred one, given the choices of other players. These rationalities are akin to the way in which philosophers like Dennett (1987) characterize rationality from the intentional stance. In view of that we can conclude that intentionality is more suitable to be the basis of preference.

If we assume the dispositional account, it is permissible to drive psychological content out of the collective level. This is because, in the dispositional interpretation, the psychological content is included in the concept of voluntariness, and the collective level does not involve any voluntariness in the ontological sense. Voluntariness only appears at the individual level and at the epistemological perspective on the collective. This shows that the dispositional account of intentionality is consistent with the primary setting of game theory, in which individual preferences and interdependence are given and from which we infer collective preferences.

One more question we have not discussed is why we can still talk about collectives in the context of game theory given that it implies individualism. The above comparison of economic rationality in game theory with our dispositional account of intentionality shows that they have much in common. For example, neither of them stipulates any independent group entities and both exclude psychological factors form collective discussion. Game theory focuses on interactions between players and need not verify players' subjective elements, and the dispositional account of intentionality considers collective intentionality as something that emerges in a developed stage and excludes psychological contents on the collective level. Furthermore, the critical point is that game theory excludes only prespecified groups, not relationships between individuals, i.e., dependencies. In other words, we can talk about the collective on the basis of the interaction between individuals because, in the relational account, the essential feature of a collective is its relational pattern and not

some independent entity or a set of individuals. On the other hand, game theory differs from our philosophical account, especially on institutional relations. Game theory cannot explain the irreducibility of such stable group relational patterns, yet in the relational account, relations as group structures are ontologically different from the inter-individual accessible relations between members. Nevertheless, while there is an ontological divide, it does not stop us from using game theory as an analytical tool to gain insights. As we discussed in Chapter 2, the difference between the collective stable relational pattern and the accessible relations among individual members is a matter of emergence of the former from the latter. The advantage of game theory is that it can depict the process of such occurrence and generation, which is needed for further clarification of the relational account. Non-cooperative games can show more details of how groups emerge from individuals and their interaction process, while cooperative games can show more details of what kind of conditions a coalition needs to meet. These two are just different perspectives, but each can bring us enough insights into the relationship between the philosophy of agency and formal theories.

4.4 Dependency: a High-order Relation

In a game-theoretical setting, given individual preferences are not enough for us to get preferences of a collective, because there are also key properties of a collective that we need to consider, namely the interactions between players. Philosophically speaking, these interrelationships are concrete concepts such as organizational structures or insititutional functions of a collective. However, in the formal system, these structures and functions are reduced to the dominant or decision relationship between one part of a collective and another; in other words, it is the dependency between various parts of a collective. In order to get a better understanding of formal theories of collective agency, we need to understand the properties of this dependency and how it relates to a individuals' preferences.

There is no doubt that dependency is a relationship that connects actions and anticipations of different agents and subgroups of agents. Nevertheless, what is the source of this relation? Just as we are interested in the source of a given preference in a formal system, we are equally interested in the source of a given dependency. From the perspective of the relata, dependency is different from preference because the relata of dependency relations are preferences of agents instead of the agents themselves or other concrete realities. When we say that the behavior of an agent depends on that of others we do not mean that

one agent depends on the other but that there is a mutual relationship between the choices or preferences of these agents. If the preference generated by an agent's dispositions with respect to different options is regarded as the basic relation, then dependency is a relation between different preference relations, that is, a second-order relation.

To illustrate this, we can take some inspiration from the discussion of intentionality. Besides basic intentionality, we also have higher order of intentionality. Here is an example. When someone is working but other people in the same building hold a boisterous party, he naturally will have the desire to yell at the party people. However, at the same time he might have a higher-order desire, such as that he wants to be a moral person. We call it a higher-order desire because it is the desire that controls other basic desires for its own purpose; if one refrains from yelling at those party animals, it shows one's high-order desire to be a civilized person.

Similar to the case of higher-order intentionality, agents in a collective can consciously accept the relations brought about by the collective structure and function or, abstractly, can accept the constraints higher than their current choices or preferences. For example, player one's initial preference was to cooperate, but when he looks at player two's options, he finds that "if he chooses to cooperate, player two will betray him and thus maxmizing his utility." Such a fact is precisely the higher-order constraints imposed by the game situation on a player's preference, forcing player one to adjust his initial preference and maximize his utility by choosing to betray the other player. From the game theoretical perspective, dependencies also satisfy the conditions of second-order relations because game theory defines players as entities with preferences, and thus dependencies that express relationships between different players are higher order relations between different individual preferences.

The fundamental qualitative difference between dependencies as second-order relations and individual preferences as first-order relations is that individual agents' specific input preferences may be arbitrary, but those interacting structures that stabilize and maintain them are not. In the relationalist explanation of collective agency, we have emphasized that the identity of a collective agent generally does not change with the variation of its members or behavioral targets, and what plays an essential role are those relatively stable relational patterns, that is, the structure and function of the collective. Dependencies in formal theory represent these relational patterns. Through the concept of dependency as a higher-order relation, we can express things such as "given other agents' current choices,

agent A chooses P," or "given other agents' current choices, agent A chooses its best option." Through these basic expressions, we can reproduce in our formal system the results of game theory, such as Pareto optimality, Nash equilibrium, core, etc., and relate them to the concept of collective agency. The idea that individual preference is regarded as a first-order relation and collective structure and function as higher-order relations is consistent with our philosophical approach of relationalist holism.

In this way, we link the philosophical theory of intentionality with the formal theory of preference. This unified picture takes the following form from bottom to top: at the bottom, each individual agent spontaneously develops a series of dispositions based on their background and current context. These dispositions refer to different types of contents, such as acts, purely psychological phenomena, or cognitive contents. Substantially different behavioral choices of agents are produced from these different dispositions. These choices made are recorded, summarized, and abstracted to form a relatively fixed ranking, and the result is the emergence of individual preferences. At the same time, the actions of an individual are also affected by their perception of the actions of others and this interactive influence gradually forms a stable dependency structure within a relatively stable period. The solidified dependency structure is what we think is the essence of collective agency, which can coordinate collective behavior internally and support its striving for the optimal welfare of the collective.

4.5 How to Represent Philosophical Ideas in a Game-theoretical Context

Having sorted out the relationship between the basic concepts of our philosophical perspective and those of game theory, let us return to the game theory context and explain what is necessary to define collective agency. From the perspective of relationalism, collective agency is embodied in the structures and functions of a collective, that is, its stable internal and external relations. Neither of these relations is directly characterized in game models. At best, the external relations are reflected in different payoffs already settled in a specific game model. Compared to this, stable aggregation functions in social choice theory are much closer to the idea of collective function (cf. List et al. (2011)). On the other hand, generally speaking, the internal relationship of a collective is actually the interaction relationship among its members, and the so-called collective structure is a fixed pattern of the interaction relationship. Game theory is adept at characterizing the inter-

action between individuals, which provides us with the possibility to further discuss such fixed patterns. Therefore, the focus of our discussion will be on collective structure, and we will try to express it within the basic setting of non-cooperative games.

How do we express collective structure in game theory? Our idea is to study the optimality of subgroups. When taking each individual group member as a special group, its Pareto optimality is the same as its individual optimality when the other players' strategies are fixed. So Nash equilibrium is, in fact, a state where each singleton of the group reaches its Pareto optimality given that the other singletons also reach their Pareto optimality. If each singleton's Pareto optimality and the whole group's Pareto optimality matter for the whole group's collective agency, how about the other subgroups' Pareto optimality? The following examples serve as guidance for our intuition on this question.

In the first example, Pareo optimality and Nash equilibrium seem to be sufficient for collective agency.

Example 2 *Three students A, B, C, live together in an apartment and share one living room. They take turns cleaning the living room every week. The following table describes the preference of each of the three students in different situations, where the row represents the choices of the first player A, the column represents the choices of second player B, and the left and right parts of the whole table represent the choices of the third player C. There*

	clean		not clean	
	clean	not clean	clean	not clean
clean	(4,4,4)	(1,3,1)	(1,1,3)	(1,2,2)
not clean	(3,1,1)	(2,2,1)	(2,1,2)	(0,0,0)

Table 4.3 Clean or not?

is only one Pareto optimal Nash equilibrium in the game, namely (clean, clean, clean). Intuitively, if the three students stick to (clean, clean, clean), they successfully team up for cleaning their living room and thus have collective agency as a group for cleaning their living room.

In the second example, we will see Pareto optimality and Nash equilibrium is insufficient for collective agency.

Example 3 *Three colleagues A, B, C, plan to go to a jazz or rock concert. For A and B, they do not care which concert to go to if only they go together. They prefer going together without C. However, C has no idea about this and prefers going together with A and B to*

not going together. The following table shows the situation modeled as a strategic game in which both (Jazz,Jazz,Jazz) and (Rock,Rock,Rock) are Pareto optimal Nash equilibria. Do the three players together as a group have collective agency at these Pareto optimal

	Jazz		Rock	
	Jazz	Rock	Jazz	Rock
Jazz	(2,2,4)	(0,0,0)	(4,4,0)	(0,0,0)
Rock	(0,0,0)	(4,4,0)	(0,0,0)	(2,2,4)

Table 4.4 Together or not

Nash equilibria? It seems unnatural to take A, B, and C as a group with collective agency. Nevertheless, what makes the Pareto optimal Nash equilibria in the game of this example different from the Pareto optimal Nash equilibrium in the game of the previous example?

An intuitive observation is that in the outcome state where the whole group satisfies equilibrium and Pareto optimality, a subgroup within it will not necessarily achieve its optimality as a subgroup. In this example, A and B prefer to choose the outcome without C, so the subgroups of A and B have reason to deviate from the collective optimal equilibrium state of (Jazz, Jazz, Jazz) and (Rock, Rock, Rock). Such a situation is clearly inconsistent with the realization of stable collective agency, suggesting we need to add additional conditions to rule out this kind of situation. The intuitive idea is that not only must collective agency of the whole satisfy Pareto optimality, but each of its subgroups must also satisfy Pareto optimality. However, that condition is too simple, since there are more complicated situations that need to be considered.

Example 4 *In a town, a small textile mill is run by C. He hired two workers A and B. After working for years, A and B both felt it necessary to have a higher wage. When they asked for it, C refused their request. So they are faced with a choice: go on strike or not? However, when C hired them, C had told each of them that if he did not join a strike initiated by the other worker, he would get compensated, and the other worker would get punished. The scenario can be modeled in a strategic game represented in table 4.5. It is easy to verify that when C keeps his word, and neither workers strike, the whole group reaches Nash equilibrium and Pareto optimality; on the other hand, when both workers strike and C does not keep his word, the whole group also reaches Nash equilibrium and Pareto optimality. Does the group has collective agency on the above two Pareto optimal Nash equilibria?*

	keep words		not keep words	
	strike	not strike	strike	not strike
strike	(2,2,4)	(0,3,5)	(2,2,4)	(2,1,5)
not strike	(3,0,5)	(1,1,6)	(1,2,5)	(1,1,6)

Table 4.5 Strike or not?

In this example, it is clear that two workers, as well as one worker and the boss, can form a coalition to pursue the subgroup's optimality. However, none of the outcome states are states in which all subgroups achieve their own optimality. In intuitive terms, this means that the condition we considerer setting after example 2 may be too strong. It excludes agencies of those groups which contain some, but not all, their subgroups' optimality. And we should note that, this kind of collective is not rare in daily life at all: there are bosses and their employees, department heads and team members, the party or the government of different levels, etc. In fact, in any collective involving institutional stratification, it is inevitable that only part of its subgroups can achieve their collective optimality, because hierarchy and the different interests generated by hierarchy lead to the inability of individuals with different interests to form a coalition that has collective agency. And note that the phenomenon is not limited to hierarchically structured institutions. It is more general because we cannot say that the interests of all the individuals and subgroups within a collective agent are strictly aligned; that would be too idealistic. The interest of a large collective is always the result of the struggle to balance the interests of the subgroups within it. From this perspective, the previously proposed condition that all subgroups achieve their optimality points to a specific case, where there is no hierarchical structure in the collective and where the interest of all (sub)groups are aligned. In other words, it has a flat structure: all members are equal, and their pursuit of interests is consistent so that any subgroup can achieve its own optimality.

We do not intend to study all specific structures of each group and give corresponding descriptions. For a standard collective classification, flat structure and hierarchical structure are enough to serve as a primary division to further our understanding of the collective structure and its impact on collective decision-making. To make this precise, we propose a logic in the next chapter and come back to give a formal characterization of the definition of collective agency in game theory in Section 5.3.3.

4.6 Summary and Ideas for a Future Investigation

In this chapter, we have connected philosophical theories of agency with game theory by discussing three concepts: intentionality, preference, and dependency, pointing out that the relational account of agency can provide a profound basis for formal research. The game theory we have considered in this chapter is limited to non-cooperative games. In fact, cooperative games, repeated games, and evolutionary games can all be related to the relational account of agency. In the following chapter, we will propose a new logic system and express the relational meaning of collective agency in contexts of both non-cooperative and cooperative games.

Since we are about to touch on the third discipline: logic, and start the analysis across the three disciplines, it is helpful to briefly describe the differences between them to draw theoretical boundaries. In the philosophical approach, theories are distinguished by being descriptive or normative. Generally speaking, a theory is descriptive if it focuses on reality and aims to restore the objective phenomenon itself as much as possible. On the contrary, a theory is normative if it starts with strict and ideal conceptual assumptions and discusses which conclusions can be derived from them in thought experiments. However, in many cases, we cannot simply call a theoretical system descriptive or normative, because most theories include abstract reasoning patterns but also are meant to be applied in specific situations. More often, characterising a theory as normative or descriptive is much more matter of locating it on a continuum. Theories are not descriptive or normative in an absolute sense, but rather some can be more normative or more descriptive than others. Comparatively speaking, the logical discussions in the next chapter are more normative than the formal theories that have been discussed in this chapter because logic can reveal more abstract relation patterns behind various essential concepts in game theory. The philosophical theories of agency and intentionality are hard to compare to logic and formal theories. The starting point of philosophical argument is always from two perspective: our intuitions about factual scenarios and the idealized presupposition of concepts. The former implies a solid descriptive flavour, while the latter is deeply connected with features of a normative system. Our discussion have covered the fields of ontology, epistemology, etc., involving a wide range of issues from the abstract to the concrete, which makes it hard to compare the results in terms of normative and descriptive.

Based on this, we can better understand the relationship from another perspective. The philosophical theory of agency and intentionality hopes to provide a versatile pack-

age of solutions, which includes the study of the whole process, from concrete cases to metaphysical speculation. Formal theories represented by game theory abstract away aspects such as ontological distinctions and redundant details in concrete instances, thus obtaining a more abstract and profound understanding of inter-individual interactions. From this perspective, such a study is a fragment of philosophical research but provides a deeper understanding of this particular fragment. The logical description goes further in the abstract sense by removing the redundant details in the formal theory and only preserving the abstract relations to get to the core of the matter, such as Pareto optimality, Nash equilibrium and other concepts. Regarding the whole research field of the philosophy of collective agency, it can be said that the following logical discussion concerns a fragment of a fragment of the philosophy, but it is the core part of the whole research.

There are other topics worth mentioning; for instance, abstraction involves one's ideological system; the phenomenon of free riding; the irrational aspects of agents, etc. Earlier, the distinction between abstraction and idealization was invoked, and we said that game theory only involves the former. However, one may argue that abstraction inevitably involves one's ideological system. When researchers want to abstract various attributes of the object, the decision is often driven by comprehensive factors such as the researchers' educational background, growth environment, cultural context, etc. Even so, game theory as an approach does not ignore other features of the agent in the ontological sense, but just abstracts them according to the research purpose. In this sense, our position remains defensible.

The collective optimum defined by Pareto optimality will leave space for agents to pay different costs for the collective, thus the problem of free riding lurks in the background. Such a phenomenon is common in daily life. For example, some agents can keep in line with collective interests without or with little transfer of personal interests, and such free riders are group members that score low in terms of belonging and responsibility. The uneven transfer of collective interests from each member is not conducive to the long-term stability of the collective. It is promising to explore and design a new type of collective agency to study this phenomena.

The basic concepts in game theory, such as equilibrium, are closely related to the rationality of agents. However, like many equilibriums cannot happen in social experiments, actual people often are irrational. Nevertheless, discussions of human irrationality are not included in this thesis; it is also almost always neglected in mainstream of metaphysics,

formal theories, and logic. How to understand irrationality and the common irrationality characteristics in collective behaviour are also very worthy of study.

The discussion in this chapter is only a small step in comparing the two fields of philosophical works and formal theories. We will follow this line of thought in the next chapter, using logic to characterize game theory concepts to gain deeper insights and connect them to philosophical perspectives.

CHAPTER 5 REASONING ABOUT DEPENDENCE, PREFERENCE AND COALITIONAL POWER

This chapter presents a logic of preference and functional dependence (LPFD) and its hybrid extension (HLPFD), both of whose sound and strongly complete axiomatization are provided. The decidability of LPFD are also proved. The application of LPFD and HLPFD to modelling non-cooperative and cooperative games in strategic and coalitional forms is explored. The resulted framework provides a unified view on Nash equilibrium, Pareto optimality and the core. The philosophical relevance of these game-theoretical notions to discussions of collective agency is made explicit. Some key connections with other logics are also revealed, for example, the coalition logic, the logic of functional dependence and the logic of ceteris paribus preference.

5.1 Introduction

Dependence, preference and coalitional power are three key concepts in game theory. There have been a lot of logical works on analyzing these three notions. To name but a few, for dependence, the dependence logic (cf. Väänänen (2007)) has been studied in various ways (c.f. Galliani (2021)) and a simple logic of functional dependence is recently proposed in Baltag et al. (2021); for coalitional power, the coalition logic (cf. Pauly (2002)) and the alternating-time temporal logic (ATL) (cf. Alur et al. (2002); Goranko et al. (2004)) are representative; for preference, good surveys can be found in Hansson (2002) and Liu (2011), Chapter 1.1. Despite not being explicitly emphasized, the concept of dependence permeates the analyses of the other two concepts, for example, in Pauly (2002) and van Benthem et al. (2009). However, as far as we know, there is hardly any logic explicitly modeling all of these three concepts, especially making dependence the hub to which the other two concepts join. In this chapter, we provide such a logic, which characterizes the interaction between the three concepts. Moreover, we show that by making the role of dependence explicit, our logical analysis leads to a unified view of several key concepts in game theory, namely Nash equilibrium, Pareto optimality and the core. We also explore a philosophical implication about collective agency of our logical analysis. We take the stability of a group to be an essential aspect of what makes it a coalition. Instead of focusing on intentionality as in the philosophical literature (cf. Roth (2017)),

we elaborate on our understanding in a game theoretical context.

Our main work in this chapter centers on introducing preference into the logic of functional dependence (cf. Baltag et al. (2021)) by adding preference relations in the original semantic model and a new modal operator in the original language for the intersection of different kinds of relations, including equivalence relations, preorders and strict preorders. By taking a game theoretic interpretation of the semantic setting, the new operator enables us to express not only Nash equilibrium but also Pareto optimality.

While Nash equilibrium is taken to be a benchmark for modern logics of games and many logics have been demonstrated to be able to express it (see (van Benthem et al. (2009), section 7.1) and the reference in it), Pareto optimality as an equally important notion in game theory① seems to receive less attention in logical literature than Nash equilibrium. As shown in this chapter, to express Pareto optimality, the new modal operator is critical. In fact, given the operator, we can express a relativized version of Nash equilibrium and Pareto optimality, that is, "given the current strategies of some players, the current strategy profile of the other players would be a Nash equilibrium/Pareto optimality." Moreover, by taking dependence relation into consideration, our logic shows that Nash equilibrium can be defined by Pareto optimality.

As Pareto optimality is rarely studied by logicians, compared to the non-cooperative game theory, the cooperative game theory (cf. Peleg et al. (2007)) seems not very salient to logicians either.② We will demonstrate that our logic of preference and functional dependence (LPFD) can also be adapted to model a qualitative version of cooperative games in strategic form (cf. Peleg et al. (2007), Section 11). We will also show that a hybrid extension of LPFD can express the core, an essential solution concept in the cooperative games analogous to Nash equilibrium in the non-cooperative games. The core characterizes a coalition's stability as a state where none of its subcoalitions has any incentive to deviate even if they can. The three concepts, dependence, preference and coalitional power, crystallize in the core. Through the lens framed by the three concepts, a unified view of the core, Nash equilibrium and Pareto optimality is revealed by our logics.

In addition to the logics and their application to a unified analysis of key game theoretical concepts, our contributions include several technical results about the logics them-

① For example, in the prisoners' dilemma, the Nash equilibrium is not Pareto optimal.

② The review on modal logic for games and information (cf. van der Hoek et al. (2007), Chapter 20) is exclusively about non-cooperative game theory; the book van Benthem (2014) touches on few issues on cooperative game theory either. The only exception we know is the work in Ågotnes et al. (2009), where two different logics are proposed to reason about cooperative games.

selves. We provide a sound and strongly complete axiomatization respectively for LPFD and its hybrid extension (HLPFD). Moreover, we also prove that the satisfiability problem of LPFD is decidable. While the proof for the completeness result of HLPFD is standard, the completeness of LPFD is much harder to prove and requires new techniques. Our proof modifies the classical unraveling method (cf. Blackburn et al. (2001), Chapter 4.5) and combines it with a special way of selecting the tree branches.

The structure of the chapter is summarized as follows. The background on the logic of functional dependence (LFD) are presented in Section 5.2. In section 5.3, we introduce the logic of preference and functional dependence and show how it can naturally express Nash equilibrium and Pareto optimality. Section 5.4 contains sound and strongly complete axiomatization of LPFD and its hybrid extension and the decidability of LPFD's satisfiability problem. For those who are not interested in the proof details, Section 5.4.3 and Section 5.4.4 can be safely skipped. In Section 5.5, we turn to our modelling of cooperative games in strategic form in LPFD and analyze the core. In Section 5.6, we show how the core can be relevant to philosophical discussions of collective agency. Before conclusion, we compare our work with the logical works in Pauly (2002), Ågotnes et al. (2009) and van Benthem et al. (2009).

Notations The following notations will be used throughout this chapter. We will use B^A to denote the set of mappings from set A to set B. Let $\mathcal{P}^{<\aleph_0}(A)$ denote the set of all finite subsets of A. We write $B \subseteq_{\aleph_0} A$ if $B \in \mathcal{P}^{<\aleph_0}(A)$. For each string $\vec{x} = (x_i : i \in I)$, we write $\mathsf{set}(\vec{x})$ for the set $\{x_i : i \in I\}$. For every language $\mathcal{L}$ and class $\mathscr{C}$ of mathematical structures, let $\mathsf{Log}_{\mathcal{L}}(\mathscr{C})$ denote the set of all valid formulas in $\mathcal{L}$ w.r.t $\mathscr{C}$.

5.2 LFD Interpreted in Games

In this section, we introduce LFD and take a game-theoretical view on it.

LFD starts with a relational vocabulary $(\mathsf{V}, \mathsf{Pred}, \mathsf{ar})$, where V is a countable set of variables, Pred is a countable set of predicate symbols and $\mathsf{ar} : \mathsf{Pred} \to \mathbb{N}$ is an arity map, associating to each predicate $P \in \mathsf{Pred}$ a natural number $\mathsf{ar}(P)$. In what follows, unless otherwise specified, the vocabulary $(\mathsf{V}, \mathsf{Pred}, \mathsf{ar})$ is the one such that $|\mathsf{V}| = \aleph_0$ and $|\{P \in \mathsf{Pred} : \mathsf{ar}(P) = n\}| = \aleph_0$ for each $n \in \omega$.

To view LFD from a game-theoretical perspective, we take the variables to represent players in games and the dependence models of LFD become models for different players' actions or strategies in static games in strategic form.

Definition 4 (Dependence models) *A model is a pair $M = (O, I)$, where O is a non-empty set of actions and I is a mapping that assigns to each predicate $P \in \mathsf{Pred}$ a subset of $O^{\mathsf{ar}(P)}$. A dependence model $\mathbf{M}$ is a pair $\mathbf{M} = (M, A)$, where $M = (O, I)$ is a model and $A \subseteq O^{\mathsf{V}}$ is a set of strategy profiles.*

For each $X \subseteq_{\aleph_0} \mathsf{V}$, we define a binary relation $=_X \subseteq A \times A$ such that $a =_X a'$ if and only if $a \restriction X = a' \restriction X$, i.e., the action of x in a is the same as her action in a' for each $x \in X$.

When $A \neq O^{\mathsf{V}}$, some strategy profiles are missing, which gives rise to restrictions on how players can act together. Suppose a strategy profile s is not in A. Then the players cannot act according to s simultaneously.①

Next, we turn to the syntax and semantics of LFD. To capture functional dependence, LFD uses two operators $\mathbb{D}$ and D in its language.

Definition 5 *The language $\mathcal{L}$ of* LFD *is given by*

$$\varphi ::= P\vec{x} \mid D_X y \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathbb{D}_X \varphi$$

where $P \in \mathsf{Pred}$, $\vec{x} = (x_1, \ldots, x_n)$ is a finite string of players of length $n = \mathsf{ar}(P)$, $X \subseteq_{\aleph_0} \mathsf{V}$ is a finite set of players and $y \in \mathsf{V}$ is a player.

$\mathbb{D}_X \varphi$ is meant to express that whenever the players in X take their current actions, φ is the case; $D_X y$ says that whenever the players in X take their current actions, y also takes its current action.

Definition 6 *Truth of a formula $\varphi \in \mathcal{L}$ in a dependence model $\mathbf{M} = (M, A)$ at a strategy profile $a \in A$ is defined as follows:*

$$\begin{array}{lll}
\mathbf{M}, a \vDash P\vec{x} & \textit{iff} & a(\vec{x}) \in I(P) \\
\mathbf{M}, a \vDash D_X y & \textit{iff} & a =_y a' \textit{ for all } a' \in A \textit{ with } a =_X a' \\
\mathbf{M}, a \vDash \neg\varphi & \textit{iff} & \mathbf{M}, a \nvDash \varphi \\
\mathbf{M}, a \vDash \varphi \wedge \psi & \textit{iff} & \mathbf{M}, a \vDash \varphi \textit{ and } \mathbf{M}, a \vDash \psi \\
\mathbf{M}, a \vDash \mathbb{D}_X \varphi & \textit{iff} & \mathbf{M}, a' \vDash \varphi \textit{ for all } a' \in A \textit{ with } a =_X a'
\end{array}$$

Note that $=_X$ is an equivalence relation on A and $a =_\varnothing a'$ holds for all $a, a' \in A$. So $\mathbb{D}_\varnothing$ is a universal operator and we define $\mathsf{A}\varphi := \mathbb{D}_\varnothing \varphi$ and $\mathsf{E}\varphi := \neg\mathsf{A}\neg\varphi$.

① Such restrictions serve as basis for representing dependence but do not necessarily imply dependence between the players. For example, for $O = \{a_1, a_2, b_1, b_2\}$ and $V = x, y$, if we take $A = \{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}$, the restriction of each player's range of actions is not necessarily due to what the other player does.

5.3 Logic of Preference and Functional Dependence

In this section, we extend LFD to LPFD.

5.3.1 Syntax and semantic for LPFD

Definition 7 (Syntax) *The language $\mathcal{L}^{\preceq}$ of LPFD is given by:*

$$\mathcal{L}^{\preceq} \ni \phi ::= P\vec{x} \mid D_X y \mid \neg\phi \mid \phi \wedge \phi \mid [\![X, Y, Z]\!]\phi$$

where $P \in \mathsf{Pred}$, $\vec{x} \in \mathsf{V}^{\mathrm{ar}(P)}$, $y \in \mathsf{V}$ and $X, Y, Z \subseteq_{\aleph_0} \mathsf{V}$.

$\mathcal{L}^{\preceq}$ only differs from the language of LFD in the new operator $[\![X, Y, Z]\!]\phi$, which is an operator for ceteris paribus group preference. In $\mathcal{L}^{\preceq}$, $\mathbb{D}_X\varphi$ is defined as $[\![X, \varnothing, \varnothing]\!]\phi$, capturing "centeris paribus". Y and Z in $[\![X, Y, Z]\!]\phi$ are used to capture group preferences. We define $\langle\!\langle X, Y, Z \rangle\!\rangle\phi := \neg[\![X, Y, Z]\!]\neg\phi$ and $D_X Y := \bigwedge_{y\in Y} D_X y$ for each $Y \subseteq_{\aleph_0} \mathsf{V}$.

Next we turn to the semantics of LPFD.

Definition 8 (PD-models) *A* preference dependence model (PD-model) *is a pair $\mathbb{M} = (\mathbf{M}, \preceq)$ in which $\mathbf{M} = (M, A)$ is a model and $\preceq\colon V \to \mathcal{P}(A \times A)$ is a mapping assigning to each $x \in \mathsf{V}$ a pre-order $\preceq_x$ on A. Let Mod denote the class of all PD-models.*

For each $x \in \mathsf{V}$, we define the binary relation $\prec_x = \{(a, b) \in \preceq_x : (b, a) \notin \preceq_x\}$. For all $a, b \in A$, we write $a \preceq_X b (a \prec_X b)$ if $a \preceq_x b (a \prec_x b)$ for each $x \in X$. We write $s \simeq_X t$ if $s \preceq_X t$ and $t \preceq_X s$.

Definition 9 *Truth of PD-formulas of the form $P\vec{x}, D_X y, \neg\phi$ or $\phi \wedge \psi$ is defined as in Definition 6. For formulas of the form $[\![X, Y, Z]\!]\phi$, we say $[\![X, Y, Z]\!]\phi$ is true at a in $\mathbb{M}$, notation: $\mathbb{M}, a \vDash [\![X, Y, Z]\!]\phi$, if*

$$\mathbb{M}, a' \vDash \varphi \text{ for all } a' \in A \text{ satisfying } a =_X a',\ a \preceq_Y a' \text{ and } a \prec_Z a'.$$

A formula $\phi \in \mathcal{L}^{\preceq}$ is valid if $\mathbb{M}, a \vDash \phi$ for all PD-model $\mathbb{M} = (M, A, \preceq)$ and $a \in A$. Let LPFD denote the set of all valid formulas, i.e., $\mathsf{LPFD} = \mathsf{Log}_{\mathcal{L}^{\preceq}}(\mathsf{Mod})$.

Note that $[\![\varnothing, \{x\}, \varnothing]\!]\varphi$ and $[\![\varnothing, \varnothing, \{x\}]\!]\varphi$ are standard modal operators defined on $\preceq_x$ and $\prec_x$ respectively. Thus $[\![X, Y, Z]\!]\varphi$ is in fact a standard modal operator defined on the intersection of the relations $=_X$, $\preceq_Y$ and $\prec_Z$. It concerns the preferences of the players in Y and Z conditional on the actions of the players in X.

There is a close connection between LPFD and the work in van Benthem et al. (2009)

on ceteris paribus preference. We will discuss this connection in Section 5.7.2. Next, we show how some key game theoretical notions can be expressed in LPFD.

5.3.2 Pareto optimality and Nash equilibrium in LPFD

Having laid out the basics of LPFD, we turn to questions concerning expressing and reasoning about Pareto optimality and Nash equilibrium in LPFD. One important assumption we will adopt is that the group of players V has to be finite. In LPFD, there is no such restriction on V. However, it is worth noting that in the language of LPFD, all subscripts in the two operators need to be finite. So to express something like $[\![-X, \varnothing, X]\!]\varphi$ in LPFD where $-X := V - X$, which is frequently referred to in game theory, we have to ensure that X and $-X$ are both finite.

We start with recalling what Nash equilibrium and weak/strong Pareto optimality mean.

Definition 10 *Let $\mathbb{M}$ be a PD-model and $X \subseteq \mathsf{V}$.*

- *s is a **Nash equilibrium** for X if for all $x \in X$ there is no $t =_{-\{x\}} s$ such that $s \prec_x t$;*
- *s is **strongly Pareto optimal** for X if there is no $t =_{-X} s$ such that (a) for all $x \in X$, $s \preceq_x t$ and (b) there is one $x \in X$ such that $s \prec_x t$;*
- *s is **weakly Pareto optimal** for X if there is no $t =_{-X} s$ such that for all $x \in X$, $s \prec_x t$.*

Note that such a way of defining the notions of Nash equilibrium, weak and strong Pareto optimality in a PD-model applies to all subgroups of V rather than only the whole group of players V.

Example 5 *Table 5.1 shows three students' preferences on staying home or going out. The row is for student a; the column is for student b; the left and right division is for student c.*

go out			stay home		
	go out	stay home		go out	stay home
go out	(4,4,4)	(1,0,1)	go out	(1,1,0)	(1,2,2)
stay home	(0,1,1)	(2,2,1)	stay home	(2,1,2)	(2,2,2)

Table 5.1

We can check that going out together and all staying home are both Nash equilibrium; going out together is also Pareto optimal. Moreover, (stay home, stay home, go out) is a

Nash equilibrium for student a and b given student c goes out, although it is not a Nash equilibrium for the whole group. (stay home, stay home, stay home) is Pareto optimal for student a and b given student c stays home, but not for student a, b and c together.

It is relatively easy to get how Nash equilibrium and weak Pareto optimality can be expressed in LPFD, as the following fact shows.

Fact 1 *Let* $\mathbb{M} = (M, A, \preceq)$ *be a PD-model and* $s \in A$*. Then*

- *s is a Nash equilibrium for* $X \subseteq \mathsf{V}$ *given that the players in* $-X$ *have acted according to s, if and only if,* $\mathbb{M}, s \vDash \bigwedge_{x\in X}[\![-\{x\}, \varnothing, \{x\}]\!]\bot$*;*
- *s is weakly Pareto optimal for* $X \subseteq \mathsf{V}$ *given that the players in* $-X$ *have acted according to s, if and only if,* $\mathbb{M}, s \vDash [\![-X, \varnothing, X]\!]\bot$*.*

In the case of weak Pareto optimality, because the truth condition of the operator $[\![-X, \varnothing, X]\!]$ depends on what formulas are satisfied on all elements in the set $\{t \in A \mid s =_{-X} t, s \prec_X t\}$, if it is an empty set and thus $\bot$ can be vacuously satisfied on all elements in it, then s is weakly Pareto optimal for X.

To express strong Pareto optimality in LPFD, we need to express the following model theoretical fact, namely, the set $\{t \in A \mid s =_{-X} t, s \preceq_X t \text{ and } t \not\preceq_X s\} = \bigcup_{x\in X}\{t \in A \mid s =_{-X} t, s \preceq_{X-\{x\}} t, s \prec_x t\}$ is empty.

Since $s \vDash [\![-X, X - \{x\}, \{x\}]\!]\bot$ iff $\{t \in A \mid s =_{-X} t, s \preceq_{X-\{x\}} t, s \prec_x t\} = \varnothing$, we can define strong Pareto optimality as follows.

Fact 2 *In a PD-model* $\mathbb{M}$*, s is strongly Pareto optimal for* $X \subseteq \mathsf{V}$ *given that the players in* $-X$ *have acted according to s iff* $\mathbb{M}, s \vDash \bigwedge_{x\in X}[\![-X, X - \{x\}, \{x\}]\!]\bot$*.*

To facilitate our discussion, we define weak and strong Pareto optimality and Nash equilibrium in LPFD as

$$\mathsf{wPa}\, X := [\![-X, \varnothing, X]\!]\bot \tag{5.1}$$

$$\mathsf{sPa}\, X := \bigwedge_{x\in X} [\![-X, X - \{x\}, \{x\}]\!]\bot \tag{5.2}$$

$$\mathsf{Na}\, X := \bigwedge_{x\in X} [\![-\{x\}, \varnothing, \{x\}]\!]\bot \tag{5.3}$$

An easy but important observation is that Nash equilibrium can be defined via Pareto optimality.

Theorem 1 $\mathsf{Na}\, X = \bigwedge_{x\in X} \mathsf{sPa}\{x\} = \bigwedge_{x\in X} \mathsf{wPa}\{x\}$.

5.3.3 Collective agency and dependence between subgroups' Pareto optimality

In this subsection, with the help of **LPFD**, we first explain what differentiates the Pareto optimal Nash equilibrium in Table 4.3 and the Pareto optimal Nash equilibria in Table 4.4. Then we bring out a structural property the Pareto optimal Nash equilibria in Table 4.5 have. At last, we summarize by proposing a definition of collective agency based on our analysis.

Recall that we intuitively believe that the Pareto optimality of subgroups also plays a role in forming collective agency. A definition of collective agency naturally arises by considering that all of its subgroups should also reach their own Pareto optimality. We call this type the complete Pareto optimality:

Definition 11 *We say that a group X reaches complete Pareto optimality at s if $s \models \bigwedge_{X' \subseteq X} \mathsf{sPa}\, X'$.*

By formalizing the games in Table 4.3 and Table 4.4 as two PD models $\mathbb{M}_1$ and $\mathbb{M}_2$ and player A, B, C within as player 1, 2, 3, respectively, it is not hard to verify that $\mathbb{M}_1, \mathrm{CCC} \models \bigwedge_{X \subseteq \{1,2,3\}} \mathsf{sPa}\, X$ where CCC is the state (clean,clean,clean); but $\mathbb{M}_2, \mathrm{JJJ} \models \neg \bigwedge_{X \subseteq \{1,2,3\}} \mathsf{sPa}\, X$ and $\mathbb{M}_2, \mathrm{RRR} \models \neg \bigwedge_{X \subseteq \{1,2,3\}} \mathsf{sPa}\, X$ where JJJ is the state (Jazz,Jazz,Jazz) and RRR is the state (Rock,Rock,Rock), because $\mathbb{M}_2, \mathrm{JJJ} \models \neg \mathsf{sPa}\{1,2\}$ and $\mathbb{M}_2, \mathrm{RRR} \models \neg \mathsf{sPa}\{1,2\}$.

Should a group's collective agency ensure the complete Pareto optimality? We do not think so. If we formalize the game in Table 4.5 as a PD model $\mathbb{M}_3$ and player A, B ,C as player 1, 2, 3, respectively, then we have $\mathbb{M}_3, \mathrm{NNK} \models \neg \bigwedge_{X \subseteq \{1,2,3\}} \mathsf{sPa}\, X$ and $\mathbb{M}_3, \mathrm{SSN} \models \neg \bigwedge_{X \subseteq \{1,2,3\}} \mathsf{sPa}\, X$ where NNK is the state (not strike,not strike,keep words) and SSN is the state (strike,strike,not keep words), because $\mathbb{M}_3, \mathrm{NNK} \models \neg \mathsf{sPa}\{1,2\}$ and $\mathbb{M}_3, \mathrm{SSN} \models \neg \mathsf{sPa}\{1,3\} \wedge \neg \mathsf{sPa}\{2,3\}$. This is similar to the Pareto optimal equilibria JJJ and RRR in Table 4.4. However, NNK and SSN, as Pareto optimal Nash equilibria, have a structural property which JJJ and RRR lack.

There are different coalitional structures in NNK and SSN. In SSN, the two workers form a union in the sense that, no matter which strategy C chooses, sticking to (strike,strike) makes the two workers as a group keep Pareto optimal. By choosing N, C makes the resulted state SSN stable and also Pareto optimal for the whole group. Using **LPFD**, $\mathbb{M}_3, \mathrm{SSN} \models \mathbb{D}_{\{1,2\}} \mathsf{sPa}\{1,2\} \wedge \mathbb{D}_{\{1,2,3\}} \mathsf{sPa}\{1,2,3\} \wedge \mathsf{Na}\{1,2,3\}$. As for NNK, the two workers A and B are separeted and each forms a coalition

with C in a similar sense to that of A and B in SSN. Using **LPFD**, $\mathbb{M}_3, \mathrm{SSN} \models \mathbb{D}_{\{1,3\}}\,\mathsf{sPa}\{1,3\} \wedge \mathbb{D}_{\{1,2,3\}}\,\mathsf{sPa}\{1,2,3\} \wedge \mathsf{Na}\{1,2,3\}$; and $\mathbb{M}_3, \mathrm{SSN} \models \mathbb{D}_{\{2,3\}}\,\mathsf{sPa}\{2,3\} \wedge \mathbb{D}_{\{1,2,3\}}\,\mathsf{sPa}\{1,2,3\} \wedge \mathsf{Na}\{1,2,3\}$.

To generaize the above observation, we define a notion called structural Pareto optimality in **LPFD**.

Definition 12 *Let X be a group of agents. A finite sequence of nonempty subsets of X, denoted by $[X_i]$ ($i \in \mathbb{N}$), is an* ordered cover *of X if $\bigcup_i X_i = X$ with $X_i \neq X$ for each X_i, and for any $i, j \in \mathbb{N}$, $X_i \neq X_j$.*

Definition 13 *We say that a group X reaches structural Pareto optimality at s, denoted by $\mathsf{Sa}\,X$, if there is an ordered cover $[X_1, \ldots, X_n]$ of X such that*

$$s \models \mathbb{D}_{(-X)\cup X_1}\,\mathsf{sPa}\,X_1 \wedge \mathbb{D}_{(-X)\cup X_1\cup X_2}\,\mathsf{sPa}(X_1 \cup X_2) \wedge \ldots \wedge \mathbb{D}_{(-X)\cup X_1\cup X_2\cup\ldots\cup X_n}\,\mathsf{sPa}(X_1 \cup X_2 \cup \ldots \cup X_n) \quad (5.4)$$

Namely,

$$\mathsf{Sa}\,X := \bigvee_{[X_i]} \bigwedge_i \mathbb{D}_{(-X)\cup X_1\cup X_2\cup\ldots\cup X_i}\,\mathsf{sPa}(X_1 \cup X_2 \cup \ldots \cup X_i)$$

where $[X_i]$ is an ordered cover of X and the first $\bigvee$ ranges over all ordered covers of X.

The formula (5.4) is not hard to decipher: fixing the actions of players in $-X$, first, no matter what the players in $(X \setminus X_1)$ do, X_1 achieves its Pareto optimality by acting according to s; then by joining X_1, X_2 together with X_1 as a whole achieves its Pareto optimality by acting according to s no matter what the players in $(X \setminus (X_1 \cup X_2))$ do; by joining one by one, a subgroup together with those subgroups coming before it as a whole achieves its Pareto optimality. In addition, $\mathsf{Sa}\,X$ has the following key property.

Proposition 1 $\models \mathsf{Sa}\,X \to \mathsf{sPa}\,X$.

This property follow immediately from Definition of $\mathsf{Sa}\,X$, since by $\bigcup_i X_i = X$, for each $[X_i]$, the last conjunct of $\mathsf{Sa}\,X$ identifies with $\mathsf{sPa}\,X$.

NNK and SSN in Table 4.5 are structurally Pareto optimal. But JJJ and RRR in Table 4.4 are not. So JJJ and RRR are neither structurally Pareto optimal nor completely Pareto optimal. This explains why we think the three colleagues in the game do not have collective agency in JJJ and RRR.

We have seen that SSN and NNK are structurally Pareto optimal but not completely Pareto optimal. How about the other direction? The following table, which is an abridged

version of the game in Table 4.4, severs as a counterexample. The state (Jazz, Jazz) is

	Jazz	Rock
Jazz	(4,4)	(0,0)
Rock	(0,0)	(2,2)

Table 5.2

completely Pareto optimal but not structurally optimal.

To capture both types of Pareto optimality, we define collective agency in **LPFD** as follows:

Definition 14 (Collective agency in LPFD)

$$\mathsf{Ca}\, X := (\mathsf{Sa}\, X \wedge \mathsf{Na}\, X) \vee \bigwedge_{X' \subseteq X} \mathsf{sPa}\, X'$$

This definition formalizes how we define collective agency, as discussed at the end of Chapter 4, where we emphasize the distinction between different kinds of collective structures in the context of game theory. In this definition, the formula on the left side of the disjunction expresses the conditions to be satisfied by the collective with a hierarchical structure to achieve collective agency, while the right side of the disjunction expresses the conditions to be satisfied by the collective with a flat structure to achieve collective agency. Next, we will introduce the axiomatization of LPFD and prove its completeness. After that, we will enter the context of cooperative game theory and try to express collective agency under cooperative games with the hybrid extension of LPFD.

5.4 Calculus of LPFD and its Hybrid Extension

In this section, a Kripke style semantics of LPFD shall be introduced. It is proved to be equivalent to the standard semantics in subsection 5.3.1. The new semantics provides us with a modal view, which facilitates our calculus $\mathsf{C}_{\mathrm{LPFD}}$ and the proof of its soundness and strongly completeness. We show that LPFD is decidable while it lacks the finite model property. Moreover, we extend it with nominals and give also a sound and complete calculus $\mathsf{C}_{\mathrm{HLPFD}}$. In Section 5.6, this hybrid extension will be useful in expressing a key game theoretic concept.

5.4.1 Kripke style semantics

In this part, we introduce the Kripke style semantics for LPFD and show the relation between this semantics and the standard one.

Definition 15 *A* relational PD-frame (RPD-frame) *is a pair* $\mathfrak{F} = (W, \sim, \leq)$*, where* W *is a non-empty set,* $\sim\colon V \to \mathcal{P}(W \times W)$ *and* $\leq\colon V \to \mathcal{P}(W \times W)$ *are maps such that* $\sim_x$ *is an equivalence relation and* $\leq_x$ *is a pre-order for all* $x \in \mathsf{V}$*. For all* $x \in \mathsf{V}$ *and* $X, Y, Z \subseteq_{\aleph_0} \mathsf{V}$*, let* $<_x = \{(w,u) \in \leq_x : (u,w) \notin \leq_x\}$ *and*

$$R(X,Y,Z) = \bigcap_{x\in X} \sim_x \cap \bigcap_{y\in Y} \leq_y \cap \bigcap_{z\in Z} <_z .$$

A relational PD-model (RPD-model) *is a pair* $\mathfrak{M} = (\mathfrak{F}, V)$ *where* $\mathfrak{F} = (W, \sim, \leq)$ *is an RPD-frame and* V *is a valuation associating to each formula of the form* $P\vec{x}$ *a subset* $V(P\vec{x})$ *of* W*. The valuation* V *is required to satisfy the following condition for all* $w, u \in W$ *and* $P \in \mathsf{Pred}$*:*

$$\text{if } w \sim_{\mathsf{set}(\vec{x})} u \text{, then } w \in V(P\vec{x}) \text{ if and only if } u \in V(P\vec{x}). \quad \textit{(Val)}$$

Let RMod *denote the class of all RPD-models. Truth of a formula* $\phi \in \mathcal{L}^{\preceq}$ *in* $\mathfrak{M} = (W, \sim, \leq, V)$ *at* $w \in W$ *is defined as follows:*

$$\begin{array}{lll}
\mathfrak{M}, w \vDash P\vec{x} & \textit{iff} & w \in V(P\vec{x}) \\
\mathfrak{M}, w \vDash D_X y & \textit{iff} & w \sim_y v \text{ for all } v \sim_X w \\
\mathfrak{M}, w \vDash \neg\varphi & \textit{iff} & \mathfrak{M}, w \nvDash \varphi \\
\mathfrak{M}, w \vDash \varphi \wedge \psi & \textit{iff} & \mathfrak{M}, w \vDash \varphi \text{ and } \mathfrak{M}, w \vDash \psi \\
\mathfrak{M}, w \vDash [\![X,Y,Z]\!]\phi & \textit{iff} & \mathfrak{M}, v \vDash \phi \text{ for all } v \in R(X,Y,Z)(w)
\end{array}$$

Validity is defined as usual. Let $\mathsf{RLPFD} = \mathsf{Log}_{\mathcal{L}^{\preceq}}(\mathsf{RMod})$*.*

We now show that the Kripke semantics is equivalent to the standard one in the sense that $\mathsf{RLPFD} = \mathsf{LPFD}$.

Definition 16 *Let* $\mathbb{M} = (O, I, A, \preceq)$ *be a PD-model. Then we define the RPD-model* $rel(\mathbb{M}) = (A, \sim, \leq, V)$ *induced by* $\mathbb{M}$ *as follows:*

- $V(P\vec{x}) = \{a \in A : a(\vec{x}) \in I(P)\}$ *for all* $P \in \mathsf{Pred}$ *and* $\vec{x} \in \mathsf{V}^{\mathsf{ar}(P)}$.
- $\sim_x = (=_x)$, $\leq_x = \preceq_x$ *for each* $x \in \mathsf{V}$.

Proposition 2 *Let* $\mathbb{M}$ *be a PD-model and* $rel(\mathbb{M})$ *the RPD-model induced by* $\mathbb{M}$*. Then for each* $a \in A$ *and formula* $\phi \in \mathcal{L}^{\preceq}$*,*

$$\mathbb{M}, a \vDash \phi \textit{ if and only if } rel(\mathbb{M}), a \vDash \phi.$$

Proof. By induction on the complexity of ϕ. □

While it is straightforward to induce an RPD-model from a PD-model, some care needs to be taken to induce a suitable PD-model from a RPD-model.

Definition 17 *Let $\mathfrak{M} = (W, \sim, \leq, V)$ be an RPD-model. Then*

- $\mathfrak{M}$ *is a* differential model*, if* $\bigcap_{x \in \mathsf{V}} \sim_x = Id_W = \{(w, w) : w \in W\}$.
- $\mathfrak{M}$ *is a* pre-differential model*, if for all* $w, u \in W$*,* $w \sim_{\mathsf{V}} u$ *implies* $w \leq_{\mathsf{V}} u$.

Let RMod_d *and* RMod_{pd} *denote the class of all differential RPD-models and pre-differential RPD-models, respectively.*

Definition 18 *Let $\mathfrak{M} = (W, \sim, \leq, V)$ be a differential model. Then we define the PD-model $dp(\mathfrak{M}) = (O, I, A, \preceq)$ induced by $\mathfrak{M}$ as follows:*

- $O = \{(x, |w|_x) : x \in \mathsf{V}, w \in W \text{ and } |w|_x = \{v \in W : w \sim_x v\}\}$.
- $A = \{w^* : w \in W\}$*, where* $w^*(x) = (x, |w|_x)$ *for each* $x \in \mathsf{V}$.
- $\preceq_x = \{(w^*, v^*) : w \leq_x v\}$ *for each* $x \in \mathsf{V}$.
- I *is the interpretation mapping each n-ary predicate* P *to the set*

$$I(P) = \{w^*(\vec{x}) : w \in W, \vec{x} \in \mathsf{V}^n \text{ and } w \in V(P\vec{x})\}.$$

$I(P)$ is well-defined for each predicate P since $x = y$ and $w \sim_x v$ whenever $w^*(x) = v^*(y)$. To see that $\preceq_x$ is a pre-order for each $x \in \mathsf{V}$, it suffices to show that $(W, \leq)$ is isomorphic to $(A, \preceq)$. Since $\mathfrak{M}$ is a differential model, for all $w, v \in W$, $w \neq v$ implies $w \nsim_y v$ for some $y \in \mathsf{V}$. Thus $w \neq v$ implies $w^* \neq v^*$ and so the function $(\cdot)^* : W \to A$ is an isomorphism. Hence $\preceq_x$ is a pre-order for all $x \in \mathsf{V}$. It is clearly that $\prec_x = \{(w^*, v^*) : w <_x v\}$ for all $x \in \mathsf{V}$.

Proposition 3 *Let $\mathfrak{M}$ be a differential RPD-model and $dp(\mathfrak{M})$ the PD-model induced by $\mathfrak{M}$. Then for each w in $\mathfrak{M}$ and formula $\phi \in \mathcal{L}^{\preceq}$,*

$$\mathfrak{M}, w \vDash \phi \text{ if and only if } dp(\mathfrak{M}), w^* \vDash \phi.$$

Proof. By induction on the complexity of ϕ. □

Theorem 2 $\mathsf{RLPFD} = \mathsf{LPFD}$.

Proof. By Proposition 2, if a formula ϕ is satisfied by some PD-model, then it is satisfied by some RPD-model, which entails $\mathsf{RLPFD} \subseteq \mathsf{LPFD}$. Let ϕ be a formula, $\mathfrak{M} = (W, \sim, \leq, V)$ an RPD-model and $w \in W$. Suppose $\mathfrak{M}, w \vDash \phi$. Since V is infinite, there is $x \in \mathsf{V}$ which does not occur in ϕ. Then let $\mathfrak{M}' = (W, \sim', \leq, V)$ be an RPD-model where

$$\sim'_y = \begin{cases} \sim_y & \text{, if } y \neq x; \\ \{(w, w) : w \in W\} & \text{, otherwise.} \end{cases}$$

Now we can readily check that $\mathfrak{M}'$ is a differential model with $\mathfrak{M}', w \vDash \phi$. By Proposition 3, we have $dp(\mathfrak{M}'), w^* \vDash \phi$. Hence $\mathsf{RLPFD} = \mathsf{LPFD}$. □

Relation between the two semantics with finite variables

The assumption $|\mathsf{V}| \geq \aleph_0$ is crucial in the proof of Theorem 2. The readers can check that the functions *rel* and dp provide a correspondence between classes of models Mod and RMod_d in the sense that for all $\mathfrak{M} \in \mathsf{RMod}_d$ and $\mathbb{M} \in \mathsf{Mod}$,

$$dp(rel(\mathbb{M})) \cong \mathbb{M} \text{ and } rel(dp(\mathfrak{M})) \cong \mathfrak{M}.$$

A simple example is given in the left part of Figure 5.1, where dotted lines stand for $\sim$ relations and arrows for preferences. Under the assumption that V is infinite, we can see from the proof of Theorem 2 that every satisfiable formula is satisfied by some differential RPD-model.

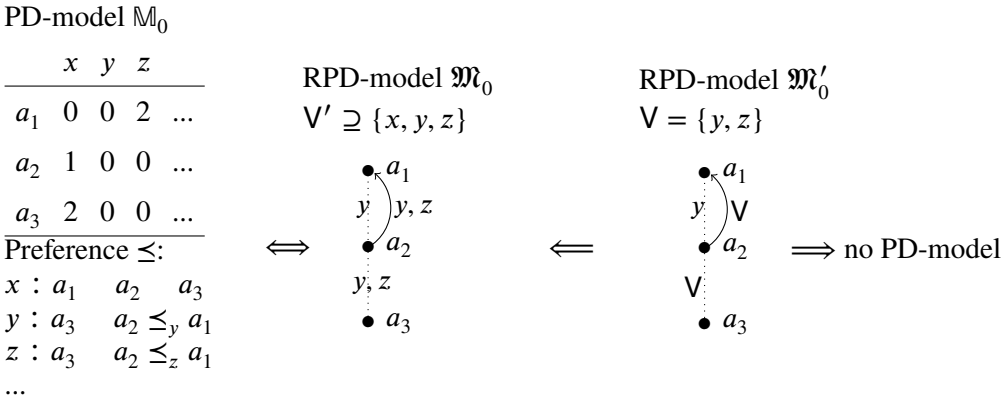


Figure 5.1 Translation between PD-models and RPD-models

However, it is not the case when V is finite. Suppose V is finite. Let $\mathfrak{M}_0'$ be the RPD-model shown in the right part of Figure 5.1. Then the readers can verify that $\mathfrak{M}_0', a_3 \Vdash [\![\varnothing, \varnothing, \mathsf{V}]\!]\bot \wedge \langle\!\langle \mathsf{V}, \varnothing, \varnothing \rangle\!\rangle \langle\!\langle \varnothing, \varnothing, \mathsf{V} \rangle\!\rangle \top$. On the other hand, for any PD-model $\mathbb{M} = (M, A, \preceq)$ and $a, b \in A$, $a =_{\mathsf{V}} b$ implies $a = b$. Thus $\phi \to [\![\mathsf{V}, \varnothing, \varnothing]\!]\phi$ is valid and so $[\![\varnothing, \varnothing, \mathsf{V}]\!]\bot \wedge \langle\!\langle \mathsf{V}, \varnothing, \varnothing \rangle\!\rangle \langle\!\langle \varnothing, \varnothing, \mathsf{V} \rangle\!\rangle \top$ cannot be satisfied by any PD-model for the vocabulary $(\mathsf{V}, \mathsf{Pred}, \mathsf{ar})$. Thus $\mathsf{RLPFD} \subsetneq \mathsf{LPFD}$.

Suppose now that V is finite. To obtain a clearer view of the relation between the two semantics, the class RMod_{pd} of RPD-models plays an important role. The readers can readily check that the following facts hold:

Fact 3 *Let $\mathfrak{M} = (W, \sim, \leq, V)$ be an RPD-model. Then*

$$\mathfrak{M} \in \mathsf{RMod}_{pd} \text{ if and only if } \mathfrak{M} \Vdash \langle\!\langle \mathsf{V}, \varnothing, \varnothing \rangle\!\rangle \phi \to \langle\!\langle \mathsf{V}, \mathsf{V}, \varnothing \rangle\!\rangle \phi.$$

Fact 4 *Let $\mathfrak{M} = (W, \sim, \leq, V)$ be a pre-differential RPD-model and we define the RPD-model $\mathfrak{M}/\sim_{\mathsf{V}} = (W', \sim', \leq', V,)$ as follows:*

- $W' = \{[w] : w \in W\}$ *where* $[w] = \{u : w \sim_{\mathsf{V}} u\}$;
- $\sim'_x = \{\langle [w],[u]\rangle : w \sim_x u\}$, $\leq'_x = \{\langle [w],[u]\rangle : w \leq_x u\}$;
- $V'(P\vec{x}) = \{[w] : w \in V(P\vec{x})\}$ *for all* $P \in \mathsf{Pred}$ *and* $\vec{x} \in \mathsf{V}^{\mathsf{ar}(P)}$.

Then $\mathfrak{M}/{\sim_{\mathsf{V}}} \in \mathsf{RMod}_d$. *Moreover, for all* $\phi \in \mathcal{L}^{\prec}$ *and* $w \in W$,

$$\mathfrak{M}, w \vDash \phi \text{ if and only if } \mathfrak{M}/{\sim_{\mathsf{V}}}, [w] \vDash \phi.$$

By Fact 3 and Fact 4, we obtain immediately that

$$\mathsf{Log}_{\mathcal{L}^{\leq}}(\mathsf{RMod}_d) = \mathsf{Log}_{\mathcal{L}^{\leq}}(\mathsf{RMod}_{pd}) = \mathsf{RLPFD} \oplus \langle\!\langle \mathsf{V}, \varnothing, \varnothing \rangle\!\rangle \phi \to \langle\!\langle \mathsf{V}, \mathsf{V}, \varnothing \rangle\!\rangle \phi.$$

Note that $\mathsf{Log}_{\mathcal{L}^{\leq}}(\mathsf{RMod}_d) = \mathsf{Log}_{\mathcal{L}^{\leq}}(\mathsf{Mod}) = \mathsf{LPFD}$, we have

Theorem 3 *If* $|\mathsf{V}| < \aleph_0$, *then* $\mathsf{RLPFD} \oplus \langle\!\langle \mathsf{V}, \varnothing, \varnothing \rangle\!\rangle \phi \to \langle\!\langle \mathsf{V}, \mathsf{V}, \varnothing \rangle\!\rangle \phi = \mathsf{LPFD}$.

5.4.2 Hilbert-style calculus $\mathsf{C}_{\mathrm{LPFD}}$

In this part, we present a calculus $\mathsf{C}_{\mathrm{LPFD}}$ of LPFD and show that $\mathsf{C}_{\mathrm{LPFD}}$ is sound, by which some key axioms are semantically explained.

(Tau) Axioms and rules for classical propositional logic;

(Nec) from ϕ infer $[\![X, Y, Z]\!]\phi$;

(K) $[\![X, Y, Z]\!](\phi \to \psi) \to ([\![X, Y, Z]\!]\phi \to [\![X, Y, Z]\!]\psi)$;

(Ord) Axioms for preference relations:

(a) $[\![X, Y, \varnothing]\!]\phi \to \phi$;

(b) $\langle\!\langle X, Y, Z \rangle\!\rangle \langle\!\langle X', Y', Z' \rangle\!\rangle \phi \to \langle\!\langle X \cap X', Y \cap Y', (Z \cap Y') \cup (Z \cap Z') \cup (Y \cap Z') \rangle\!\rangle \phi$;

(c) $[\![X, Y, Z]\!]\phi \to [\![X', Y', Z']\!]\phi$, provided $X \subseteq X'$, $Y \subseteq Y'$ and $Z \subseteq Z'$.

(d) $\langle\!\langle X, Y, Z \rangle\!\rangle \phi \to \langle\!\langle X, Y \cup Z, Z \rangle\!\rangle \phi$;

(e) $(\phi \wedge \langle\!\langle X, Y, Z \rangle\!\rangle \psi) \to \langle\!\langle X, Y, Z \rangle\!\rangle (\psi \wedge \langle\!\langle X, Y, \varnothing \rangle\!\rangle \phi) \vee \bigvee_{y \in Y} \langle\!\langle X, Y, Z \cup \{y\} \rangle\!\rangle \psi$.

(Dep) Axioms and rules for dependence:

(a) $D_X X$;

(b) $\phi \to \mathbb{D}_X \phi$, provided $\phi \in \mathsf{Atom}(X) = \{P\vec{x} : \mathsf{set}(\vec{x}) \subseteq X\} \cup \{D_Y z : Y \subseteq X\}$;

(c) $D_X S \wedge D_S T \to D_X T$;

(d) $D_X S \wedge [\![S, Y, Z]\!]\phi \to [\![X, Y, Z]\!]\phi$.

In what follows, we write C for $\mathsf{C}_{\mathrm{LPFD}}$ if there is no danger of confusion.

Theorem 4 (Soundness) *For each* $\phi \in \mathcal{L}^{\preceq}$, $\vdash_{\mathsf{C}} \phi$ *implies* $\phi \in \mathsf{LPFD}$.

Proof. We take (Ord,b) and (Ord,e) as two examples, showing their validity and giving some intuitions. Other axioms and rules can be easily checked to be valid. Let $\mathfrak{M} = (W, \sim, \leq, V)$ be a RPD-model and $w \in W$ a point.

For (Ord,b), it characterizes some kind of generalized transitivity. Suppose $\mathfrak{M}, w \vDash \langle\!\langle X, Y, Z\rangle\!\rangle\langle\!\langle X', Y', Z'\rangle\!\rangle\phi$. Then there are points $u, v \in W$ such that $u \in R(X, Y, Z)(w)$, $v \in R(X', Y', Z')(u)$ and $\mathfrak{M}, v \vDash \phi$. Let $T = (Z \cap Y') \cup (Z \cap Z') \cup (Y \cap Z')$. It is obvious that $w \sim_{X \cap X'} v$ and $w \leq_{Y \cap Y'} v$ hold. It suffices to show that $w <_T v$. Suppose $x \in Z \cap Y'$. Then $w \leq_x u$, $u \not\leq_x w$ and $u \leq_x v$. By the transitivity of $\leq_x$, we see $w \leq_x v$ and $v \not\leq_x w$, i.e., $w <_x v$. Similarly, we see $w <_x v$ whenever $x \in Y \cap Z'$ or $x \in Z \cap Z'$. Hence $\mathfrak{M}, w \vDash (Ord, b)$.

For (Ord,e), it characterizes to some degree the definition of $<$. Suppose $\mathfrak{M}, w \vDash \phi \wedge \langle\!\langle X, Y, Z\rangle\!\rangle\psi$. Then there is a point $u \in R(X, Y, Z)(w)$ such that $\mathfrak{M}, u \vDash \psi$. If $u \leq_Y w$, then clearly $\mathfrak{M}, u \vDash \psi \wedge \langle\!\langle X, Y, \varnothing\rangle\!\rangle\phi$, which entails $\mathfrak{M}, w \vDash \langle\!\langle X, Y, Z\rangle\!\rangle(\psi \wedge \langle\!\langle X, Y, \varnothing\rangle\!\rangle\phi)$. Suppose $u \not\leq_Y w$. Then there is $y \in Y$ such that $u \not\leq_y w$ and so $w <_y u$. Recall that $u \in R(X, Y, Z)(w)$, we obtain $u \in R(X, Y, Z \cup \{y\})$ and so $\mathfrak{M}, w \vDash \langle\!\langle X, Y, Z \cup \{y\}\rangle\!\rangle\psi$. Hence $\mathfrak{M}, w \vDash (Ord, e)$. □

5.4.3 Strong completeness of $\mathsf{C}_{\mathrm{LPFD}}$

For the proof of completeness, a special kind of unraveling method is used. The main reason we take such a method is that the 'canonical model' need not be an RPD-model, and modification is needed. To construct an RPD-model satisfying some given consistent set of formulas, we first pick out those so-called saturated formulas, which are sufficient to determine the preference relations in the model. Then we take 'paths' as the domain of the desired model instead of using just maximal consistent sets, which helps us deal with the intersections of relations. The relations in this model are closures of some 'one-step' relations, which help solve the problems that arise from dependence formulas. With such a model, we prove the Truth Lemma and so the Completeness Theorem.

To define a model for some satisfiable set of formulas Γ, we first define the canonical quasi-frame and investigate some properties of it:

Definition 19 (Canonical Quasi PD-Frame) *Let Δ be a set of $\mathcal{L}^{\preceq}$-formulas. We say that Δ is consistent if $\Delta \nvdash \bot$. We say that Δ is a maximal consistent set (MCS) if Δ is consistent and every proper extension of Δ is not consistent. The canonical Quasi PD-frame $\mathfrak{F}^q = (W^q, R^q)$ of C is defined as follows:*

- *W^q is the set of all MCSs;*
- *for all $X, Y, Z \subseteq_{\aleph_0} \mathsf{V}$, we define $R^q(X, Y, Z) \subseteq W^q \times W^q$ by:*

$$wR^q(X, Y, Z)u \text{ if and only if } \{\phi \in \mathcal{L}^{\preceq} : [\![X, Y, Z]\!]\phi \in w\} \subseteq u.$$

Proposition 4 *For all $\Delta_1, \Delta_2, \Delta_3 \in W^q$ and $X, Y, Z \subseteq_{\aleph_0} \mathsf{V}$:*

(1) $R^q(X, Y, \varnothing)$ is reflexive;

(2) If $\Delta_1 R^q(X, Y, Z)\Delta_2$, then $\Delta_1 R^q(X', Y', Z')\Delta_2$ for all $X' \subseteq X$, $Y' \subseteq Y \cup Z$ and $Z' \subseteq Z$;

(3) For all $Z' \subseteq_{\aleph_0} \mathsf{V}$, if $Z, Z' \subseteq Y$, $\Delta_1 R^q(X, Y, Z)\Delta_2$ and $\Delta_2 R^q(X, Y, Z')\Delta_3$, then $\Delta_1 R^q(X, Y, Z \cup Z')\Delta_3$;

(4) If $D_X S \in \Delta_1$ and $\Delta_1 R^q(X, Y, Z)\Delta_2$, then $\Delta_1 R^q(S, Y, Z)\Delta_2$ and $D_X S \in \Delta_2$.

Proof. (1) follows form axiom (Ord,a), (2) follows from Axiom (Ord,c,d), (3) follows from axiom (Ord,b) and (4) follows from axiom (Dep,b,d) immediately. □

Definition 20 *Let Σ be a MCS and $\langle\!\langle X, Y, Z\rangle\!\rangle\phi \in \Sigma$. We say that $\langle\!\langle X, Y, Z\rangle\!\rangle\phi$ is a saturated formula in Σ if $\bigvee_{y \in Y} \langle\!\langle X, Y, Z \cup \{y\}\rangle\!\rangle\phi \notin \Sigma$ and $Y \cap Z = \varnothing$. Let $S(\Sigma)$ denote the set of all saturated formulas in Σ.*

Lemma 1 *Let $\Sigma \in W^q$ be a MCS, $\langle\!\langle X, Y, Z\rangle\!\rangle\phi \in \Sigma$ and $S = Y \cup Z$. Then there is $T \in \mathcal{P}(\mathsf{V})$ such that $\langle\!\langle X, T, (Y \cup Z) \setminus T\rangle\!\rangle\phi \in S(\Sigma)$.*

Proof. The proof proceeds by induction on the size n of $Y \setminus Z$. When $n = 0$, one obtains $Z = Y \cup Z$. By axiom (Ord,c), $\langle\!\langle X, \varnothing, Z\rangle\!\rangle\phi \in \Sigma$. Note that $\bigvee \varnothing = \bot \notin \Sigma$, $\varnothing$ is the desired set. Suppose $n > 0$ and $Y \setminus Z = \{y_0, \cdots, y_{n-1}\}$. If $\langle\!\langle X, Y, Z \cup \{y_i\}\rangle\!\rangle\phi \notin \Sigma$ for any $i < n$, then $T = Y \setminus Z$ satisfies the requirement. Suppose $\langle\!\langle X, Y, Z \cup \{y_i\}\rangle\!\rangle\phi \in \Sigma$ for some $i < n$. Then we see $|Y \setminus (Z \cup \{y_i\})| < n$ and by induction hypothesis, there is $T \in \mathcal{P}(\mathsf{V})$ such that $\langle\!\langle X, T, (Y \cup Z) \setminus T\rangle\!\rangle\phi \in S(\Sigma)$. Since $Y \cup Z = Y \cup (Z \cup \{y_i\})$, T satisfies the requirement. □

Lemma 2 *Let $\Sigma \in W^q$ and $\langle\!\langle X, Y, Z\rangle\!\rangle\phi \in S(\Sigma)$. Then there is a MCS $\Delta \in R^q(X, Y, Z)(\Sigma)$ such that $\phi \in \Delta$ and $\Delta R^q(X, Y, \varnothing)\Sigma$.*

Proof. We write $\Box$ for $[\![X, Y, Z]\!]$ and $\blacklozenge$ for $\langle\!\langle X, Y, \varnothing\rangle\!\rangle$ in this proof. It is sufficient to show that $\Delta_0 = \{\psi : \Box\psi \in \Sigma\} \cup \{\blacklozenge\gamma : \gamma \in \Sigma\} \cup \{\phi\}$ is consistent. Otherwise, there are formulas $\Box\psi_1, \cdots, \Box\psi_n, \gamma_1, \cdots, \gamma_m, \in \Sigma$ such that

$$\vdash \psi_1 \wedge \cdots \wedge \psi_n \wedge \blacklozenge\gamma_1 \wedge \cdots \wedge \blacklozenge\gamma_m \wedge \phi \to \bot.$$

Let $\gamma = \gamma_1 \wedge \cdots \wedge \gamma_m$ and $\psi = \psi_1 \wedge \cdots \wedge \psi_n$. Clearly, $\gamma \in \Sigma$. By axiom (Nec) and

(K), we have $\vdash \blacklozenge\gamma \to (\blacklozenge\gamma_1 \wedge \cdots \wedge \blacklozenge\gamma_m)$. Thus $\vdash \psi \wedge \blacklozenge\gamma \wedge \phi \to \bot$, which entails $\vdash \Box\psi \to \neg\Diamond(\phi \wedge \blacklozenge\gamma)$. Note that $\Box\psi \in \Sigma$, we have $\neg\Diamond(\phi \wedge \blacklozenge\gamma) \in \Sigma$. Since $\gamma \wedge \Diamond\phi \in \Sigma$ and $(\gamma \wedge \Diamond\phi) \to \Diamond(\phi \wedge \blacklozenge\gamma) \vee \bigvee_{y\in Y}\langle\!\langle X, Y, Z \cup \{y\}\rangle\!\rangle\phi$ is an instant of axiom (Ord,e), we obtain $\bigvee_{y\in Y}\langle\!\langle X, Y, Z \cup \{y\}\rangle\!\rangle \in \Sigma$, which contradicts that $\langle\!\langle X, Y, Z\rangle\!\rangle\phi \in S(\Sigma)$. □

With the help of Lemma 1 and Lemma 2, we are now able to define the paths in W^q, which constitute the domain of our desired model.

Definition 21 *A path in W^q is a sequence $\pi = \langle\Sigma_0, \psi_0, \cdots, \Sigma_{n-1}, \psi_{n-1}, \Sigma_n\rangle$ in which the following conditions hold for all $i < n$:*

- *$\psi_i = \langle\!\langle X_i, Y_i, Z_i\rangle\!\rangle\phi_i \in S(\Sigma_i)$ is a saturated formula in $\Sigma_i \in W^q$;*
- *$\phi_i \in \Sigma_{i+1} \in W^q$, $\Sigma_{i+1}R^q(X_i, Y_i, \varnothing)\Sigma_i$ and $\Sigma_i R^q(X_i, Y_i, Z_i)\Sigma_{i+1}$.*

We denote Σ_0 by start(π)*, Σ_n by* last(π) *and the set of all paths by* Path.

In what follows, let Γ be some fixed consistent set. Without loss of generality, suppose Γ is a MCS. We now construct a model for Γ.

Definition 22 (Γ-Canonical PD-model) *The Γ-canonical PD-model $\mathfrak{M}^c_\Gamma = (\mathfrak{F}^c_\Gamma, V^c)$, in which $\mathfrak{F}^c_\Gamma = (W^c_\Gamma, \leq^c, \sim^c)$, is defined as follows:*

- *$W^c_\Gamma = \{\pi \in \text{Path} : \text{start}(\pi) = \Gamma\}$, and we write W^c for W^c_Γ in what follows;*
- *for all $y \in \mathsf{V}$ and $\pi, \pi' \in W^c$, $\pi \leq_y \pi'$ iff one of the following holds:*
  - *$\pi' = \langle\pi, \langle\!\langle X, Y, Z\rangle\!\rangle\phi, \Sigma\rangle$ and $y \in Y \cup Z$;*
  - *$\pi = \langle\pi', \langle\!\langle X, Y, Z\rangle\!\rangle\phi, \Sigma\rangle$ and $y \in Y$;*
  - *$\pi = \pi'$.*

  Let $\leq^c_y$ be the transitive closure of $\leq_y$.
- *for all $s \in \mathsf{V}$ and $\pi, \pi' \in W^c$, $\pi \rightharpoonup_s \pi'$ if and only if $\pi' = \langle\pi, \langle\!\langle X, Y, Z\rangle\!\rangle\phi, \Sigma\rangle$ and $D_X s \in \text{last}(\pi)$. Let $\rightleftharpoons_s$ be the reflexive-symmetric closure of $\rightharpoonup_s$.*

  Let $\sim^c_s$ be the transitive closure of $\rightleftharpoons_s$.
- *for all $P\vec{x} \in \mathcal{L}$, $V^c(P\vec{x}) = \{\pi \in W^c : P\vec{x} \in \text{last}(\pi)\}$.*

For all $X, Y, Z \subseteq_{\aleph_0} \mathsf{V}$, the binary relations $R^c(X, Y, Z)$, $\sim^c_X$, $\leq^c_Y$ and $<^c_Z$ are defined in the natural way. By Axiom (Dep,a), $D_X X$ always holds. Thus for each $\pi \in W^c_\Gamma$ and $\pi' = \langle\pi, \langle\!\langle X, Y, Z\rangle\!\rangle\phi, \Delta\rangle$, we have $\pi R^c(X, Y, Z)\pi'$.

To characterize the structure of W^c, we define $T \subseteq W^c \times W^c$ as follows:

$$\pi T \pi' \text{ if and only if } \pi' \text{ is of the form } \langle\pi, \langle\!\langle X, Y, Z\rangle\!\rangle\phi, \Sigma\rangle.$$

It is clear that (W^c, T) is a tree. Then for all $\pi, \pi' \in W^c$, there is a shortest T-sequence $\langle\pi_0, \cdots, \pi_n\rangle$ such that $\pi = \pi_0$, $\pi' = \pi_n$ and for all $i < n$, $\pi_i T \pi_{i+1}$ or $\pi_{i+1} T \pi_i$. We denote

the shortest sequence by $T^{\pi}_{\pi'}$.

Fact 5 *Let $\pi, \pi' \in W^c$, $T^{\pi}_{\pi'} = \langle \pi_0, \cdots, \pi_n \rangle$ and $y, s \in \mathsf{V}$. Then*

(1) $\pi \sim^c_s \pi'$ iff $\pi_i \rightleftharpoons_s \pi_{i+1}$ for all $i < n$.

(2) $\pi \leq^c_y \pi'$ iff $\pi_i \leq_y \pi_{i+1}$ for all $i < n$.

Proof. Since $\rightleftharpoons_s, \leq_y \subseteq (T \cup T^{-1})$, $\sim^c_s$ is the transitive closure of $\rightleftharpoons_s$ and $\leq^c_y$ the transitive closure of $\leq_y$, the proof can be done by induction on n easily. □

In what follows, we show that the relations $R^c(X, Y, Z)$ are consistent with the relations $R^q(X, Y, Z)$.

Lemma 3 *Let $\pi, \pi' \in W^c$, $X, Y, Z \subseteq \mathsf{V}$, $\pi \rightleftharpoons_X \pi'$, $\pi \leq_Y \pi'$ and $\pi <_Z \pi'$. Then*

(1) $\mathrm{last}(\pi) R^q(X, Y, Z) \mathrm{last}(\pi')$.

(2) if $D_X S \in \mathrm{last}(\pi)$, then $\pi \rightleftharpoons_S \pi'$.

Proof. Suppose $\pi \rightleftharpoons_X \pi'$, $\pi \leq_Y \pi'$ and $\pi <_Z \pi'$. Then we have three cases:

- $\pi = \pi'$. Then $Z = \emptyset$. By Proposition 4(1), $R^q(X, Y, \emptyset)$ is reflexive and $\mathrm{last}(\pi) R^q(X, Y, Z) \mathrm{last}(\pi')$.
- $\pi = \langle \pi', \langle\!\langle X', Y', Z' \rangle\!\rangle \psi, \Delta \rangle$. Then $Z = \emptyset$, $Y \subseteq Y'$ and $D_{X'} X \in \mathrm{last}(\pi')$. Clearly, $\mathrm{last}(\pi') R^q(X', Y', Z') \mathrm{last}(\pi)$. by Proposition 4(4), $D_{X'} X \in \mathrm{last}(\pi)$. Recall that one has $\mathrm{last}(\pi) R^q(X', Y', \emptyset) \mathrm{last}(\pi')$, by Proposition 4(2,4), we see $\mathrm{last}(\pi) R^q(X, Y, Z) \mathrm{last}(\pi')$.
- $\pi' = \langle \pi, \langle\!\langle X', Y', Z' \rangle\!\rangle \psi, \Delta \rangle$. Then $Z \subseteq Z'$, $Y \subseteq Y' \cup Z'$ and $D_{X'} X \in \mathrm{last}(\pi)$. Note that $\mathrm{last}(\pi) R^q(X', Y', Z') \mathrm{last}(\pi')$, by Proposition 4(2,4), we see $\mathrm{last}(\pi') R^q(X, Y, Z) \mathrm{last}(\pi')$.

Hence $\mathrm{last}(\pi) R^q(X, Y, Z) \mathrm{last}(\pi')$ and (1) holds.

For (2), suppose $D_X S \in \mathrm{last}(\pi)$. Then we have also three cases:

- $\pi = \pi'$. Note that $\rightleftharpoons_S$ is reflexive, $\pi \rightleftharpoons_S \pi'$.
- $\pi \rightharpoonup_X \pi'$. Then π' is of the form $\langle \pi, \langle\!\langle X', Y', Z' \rangle\!\rangle \psi, \Delta \rangle$ and $D_{X'} X \in \mathrm{last}(\pi)$. By axiom (Dep,c), $D_{X'} S \in \mathrm{last}(\pi)$. Thus $\pi \rightharpoonup_S \pi'$.
- $\pi' \rightharpoonup_X \pi$. Then π is of the form $\langle \pi', \langle\!\langle X', Y', Z' \rangle\!\rangle \psi, \Delta \rangle$ and $D_{X'} X \in \mathrm{last}(\pi')$. By (1), $D_X S \in \mathrm{last}(\pi')$. By axiom (Dep,c), $D_{X'} S \in \mathrm{last}(\pi)$. Thus $\pi' \rightharpoonup_S \pi$.

Hence $\pi \rightleftharpoons_S \pi'$ and (2) holds. □

Lemma 4 *Let $\pi, \pi' \in W^c$, $X, Y, Z \subseteq_{\aleph_0} \mathsf{V}$ and $\pi R^c(X, Y, Z) \pi'$. Then*

(1) $\mathrm{last}(\pi)R^q(X,Y,Z)\mathrm{last}(\pi')$.

(2) $D_XS \in \mathrm{last}(\pi)$ *implies* $\pi R^c(S,Y,Z)\pi'$.

Proof. Suppose $\pi R^c(X,Y,Z)\pi'$. Then $\pi \sim^c_X \pi'$, $\pi \leq^c_{Y\cup Z} \pi'$ and $\pi <^c_Z \pi'$. Let $T^{\pi}_{\pi'} = \langle \pi_0, \cdots, \pi_n\rangle$. By Fact 5, for all $i < n$, $\pi_i \rightleftharpoons_X \pi_{i+1}$ and $\pi_i \leq_{Y\cup Z} \pi_{i+1}$. Moreover, for each $z \in Z$, there is $i_z \in n$ such that $\pi_{i_z} <_z \pi_{i_z+1}$. Then by Lemma 3(1), $\mathrm{last}(\pi_i)R^q(X, Y \cup Z, \varnothing)\mathrm{last}(\pi_{i+1})$ for all $i \in n$ and for all $z \in Z$, $\mathrm{last}(\pi_{i_z})R^q(X, Y \cup Z, \{z\})\mathrm{last}(\pi_{i_z+1})$. Then by Proposition 4(2,3), we see $\mathrm{last}(\pi)R^q(X,Y,Z)\mathrm{last}(\pi')$ and (1) holds. Suppose $D_XS \in \mathrm{last}(\pi)$. Note that $\pi \sim^c_X \pi_i$ for all $i \leq n$, by (1), $D_XS \in \mathrm{last}(\pi_i)$ for all $i \leq n$. Then by Lemma 3(2), $\pi_i \rightleftharpoons_S \pi_{i+1}$ for all $i \in n$, which entails $\pi \sim^c_S \pi'$. □

The final step is to show that $\mathfrak{M}^c$ is a PD-model in which Γ is satisfiable.

Lemma 5 $\mathfrak{M}^c$ *is a PD-model.*

Proof. It suffices to show that V^c satisfies (Val). Let $\pi, \pi' \in W^c$ be points such that $\pi \sim^c_X \pi'$. By Lemma 4, $\mathrm{last}(\pi)R^q(X,\varnothing,\varnothing)\mathrm{last}(\pi')$. Assume $P\vec{x} \in \mathrm{last}(\pi)$, then by axiom (Dep,b), $\mathbb{D}_XP\vec{x} \in \mathrm{last}(\pi)$, which entails $P\vec{x} \in \mathrm{last}(\pi')$. Similarly, we can verify that $P\vec{x} \in \mathrm{last}(\pi')$ implies $P\vec{x} \in \mathrm{last}(\pi)$. Thus V^c satisfies (Val) and so $\mathfrak{M}^c_\Gamma$ is a PD-model. □

Lemma 6 (Truth Lemma) *For each formula* $\phi \in \mathcal{L}^{\leq}$ *and path* $\pi \in W^c$, $\mathfrak{M}^c, \pi \vDash \phi$ *if and only if* $\phi \in \mathrm{last}(\pi)$.

Proof. The proof proceeds by induction on the complexity of ϕ. The case when ϕ is of the form $P\vec{x}$ is trivial. The Boolean cases are also trivial. Let ϕ be of the form D_Xs. Suppose $D_Xs \in \mathrm{last}(\pi)$. Let $\pi' \in W^c$ such that $\pi \sim^c_X \pi'$. By Lemma 4, $\mathrm{last}(\pi)R^q(X,\varnothing,\varnothing)\mathrm{last}(\pi')$. Then by Proposition 4(2,4), $\pi \sim^c_s \pi'$. Thus $\mathfrak{M}^c, \pi \vDash D_Xs$. Suppose $D_Xs \notin \mathrm{last}(\pi)$. Let $\pi' = \langle \pi, \langle\!\langle X,\varnothing,\varnothing\rangle\!\rangle \top, \mathrm{last}(\pi)\rangle$. Then $\pi \not\sim_s \pi'$ and so $\pi \not\leq_s \pi'$. Clearly, $T^{\pi}_{\pi'} = \langle \pi, \pi'\rangle$. By Fact 5, $\pi \not\sim^c_s \pi'$. Note that $\pi \sim^c_X \pi'$, we see $\mathfrak{M}^c, \pi \nvDash D_Xs$. Let $\phi = \langle\!\langle X,Y,Z\rangle\!\rangle\psi$. Suppose $\mathfrak{M}^c, \pi \vDash \phi$. Then there is $\pi' \in R^c(X,Y,Z)\pi$ such that $\mathfrak{M}^c, \pi' \vDash \psi$. By induction hypothesis, $\psi \in \mathrm{last}(\pi')$. By Lemma 4, $\mathrm{last}(\pi)R^q(X,Y,Z)\mathrm{last}(\pi')$. Then $\phi \in \mathrm{last}(\pi)$. Suppose $\phi \in \mathrm{last}(\pi)$. Without loss of generality, assume that $\phi \in S(\mathrm{last}(\pi))$. Then by Lemma 2, there is a Δ such that $\pi' = \langle \pi, \phi, \Delta\rangle$ is a path with $\psi \in \mathrm{last}(\pi')$. By induction hypothesis, $\mathfrak{M}^c, \pi' \vDash \psi$. Note that $\pi R^c(X,Y,Z)\pi'$, we have $\mathfrak{M}^c, \pi \vDash \phi$. □

Theorem 5 *For each* $\Gamma \subseteq \mathcal{L}^{\leq}$, *if* Γ *is consistent, then* Γ *is satisfiable.*

5.4.4 Properties of LPFD

In this part, we prove that LPFD lacks the finite model property. The decidability of LPFD shall also be shown.

Theorem 6 *LPFD lacks the finite model property; that is, some formula $\phi \in \mathcal{L}^{\preceq}$ is only satisfiable in infinite RPD-models.*

Proof. Let $\phi = \neg([\![\varnothing, \varnothing, \{z\}]\!]\bot \vee \langle\!\langle \varnothing, \varnothing, \{z\}\rangle\!\rangle [\![\varnothing, \varnothing, \{z\}]\!]\bot)$. Note that for each PD-frame $\mathfrak{F} = (W, \sim, \leq)$ and $z \in \mathsf{V}$, $<_z$ is irreflexive and transitive. Thus for each finite PD-frame $\mathfrak{G}$, we have $\mathfrak{G} \vDash [\![\varnothing, \varnothing, \{z\}]\!]\bot \vee \langle\!\langle \varnothing, \varnothing, \{z\}\rangle\!\rangle [\![\varnothing, \varnothing, \{z\}]\!]\bot$. Clearly, ϕ is satisfiable in $(\omega, \sim, \leq)$, where $\leq_z$ is the usual $\leq$ relation on ω. □

In what follows, let α be some fixed formula, V_α the set of variables occur in α and Pred_α the set of predicates occur in α. Without loss of generality, we assume that the modal depth of α is not 0. Then we define $\mathrm{Vo} = (V_\alpha, \mathsf{Pred}_\alpha, ar \restriction V_\alpha)$ as the vocabulary restricted to α. Let $\mathcal{L}_\alpha$ be the fragment of $\mathcal{L}^{\preceq}$ based on Vo, in which every formula is of modal degree no more than α. It can be easily verified that up to modal equivalence, $\mathcal{L}_\alpha$ contains only finitely many formulas.

Definition 23 *A set Γ of $\mathcal{L}_\alpha$-formulas is said to be a $\mathcal{L}_\alpha$-maximal consistent set if $\Gamma \nvdash \bot$ and $\Gamma' \vdash \bot$ for all Γ' such that $\Gamma \subsetneq \Gamma' \subseteq \mathcal{L}_\alpha$. Let MCS_α denote the set of all $\mathcal{L}_\alpha$-maximal consistent sets. For all $X, Y, Z \subseteq V_\alpha$ and $\Delta, \Sigma \in \mathrm{MCS}_\alpha$, we write $\Delta R^p_\alpha(X, Y, Z)\Sigma$ if $\{\langle\!\langle X \cap X', Y \cap Y', (Z \cap Y') \cup (Z' \cap Y) \cup (Z \cap Z')\rangle\!\rangle \phi \in \mathcal{L}_\alpha : \langle\!\langle X', Y', Z'\rangle\!\rangle \phi \in \Sigma\} \subseteq \Delta$.*

One may find that the definition of $R^p_\alpha(X, Y, Z)$ is modified from the Lemmon filtration. Given that $\langle\!\langle X, Y, Z\rangle\!\rangle \phi \in \mathcal{L}_\alpha$ and $\Delta R^p_\alpha(X, Y, Z)\Sigma$, we see $\phi \in \Sigma$ implies $\langle\!\langle X, Y, \varnothing\rangle\!\rangle \phi \in \Sigma$ and so $\langle\!\langle X, Y, Z\rangle\!\rangle \phi \in \Delta$. Then we have the following proposition:

Proposition 5 *For all $\Delta_1, \Delta_2, \Delta_3 \in \mathrm{MCS}_\alpha$ and $X, Y, Z \subseteq V_\alpha$,*

(1) $R^p_\alpha(X, Y, \varnothing)$ is reflexive;

(2) If $\Delta_1 R^p_\alpha(X, Y, Z)\Delta_2$, then $\Delta_1 R^p_\alpha(X', Y', Z')\Delta_2$ for all $X' \subseteq X$, $Y' \subseteq Y \cup Z$ and $Z' \subseteq Z$;

(3) If $D_X S \in \Delta_1$ and $\Delta_1 R^p_\alpha(X, Y, Z)\Delta_2$, then $\Delta_1 R^p_\alpha(S, Y, Z)\Delta_2$ and $D_X S \in \Delta_2$

(4) For all $Z' \subseteq_{\aleph_0} V_\alpha$, if $Z, Z' \subseteq Y$, $\Delta_1 R^p_\alpha(X, Y, Z)\Delta_2$ and $\Delta_2 R^p_\alpha(X, Y, Z')\Delta_3$, then $\Delta_1 R^p_\alpha(X, Y, Z \cup Z')\Delta_3$.

Proof. (1) and (2) are trivial. For (3), $\Delta_1 R^p_\alpha(S, Y, Z)\Delta_2$ follows from axiom (Dep,d). Recall that the modal depth of α is not 0, we see $\mathbb{D}_X D_X S \in \Delta_1$ and so $D_X S \in \Delta_2$. For (4), suppose $\langle\!\langle X_0, Y_0, Z_0\rangle\!\rangle \phi \in \Delta_3$. Then $\langle\!\langle X \cap X_0, Y \cap Y_0, (Z' \cap Y_0) \cup (Z_0 \cap Y) \cup (Z' \cap Z_0)\rangle\!\rangle \phi \in$

Δ_2. Recall that $Z, Z' \subseteq Y$, it follows that

$$\langle\!\langle X \cap X_0, Y \cap Y_0, ((Z \cup Z') \cap Y_0) \cup (Z_0 \cap Y) \cup ((Z \cup Z') \cap Z_0)\rangle\!\rangle \phi \in \Delta_1.$$

Thus $\Delta_1 R^p_\alpha(X, Y, Z \cup Z')\Delta_3$ and (4) holds. □

Definition 24 ($\mathcal{L}_\alpha$-Pre-model) *An $\mathcal{L}_\alpha$-pre-model is a set F of $\mathcal{L}_\alpha$-MCSs such that for all $X, Y, Z \subseteq V_\alpha$ and $\Delta \in F$, the following statement holds:*

(†) If $\langle\!\langle X, Y, Z\rangle\!\rangle \phi$ is a saturated formula in Δ, then there is $\Sigma \in F$ such that $\Delta R^p_\alpha(X, Y, Z)\Sigma$, $\phi \in \Sigma$ and $\Sigma R^p_\alpha(X, Y, \varnothing)\Delta$.

We say ϕ is satisfied in F if there is some $\Delta \in F$ such that $\phi \in \Delta$.

Lemma 7 *For each satisfiable $\phi \in \mathcal{L}^{\leq}$, ϕ is satisfied in some pre-model.*

Proof. Let $\mathfrak{M} = (W, \sim, \leq, V)$ be a RPD-model and $w \in W$ such that $\mathfrak{M}, w \vDash \phi$. Then we define $F_{\mathfrak{M}=\{\Delta_w : w \in \mathfrak{M} \text{ and } \Delta_w = \{\phi \in \mathcal{L}_\alpha : \mathfrak{M}, w \vDash \phi\}\}}$. It suffices to show that $F_{\mathfrak{M}}$ satisfies (†). Suppose $\langle\!\langle X, Y, Z\rangle\!\rangle \phi$ is a saturated formula in Δ_w. Then $\mathfrak{M}, w \vDash \langle\!\langle X, Y, Z\rangle\!\rangle \phi$ and there is $u \in R(X, Y, Z)(w)$ such that $\mathfrak{M}, u \vDash \phi$. Note that $\langle\!\langle X, Y, Z\rangle\!\rangle \phi$ is a saturated formula, we have $w \in R(X, Y, \varnothing)(u)$. Then it is not hard to verify that $\Delta_w R^p_\alpha(X, Y, Z)\Delta_u$ and $\Delta_u R^p_\alpha(X, Y, \varnothing)\Delta_w$. Recall that $\phi \in \Delta_u$, we see that (†) holds for $F_{\mathfrak{M}}$. □

Definition 25 (Induced Model) *Let F be a pre-model. An F-path is a tuple $\langle \Sigma_0, \psi_0, \cdots, \Sigma_{n-1}, \psi_{n-1}, \Sigma_n \rangle$ where the following conditions hold for all $i < n$:*

- *$\psi_i = \langle\!\langle X_i, Y_i, Z_i\rangle\!\rangle \phi_i$ is a saturated formula in $\Sigma_i \in F$;*
- *$\phi_i \in \Sigma_{i+1} \in F$, $\Sigma_{i+1} R^p_\alpha(X_i, Y_i, \varnothing)\Sigma_i$ and $\Sigma_i R^p_\alpha(X_i, Y_i, Z_i)\Sigma_{i+1}$.*

The RPD-model $\mathfrak{M}^F = (W^F_\Gamma, \leq^F, \sim^F, V^F)$ induced by $\Gamma \in F$ is defined by:

- *W^F_Γ is the set of all paths in F begins with ϕ.*
- *for all $y \in V_\alpha$ and $\pi, \pi' \in W^F_\Gamma$, $\pi \leq_y \pi'$ iff one of the following holds:*
  - *$\pi' = \langle \pi, \langle\!\langle X, Y, Z\rangle\!\rangle \phi, \Sigma \rangle$ and $y \in Y \cup Z$;*
  - *$\pi = \langle \pi', \langle\!\langle X, Y, Z\rangle\!\rangle \phi, \Sigma \rangle$ and $y \in Y$.*

  Let $\leq^F_y$ be the reflexive-transitive closure of $\leq_y$.
- *for all $s \in V_\alpha$ and $\pi, \pi' \in W^F_\Gamma$, $\pi \rightharpoonup_s \pi'$ if and only if $\pi' = \langle \pi, \langle\!\langle X, Y, Z\rangle\!\rangle \phi, \Sigma \rangle$ and $D_X s \in \text{last}(\pi)$. Let $\sim^F_s$ the reflexive-symmetric-transitive closure of $\rightleftharpoons_s$.*
- *for all $P\vec{x} \in \mathcal{L}_\alpha$, $V^F(P\vec{x}) = \{\pi \in W^F : P\vec{x} \in \text{last}(\pi)\}$.*

One may notice now that the construction of the desired model is almost the same as the one we used in the proof of Completeness Theorem. And similar to the proof of

Completeness Theorem, with the help of Fact 5 and the definition of pre-models, we can verify that the following lemma holds:

Lemma 8 (Truth Lemma) *For each formula $\phi \in \mathcal{L}_\alpha$ and path $\pi \in W^F$,*

$$\mathfrak{M}^F, \pi \vDash \phi \text{ if and only if } \phi \in \text{last}(\pi).$$

As a consequence, for each $\phi \in \mathcal{L}^{\preceq}$, ϕ is satisfiable if and only if ϕ is satisfied in some $\mathcal{L}_\phi$-pre-model. Recall that up to modal equivalence, $\mathcal{L}_\phi$ contains finitely many formulas, MCS_ϕ is finite for each $\phi \in \mathcal{L}^{\preceq}$, we obtain the following theorem:

Theorem 7 *The satisfiability problem of LPFD is decidable.*

5.4.5 The hybrid extension of LPFD

In this subsection, we extend LPFD with nominals. We will use this hybrid extension to express an important solution concept for cooperative games – the core – in Section 5.5.

By a vocabulary with nominals we mean a tuple $(\mathsf{V}, \mathsf{Pred}, \mathsf{Nom}, \mathsf{ar})$ where $(\mathsf{V}, \mathsf{Pred}, \mathsf{ar})$ is a vocabulary and $\mathsf{Nom} = \{i_k : k \in \omega\}$ a denumerable set of nominals. The language $\mathcal{L}^{\preceq}_{\mathsf{Nom}}$ with nominals is given by:

$$\mathcal{L}^{\preceq}_{\mathsf{Nom}} \ni \phi ::= P\vec{x} \mid D_X y \mid i \mid \neg\phi \mid \phi \wedge \phi \mid [\![X, Y, Z]\!]\phi,$$

which only differs from the language of LPFD in those nominals.

We modify the valuation V in a RPD-model $\mathfrak{M} = (W, \sim, \leq, V)$ correspondingly such that $V \restriction \mathsf{Nom}$ is a partial function from Nom to W. The resulted RPD-models are called RPDN-models. The semantic truth of the nominals in an RPDN-model is defined as follows:

$$\mathfrak{M}, w \vDash i \quad \text{if and only if} \quad w = V(i)$$

As usual, we call LPFD with nominals 'hybrid LPFD', abbreviated to HLPFD. Let Nom be a fixed set of nominals. We present here the calculus $\mathsf{C}_{\mathsf{Nom}}$ for HLPFD and show its soundness and completeness. Let $X, Y, Z \in \mathcal{P}^{<\aleph_0}(\mathsf{V})$, $\phi, \psi \in \mathcal{L}^{\preceq}_{\mathsf{Nom}}$, $i, j \in \mathsf{Nom}$, $P \in \mathsf{Pred}$ and $v \in \mathsf{V}$. The axioms and rules of $\mathsf{C}_{\text{HLPFD}}$ are as follows:

(Tau) Axioms and rules for classical propositional logic;

(Nec) from ϕ infer $[\![X, Y, Z]\!]\phi$;

(K) $[\![X, Y, Z]\!](\phi \to \psi) \to ([\![X, Y, Z]\!]\phi \to [\![X, Y, Z]\!]\psi)$;

(Dep) $\phi \to \mathbb{D}_X\phi$, provided $\phi \in \mathsf{Atom}(X) = \{P\vec{x} : \mathsf{set}(\vec{x}) \subseteq X\}$;

(Nom) $@_i\phi \to [\![\varnothing, \varnothing, \varnothing]\!](i \to \phi)$, provided $i \in \mathsf{Nom}$;

(Name) from $i \to \phi$ infer ϕ, provided that $i \notin \phi$, i.e., i does not occur in ϕ;

(Paste) from $@_i\langle\!\langle X, Y, Z\rangle\!\rangle j \to @_j\phi$ infer $@_i[\![X, Y, Z]\!]\phi$, provided $i \neq j$ and $j \notin \phi$;

(DD) Axioms and rules for $[\![\ ]\!] - D$ interaction:

(1) $D_X s \wedge [\![\{s\}, \varnothing, \varnothing]\!]\phi \to [\![X, \varnothing, \varnothing]\!]\phi$;

(2) $i \wedge \neg D_X s \to \langle\!\langle X, \varnothing, \varnothing\rangle\!\rangle [\![s, \varnothing, \varnothing]\!]\neg i$.

(Ord) Axioms for the preference orders:

(1) $[\![X, Y, \varnothing]\!]\phi \to \phi$;

(2) $\phi \to [\![\{v\}, \varnothing, \varnothing]\!]\langle\!\langle \{v\}, \varnothing, \varnothing\rangle\!\rangle \phi$;

(3) $\langle\!\langle X, Y, Z\rangle\!\rangle\langle\!\langle X', Y', Z'\rangle\!\rangle \phi \to \langle\!\langle X \cap X', Y \cap Y', (Z \cap Y') \cup (Z \cap Z') \cup (Y \cap Z')\rangle\!\rangle \phi$;

(4) $@_i\langle\!\langle \varnothing, \varnothing, \{v\}\rangle\!\rangle j \leftrightarrow @_i\langle\!\langle \varnothing, \{v\}, \varnothing\rangle\!\rangle j \wedge @_j\neg\langle\!\langle \varnothing, \{v\}, \varnothing\rangle\!\rangle i$, provided $i, j \in \mathsf{Nom}$;

(5) $\langle\!\langle X, Y, Z\rangle\!\rangle i \wedge \langle\!\langle X', Y', Z'\rangle\!\rangle i \leftrightarrow \langle\!\langle X \cup X', Y \cup Y', Z \cup Z'\rangle\!\rangle i$, provided $i \in \mathsf{Nom}$.

Comparing $\mathsf{C}_{\mathsf{Nom}}$ with C, in addition to the standard axioms and rules for nominals, axioms (Ord,4,5) and (DD,2) are new, which characterize RPD-models in a more refined way. Note also that some old axioms in C are presented in $\mathsf{C}_{\mathsf{Nom}}$ in a different way. For example, axiom (DD,1) in $\mathsf{C}_{\mathsf{Nom}}$ are bottom-up versions of axioms (Dep,d) in C.

With the above mentioned changes in $\mathsf{C}_{\mathsf{Nom}}$ due to the addition of nominals, the completeness of $\mathsf{C}_{\mathsf{Nom}}$ can be proved by directly using the canonical model, which is a standard method and relatively routine.

Theorem 8 $\mathsf{C}_{\mathsf{Nom}}$ *is sound and strongly complete.*

Lemma 9 *Let* Γ *be a* $\mathsf{C}_{\mathsf{Nom}}$*-consistent set and* $\mathsf{Nom}' = \mathsf{Nom} \cup \{j_n : n \in \omega\}$. *Then* Γ *can be extended to a maximal* $\mathsf{C}_{\mathsf{Nom}'}$*-consistent set* Γ^+ *of formulas satisfying the following conditions:*

(Named) $\Gamma^+ \cap \mathsf{Nom}' \neq \varnothing$;

(Pasted) For all $@_i\langle\!\langle X, Y, Z\rangle\!\rangle \phi \in \Gamma$, *there is a nominal* $j \in \mathsf{Nom}'$ *such that* $@_i\langle\!\langle X, Y, Z\rangle\!\rangle j \wedge @_j\phi \in \Gamma$.

The proof of Lemma 9 is standard.

Fact 6 *Let* Γ *be a named and pasted maximal* $\mathsf{C}_{\mathsf{Nom}}$*-consistent set. For each* $i \in \mathsf{Nom}$ *such that* $@_i\top \in \Gamma$, *let* $\Delta_i = \{\phi : @_i\phi \in \Gamma\}$. *Then for all* $i, j \in \mathsf{Nom}$,

(1) Δ_i *is a maximal* $\mathsf{C}_{\mathsf{Nom}}$*-consistent set.*

(2) $i \in \Delta_j$ *if and only if* $\Delta_i = \Delta_j$.

Definition 26 *Given a named and pasted maximal* $\mathsf{C}_{\mathsf{Nom}}$*-consistent set* Γ*, we define the canonical model* $\mathfrak{M}_\Gamma = (W_\Gamma, \sim_\Gamma, \leq_\Gamma, V_\Gamma)$ *for* Γ *as follows:*

- $W_\Gamma = \{\Delta_i : @_i\top \in \Gamma$ *and* $\Delta_i = \{\phi : @_i\phi \in \Gamma\}\}$*;*
- *for each* $v \in \mathsf{V}$*,* $\Delta_i \sim_v \Delta_j$ *if and only if* $@_i\langle\!\langle\{v\}, \varnothing, \varnothing\rangle\!\rangle j \in \Gamma$*;*
- *for each* $v \in \mathsf{V}$*,* $\Delta_i \leq_v \Delta_j$ *if and only if* $@_i\langle\!\langle\varnothing, \{v\}, \varnothing\rangle\!\rangle j \in \Gamma$*;*
- $V(P\vec{x}) = \{\Delta_i : @_iP\vec{x} \in \Gamma\}$ *and* $V(i) = \Delta_i$*.*

Lemma 10 *Let* Γ *be a named and pasted maximal* $\mathsf{C}_{\mathsf{Nom}}$*-consistent set. Then* $\mathfrak{M}_\Gamma = (W, \sim, \leq, V)$ *is an RDPN-model.*

Proof. Let $v \in \mathsf{V}$. By axiom (Ord,1,2,3), $\sim_v$ is a pre-order and $\leq_v$ is an equivalence relation. Then $(W, \sim, \leq)$ is an RPD-frame. Note that $V(i) \in W$ for each $i \in \mathsf{Nom} \cap \mathrm{dom}(V)$. To show that $\mathfrak{M}_\Gamma$ is a RPDN-model, it suffices to show that V satisfies (Val). Let $\vec{x} = (x_1, \cdots, x_n)$. Suppose $\Delta_i \sim_{\mathsf{set}(\vec{x})} \Delta_j$ and $\Delta_i \in V(P\vec{x})$. Then $P\vec{x} \in \Delta_i$. By (Dep), $\mathbb{D}_X P\vec{x} \in \Delta_i$, which entails $P\vec{x} \in \Delta_j$. □

Lemma 11 *Let* Γ *be a named and pasted maximal* $\mathsf{C}_{\mathsf{Nom}}$*-consistent set,* $\mathfrak{M}_\Gamma = (W, \sim, \leq, V)$*,* $i \in \mathsf{Nom}$ *and* $\Delta_i \in W$*. Then*

(1) If $\langle\!\langle X, Y, Z\rangle\!\rangle j \in \Delta_i$*, then* $\Delta_i R(X, Y, Z)\Delta_j$*;*

(2) If $@_i\langle\!\langle X, Y, Z\rangle\!\rangle\phi \in \Gamma$*, then there is* $j \in \mathsf{Nom}$ *with* $\phi \in \Delta_j$ *and* $\Delta_i R(X, Y, Z)\Delta_j$*.*

(3) $D_X s \in \Delta_i$ *if and only if* $\mathfrak{M}_\Gamma, \Delta_i \vDash D_X s$*.*

(4) For all $\phi \in \mathcal{L}_{\mathsf{Nom}}$*,* $\phi \in \Delta_i$ *if and only if* $\mathfrak{M}_\Gamma, \Delta_i \vDash \phi$*.*

Proof. For (1), suppose $\langle\!\langle X, Y, Z\rangle\!\rangle j \in \Delta_i$. By axiom (Ord,5), we see $\langle\!\langle\{x\}, \varnothing, \varnothing\rangle\!\rangle j$, $\langle\!\langle\varnothing, \{y\}, \varnothing\rangle\!\rangle j, \langle\!\langle\varnothing, \varnothing, \{z\}\rangle\!\rangle j \in \Delta_i$ for all $x \in X$, $y \in Y$ and $z \in Z$, which entails by axiom (Ord,4) that $\Delta_i \sim_X \Delta_j$, $\Delta_i \leq_Y \Delta_j$ and $\Delta_i <_Z \Delta_j$. Thus $\Delta_i R(X, Y, Z)\Delta_j$.

For (2), suppose $@_i\langle\!\langle X, Y, Z\rangle\!\rangle\phi \in \Gamma$. Since Γ is pasted, there is $j \in \mathsf{Nom}$ such that $@_i\langle\!\langle X, Y, Z\rangle\!\rangle j \wedge @_j\phi \in \Gamma$. Thus $\phi \in \Delta_j$ and $\Delta_i R(X, Y, Z)\Delta_j$.

For (3), suppose $D_X s \in \Delta_i$ and $\Delta_i \sim_X \Delta_j$. We show that $\langle\!\langle\{s\}, \varnothing, \varnothing\rangle\!\rangle j \in \Delta_i$. Assume $\langle\!\langle\{s\}, \varnothing, \varnothing\rangle\!\rangle j \notin \Delta_i$. Then by axiom (DD,1), we see $\mathbb{D}_X\neg j \in \Delta_i$, which contradicts to $\Delta_i \sim_X \Delta_j$. Thus $\mathfrak{M}_\Gamma, \Delta_i \vDash D_X s$. Suppose $D_X s \notin \Delta_i$. Then $i \wedge \neg D_X s \in \Delta_i$. By axiom (DD,2), we see $@_i\langle\!\langle X, \varnothing, \varnothing\rangle\!\rangle\mathbb{D}_s\neg i \in \Gamma$. Since Γ is pasted, there is $j \in \mathsf{Nom}$ such that $@_i\langle\!\langle X, \varnothing, \varnothing\rangle\!\rangle j \wedge @_j\mathbb{D}_s\neg i \in \Gamma$. Thus $\Delta_i \sim_X \Delta_j$ and $\Delta_i \nsim_s \Delta_j$. Note that $\sim_s$ is symmetric, $\Delta_j \nsim_s \Delta_i$. Thus $\mathfrak{M}_\Gamma, \Delta_i \nvDash D_X s$.

For (4), the proof proceeds by induction on the complexity of ϕ. The case when $\phi = D_X s$ follows from (3) immediately. The case $\phi = P\vec{x}$ or $\phi \in \mathsf{Nom}$ is trivial.

The Boolean cases are also trivial. Let $\phi = [\![X, Y, Z]\!]\psi$. Assume $[\![X, Y, Z]\!]\psi \notin \Delta_i$. Then $\langle\!\langle X, Y, Z\rangle\!\rangle\neg\psi \in \Delta_i$ and so $@_i\langle\!\langle X, Y, Z\rangle\!\rangle\neg\psi \in \Gamma$. By (2), $\neg\psi \in \Delta_j$ for some $\Delta_j \in R(X, Y, Z)(\Delta_i)$. Then $\psi \notin \Delta_j$ and by induction hypothesis, $\mathfrak{M}_\Gamma, \Delta_j \not\models \psi$, which entails $\mathfrak{M}_\Gamma, \Delta_i \not\models [\![X, Y, Z]\!]\psi$. Assume that $\mathfrak{M}_\Gamma, \Delta_i \not\models [\![X, Y, Z]\!]\psi$. Then there is $\Delta_j \in R(X, Y, Z)(\Delta_i)$ such that $\mathfrak{M}_\Gamma, \Delta_j \not\models \psi$. By induction hypothesis, $\psi \notin \Delta_j$ and so $\neg\psi \wedge j \in \Delta_j$. Note that $\langle\!\langle X, Y, Z\rangle\!\rangle j \in \Delta_i$, we see $\langle\!\langle X, Y, Z\rangle\!\rangle\neg\psi \in \Delta_i$, which entails $[\![X, Y, Z]\!]\psi \notin \Delta_i$. □

Theorem. $\mathsf{C}_{\mathsf{Nom}}$ *is sound and strongly complete.*

Proof. Soundness is not hard to verify. Let $\Gamma^- \subseteq \mathcal{L}_{\mathsf{Nom}}$ be any consistent set of formulas. By Lemma 9, Γ^- can be extended to a named and pasted maximal $\mathbf{LPFD}_{\mathbf{Nom}'}$-consistent set Γ. By Lemma 10, the triple $\mathfrak{M}_\Gamma = (W_\Gamma, R_\Gamma, V_\Gamma)$ defined in Definition 26 is a DP-model with nominals. By Lemma 11, we see $\mathfrak{M}_\Gamma \models \Gamma^-$. Then $\mathfrak{M}_\Gamma{\restriction}\mathsf{Nom}$ is a RPDN-model satisfying Γ^-. □

Note that the equivalence between PD-models with nominals (where $I(i) \in A$ for $i \in \mathsf{Nom}$) and RPDN-models with respect to $\mathcal{L}^{\leq}_{\mathsf{Nom}}$ can be established as in Section 5.4.1. So for the class of PD-models with nominals we also have the soundness and strong completeness of $\mathsf{C}_{\mathsf{Nom}}$.

As for the decidability of HLPFD, we cannot prove it by directly following the strategy used in the proof of LPFD's decidability. We will not attack this problem in the thesis but rather leave it for future work.

Remark 1 *For the case when* V *is finite, the class* RMod_d *is characterized by the formula* $\phi_d = i \to [\![\mathsf{V}, \varnothing, \varnothing]\!]i$. *It is not hard to verify that the equation* $\mathsf{Log}_{\mathcal{L}^{\leq}_{\mathsf{Nom}}}(\mathsf{Mod}) = \mathsf{Log}_{\mathcal{L}^{\leq}_{\mathsf{Nom}}}(\mathsf{RMod}_d)$ *holds. Thus* $\mathsf{C}_{\mathsf{Nom}} \oplus \phi_d$ *is the desired calculus for such case.*

5.5 Cooperative Games and the Core in HLPFD

Should there be any difference between a coalitional action and an agglomeration of actions? This is a key issue in the philosophical analysis of collective agency (cf. Roth (2017)). In this section, we provide a game theoretical perspective on this issue by modelling cooperative games in strategic and coalitional form (cf. Peleg et al. (2007), Section 11) in LPFD and characterizing one of its solution concepts, the core, in HLPFD.

5.5.1 Cooperative games in LPFD

Different from non-cooperative games, in cooperative games in strategic form (cf. Peleg et al. (2007), Section 11), players can not only act individually but also choose to join a coalition and act as a part of the coalition. In such games, the players in a coalition can do something together in agreement rather than separately. So collective actions and power are different from an agglomeration of individual actions and its effectiveness. In this part, we propose a framework based on LPFD to represent cooperative games and make the difference explicit. For simplicity, we restrict ourselves to the cases where the set of the players is finite.

To explicitly model coalitions as a different part of each player's choices from strategies, we distinguish between the terms "strategy"(or equivalently "action") and "choice".

Definition 27 (Players' Choices and Choices Merging)

- ***Players' Choices:*** *The set of the players' choices is defined as follows:*

$$O := \{f : I \to \Sigma \mid I \subseteq \mathsf{V}\}$$

  where Σ *is the set of all possible strategies of all players.*

- ***Choices Merging:*** *For* $f, f' \in O$ *with* $\mathsf{dom}(f) \cap \mathsf{dom}(f') = \varnothing$, $f \oplus f' := f \cup f'$.

For example, given three players $\mathsf{V} = \{1, 2, 3\}$ and the players' possible strategies in $\Sigma = \{\alpha, \beta\}$, $f = \{(1, \alpha), (3, \beta)\} \in O$ denotes a possible choice of the players 1 and 3 as a coalition; $f' = \{(2, \alpha)\} \in O$ denotes a possible choice of the player 2. Then $f \oplus f' = \{(1, \alpha), (2, \alpha), (3, \beta)\}$.

In a PD-model, there is no requirement on $A \subseteq O^{\mathsf{V}}$. This is not the case any longer when the players' choices concern forming coalitions. We impose three conditions on a *realizable* choice profile. First of all, a player cannot choose to form a coalition she is not in. Second, a player cannot choose to form a coalition without the others in the coalition making the same choice. Third, once a coalition forms, it acts as a whole, which means that its members act according to a unique strategy sequence. This strategy sequence can be seen as a collective plan which is made effective by common consent.

To make the definition of realizable choice profiles precise, we make use of the following notations.

Notation

- $\Pi(\mathsf{V})$ is the set of all partitions of V. ①
- Given $a \in O^{\mathsf{V}}$,

① A partition of V is a set of non-empty subsets of V whose union is V and which do not intersect each other.

- a_i denotes the ith element of a, which is a function;
- $a_{\mathsf{rng}} := \{a_i \in O \mid i \in \mathsf{V}\}$;
- $a_{\mathsf{dom}} := \{\mathsf{dom}(a_i) \subseteq \mathsf{V} \mid i \in \mathsf{V}\}$;

Definition 28 (Realizable Choice and Strategy Profiles)

- *A choice profile $a \in O^N$ is realizable if and only if it satisfies the following three conditions:*
  1. $i \in \mathsf{dom}(a_i)$;
  2. $a_{\mathsf{dom}} \in \Pi(\mathsf{V})$
  3. $\mathsf{dom}(a_i) = \mathsf{dom}(a_j)$ *implies that* $a_i = a_j$ *for all* $i, j \in \mathsf{V}$.
- *Let Ξ denote the set of all realizable choice profiles.*

  Let $a_{\mathsf{merge}} := \bigoplus_{f \in a_{\mathsf{rng}}} f$ for $a \in \Xi$. Given $A \subseteq \Xi$, the set of all realizable strategy profiles of a partition $\pi \in \Pi(\mathsf{V})$ in A is

$$\sigma_A(\pi) := \{a_{\mathsf{merge}} \mid a \in A \text{ and } a_{\mathsf{dom}} = \pi\} \ .$$

  When there is no danger of ambiguity, we will leave out the subscript A.

Having defined O and Ξ, we define a class of PD-models we will work with.

Definition 29 (Coalition-preference-dependence (CPD) models) *A coalition-preference-dependence model is a PD-model $\mathbb{M} = ((O, I), A)$ in which O is defined in Definition 27 and A and $\preceq_i$ satisfy the following conditions:*

1. $A \subseteq \Xi$;
2. $\{a_{\mathsf{dom}} \mid a \in A\} = \Pi(\mathsf{V})$;
3. *if $\pi \in \Pi(\mathsf{V})$ is finer than $\pi' \in \Pi(\mathsf{V})$,*[①] *then $\sigma_A(\pi) \subseteq \sigma_A(\pi')$;*
4. *if $a_{\mathsf{merge}} = a'_{\mathsf{merge}}$, then $a \simeq_i a'$ for all $i \in \mathsf{V}$;*
5. *$\preceq_i$ is total for all $i \in \mathsf{V}$.*

The first condition says that A should contain realizable choice profiles. The second condition says that the players can form coalitions according to all possible partitions of N. The third condition requires bigger coalitions to have no less strategies than smaller coalitions. The fourth condition requires that the players' preference relations depend directly on strategy profiles. The players' choices of coalitions can only influence the players' preferences by affecting their strategies. The last condition requires the players' preference relations to be total, which is a standard assumption in game theory.

The following example illustrates our notations and the CPD-models.

① That is, for all $X \in \pi$ there is $X' \in \pi'$ such that $X \subseteq X'$.

Example 6 *Let* $\mathsf{V} = \{1,2,3\}$ *and* $\Sigma = \{\alpha, \beta, \gamma\}$. *A is given in Table 5.3. According to*

	1	2	3
a	$\{(1,\alpha)\}$	$\{(2,\beta)\}$	$\{(3,\alpha)\}$
a'	$\{(1,\alpha),(2,\beta)\}$	$\{(1,\alpha),(2,\beta)\}$	$\{(3,\alpha)\}$
a''	$\{(1,\alpha),(2,\gamma)\}$	$\{(1,\alpha),(2,\gamma)\}$	$\{(3,\beta)\}$
$a^{3\prime}$	$\{(1,\alpha)\}$	$\{(2,\beta),(3,\alpha)\}$	$\{(2,\beta),(3,\alpha)\}$
$a^{4\prime}$	$\{(1,\beta)\}$	$\{(2,\beta),(3,\gamma)\}$	$\{(2,\beta),(3,\gamma)\}$
$a^{5\prime}$	$\{(1,\alpha),(3,\alpha)\}$	$\{(2,\beta)\}$	$\{(1,\alpha),(3,\alpha)\}$
$a^{6\prime}$	$\{(1,\gamma),(3,\alpha)\}$	$\{(2,\alpha)\}$	$\{(1,\gamma),(3,\alpha)\}$
$a^{7\prime}$	$\{(1,\alpha),(2,\beta),(3,\alpha)\}$	$\{(1,\alpha),(2,\beta),(3,\alpha)\}$	$\{(1,\alpha),(2,\beta),(3,\alpha)\}$
$a^{8\prime}$	$\{(1,\alpha),(2,\gamma),(3,\beta)\}$	$\{(1,\alpha),(2,\gamma),(3,\beta)\}$	$\{(1,\alpha),(2,\gamma),(3,\beta)\}$
$a^{9\prime}$	$\{(1,\beta),(2,\beta),(3,\gamma)\}$	$\{(1,\beta),(2,\beta),(3,\gamma)\}$	$\{(1,\beta),(2,\beta),(3,\gamma)\}$
$a^{10\prime}$	$\{(1,\gamma),(2,\alpha),(3,\alpha)\}$	$\{(1,\gamma),(2,\alpha),(3,\alpha)\}$	$\{(1,\gamma),(2,\alpha),(3,\alpha)\}$
$a^{11\prime}$	$\{(1,\gamma),(2,\gamma),(3,\gamma)\}$	$\{(1,\gamma),(2,\gamma),(3,\gamma)\}$	$\{(1,\gamma),(2,\gamma),(3,\gamma)\}$

Table 5.3 A in Example 6

our notation,

- $a_{\mathsf{merge}} = a'_{\mathsf{merge}} = a^{3\prime}_{\mathsf{merge}} = a^{5\prime}_{\mathsf{merge}} = a^{7\prime}_{\mathsf{merge}} = \{(1,\alpha),(2,\beta),(3,\alpha)\}$;
- $\sigma(\{\{1\},\{2\},\{3\}\}) = \{\{(1,\alpha),(2,\beta),(3,\alpha)\}\}$
  and $\sigma(\{\{1,2\},\{3\}\}) = \{\{(1,\alpha),(2,\beta),(3,\alpha)\},\{(1,\alpha),(2,\gamma),(3,\beta)\}\}$.

As the readers can verify, all the requirements of a CPD-model concerning A are satisfied here. For example, $\sigma(\{\{1\},\{2\},\{3\}\}) \subseteq \sigma(\{\{1,2\},\{3\}\}) \subseteq \sigma(\{N\})$. *To make sure* $\preceq_i$ *satisfy the requirements,* $a \simeq_i a' \simeq_i a^{3\prime} \simeq_i a^{5\prime} \simeq_i a^{7\prime}$ *needs to be the case.*

As can be easily spotted in the above example, coalitions are explicitly incorporated into the players' choices in the CPD-models. Once a coalition forms, the players in it act as a whole. Moreover, a coalition could possibly do more than its constituent parts. ①

① CPD-models are similar to cooperative games in strategic form (Peleg et al. (2007), Definition 11.1.1) but not exactly the same. While cooperative games in strategic form use a function to assign to each coalition its available strategies, we start with each player's possible choices of coalitions and actions.

Conceptually, starting with individual choices rather than coalitional strategies leaves room for an analysis of the relationship between interaction between individual choices and collective agency. Technically, CPD-models impose less restrictions on the strategy profiles of coalitions. In Example 6, according to cooperative games in strategic form, $B = \{\{(1,\alpha),(2,\beta)\},\{(1,\alpha),(2,\gamma)\}\}$ includes the available strategies for coalition $\{1,2\}$; $C = \{\{(3,\alpha)\},\{(3,\beta)\}\}$ includes the available strategies for coalition $\{3\}$. But unlike cooperative games in strategic form, we do not require that $B \times C$ should be included in the available strategies for $\{1,2,3\}$. For example, $\{(1,\alpha),(2,\beta),(3,\beta)\}$ is not in it.

As for cooperative games in characteristic function form (also called coalitional games with transferable utility) (Peleg et al. (2007), Definition 2.1.1), its framework centers around a function which associates a real number with each group of players (called the worth of the group). The theory of coalitional games is then largely about each

The coalition partition formed in a game directly affects each player's strategy. Hence it has a substantial influence on the final outcome of the game. Can the language of LPFD express what partition is formed in a realizable choice profile? The following proposition gives a partially positive answer.

Proposition 6 *Let* $\mathbb{M} = ((M, A), \leq)$ *be a CPD-model with* $\mathbb{M}, a' \vDash \neg D_X(-X)$ *for all* $a' \in A$ *satisfying* $a'_{\mathsf{dom}} = \{X, -X\}$. *Then for all* $a \in A$ *and non-empty subset* $X \subseteq \mathsf{V}$, *the following two are equivalent:*

1. $X \in a_{\mathsf{dom}}$;
2. $\mathbb{M}, a \vDash \bigwedge_{i \in X} D_i X \wedge \bigwedge_{j \notin X} \neg D_X j$.

Proof. **From 1 to 2.**

Assume $X \in a_{\mathsf{dom}}$. Suppose $a' \in A$ and $a =_i a'$ for some $i \in X$. Then $X \in a'_{\mathsf{dom}}$. Since $A \subseteq \Xi$, $a_i = a_j$ and $a'_i = a'_j$ for all $i, j \in X$. Note that $a =_i a'$ for some $i \in X$, we see $a_j = a_i = a'_i = a'_j$ for all $j \in X$, i.e. $a =_X a'$. Thus $\mathbb{M}, a \vDash D_i X$. By the arbitrariness of $i \in X$, we see $\mathbb{M}, a \vDash \bigwedge_{i \in X} D_i X$.

When $X = \mathsf{V}$, we see that $\bigwedge_{j \notin X} \neg D_X j$ is $\top$ and $\mathbb{M}, a \vDash \bigwedge_{j \notin X} \neg D_X j$. Suppose $X \neq \mathsf{V}$. Take an arbitrary $j \notin X$. Then we have the following cases:

- $a_{\mathsf{dom}} \neq \{X, -X\}$. Let $\pi = \{X, -X\}$. Note that $\sigma_A(a_{\mathsf{dom}}) \subseteq \sigma_A(\pi)$, there must be $b \in A$ such that $b_{\mathsf{dom}} = \pi$ and $a_{\mathsf{merge}} = b_{\mathsf{merge}}$. Then it must be the case that $\mathsf{dom}(b_j) = -X \neq \mathsf{dom}(a_j)$ and so $a \neq_j b$.
- $a_{\mathsf{dom}} = \{X, -X\}$. Since $\mathbb{M}, a \vDash \neg D_X(-X)$, there must be $b \in A$ such that $a =_X b$ and $a \neq_{-X} b$. If $a_{\mathsf{dom}} \neq b_{\mathsf{dom}}$, then $\mathsf{dom}(a_j) = -X \neq \mathsf{dom}(b_j)$ and so $a \neq_j b$. Suppose $a_{\mathsf{dom}} = b_{\mathsf{dom}}$. Then a_k are all the same for $k \in -X$ and b_h are all the same for $h \in -X$. Since $a \neq_{-X} b$, we see $a_j \neq b_j$.

Hence $\mathbb{M}, a \vDash \neg D_X j$. By the arbitrariness of j, we see $\mathbb{M}, a \vDash \bigwedge_{j \notin X} \neg D_X j$.

From 2 to 1.

Assume that $X \notin a_{\mathsf{dom}}$ and $\mathbb{M}, a \vDash \bigwedge_{i \in X} D_i X \wedge \bigwedge_{j \notin X} \neg D_X j$. Let $x \in X$.

- $X \subsetneq \mathsf{dom}(a_x)$. Then there is $j \in \mathsf{dom}(a_x) \backslash X$ such that $a_j = a_i$ for all $i \in \mathsf{dom}(a_x)$. So for all $a' =_X a$, $a'_j = a_i = a'_i$ for all $i \in X$. Then we have $\mathbb{M}, a \vDash D_X j$ where $j \notin X$. Contradiction!
- Otherwise, there is $j \in X \setminus \mathsf{dom}(a_x)$. Since $\mathbb{M}, a \vDash D_x X$, we see $\mathbb{M}, a \vDash D_{\mathsf{dom}(a_x)} j$. By the direction we have proved above, $\mathbb{M}, a \vDash \neg D_{\mathsf{dom}(a_x)} j$, which is a contradiction.

player's possible choices of coalitions and how the coalition's worth should be allocated to keep the coalition stable. It abstracts away how each player acts in a coalition, which is different from CPD-models. The explicit representation of each player's choices in CPD-models is key to our later unified treatment of Nash equilibrium and the core.

□

The assumption of the above proposition that $\mathbb{M}, a \vDash \neg D_X(-X)$ for all $a \in A$ satisfying $a'_{\mathsf{dom}} = \{X, -X\}$ requires that no coalition can completely decides what its complementary coalition chooses to do. If X can completely control what $-X$ chooses, then the division of X and $-X$ is senseless, because $\mathbb{M}, a \vDash D_X\bigvee$ follows from $\mathbb{M}, a \vDash D_X(-X)$. As the readers can verify, the CPD-model in Example 6 does not satisfy the assumption at $a'', a^{4\prime}, a^{6\prime}$.

To avoid vacuous coalitions division, we will work with the CPD-models with the above assumption.

Definition 30 (Real CPD-models) *A real CPD-model (RCPD-model)* $\mathbb{M}$ *is a CPD-model that satisfies the assumption that* $\mathbb{M}, a \vDash \neg D_X(-X)$ *for all* $a \in A$ *satisfying* $a_{\mathsf{dom}} = \{X, -X\}$.

In an RCPD-model, $\bigwedge_{i\in X} D_i X \wedge \bigwedge_{j\notin X} \neg D_X j$ expresses that X is in the coalition partition. We will use the abbreviation

$$p_X := \bigwedge_{i\in X} D_i X \wedge \bigwedge_{j\notin X} \neg D_X j$$

for convenience.

5.5.2 The core in HLPFD

Having set up the LPFD framework for representing cooperative games in strategic and coalitional form, in this part, we show that the core, an important solutions concept in the cooperative game theory, can be expressed in HLPFD. Moreover, by considering functional dependence explicitly, we generalize the core and show how it is related to Nash equilibrium and Pareto optimality.

Just as Nash equilibrium in non-cooperative games captures stability of a strategy profile, the concept of the core, as a basic solution concept in cooperative games, also captures stability of a strategy profile in cooperative games. The difference is that the core takes the stability of a coalition into consideration. There are other notions for characterizing stability in cooperative games, for example, stable set, bargaining set and so on. We focus on the core.

The concept of the core is formulated in CPD-models as follows. ①

Definition 31 (Core in CPD-Model) *Given a CPD-model* $\mathbb{M}$*, a choice profile* $a \in A$ *is*

① The definition of the core can vary in different settings. Our definition is based on (Conzalez et al. (2021), Definition 2.2), which is a relatively general version.

in the core of $\mathbb{M}$ *if and only if*

1. $a_{\mathsf{dom}} = \{\mathsf{V}\}$*; and*
2. *there is no* $X \subseteq \mathsf{V}$ *and* $a' \in A$ *such that*
   (a) $X \in a'_{\mathsf{dom}}$*; and*
   (b) *for all* $a'' =_X a'$ *and all* $i \in X$, $a \prec_i a''$.

Let $Co_{\mathbb{M}}$ *denote the core of* $\mathbb{M}$.

If the set of the players V arrives at a choice profile a, which is in the core, then no $X \subset \mathsf{V}$ has any incentive to deviate from the coalition V, because forming the coalition X cannot guarantee all players in X end up with a better outcome. Coalitional power plays a key role in the basic idea of the core, because whether X has any incentive to deviate depends on whether X as a coalition can force a choice profile that all of its members prefer to the current choice profile.

Note that according to the definition of the core, if $X = \mathsf{V}$, there is no other choice profile with the coalition partition $\{\mathsf{V}\}$ which is strictly preferred by every player in V. Namely, a is weakly Pareto optimal among the choice profiles with the coalition partition $\{\mathsf{V}\}$. In fact, the following proposition holds.

Proposition 7 *Given a CPD-model* $\mathbb{M}$*, if a choice profile* $a \in A$ *is in the core of* $\mathbb{M}$ *then* a *is weakly Pareto optimal.*

Proof. Since $\mathbb{M}$ satisfies the condition that $\sigma_A(\pi) \subseteq \sigma_A(\{\mathsf{V}\})$ for all $\pi \in \Pi(\mathsf{V})$, by the fourth condition of Definition 29, the weak Pareto optimality of a within the choice profiles having $\{\mathsf{V}\}$ as their coalition partition can be generalized trivially to all choice profiles. □

The following example illustrates the concept of core and how it differs from Nash equilibrium and Pareto optimality.

Example 7 *Let* $\mathsf{V} = \{1, 2\}$ *and* $\Sigma = \{\alpha, \beta\}$*. A and the preference relations are given in Table 5.4. The preference relations are given in the form of a pair of ordinal utilities where the first element is for player 1 and the second for player 2. Readers familiar with game theory can recognize that without the last four rows the table represents the prisoners' dilemma.* $a^{3\prime}$ *is a Nash equilibrium but a is not as in the original prisoners' dilemma. Now our coalitional version allows player 1 and player 2 to form a coalition by whatever means, for example, a binding agreement or switching to the mode of team reasoning simultaneously. So there are four extra profiles in which both players explicitly choose to join the coalition. Among these four extra profiles, although all of them are trivially Nash*

	1	2	Ordinal Utility
a	$\{(1,\alpha)\}$	$\{(2,\alpha)\}$	(9,9)
a'	$\{(1,\alpha)\}$	$\{(2,\beta)\}$	(0,10)
a''	$\{(1,\beta)\}$	$\{(2,\alpha)\}$	(10,0)
$a^{3\prime}$	$\{(1,\beta)\}$	$\{(2,\beta)\}$	(1,1)
$a^{4\prime}$	$\{(1,\alpha),(2,\alpha)\}$	$\{(1,\alpha),(2,\alpha)\}$	(9,9)
$a^{5\prime}$	$\{(1,\alpha),(2,\beta)\}$	$\{(1,\alpha),(2,\beta)\}$	(0,10)
$a^{6\prime}$	$\{(1,\beta),(2,\alpha)\}$	$\{(1,\beta),(2,\alpha)\}$	(10,0)
$a^{7\prime}$	$\{(1,\beta),(2,\beta)\}$	$\{(1,\beta),(2,\beta)\}$	(1,1)

Table 5.4 A in Example 7

equilibria, $a^{4\prime}$ is the only element in the core.

Note that in the example $\{1,2\}$ as a coalition does not expand what each of the players can choose, namely $\sigma(\{1,2\}) = \sigma(\{\{1\},\{2\}\})$. But the coalition still makes some difference. Each member of the coalition anchors their actions to the coalition, which may bring extra stability.

Next, we show that the core can be expressed in HLPFD with respect to the class of RCPD-models (Definition 30) with nominals.

Proposition 8 *Given a RCPD-model $\mathbb{M}$ with nominals Nom, the current choice profile a with name i, i.e., $a = I(i) \in A$, is in the core of $\mathbb{M}$, if and only if*

$$\mathbb{M}, a \vDash i \wedge p_{\mathsf{V}} \wedge \bigwedge_{\varnothing \neq X \subseteq \mathsf{V}} \mathsf{A}(p_X \to \langle\!\langle X, \varnothing, \varnothing \rangle\!\rangle \bigvee_{x \in X} \langle\!\langle \varnothing, \{x\}, \varnothing \rangle\!\rangle i).$$

In the above HLPFD formula, p_{V} specifies the first condition of the core. The second condition is specified by the big conjunction, which says that for any subgroup X of V, no matter what X as a coalition chooses to do, it cannot guarantee that everyone in X end up with a better outcome. In other words, given the choice of X, there is always a possibility where someone in X would not become better than he does currently as a member of the coalition V in the choice profile i.

To generalize the concept of the core, we can have its relativized version as in the case of Nash equilibrium and Pareto optimality.

$$\mathsf{Core}_X i := i \wedge p_X \wedge \bigwedge_{\varnothing \neq C \subseteq X} \mathbb{D}_{-X}(p_C \to \langle\!\langle -X \cup C, \varnothing, \varnothing \rangle\!\rangle \bigvee_{c \in C} \langle\!\langle -X, \{c\}, \varnothing \rangle\!\rangle i)$$

Note that when taking $X = \mathsf{V}$, we get the original definition of the core as expressed in Proposition 8. The relativized version of the core enables us to express some interesting

relationships between coalitions. For example,

$$\mathsf{Core}_X i \wedge \mathsf{Core}_{-X} i$$

which says that in the current choice profile i, both X and $-X$ form coalitions and are in their relativized cores.

More generally, we can define the following concept:

$$\mathsf{Core}_\pi i := \bigwedge_{X \in \pi} \mathsf{Core}_X i$$

where π is a partition of N. It characterizes the stability of a collection of coalitions at a choice profile i. The core is a special case of it where $\pi = \{\mathsf{V}\}$. Moreover, Nash equilibrium $\mathsf{Na}\,\mathsf{V}$ is also a special case of it where $\pi = \{\{1\}, \{2\}, \ldots, \{n\}\}$.

Theorem 9 *Given a RCPD-model* $\mathbb{M}$ *with nominals* $i \in \mathsf{Nom}$*, and* $a \in A$ *with* $a_{\mathsf{dom}} = \pi = \{\{1\}, \{2\}, \ldots, \{n\}\}$*,*

$$\mathbb{M}, a \vDash \mathsf{Core}_\pi i \leftrightarrow (i \wedge \mathsf{Na}\,\mathsf{V}) \ .$$

As a corollary to this proposition, we see that unlike the core, $\mathsf{Core}_\pi i$ does not necessarily imply the weak Pareto optimality of i for V. But the following generalization of Proposition 7 holds.

Theorem 10 *Given a RCPD-model* $\mathbb{M}$ *with nominals* $i \in \mathsf{Nom}$ *and* $a \in A$*,*

$$\mathbb{M}, a \vDash \mathsf{Core}_\pi i \rightarrow \bigwedge_{X \in \pi} \mathsf{wPa}\, X \ .$$

Therefore, in the sense of the above two theorems, our generalization of the core can be seen as a notion that unifies the core, Nash equilibrium and Pareto optimality. Moreover, it is worthwhile to emphasize that $\mathsf{Core}_\pi i$ is more than a simple combination of each coalition's stability, because each coalition's stability is actually depends on other coalitions' stability. The overall stability reflected in $\mathsf{Core}_\pi i$ lies in the interdependence of each coalition's stability.

5.6 Stability, Coalitional Power and Collective Agency

In this section, we show how the CPD-models can help clarify issues on collective agency and explore some philosophical implications from the game-theoretical perspective on collective agency in CPD-models.

As mentioned in previous chapters, philosophical discussions about collective agency have flourished in recent decades. Despite disagreements on the detailed definition of col-

lective agency, most theories share the idea that joint actions by a group with collective agency are more than simply a coordination or cooperation between its members (cf. Bratman (2014); Gilbert (2006); List et al. (2011); Searle (2010); Tollefsen (2002); Tuomela (2013)). Nevertheless, the conundrum is where does this essential difference lie. Gilbert (2006), Searle (2010), and Tuomela (2013) admit an irreducible concept of a collective in a methodological sense. In Gilbert, it is a unique type of commitment of will: "joint commitment"; in Searle, it is a special kind of intention: "we-intention"; and in Tuomela, it is a complex of "we-intention" and a particular form of attitude: "we-mode." They all try to start from an irreducible concept of a collective to capture the extras of joint actions. In a similar sense, List et al. (2011) emphasize that a group with agency must have a procedure to ensure that its decision-making process meets the necessary functional conditions of an agent, such as manifest rationality; Tollefsen (2002) highlights that a group with agency must contain stable structures in order to conform to the basic phenomena that can be reasonably explained by the observer. Even for Bratman (2014), who famously argues that we can explain collective agency without any irreducible concept of a collective, he still claims the critical role of the interdependent relations and the mesh of individual plans between members in forming collective intentions.

As Bratman emphasized, we also pay particular attention to the critical role of interdependence in forming a collective agent. Following basic abstraction and restriction in game theory, we focus on the non-psychological aspects that make joint action collective. We only presuppose individual preferences and choices as the starting point without assuming any initial concept of a collective or its intention. This approach abstracts individual intentionality, but it also draws boundaries for our formal analysis. On the one hand, philosophical or psychological discussions are helpful only in questioning the sources of individual preferences and applying formal analysis to specific situations. On the other hand, by dividing this borderline, our analysis covers only facts, i.e., individual preferences and interdependency between them as the base facts and the collective decisions derived from them. In this way, the mystical and spiritual elements in the collective context are removed, thus ensuring the objectivity of formal analysis. Along this way we explore the stability of interdependence as embodied in the core.

In CPD-models, coalitions are taken explicitly as a part of each individual player's choices. That is, each player chooses which coalitions to join. This makes it possible to distinguish between a group action and a set of individual actions. In Example 7, although

$a^{4\prime}$ and a have the same strategy profile, namely $a^{4\prime}_{\mathsf{merge}} = a_{\mathsf{merge}}$, acting together ($a^{4\prime}_{\mathsf{dom}} = \{\{1,2\}\}$) or acting individually ($a_{\mathsf{dom}} = \{\{1\},\{2\}\}$) make two totally different choice profiles. However, condition 4 in Definition 29 stipulates that once their strategy profile keeps the same, no players would prefer one to the other. This means that $a^{4\prime}$ and a make no difference to both players' preference relations. Then what can the difference between $a^{4\prime}$ and a bring about to the players? The critical observation is that $a^{4\prime}$ is in the core while a is not even a Nash equilibrium. That is, although their strategy profiles are the same, one as a joint action of the group is stable (in the sense of the core) while the other as cooperation of two parties is not stable (in the sense of the Nash equilibrium). This suggests that the stability of acting together is essential for understanding collective agency.

To elaborate on the above claim, we first make the following clarification about condition 3 in Definition 29. It does *not* require that by choosing to join the same coalition together, the players in the coalition should have *more* strategies than they have when acting separately, but only *no less than*. We leave it open whether it is a strict inclusion ($\subset$) or an equation ($=$). As we can see in Example 7, the coalition $\{1,2\}$ does not have extra strategies. The special status of $a^{4\prime}$ does not rely on having a strategy profile that cannot be realized by the players separately. That is to say, we do not need to presuppose redundant action profiles to describe collective agency. Coalitions are those stable and undeviated states of choice, which are part of all the possible states of choice.

Moreover, we may further clarify our viewpoint by reinterpreting Example 7. Suppose players 1 and 2 are now a couple, facing a drastic change in their life: bitter divorce ①. Each chooses to cooperate when they are a couple, and we can denote those joint actions as $a^{4\prime}$ - $a^{7\prime}$. Among which all are trivially equilibrium and only $a^{4\prime}$ is in the core. We could say they were constantly living in the stable state $a^{4\prime}$ without any incentive to change. However, now they get divorced. Although nothing happens to the utility of each strategies, since they no longer want to choose act jointly, the possible action profiles now have only a - $a^{3\prime}$. And then, the current state switch from $a^{4\prime}$ to a, since their strategies and ordinal utility are the same. Furthermore, in a non-cooperative context, state a is only Pareto optimal but not a Nash equilibrium, which means such a state is not stable, and they can only swith to $a^{3\prime}$ as the only Nash equilibrium in the non-cooperative context. Therefore we have a concrete example that given ceteris paribus, only because players

① We appreciate the examples provided by the pre-Defense Committee.

do not want to cooperate, their stable action results suffer a drastic change and slide towards another different ending. This example combines the concepts of cooperative and non-cooperative games to illustrate the role of collective agency visually. At the same time, this also breaks the cooperative-competitive classification of game theory, and again demonstrates our philosophical view that collective agency as a set of relational eventualities can be generated, changed and vanished, and that the change of the cooperative to non-cooperative context is an instantiation.

As the above discussion implies, RCPD-models enable us to distinguish the following three levels of group actions. First, coalitional effectiveness, that is, what an agglomeration of a group of players' actions can force. We can use $\mathbb{D}_X\varphi$ in our logic to express that a group of agent X can force φ to be the case if they act according to their current choices. Second, collective effectiveness, expressed by $p_X \wedge \mathbb{D}_X\varphi$. Different from coalitional effectiveness, collective effectiveness requires that each player in X explicitly chooses to join coalition X and act as its member. Third, the core effectiveness, $\mathsf{Core}_X i \wedge \mathbb{D}_X\varphi$. Compared with collective effectiveness, it requires not only that each player of X acts as a member of coalition X but also that the collective action should be sustainable or stable given what the players in $-X$ choose to do. We contend that collective agency would emerge at no lower level than the core effectiveness.

It reveals that collective agency should be a binding power that makes a coalition and its joint action stable. This binding power may come from different sources and be present in different forms over which various theories on collective agency debate. No matter which source it comes from and which form it takes, the binding power should come with the stability of what it binds together. Regarding stability, we share the same spirit with Gold et al. (2007); Sugden (2003); Tollefsen (2002); Tuomela (2013), in which they also directly or indirectly take stability as a condition for the formation of a collective agent. Moreover, suppose we further abstractly understand the concept of the core as a specific pattern for inter-sub-coalition relations within a coalition. In that case, our interpretation highlights the understanding of collective agency as a relatively stable state of relations rather than an imagined conceptual entity. In this sense, we are in line with the call for a relationalist account (cf. Baier (1997); Meijers (2003); Schmid (2003)).

Game theory, especially the cooperative game theory, is a powerful tool for analyzing the kind of stability we consider essential for collective agency. The concept of the core is not the only solution concept in cooperative game theory. A lot of other solution

concepts have been proposed, taking different issues related to coalitional stability into consideration. Abstracting and logically fusing these concepts into a unified framework will bring more insights into the philosophical discussion of collective agency. Our analysis by CPD-models serves as a first attempt to make this connection explicit by testing collective agency in games.

5.7 Related Works and Summary

Before summary, we compare our work with three closely related works, the colitional logic in Pauly (2002), the modal coalitional game logic (MCGL) in Ågotnes et al. (2009)[①] and the logic of ceteris paribus preference (LCP) in van Benthem et al. (2009).

5.7.1 Comparison with the coalition Logic

The coalitional effectiveness that the coalition logic aims to reason about is formally characterized by an effectivity function E_G. Based on this effectivity function, the main operator of the coalition logic $[C]\varphi$ is defined, expressing that the set of agents C can force φ to be the case at their current state.

The effective function, when adapted in a dependence model $\mathbf{M} = (M, A)$, can be defined as $E_{\mathbf{M}} : \mathcal{P}^{<\aleph_0}(\mathsf{V}) \to \mathcal{P}(\mathcal{P}(A))$ satisfying

$$S \in E_{\mathbf{M}}(X) \text{ iff } \exists a \in A, \forall a' \in A \text{ if } a' =_X a \text{ then } a' \in S \ .$$

Here, $S \in E_{\mathbf{M}}(X)$ means that the coalition X can force the game to be in S. We can express $S \in E_{\mathbf{M}}(X)$ in LFD as $\mathsf{E}\mathbb{D}_X\varphi$ assuming that $S = [\![\varphi]\!]$, because

$$\mathbf{M} \vDash \mathsf{E}\mathbb{D}_X\varphi \text{ iff } [\![\varphi]\!] \in E_{\mathbf{M}}(X) \ .$$

The operator $[C]\varphi$ in the coalition logic essentially has the same semantic meaning despite being interpreted in the neighborhood semantics.

We will not go into a detailed comparison between LFD and the coalition logic, but only point out a substantial difference between $\mathsf{E}\mathbb{D}_X\varphi$ and $[C]\varphi$ with regard to the characteristic axiom of the coalition logic, superadditivity:

$$([C_1]\varphi_1 \wedge [C_2]\varphi_2) \to [C_1 \cup C_2](\varphi_1 \wedge \varphi_2) \text{ where } C_1 \cap C_2 = \varnothing \ .$$

Superadditivity fails for $\mathsf{E}\mathbb{D}_X\varphi$, because in a dependence model it is possible that there is

① There are two logics in Ågotnes et al. (2009). MCGL is the second one. The first one is more customized and limited than the second one. For example, it only considers finite games where both players and states need to be finite.

$a, a' \in A$ such that, for $C_1 \cap C_2 = \varnothing$, there is no $a'' \in A$ satisfying both $a =_{C_1} a''$ and $a' =_{C_2} a''$.

As the readers who are familiar with the coalition logic can verify, except for super-additivity, its other axioms are all valid for $\mathsf{E}\mathbb{D}_X$ in LFD.

5.7.2 Comparison with LCP and MCGL

Both LCP and MCGL use modal operators for characterizing preorders. Given a preorder $\preceq$ in its semantic model, LCP only includes one modal operator for $\preceq$ and one for $\prec$. MCGL concerns a multi-agent setting where for each agent there is a preorder. Besides modal operators for individual agents, MCGL includes group operators, one for the intersections of a set of preorders and one for the intersection of a set of strict preorders. It also includes modal operators for the inverse of the preorders and a difference operator. Nevertheless, it does not have any operator for the intersection of strict and non-strict preorders. Our logic has such operators and we show that they are critical for expressing strong Pareto optimality.

Next, with each of these two other logics, the comparison will focus on different aspects.

Comparison with Ågotnes et al. (2009) on different formulations of the core

It is shown in Ågotnes et al. (2009) that MCGL can express not only the core in coalitional games but also the stable set and the bargaining set. However, the setting they adopt for representing coalitional games is not general enough to model the coalitional games formalized by the CPD-models. The limitation is due to their way of defining the cooperative effective function or the characteristic function as they call it. In a CPD-model $\mathbb{M}$, their characteristic function can be understood as $f : 2^N \setminus \{\varnothing\} \to \mathcal{P}(A)$, a function assigning a set of choice profiles to each coalition. Their formulation of the core only requires that the current choice profiles are strictly preferred to all the choice profiles in $f(X)$ for all $X \subseteq \mathsf{V}$. But in our formulation of the core in Definition 31, what matters is the following set for each $X \subseteq \mathsf{V}$

$$E(X) := \{a(X) \subseteq A \mid a \in A \text{ and } X \in a_{\mathsf{dom}}\}$$

where $a(X) := \{a' \in A \mid a =_X a'\}$. $E : 2^N \to \mathcal{P}(\mathcal{P}(A))$ is a function assigning to each coalition a set of sets of choices profiles. Our formulation of the core requires a comparison between the current choice profile and each of the set in $E(X)$. Note that the compartmentalization of what a coalition X can force is essential for our formulation

of the core, because what a coalition X can enforce depends on what X would do. This subtlety is not captured by the characteristic function in Ågotnes et al. (2009).

Comparison with van Benthem et al. (2009) on different ways of characterizing dependence

We have seen that in LPFD variables are taken to partition the space of possible assignments according to their possible values. The dependence relation is the relation between different partitions. In LCP, what partitions the space of possible states are all possible sets of formulas of its base language. If we think of a formula as a binary variable with its values 0 or 1, then the operators $[\Gamma]$, $[\Gamma]^{\preceq_x}$ and $[\Gamma]^{\prec_x}$ in LCP correspond to our operators $[\![\Gamma, \varnothing, \varnothing]\!]$, $[\![\Gamma, x, \varnothing]\!]$ and $[\![\Gamma, \varnothing, x]\!]$ respectively. This raises an interesting question: if we only allow binary variables, what is the difference between using variables (as in LFD) and formulas (as in LCP) to capture the functional dependence between variables? Furthermore, do we really lose anything in LFD if we only allow binary variables? A systematic study of these two questions would require future work.

5.7.3 Summary and ideas for a future investigation

We have proposed two logics by extending LFD and studying their axiomatizations and other properties. We have also demonstrated how our logics can help reason about the notions of dependence, preference and coalitional power in a game theoretical setting and provide a unified view on three key concepts in game theory, i.e., Nash equilibrium, Pareto optimality and the core. On the basis of the two logics, we bring insights to the general discussion on collective agency.

The core effectiveness we have shown in subsection 5.6 highlights the stability of both the coalition and its collective action, which is consistent with relational account of collective agency. More work on collective agency from a cooperative-game-theoretical perspective needs to be done as we have instigated.

The connection between LFD and the coalition logic we have revealed indicates that it may be fruitful to explore the relationship between LPFD and ATL (cf. Goranko et al. (2004)). Some work has been done on exploring the temporal dimension of dependence (cf. Baltag et al. (2022)). Further work in these directions could make a logical analysis of extensive games more full-fledged.

CHAPTER 6 CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusions

We have reviewed several mainstream accounts of collective agency and collective intentionality, and we have pointed out that they all manifest the tendency of physicalistic individualism resulting from long-term Cartesian influence. We have shown that the foundation of such physicalistic individualism is fragile and that theories that subscribe to it cannot adequately explain the source of irreducible concepts of a collective. We then have developed a relational account, which regards the subject of collective intentionality as the relational identity composed of all the relative relational eventualities. An individual identifies and acts in a relational identity of a collective through their social relations, thus manifesting collective agency. Collective agency has both internal relations (collective structures) and external relations (collective functions). This confirms that interpretationism and functionalism about agency can be integrated into a relational, holistic account.

Next, we pointed out that the irreducibility of the concept of collective has a fundamental incompatibility with the concept of intentionality. This incompatibility forces us to review the standard criterion of intentionality and gradually screen out which one is incompatible with the collective. We found that it is ambiguous to insist on a mental constraint. In this way, we study the relationship between intentionality and disposition, and bring both individual and collective intentionality into a unified disposition-based account, which can explain their similarity naturally. With this account, we demonstrate the existence of collective intentionality by analyzing different forms of attributive judgments and by introducing the indispensable collective responsibility.

On the basis of ontological relationalism of agency and dispositional account of intentionality, we have established a unified foundation for discussing individual and collective. Upon which we are able to explore a broader range of topics; for example, linking philosophy of action or social ontology with general social sciences such as game theory or social choice theory, or with general social philosophy or phenomenological explanations of intentionality, and comparing the similarities and differences between their underlying concepts of concern. In a unified framework, these comparisons can become more explicit and facilitate a macro understanding of the various parallel studies on collective.

Through the dispositional account, we can view the relationship between an individual and the world and the one between people in a more unified way.

Then we connect philosophical theories of intentionality with formal theories of preferences by discussing three concepts of intentionality, preference, and dependencies, pointing out that philosophical theories can provide a solid basis for formal research. The game theory we cover here is limited to non-cooperative games. In fact, cooperative games, repeated games, and evolutionary games can all be related to the philosophy of intentionality to carry out meaningful discussions. In future work, we will propose a new logic system and express the meaning of collective agency in the context of non-cooperative games and cooperative games.

Regarding the logical part, we have proposed two logics by extending LFD and studying their axiomatizations and other properties. We have also demonstrated how our logics can help reason about the notions of dependence, preference, and coalitional power in a game theoretical setting and provide a unified view on three key concepts in game theory, i.e., Nash equilibrium, Pareto optimality, and the core. On the basis of the two logics, we bring novel insights to the general discussion on collective agency, where we consider agency of a collective as a stable state that is constituted by each member's preference and the interdependency between them. Such a state is stabilized by the binding power of a collective, where each member is stable in her current choice and has no incentive to deviate from their joint action. Through the provided unified perspective, we can clearly understand the connotation of this binding power, and we claim that it is minimally an essential feature for collective agency.

6.2 Future Directions

The Position of our overall study is between normative and descriptive. Compared to the previous philosophical theories, our way of interpreting is relationalist and dispositionalist. Compared to previous logical systems, our way of formalizing highlights the interactions inside a group. Although we have been trying to discuss the topic in an argumentative rather than a position way, many position premises inevitably accompany the process. The relaxation of these position premises will open up a wealth of avenues of inquiry, some of the most promising of which will be listed here.

Further problems regarding collective agency

In general, our account holds that the rationality of a collective agent can be guar-

anteed by its individual members' rationality and the stable structure of the collective. However, the collective agent still runs the risk of losing its rationality. These risks arise from the two aspects mentioned above: individual rationality and collective structure.

People are not always rational; instead, our decisions also always come from unconscious processes, feelings, emotions, peer pressures, etc. So far, we have only mentioned the rational part of an individual since our purpose is to explain collective agency theoretically. Nevertheless, we cannot neglect the irrational aspect of human beings (cf. Barrett (1962)). Individuals' irrationality hampers almost every essential concept in our framework: agent, agency, collective agency, equilibrium, optimality, the core, etc., since all of these concepts are discussed in an ideal context. Therefore, once irrationality is concerned, the so-called collective agent may act irrationally. For instance, a dictatorship collective may act extremely irrationally due to the irrationality of the dictator and thus can no longer qualify as an agent. Therefore, relaxing the requirement of individual rationality and further integrating individual irrationality into our explanatory framework to study the interaction between irrationality and stable collective structure is a promising way of research.

On the other hand, even if the assumption of individual rationality is not relaxed, the stable collective structure will still be accompanied by some not-so-good phenomena; a prominent example is free riding. Different members have different individual rights transferred for the collective optimal. Extreme situations may occur: some people sacrifice their own rights for the collective, and some people can enjoy the convenience of the collective without transferring any personal interests. Within the framework of individual rationality and collective structure, how to further improve the collective setting to cope with these phenomena is also a promising study.

As relaxing individual rationality has been mentioned above, there are also research directions based on relaxing collective structure. All the discussions in this thesis are based on the assumption that a collective has structural characteristics. Structures or functions of a collective guarantee minimum relational eventualities that constitute the relational identity of a collective. However, groups are not always structural in reality; for example, people who suddenly run together due to an alarm; who happen to pass by and offer assistance; and who walk together for no reason; etc. Various social phenomena need to be explained. For groups with opaque structures that do have constituted identities of collectives, it may be appropriate to further talk about particular groups' pre-structural

or quasi-structural characteristics on a looser basis.

Phenomenological Interpretation

There is one kind of theorizing in the area of collective agency and collective intentionality that we have not said anything about in the chapter so far, which is that of phenomenological interpretation. In addition, the trend of phenomenological interpretation has gradually become notable. Starting from Husserl, many phenomenological interpretations of intentionality have been developed, and some of them have been integrated into contemporary accounts of collective intentionality. For example, Schmid (2014) makes use of a phenomenological concept: pre-reflexive self-awareness can be applied to multi-agent situations. According to Schmid, an agent cannot engage in reflexive reflection without being aware of their own identity. Only based on pre-reflexive awareness can we consider our options for action and make decisions in our own situation without doubting our actions and ourselves. Therefore, when it applies to the collective level, collective agency will be considered as an adoption of pre-reflexive self-awareness of the collective identity. An investigation of the relationship between this view and our relational account is, unfortunately, beyond the scope of this chapter. Nevertheless, there is a possible alternative explanation of collective agency: the individual agent themselves are already able to recognize others as agents before their own pre-reflexive awareness (see Wittgenstein (2010), a view that, arguably, finds support in cognitive psychological and anthropological research, as again in Tomasello et al. (2005)). Comparing these two viewpoints and further developing the relational account seems a promising direction for further research.

Supervenience

Supervenience is a core concept in functionalism and has also been applied to the study of collective agency (cf. List et al. (2011)). Formally, supervenience refers to that fixing one set of facts A will necessarily lead to the fix of another set of facts B, and we call it the set of facts B supervenes on the set of facts A. List et al. (2011) further holds that the attitudes and actions of a group agent supervene on the contributions of its members. Thus in the framework of social choice theory, we can understand different aggregate functions of individual preferences as embodiments of such supervenience.

Although we have not said anything about such a concept, it is deeply related to at least two directions discussed in this thesis. One is the relational account of agency, according to which a collective in the ontological sense is the relational pattern that constitutes it. Then, the concept of supervenience can be further applied to the analysis between the

relational pattern and the properties that supervene on it. Another is the logic of functional dependence, where the formal understanding of dependency and that of supervenience are exactly the same. How to further explore the differences between these two concepts is a promising direction. One possible idea is to use typological thinking to limit the applied scope of dependency and supervenience.

Open formal problems

The connection between LFD and the coalition logic we have revealed indicates that exploring the relationship between LPFD and ATL (cf. Goranko et al. (2004)) may be fruitful. Some work has been done on exploring the temporal dimension of dependence (cf. Baltag et al. (2022)). Further work in these directions could make a logical analysis of extensive games more full-fledged.

There have been other studies linking intentionality to game theory. For instance, Roy (2008) detailed studies on the key role of intention in coordination games, and proposes an extension of epistemic logic to characterize the concept of intention formally. In contrast, this thesis starts from the philosophical relationalism of collective agency, and thus pays more attention to the interdependence relations of groups by extending the logic of functional dependence. Our work can complement the epistemic approach of collective intentionality research.

This thesis only uses logic to discuss the connection between philosophy and game theory, especially single non-cooperative and cooperative games. Besides, many similar combined studies can be carried out, such as combining agency research with social choice theory, repeated game theory, evolutionary game theory and other directions. Moreover, using logical language to reveal the correlation behind the significant concepts in these fields is also insightful.

REFERENCES

Ågotnes T, van der Hoek W, Wooldridge M, 2009. Reasoning about coalitional games[J/OL]. Artificial Intelligence, 173(1): 45-79. DOI: https://doi.org/10.1016/j.artint.2008.08.004.

Ågotnes T, van der Hoek W, Wooldridge M, 2011. On the logic of preference and judgment aggregation [J]. Autonomous Agents and Multi-Agent Systems, 22(1): 4-30.

Alur R, Henzinger T, Kupferman O, 2002. Alternating-time temporal logic[J/OL]. Journal of the ACM, 49: 672-713. DOI: 10.1145/585265.585270.

Anscombe G E M, 1957. Intention[M]. Oxford: Blackwell.

Anscombe G E M, 1965. The intentionality of sensation: A grammatical feature[M]//Butler R J. Analytic Philosophy. Oxford: Blackwell: 158-80.

Arrow K J, 1951. Social choice and individual values[M]. New York: Wiley.

Audi R, 1994. Dispositional beliefs and dispositions to believe[J/OL]. Noûs, 28(4): 419-434. http://www.jstor.org/stable/2215473.

Audi R, 2019. Faith, belief, and will: Toward a volitional stance theory of faith[J/OL]. Sophia, 58(3): 409-422. DOI: 10.1007/s11841-018-0653-x.

Austen J, 2004. Sense and sensibility[M]. Oxford: Oxford University Press.

Bach E, 1986. The algebra of events[J]. Linguistics and Philosophy, 9(1): 5-16.

Bacharach M, 1999. Interactive team reasoning: A contribution to the theory of co-operation[J/OL]. Research in Economics, 53: 117-147. DOI: 10.1006/reec.1999.0188.

Bacharach M, Gold N, Sugden R, 2006. Beyond individual choice : teams and frames in game theory [M]. Princeton: Princeton University Press.

Baier A C, 1997. Doing things with others: The mental commons[M]//Alanen L, Heinämaa S, Wallgren T. Commonality and Particularity in Ethics. London: Palgrave Macmillan: 15-44.

Bakeman R, Adamson L B, 1984. Coordinating attention to people and objects in mother-infant and peer-infant interaction[J]. Child Development: 1278-1289.

Baltag A, Moss L S, 2004. Logics for epistemic programs[J]. Synthese, 139(2): 165-224.

Baltag A, van Benthem J, 2021. A simple logic of functional dependence[J/OL]. Journal of Philosophical Logic, 50: 939-1005. DOI: https://doi.org/10.1007/s10992-020-09588-z.

Baltag A, van Benthem J, Li D, 2022. A Logical Analysis of Dynamic Dependence[A]. arXiv preprint: 2204.07839.

Barrett W, 1962. Irrational man: A study in existential philosophy: volume 321[M]. NYC: Anchor.

Bauer W A, 2016. Physical intentionality, extrinsicness, and the direction of causation[J/OL]. Acta Analytica, 31(4): 397-417. DOI: 10.1007/s12136-016-0283-2.

Belnap N, Perloff M, 1990. Seeing to it that: A canonical form for agentives[M/OL]. Dordrecht: Springer Netherlands: 167-190. https://doi.org/10.1007/978-94-009-0553-5_7.

Binmore K G, 1994. Game theory and the social contract: Playing fair[M/OL]. Cambridge: MIT Press. https://books.google.com.hk/books?id=8cDiGo2REBIC.

Bird A, 2007. Nature's metaphysics: Laws and properties[M]. Oxford: Oxford University Press.

Blackburn P, de Rijke M, Venema Y, 2001. Modal logic[M]. Cambridge Tracts in Theoretical Computer Science. Cambridge: Cambridge University Press.

Bourget D, 2010. Consciousness is underived intentionality[J/OL]. Noûs, 44(1): 32-58. DOI: https://doi.org/10.1111/j.1468-0068.2009.00730.x.

Bourget D, Mendelovici A, 2019. Phenomenal Intentionality[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Fall 2019 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/phenomenal-intentionality/.

Bratman M E, 1987. Intention, plans, and practical reason[M]. Cambridge: Cambridge University Press.

Bratman M E, 1999. Faces of intention: Selected essays on intention and agency[M]. Cambridge: Cambridge University Press.

Bratman M E, 2006. Dynamics of sociality[J]. Midwest Studies in Philosophy, 30: 1-15.

Bratman M E, 2007. Structures of agency: Essays[M]. Oxford: Oxford University Press.

Bratman M E, 2014. Shared agency: A planning theory of acting together[M]. Oxford: Oxford University Press.

Bratman M E, 2018. Planning, time, and self-governance: Essays in practical rationality[M]. Oxford: Oxford University Press.

Bratman M E, Israel D J, Pollack M E, 1988. Plans and resource-bounded practical reasoning[J]. Computational intelligence, 4(3): 349-355.

Brentano F, 1874. Psychology from an empirical standpoint (psychologie vom empirischen standpunkt) [M]. London: Routledge and Kegan Paul.

Broersen J M, 2011. Making a start with the stit logic analysis of intentional action[J]. Journal of philosophical logic, 40(4): 499-530.

Broome J, 1991. Desire, belief and expectation[J]. Mind, 100(2): 265-267.

Byrne A, Hájek A, 1997. David hume, david lewis, and decision theory[J]. Mind, 106(423): 411-728.

Chatain O, 2016. Cooperative and non-cooperative game theory[M/OL]. London: Palgrave Macmillan: 1-3. https://doi.org/10.1057/978-1-349-94848-2_468-1.

Chisholm R M, 1957. Perceiving: A philosophical study[M]. Ithaca: Cornell University Press.

Choi S, Fara M, 2021. Dispositions[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Spring 2021 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/dispositions/.

Clark A, 2020. Beyond desire? agency, choice, and the predictive mind[J/OL]. Australasian Journal of Philosophy, 98(1): 1-15. https://doi.org/10.1080/00048402.2019.1602661.

Cohen P R, Levesque H J, 1990. Intention is choice with commitment[J]. Artificial intelligence, 42 (2-3): 213-261.

Collins S, 2019. Group duties: Their existence and their implications for individuals[M]. Oxford: Oxford University Press.

Conzalez S, Lardon A, 2021. Mathematical social science[J/OL]. Mathematical Social Sciences, 114: 28-38. DOI: 10.1061/j.mathsocsci.2021.09.001.

Cooper D E, 1968. Collective responsibility[J/OL]. Philosophy, 43(165): 258-268. DOI: 10.1017/S0031819100009220.

Copp D, 2007. The collective moral autonomy thesis[J]. Journal of Social Philosophy, 38(3): 369-388.

Crane T, 1998. Intentionality as the mark of the mental[J/OL]. Royal Institute of Philosophy Supplement, 43: 229–251. DOI: 10.1017/S1358246100004380.

Darwall S, 1996. The second person standpoint: Morality, respect, and accountability[M]. Cambridge: Harvard University Press.

Davidson D, 1963. Actions, reasons, and causes[J/OL]. The Journal of Philosophy, 60: 685. DOI: 10.2307/2023177.

Davidson D, 2001. Essays on actions and events: Philosophical essays volume 1[M]. Oxford: Clarendon Press.

Davis W A, 1982. A causal theory of enjoyment[J]. Mind, 91(362): 240-256.

de Condorcet M, 1785. Essay on the application of analysis to the probability of plurality decisions (essai sur l'application de l'analyse à la probabilité des décisions, rendues à la pluralité des voix)[M]. A Paris: De l'imprimerie Royale.

de Haan N, 2021. On the relation between collective responsibility and collective duties[J]. Philosophy, 91(1): 99-133.

Dennett D C, 1987. The intentional stance[M]. Cambridge: MIT press.

Dretske F, 1991. Explaining behavior: Reasons in a world of causes[M]. Cambridge: MIT press.

Endriss U, 2011. Logic and social choice theory[J]. Logic and philosophy today: 2:333-377.

Fagin R, Halpern J Y, Moses Y, et al., 1995. Reasoning about knowledge[M]. Cambridge: MIT Press.

French P A, 1984. Collective and corporate responsibility[M]. NYC: Columbia University Press.

Galliani P, 2021. Dependence Logic[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Summer 2021 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/logic-dependence/.

Galliers J R, 1989. A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict[D]. Open University (United Kingdom).

Gertler B, 2021. Self-Knowledge[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Winter 2021 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/self-knowledge/.

Gilbert M, 1992. On social facts[M]. Princeton: Princeton University Press.

Gilbert M, 2000. Sociality and responsibility: New essays in plural subject theory[M]. Washington DC: Rowman & Littlefield.

Gilbert M, 2002. Acting together[M]//Meggle G. Social Facts and Collective Intentionality. Philosophische Forschung. Frankfurt a. Main: Dr. Hänsel-Hohenhausen.

Gilbert M, 2005. Shared values, social unity, and liberty[J]. Public Affairs Quarterly, 19(1): 25-49.

Gilbert M, 2006. A theory of political obligation: Membership, commitment, and the bonds of society [M]. Oxford: Oxford University Press.

Gillies D B, 1959. Solutions to general non-zero-sum games[J]. Contributions to the Theory of Games, 4: 47-85.

Godfrey-Smith P, 2009. Abstractions, idealizations, and evolutionary biology[M]//Barberousse A, Morange M, Pradeu T. Mapping the future of biology. Dordrecht: Springer Netherlands: 47-56.

Gold N, Sugden R, 2007. Collective intentions and team agency[J/OL]. Journal of Philosophy, 104: 109-137. DOI: 10.5840/jphil2007104328.

Goranko V, Jamroga W, 2004. Comparing semantics of logics for multi-agent systems[J/OL]. Synthesis, 139: 241-280. DOI: 10.1061/j.mathsocsci.2021.09.001.

Hansson S O, 2002. Preference logic[M/OL]. Dordrecht: Springer Netherlands: 319-393. https://doi.org/10.1007/978-94-017-0456-4_4.

Harsanyi J C, 1990. Interpersonal utility comparisons[M]//Eatwell J, Milgate M, Newman P. Utility and Probability. London: Palgrave Macmillan: 128-133.

Heckhausen J, 2007. The motivation-volition divide and its resolution in action-phase models of developmental regulation[J]. Research in Human Development, 4(3-4): 163-180.

Heil J, 2003. From an ontological point of view[M]. Oxford: Oxford University Press.

Hindriks F, 2018. Collective agency: Moral and amoral[J]. Dialectica, 72(1): 3-23.

Hodgson D, 1967. Consequences of ultitarianism : a study in normative ethics and legal theory.[M]. Oxford: Clarendon Press.

Huebner B, Hedhal M, 2012. Collective values[M]//Kaldis B. Encyclopedia of Philosophy and the Social Sciences. London: Sage publications.

Humberstone I L, 1997. Two types of circularity[J]. Philosophy and Phenomenological Research: A Quarterly Journal: 249-280.

Hurley S, 2005. Teamwork : multi-disciplinary perspectives: Rational agency, cooperation and mind-reading.[M]. London: Palgrave Macmillan: 200-215.

Jacob P, 2019. Intentionality[M]//Zalta E N. The Stanford Encyclopedia of Philosophy. Winter 2019 ed. Metaphysics Research Lab, Stanford University.

Jamroga W, 2008. A temporal logic for markov chains[C]//Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2. Citeseer: 697-704.

Jones M R, 2005. Idealization and abstraction: A framework[M]//Idealization XII: Correcting the model. Leiden: Brill: 173-217.

Karpus J, Radzvilas M, 2017. Team reasoning and a measure of mutual advantage in games[J/OL]. Economics and Philosophy, 34: 1-30. DOI: 10.1017/s0266267117000153.

Kneale W, 1968. Symposium: Intentionality and intensionality[J/OL]. Proceedings of the Aristotelian Society, Supplementary Volumes, 42: 73-90[2022-06-11]. http://www.jstor.org/stable/4106586.

Kriegel U, 2011. The sources of intentionality[M]. Oxford: Oxford University Press.

Levesque H J, Cohen P R, Nunes J H, 1990. On acting together[M]. Menlo Park: SRI International.

Levin J, 2021. Functionalism[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Winter 2021 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/functionalism/.

Lewis D, 1972. Psychophysical and theoretical identifications[J]. Australasian Journal of Philosophy, 50(3): 249-258.

Lewis H D, 1948. Collective responsibility[J]. Philosophy, 23(84): 3-18.

List C, 2022. Social Choice Theory[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Spring 2022 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/social-choice/.

List C, Pettit P, 2011. Group agency: The possibility, design, and status of corporate agents[M]. Oxford: Oxford University Press.

Liu F, 2009. Diversity of agents and their interaction[J]. Journal of Logic, Language and information, 18(1): 23-53.

Liu F, 2011. Reasoning about preference dynamics[M]. Dordrecht: Springer Netherlands.

Liu F, Seligman J, Girard P, 2014. Logical dynamics of belief change in the community[J]. Synthese, 191(11): 2403-2431.

Lorini E, Sartor G, 2016. A stit logic for reasoning about social influence[J]. Studia logica, 104(4): 773-812.

MacBride F, 2020. Relations[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Winter 2020 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/relations/.

Marcus R B, 1990. Some revisionary proposals about belief and believing[J/OL]. Philosophy and Phenomenological Research, 50: 133-153. http://www.jstor.org/stable/2108036.

Martin C B, 2007. The mind in nature[M]. Oxford: Oxford University Press.

Marx K, 1970. Theses on feuerbach (1845)[M]//Karl Marx and Frederick Engels, Selected Works. Maryland: Wildside Press.

May K O, 1952. A set of independent necessary and sufficient conditions for simple majority decision [J]. Econometrica: Journal of the Econometric Society: 680-684.

McDowell J, McFetridge I G, 1978. Are moral requirements hypothetical imperatives?[J]. Proceedings of the Aristotelian Society, Supplementary Volumes, 52: 13-42.

Meijers A M, 2003. Can collective intentionality be individualized?[J]. American Journal of Economics and Sociology, 62(1): 167-183.

Mele A R, 1995. Autonomous agents: From self control to autonomy[M]. Oxford: Oxford University Press.

Menzies P, Beebee H, 2020. Counterfactual Theories of Causation[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Winter 2020 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/causation-counterfactual/.

Meyer J J C, van der Hoek W, 1995. Epistemic logic for ai and computer science[M]. Cambridge: Cambridge University Press.

Ministerie van Onderwijs, Cultuur en Wetenschap, 2019. What we do[EB/OL]. (2019-07-29)[2022-04-29]. https://english.cultureelerfgoed.nl/about-us/what-we-do.

Molnar G, 2003. Powers: A study in metaphysics[M]. Oxford: Oxford University Press.

Mumford S, 1999. Intentionality and the physical: A new theory of disposition ascription[J/OL]. Philosophical Quarterly, 49(195): 215-225. DOI: 10.1111/1467-9213.00138.

Mumford S, Anjum R L, 2011. Getting causes from powers[M]. Oxford: Oxford University Press.

Nash Jr J F, 1950. Equilibrium points in n-person games[J]. Proceedings of the national academy of sciences, 36(1): 48-49.

Page S, Miller J H, 2009. Complex adaptive systems: An introduction to computational models of social life[M]. Princeton: Princeton University Press.

Pauly M, 2002. A modal logic for coalitional power in games[J/OL]. Journal of Logic and Computation, 12: 149-166. DOI: 10.1093/logcom/12.1.149.

Pauly M, 2001. Logic for social software[D]. Universiteit van Amsterdam.

Peleg B, Sudhölter P, 2007. Introduction to the theory of cooperative games[M]. Heidelberg: Springer Berlin.

Petersson B, 2007. Collectivity and circularity[J]. The Journal of Philosophy, 104(3): 138-156.

Pettit P, 2007. Rationality, reasoning and group agency[J]. Dialectica, 61(4): 495-519.

Pettit P, 2009a. Corporate responsibility revisited[J]. Netherlands Journal of Legal Philosophy, 38(2): 159-176.

Pettit P, 2009b. The reality of group agents[J]. Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice: 67-91.

Place U T, 1996. Intentionality as the mark of the dispositional[J]. Dialectica, 50(2): 91-120.

Place U T, 1999. Intentionality and the physical: a reply to mumford[J]. The Philosophical Quarterly, 49: 225-230.

Plaza J, 1989. Logics of public announcements[C]//In Proceedings 4th International Symposium on Methodologies for Intelligent Systems. 201-216.

Pollock J L, 2006. Thinking about acting: Logical foundations for rational decision making[M]. Oxford: Oxford University Press.

Popper K R, 1962. The open society and its enemies, 2 vols., reprint[M]. Princeton: Princeton University Press.

Price H, 1989. Defending desire-as-belief[J]. Mind, 98(389): 119-127.

Putnam H, 1963. Brains and behavior[M]//Butler R J. Analytical Philosophy: Second Series. Oxford: Blackwell.

Rao A S, Georgeff M P, 1998. Decision procedures for bdi logics[J]. Journal of logic and computation, 8(3): 293-343.

Ross D, 2014. Theory of conditional games[M]. Milton Park: Taylor & Francis.

Ross D, 2021. Game Theory[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Fall 2021 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/game-theory/.

Ross H S, Lollis S P, 1987. Communication within infant social games.[J]. Developmental Psychology, 23(2): 241.

Roth A S, 2017. Shared Agency[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Summer 2017 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/shared-agency/.

Rovane C, 1997. The bounds of agency[M]. Princeton: Princeton University Press.

Rovane C, 2014. Group agency and individualism[J]. Erkenntnis, 79(9): 1663-1684.

Roy O, 2008. Thinking before acting: intentions, logic, rational choice[D]. University van Amsterdam.

Ryle G, 1949. The concept of mind[M]. Chicago: University of Chicago Press.

Samuelson P A, 1938. A note on the pure theory of consumer's behaviour[J]. Economica, 5(17): 61-71.

Scanlon T, 2000. What we owe to each other[M]. Cambridge: Belknap Press.

Schmid H B, 2003. Can brains in vats think as a team?[J]. Philosophical Explorations, 6(3): 201-217.

Schmid H B, 2009. Plural action: Essays in philosophy and social science[M]. Dordrecht: Springer Netherlands.

Schmid H B, 2014. Plural self-awareness[J]. Phenomenology and the Cognitive Sciences, 13(1): 7-24.

Schmid H B, Seddone G, 2008. Wir-intentionalitat. kritik des ontologischen individualismus und rekonstruktion der gemeinschaft[J]. Rivista di Storia Della Filosofia, 63(1): 201.

Schroeder T, 2004. Three faces of desire[M]. Oxford: Oxford University Press.

Schweikard D P, Schmid H B, 2021. Collective Intentionality[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Fall 2021 ed. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality/.

Schwitzgebel E, 2002. A phenomenal, dispositional account of belief[J/OL]. Noûs, 36(2): 249-275. DOI: https://doi.org/10.1111/1468-0068.00370.

Searle J, 1983. Intentionality: An essay in the philosophy of mind[M]. Cambridge: Cambridge University Press.

Searle J, 1990. Collective intentions and actions[J]. Intentions in Communication, 195: 220.

Searle J, 1992. The rediscovery of the mind[M]. Cambridge: MIT press.

Searle J, 2010. Making the social world: The structure of human civilization[M]. Oxford: Oxford University Press.

Seligman J, Liu F, Girard P, 2013. Facebook and the epistemic logic of friendship[J]. preprint arXiv: 1310.6440.

Sellars W, 1963. Imperatives, intentions, and the logic of "ought"[M]. Detroit: Wayne State University Press.

Sen A, 1970. Collective choice and social welfare[M]//Collective Choice and Social Welfare. Cambridge: Harvard University Press.

Shoham Y, Tennenholtz M, 1992a. Emergent conventions in multi-agent systems[C]//Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning. 225-231.

Shoham Y, Tennenholtz M, 1992b. On the synthesis of useful social laws for artificial agent societies [C]//AAAI. 276-281.

Skyrms B, 1990. The dynamics of rational deliberation[M]. Cambridge: Harvard University Press.

Smith M, 1987. The humean theory of motivation[J]. Mind, 96(381): 36-61.

Smith M R, 1994. The moral problem[M]. Oxford: Blackwell.

Stalnaker R C, 1984. Inquiry[Z].

Stirling W C, 2012. Theory of conditional games[M]. Cambridge: Cambridge University Press.

Stokhof M, Van Lambalgen M, 2011. Abstractions and idealisations: The construction of modern linguistics[J]. De Gruyter Mouton.

Stoutland F, 1997. Why are philosophers of action so anti-social?[M]//Alanen L, Heinämaa S, Wallgren T. Commonality and Particularity in Ethics. London: Palgrave Macmillan UK: 45-74.

Strawson G, 1994. Mental reality[M]. Cambridge: MIT Press.

Strawson P, 1962. Freedom and resentment[J]. Proceedings of the British Academy, 48: 187-211.

Sugden R, 2003. The logic of team reasoning[J/OL]. Philosophical Explorations, 6: 165-181. DOI: 10.1080/10002003098538748.

Sverdlik S, 1987. Collective responsibility[J/OL]. Philosophical Studies, 51(1): 61-76. DOI: 10.100 7/BF00353963.

Tahko T E, Lowe E J, 2020. Ontological Dependence[M/OL]//Zalta E N. The Stanford Encyclopedia of Philosophy. Fall 2020 ed. Metaphysics Research Lab, Stanford University. https://plato.stanfo rd.edu/archives/fall2020/entries/dependence-ontological/.

Thomson W, 2001. On the axiomatic method and its recent applications to game theory and resource allocation[J]. Social Choice and Welfare, 18(2): 327-386.

Tollefsen D, 2002. Organizations as true believers[J]. Journal of Social Philosophy, 33(3): 395-410.

Tomasello M, Carpenter M, Call J, et al., 2005. Understanding and sharing intentions: The origins of cultural cognition[J]. Behavioral and Brain Sciences, 28(5): 675-691.

Tooming U, 2019. Active desire[J/OL]. Philosophical Psychology, 32(6): 945-968. DOI: 10.1080/09 515089.2019.1629406.

Tuomela R, 2005. We-intentions revisited[J]. Philosophical Studies, 125(3): 327-369.

Tuomela R, 2006. Joint intention, we-mode and i-mode[J]. Midwest studies in philosophy, 30: 35-58.

Tuomela R, 2007. The philosophy of sociality: The shared point of view[M]. Oxford: Oxford University Press.

Tuomela R, 2013. Social ontology: Collective intentionality and group agents[M]. Oxford: Oxford University Press.

Tuomela R, Miller K, 1988. We-intentions[J]. Philosophical Studies, 53(3): 367-89.

Väänänen J, 2007. Dependence logic: A new approach to independence friendly logic[M]. Cambridge: Cambridge University Press.

van Benthem J, 1989. Semantic parallels in natural language and computation[M]//Studies in Logic and the Foundations of Mathematics: volume 129. Elsevier: 331-375.

van Benthem J, 2014. Logic in games[M]. Cambridge: Mit Press.

van Benthem J, Girard P, Roy O, 2009. Everything else being equal: A modal logic for ceteris paribus preferences[J]. Journal of philosophical logic, 38(1): 83-125.

van der Hoek W, Pauly M, 2007. 20 modal logic for games and information[M/OL]//Blackburn P, Van Benthem J, Wolter F. Studies in Logic and Practical Reasoning: volume 3 Handbook of Modal Logic. Elsevier: 1077-1148. DOI: https://doi.org/10.1016/S1570-2464(07)80023-1.

van Ditmarsch H, van Der Hoek W, Kooi B, 2007. Dynamic epistemic logic[M]. Dordrecht: Springer Netherland.

von Neumann J, Morgenstern O, 1947. Theory of games and economic behavior, 2nd rev[M]. Princeton: Princeton university press.

von Wright G H, 1951. Deontic logic[J]. Mind, 60(237): 1-15.

Wall D, 2009. Are there passive desires?[J/OL]. Dialectica, 63(2): 133-155. DOI: 10.1111/dltc.2009.63.issue-2.

Wang Y, Stokhof M, 2022. A relational perspective on collective agency[J/OL]. Philosophies, 7(3). DOI: 10.3390/philosophies7030063.

Watson G, 1996. Two faces of responsibility[J/OL]. Philosophical Topics, 24(2): 227-248. DOI: 10.5840/philtopics199624222.

Weber M, 1978. Economy and society: An outline of interpretive sociology: volume 2[M]. Oakland: University of California Press.

Wittgenstein L, 2010. Philosophical investigations[M]. Hoboken: John Wiley & Sons.

Wooldridge M, 2003. Reasoning about rational agents[M]. Cambridge: MIT press.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my esteemed supervisors at Tsinghua and ILLC: Fenrong Liu, Martin Stokhof, and Sonja Smets. Their exceptional expertise in philosophy and logic makes them an ideal match for my doctoral research. Particularly, I extend my profound appreciation to Fenrong Liu, whose unwavering commitment to academia and joint career pursuits has opened up numerous possibilities for me. Her passion and dedication have left a significant impact on my academic journey. I am equally immensely thankful to Martin Stokhof, whose meticulous mentorship has played a pivotal role in my progress. His weekly feedback and step-by-step guidance have empowered me to overcome seemingly insurmountable challenges with confidence and determination. Additionally, I extend my heartfelt gratitude to Sonja Smets for her invaluable assistance and a warm welcome to the research team at ILLC.

I am humbled and honoured by the willingness of Robert van Rooij, Huub Dijstelbloem, Federica Russo, Johan van Benthem, Hao Tang, Donghua Zhu, and Asher Jiang to serve on my thesis committee.

I wish to acknowledge and extend special thanks to my co-authors, Martin Stokhof, Chenwei Shi, and Qian Chen, whose collaborative efforts have been immensely instructive. Many of the diverse ideas explored in this dissertation were born out of insightful discussions with them.

My gratitude also extends to the professors I encountered at Tsinghua and ILLC: Jeremy Seligman, Dag Westerstal, Lu Wang, Qi Feng, Junhua Yu, Alexandru Baltag, Ulle Endriss, Soroush Rafiee Rad, Anthonie Meijers, Marc Slors and numerous others have significantly contributed to my growth through their courses or engaging discussions.

I am especially thankful to my fellow students, Jialiang Yan and Lei Li, for their companionship and mutual support during our studies in Beijing and Amsterdam. I also wish to extend my appreciation to all those who shared their inspirations with me, including Dazhu Li, Zhiqiang Sun, Kaibo Xie, Fei Xue, Kexi Chen, and many more.

Lastly, I am deeply indebted to my beloved wife, Yu Shi, and my parents, as well as my parents-in-law, for their unwavering support and understanding throughout my journey. My wife's encouragement has been my pillar of strength during times of mental stress, and without her, completing my PhD would not have been possible.

RÉSUMÉ AND ACADEMIC ACHIEVEMENTS

Résumé

Yiyan Wang was born on the 14th of December 1991 in Xinzhou, Shanxi, China.

In September 2009, He began his bachelor's study in the Department of Philosophy, Sun Yat-sen University, majoring in philosophy, and got a Bachelor of Philosophy degree in July 2013.

In September 2013, He began his master's study in the Department of Philosophy, Tsinghua University, and got a Master of Philosophy degree in Logic in July 2016.

In September 2019, he started to pursue a doctorate in the Department of Philosophy at Tsinghua University. In March 2020, he was admitted to the jointly awarded doctorate program of Tsinghua University and the Institute for Logic, Language and Computation, University of Amsterdam.

Academic Achievements

[1] Wang, Yiyan. Intentionality as Disposition. In: Liao Beishui, Wáng Yi (eds) Context, Conflict and Reasoning. Logic in Asia: Studia Logica Library. Springer, Singapore. 2020.

[2] Wang, Yiyan, and Stokhof, Martin. A Relational Perspective on Collective Agency. Philosophies 7 (3):63. 2022.

[3] Chen, Qian, Shi, Chenwei, and Wang, Yiyan. Reasoning about Dependence, Preference, and Coalitional Power. Under review. 2023.

[4] Wang, Yiyan. Towards a Dispositional Account of Intentionality. Under review. 2023.

[5] Wang, Yiyan. Formal and Natural, Self and No-self. Publisher. 2023(6). Forthcoming. 2023. (published in Chinese)

[6] Wang, Yiyan. Intentionality, Preference and Dependency. Manuscript. 2023.

Titles in the ILLC Dissertation Series:

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic

ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility

ILLC DS-2020-16: **Ilaria Canavotto**
Where Responsibility Takes You: Logics of Agency, Counterfactuals and Norms

ILLC DS-2020-17: **Francesca Zaffora Blando**
Patterns and Probabilities: A Study in Algorithmic Randomness and Computable Learning

ILLC DS-2021-01: **Yfke Dulek**
Delegated and Distributed Quantum Computation

ILLC DS-2021-02: **Elbert J. Booij**
The Things Before Us: On What it Is to Be an Object

ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices

ILLC DS-2021-04: **Sophie Arnoult**
Adjunction in Hierarchical Phrase-Based Translation

ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism

ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions

ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners

ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories

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-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-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