TU WIEN Informatics

Agents' Knowledge and Its Limits in Byzantine Fault-Tolerant Distributed Systems

DISSERTATION

zur Erlangung des akademischen Grades

Doktorin der Technischen Wissenschaften

eingereicht von

Krisztina Fruzsa
Matrikelnummer 11834266

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof. Dr. Ulrich Schmid
Zweitbetreuung: Dr. Roman Kuznets, Dr. Hans van Ditmarsch

Diese Dissertation haben begutachtet:

Nick Bezhanishvili	Thomas Studer

Wien, 30. Jänner 2024

Krisztina Fruzsa

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Agents' Knowledge and Its Limits in Byzantine Fault-Tolerant Distributed Systems

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktorin der Technischen Wissenschaften

by

Krisztina Fruzsa
Registration Number 11834266

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Dr. Ulrich Schmid
Second advisor: Dr. Roman Kuznets, Dr. Hans van Ditmarsch

The dissertation has been reviewed by:

Nick Bezhanishvili

Thomas Studer

Vienna, 30th January, 2024

Krisztina Fruzsa

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Krisztina Fruzsa

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 30. Jänner 2024

Krisztina Fruzsa

Acknowledgements

First and foremost, I want to express my gratitude to my supervisor, Ulrich Schmid. I feel truly privileged for having had the opportunity to be influenced by him both professionally and personally over these last several years. His multifaceted research interest and genuine curiosity mixed with, what seems to me, a deep trust in the whole process of scientific discovery have been and continue to be a great source of inspiration to me to this day.

Next, I want to thank my co-supervisors Roman Kuznets and Hans van Ditmarsch. Thanks to Roman, with whom I have spent countless hours discussing logic-related matters, I have deepened my knowledge greatly — to the point I would, for sure, never find out was possible for me. He was the one who knew how to push me to the limits of my knowledge (pun intended) and help me overcome them. Hans has been a role model to me. Thanks to his great mentorship, I managed to reconnect with the joy of figuring things out when I, in retrospect, needed it the most. I will forever be grateful to him, especially for that.

Moreover, I also thank Nick Bezhanishvili and Thomas Studer for agreeing to review my thesis[1] and for all their insightful comments, which helped me improve various parts of it.

I have met so many amazing people on my PhD journey. Firstly, I want to thank my master's thesis supervisor, Rozália Madarász Szilágyi, who encouraged me to pursue a PhD degree in the first place. She was the one who informed me about the LogiCS doctoral college. She was also the one who helped me discover my passion for mathematical logic via her highly engaging way of lecturing. It is fair to say that I would not be where I am without her. Secondly, I want to thank Anna Prianichnikova, the coordinator of LogiCS. She, by her very nature, made me feel immediately welcome from the very start and has been helping immensely to make this journey go smoothly. A special thanks in this respect goes to Marijana Lazić, a LogiCS alumna, for her generous practical help as well as mental support when I first arrived in Vienna. A big thanks also goes to the secretaries Beatrix Buhl and Traude Sommer, for their huge help with various, sometimes even personal, administrative issues.

[1]This work has been supported by the Austrian Science Fund FWF mainly through the LogiCS project (W1255-N23), but also through the projects DMAC (P32431) and ByzDEL (P33600).

Thanks to the generous financial support from the Austrian Science Fund FWF (through projects W1255-N23, P32431, and P33600), I was fortunate enough to experience several research stays abroad during my PhD studies. To this end, I thank Hans van Ditmarsch for welcoming me not once, but twice at the Open University of Netherlands, where I had the opportunity to reflect more deeply about the research problems I was tackling as well as to significantly improve my skills through a very active collaboration. I also thank Thomas Studer for welcoming me at the University of Bern, where I had, thanks to him and the truly amazing people at his institute, the best time possible — I have researched productively, spent a lot of quality time with the people there, and I have been mostly outside, in beautiful nature.

I am grateful to all my colleagues from the ECS group for having enriched this journey of mine. Specifically, I want to thank Florian, Heinz, Josef, Kyrill, Manfred, Martin, and Thomas J., for having made me feel completely welcome when I joined ECS. Thinking back at all our entertaining lunch breaks will always warm my heart. Next, a big thanks goes to the members of the ByzDEL group: Giorgio, Hugo, Rojo, Stephan, and Thomas S., for all the inspiring meetings as well as for the, very much needed, relaxing movie nights. For brightening up my days at the office, a special thanks goes to Arman. Last but not least, I want to thank Raghda for regularly checking in with me and for managing to put me in a good mood, even on days I felt absolutely stuck.

I consider myself lucky for having been part of the LogiCS community. A special thanks goes to Davide and Miroslav for all the carefree adventures over the years. I am so happy that life (LogiCS, to be precise :) brought us together. A special thanks also goes to Sanja for initiating and organizing numerous birthday get-togethers.

Furthermore, I want to thank all the people in my life who have been in one way or another a part of this journey. I thank all my friends, especially for their patience as I know I have been too busy sometimes (shout-out to Bane, Kaća, and Tea). A special thanks goes to Florina and Roman B., my aunt and uncle, as well as to my cousin Florin and uncle Oszkar, who have been an important part of my support system for as long as I remember.

The importance of my family in all of this cannot be overstated. From the bottom of my heart, I thank my parents, Dojna and Tibor, for always being there for me and my brother, Denis, for celebrating all my achievements so proudly! His visits to me, which always recharge me, are what often keeps me going.

The list of things for which I am grateful to my husband, Nenad, is too long to fit into any paragraph (had to be said :). He has been with me through it all — from the first day of my bachelor's studies in mathematics (already more than a decade ago) all the way to finishing writing this thesis... No matter how much I try, there is no convincing this man that I am incapable of something! Thank you for stubbornly believing in me.

Finally, this list would not be complete if I did not mention my dog, Ludwig, who has been (let us be honest) one of my main helpers while writing this thesis.

Kurzfassung

In dieser Arbeit veranschaulichen wir, wie (temporal-)epistemische Logik angewendet werden kann, um byzantinische fehlertolerante asynchrone verteilte Systeme zu untersuchen, in denen sich Agenten auf willkürliche ("byzantinische") Weise falsch verhalten können. Basierend auf unserem Framework zur Modellierung solcher Systeme beginnen wir damit, herauszufinden, was Agenten immer wissen können und was Agenten in Gegenwart byzantinisch-fehlerhafter Agenten niemals wissen können. Da, wie wir zeigen, Standardwissen für Agenten in den meisten Fällen von Interesse nicht erreichbar ist, untersuchen wir, wie ihre epistemischen Zustände in verschiedenen Szenarien am besten erfasst werden können. Im Zuge dessen stoßen wir auf verschiedene epistemische Modalitäten, insbesondere die “hope” Modalität, und untersuchen sie von einem rein logischen Standpunkt aus. Genauer gesagt suchen wir nach geeigneten Axiomatisierungen (d.h. korrekten und vollständigen Axiomatisierungen) für die betreffenden epistemischen Modalitäten und untersuchen, wie sie miteinander interagieren.

Unser oberstes Ziel ist es jedoch, Einblicke in den Entscheidungsprozess von Agenten in byzantinischen fehlertoleranten Systemen zu erhalten. Daher verwenden wir (temporal-) epistemische Logik, basierend auf einigen der neu eingeführten epistemischen Modalitäten, um ein kanonisches verteiltes Problem namens Firing Rebels with Relay (FRR) innerhalb des byzantinischen fehlertoleranten asynchronen Modells zu analysieren. Das FRR-Problem ist im Wesentlichen ein Vereinbarungsproblem, welches erfordert, dass jeder korrekte Agent eine Aktion namens FIRE ausführt, und zwar auf eine Alles-oder-Nichts-Weise (aber nicht notwendigerweise gleichzeitig) und nur, wenn mindestens ein korrekter Agent lokal ein Triggerereignis namens START beobachtet hat. Es ist in der Distributed Computing Community bekannt, dass im Falle von milden Fehlern (z. B. solche, bei denen Agenten einfach nichts mehr tun oder Nachrichten verlieren) das Erreichen einer Einigung mit bestimmten Formen des allgemeinen Standardwissens verbunden ist. Interessanterweise stellt sich heraus, dass ein temporal-epistemischer Gruppenbegriff “hope”, nämlich common eventual hope, im Mittelpunkt jeder Lösung des FRR-Problems steht.

Abstract

In this thesis, we illustrate how (temporal-)epistemic logic can be applied to study byzantine fault-tolerant asynchronous message-passing distributed systems, in which agents may misbehave in an arbitrary ("byzantine") way. Based on our framework for modeling such systems, we start by establishing what agents can always know and what agents can never know in the presence of byzantine faulty agents. Since, as we show, standard knowledge is not achievable by agents in most cases of interest, we explore how to best capture their epistemic states in various scenarios. On that journey, we encounter different epistemic modalities, in particular, the hope modality, and study them from a purely logical point of view. More precisely, we search for appropriate axiomatizations (meaning, sound and complete axiomatizations) for the encountered epistemic modalities and investigate how they interact with each other.

Our ultimate goal is gaining insight into agents' decision-making process in byzantine fault-tolerant systems, however. Therefore, we use (temporal-)epistemic logic based on some of the newly introduced epistemic modalities to analyze a canonical distributed computing problem called Firing Rebels with Relay (FRR) within the byzantine fault-tolerant asynchronous model. The FRR problem is, essentially, an agreement problem requiring that every correct agent performs an action called FIRE, in an all-or-none fashion (though not necessarily simultaneously), and only if at least one correct agent locally observed a trigger event called START. It is well-known in the distributed computing community that, in case of benign faults (like the ones when agents just stop operating or lose messages), reaching agreement is connected with certain forms of standard common knowledge. Interestingly, it turns out that a temporal-epistemic group notion of hope, namely, common eventual hope, is at the heart of any solution of the FRR problem.

Contents

Kurzfassung **ix**

Abstract **xi**

Contents **xiii**

1 Introduction **1**
- 1.1 Epistemic modeling of distributed systems . . . 3
- 1.2 Motivation . . . 5
- 1.3 Contributions and outline of the thesis . . . 6

2 Preliminaries **11**
- 2.1 Multi-agent epistemic logic – basic concepts . . . 11
- 2.2 Modeling byzantine fault-tolerant asynchronous message-passing systems – basic concepts . . . 17

3 Epistemic analysis **37**
- 3.1 Obtaining Kripke models . . . 38
- 3.2 Modeling the brain-in-a-vat scenario . . . 41
- 3.3 Relativizing the preconditions of actions . . . 49
- 3.4 Related work . . . 52

4 A modal logic of hope **53**
- 4.1 Towards a logic of hope . . . 54
- 4.2 Soundness and completeness results . . . 55
- 4.3 Finite model property and decidability . . . 73
- 4.4 Related work . . . 75

5 A new hope **77**
- 5.1 Axiomatizing hope (again) . . . 78
- 5.2 Individual hope and individual knowledge . . . 81
- 5.3 Common hope and common knowledge . . . 93
- 5.4 Soundness and completeness of $\mathscr{KHC}$ with respect to KH . . . 99
- 5.5 Finite model property and decidability . . . 116

5.6 Defining common eventual hope 119
5.7 Related work 126

6 The Firing Rebels with Relay **127**
6.1 Formulating the problem 128
6.2 Modeling via interpreted systems 129
6.3 Necessary and sufficient conditions 133
6.4 Related work 142

7 Summary of our accomplishments and follow-up/future work **145**
7.1 Summary of our accomplishments 145
7.2 Follow-up/future work 146

List of Figures **149**

Bibliography **151**

CHAPTER 1

Introduction

A distributed system is a collection of networked agents, viewed as abstract processes (representing computers, for example), that need to actively collaborate in order to achieve a common goal. A distributed algorithm makes the collaboration among agents possible, by dictating each and every step agents should take towards their desired goal. In reality, however, various sources of uncertainty like varying execution speeds, unpredictable message delays and all kinds of failures make it difficult for agents to succeed. Moreover, different combinations of distributed systems model assumptions can lead to vastly different algorithms and, as such, they usually require separate consideration. Consequently, the literature on distributed algorithms is abundant [Lyn96, AW04].

There are three major categories of different model assumptions considered in the literature.

I.) Assumptions about the communication means:

a.) *message-passing* systems (see e.g. [LSP82]), where agents communicate by sending messages to each other over communication channels;

b.) *shared-memory* systems (see e.g. [HS99]), where agents store their data in a joint memory.

II.) Timing assumptions:

a.) *asynchronous* systems (see e.g. [FLP85, ASW88]), where both agents' execution speeds and message delays are unbounded;

b.) *synchronous* systems (see e.g. [WSS19]), where both agents' execution speeds and message delays are bounded (so, in this case agents progress in rounds — a round is an interval of time during which all agents first send their messages, wait to

receive messages sent by other agents in the same round, and then change their local memories accordingly);

c.) *partially synchronous* systems (see e.g. [DLS88, WS07]), which are systems that behave asynchronously for some finite (but unknown) period of time, and synchronously otherwise.

III.) Assumptions about agent faults:

a.) *crash* failures (see e.g. [FLP85, HS99]), where (some) agents may stop operating (possibly without completing their last operating step);

b.) *omission* failures (see e.g. [PT86]), where (some) agents may fail to send/receive some messages;

c.) *byzantine* failures (see e.g. [LSP82]), where (some) agents may misbehave in an arbitrary way, e.g., by sending inconsistent information to different agents.

The maximal number of faulty agents that can occur during a single execution (run) of the system is usually denoted by f, while the total number of agents in the system is usually denoted by n. In general, when allowing faults to happen, $0 < f < n$ is assumed, as, obviously, not much can be guaranteed to happen if all agents can become faulty during any execution. Note that this does not mean that we necessarily treat different agents differently — what is usually assumed is that any of them can, in principle, become faulty, but there can never be more than f of them in total during a single execution.

Additional model assumptions include, for example:

- assumptions about the topology of the communication network: *fully connected* (see e.g. [LSP82]) versus *partially connected* communication network (see e.g. [ASW88, WSS19]);
- assumptions about communication channels: *reliable* (see e.g. [FLP85]) versus *unreliable* communication channels (see e.g. [WSS19]);
- assumptions about agents' memory capacity: *perfect recall* (see e.g. [FLP85, HS99]) versus *partial memory* (see e.g. [ASW88, WSS19]);
- assumptions about the nature of agents: *homogenous* (which is usually assumed) versus *heterogenous* collection of agents (see e.g. [Dij74]).

In this thesis we focus entirely on the byzantine fault-tolerant asynchronous message-passing model.

1.1 Epistemic modeling of distributed systems

At least since the groundbreaking work by Halpern and Moses [HM90], it is well-known in the distributed computing community that knowledge [Hin62] is a powerful conceptual abstraction for studying distributed systems. In this section, we will provide an overview of some of the most important results in the area.

The standard relational semantics of epistemic logic relies on the notion of a Kripke model M, which consists of all "possible worlds" agents can be in along with accessibility relations R_i for each agent i (e.g., wR_iw' means that, when in world w, agent i considers world w' possible) and a valuation function determining in which worlds which atomic propositions (of the logical language being interpreted in the semantics) are true. Finally, agent i's knowledge of some formula φ in a world w is captured by a modal knowledge operator K_i and defined to hold in the following way:

$$M, w \models K_i\varphi \text{ iff for every } w' \text{ with } wR_iw' \text{ it holds that } M, w' \models \varphi.$$

So, we say that agent i *knows* formula φ in the world w, if formula φ holds in every world w' that is accessible to agent i from the world w.

In the *interpreted runs-and-systems* framework [HM90, FHMV95], used for analyzing distributed systems using (temporal-)epistemic logic, the set of all possible runs R of a system together with a valuation function π for atomic propositions, i.e., an *interpreted system* $\mathcal{I} = (R, \pi)$, determines a Kripke model, consisting of pairs (r, t) of a run $r \in R$ and a time $t \in \mathbb{N}_0$ (representing global states $r(t)$ of the underlying system). Two pairs (r, t) and (r', t') are defined to be indistinguishable for agent i if and only if i has the same local state in both global states represented by those pairs, formally, if $r_i(t) = r'_i(t')$. Thus, the indistinguishability relations play the role of accessibility relations in the corresponding Kripke semantics. We write

$$(I, r, t) \models K_i\varphi,$$

iff for every $r' \in R$ and for every $t' \in \mathbb{N}_0$ with $r_i(t) = r'_i(t')$ it holds that $(I, r', t') \models \varphi$. Again, we say that agent i *knows* formula φ at time t in run r (i.e., in (r, t)), if formula φ holds in every pair (r', t') that is indistinguishable from the pair (r, t) for agent i.

Important additional epistemic modalities often used for analyzing distributed systems are:

1. *mutual knowledge* $E_G\varphi := \bigwedge_{i \in G} K_i\varphi$, which states that every agent in group G knows φ,

2. *common knowledge* $C_G\varphi$, informally expressed as the infinite conjunction $E_G\varphi \wedge E_GE_G\varphi \wedge \ldots$, which states that every agent in group G knows φ, and every agent in group G knows that every agent in group G knows φ, and so on ad infinitum.

The knowledge-based approach has provided a number of fundamental insights about distributed systems. Most of them rely on the *Knowledge of Preconditions* principle (KoP), which has been crisply formulated by Moses [Mos15] as:

KoP: if formula φ is a necessary condition for agent i to perform action α, then $K_i\varphi$ is also a necessary condition for i to perform α.

Furthermore, using KoP, it can be shown that performing simultaneous actions requires common knowledge [HM90, DM90, FHMV95, BZM10, Mos15] and performing actions in a linear temporal order requires *nested knowledge* [BZM14, Mos15].

Given a distributed computing problem and a candidate protocol for solving it, showing that agents act without having attained all the respective necessary knowledge requirements can, hence, be used for effectively proving protocol incorrectness. Moreover, KoP holds regardless of the underlying model of distributed systems, so it is widely applicable.

Over the years, epistemic reasoning has been successfully applied for analyzing various distributed computing problems, primarily in fault-free systems and systems with crash and omission failures, however.

In asynchronous message-passing systems, where the absence of communication is indistinguishable from delayed communication, agents can gain new knowledge only based on the messages they receive, as precisely captured by *message chains* [CM86]. In synchronous message-passing systems, where message delays are upper-bounded, agents can gain new knowledge also from the absence of communication (communication-by-time). This has been already observed by Chandy and Misra in [CM86]: "If there is a global clock common to all processes, then processes may learn or forget merely by the passage of time."

In [BZM14], Ben-Zvi and Moses analyze, using epistemic logic, the *Ordered Response problem* (OR), in which temporal linear ordering of agents' actions is required in response to a triggering event. They show that nested knowledge about the triggering event plays a crucial role in achieving such a coordination. They also introduce a causal structure called *centipede*, and show that a centipede must exist in every execution of a protocol solving OR (because its existence is necessary for achieving the required nested knowledge). Centipedes are defined using two relations: *syncausality* and *bound guarantee*. The former is obtained by augmenting Lamport's happened-before relation [Lam78] by causal links indicating no communication within the message delay upper bound to also capture causality induced via communication-by-time, and the latter is based on the message delivery time bounds. In the conference version [BZM10] of [BZM14], Ben-Zvi and Moses also analyze the *Simultaneous Response problem* (SiR), in which agents must act simultaneously in response to a triggering event. They show that common knowledge plays a crucial role in achieving such a coordination. In addition, a variation on the centipede structure called *centibroom* is introduced. A centibroom must exist in every execution of a protocol solving SiR (because its existence is necessary for achieving the

required common knowledge). The corresponding epistemic analyses in both [BZM10] and [BZM14] are performed within the fault-free synchronous model of distributed systems with reliable communication.

In [GM13], Gonczarowski and Moses analyzed *timely coordination*, in which explicit bounds on the relative times at which actions are performed are specified. By defining the notion of *timely common knowledge* as a vectorial fixpoint, they characterize both solvability and optimal solutions of a general class of timely coordination tasks.

In [CGM14], Castañeda, Gonczarowski, and Moses derive the very first unbeatable protocol for solving *consensus* by epistemic reasoning. The corresponding epistemic analysis is performed within the synchronous model of distributed systems restricted to crash failures.

Simultaneous Byzantine Agreement (SBA) has been studied in [DM90] by Dwork and Moses. Using common knowledge, they derive an optimal protocol for SBA. The corresponding epistemic analysis is performed within the synchronous model of distributed systems restricted to crash failures.

A general class of problems involving performing simultaneous actions has been introduced in [MT86, MT88]. In [MT86, MT88], Moses and Tuttle use common knowledge to study efficient protocols for solving such problems. The corresponding epistemic analysis is performed within the synchronous model of distributed systems restricted to omission failures.

Eventual Byzantine Agreement (EBA) has been studied in [HMW01]. The authors characterize optimal EBA protocols using *continual common knowledge*. Interestingly, continual common knowledge is a stronger group notion of knowledge than common knowledge. The corresponding epistemic analysis is performed within the synchronous model of distributed systems restricted to crash and omission failures.

The uncertainty added by byzantine faults severely complicates the already challenging design and analysis of distributed algorithms. Even though several of the above papers have “byzantine” in their title, it is nevertheless the case that they solely consider benign faults, such as crash and omission failures. We are not aware of any attempt to extend epistemic reasoning to systems with truly byzantine faults, except for [Mic89]. However, faulty agents considered in [Mic89] may not really behave arbitrarily either, as they are not allowed to exhibit a behaviour that cannot be observed in some correct execution as well.

1.2 Motivation

In this thesis, we aim to develop a sound understanding of how agents “reason” and, thus, make decisions in the presence of byzantine faulty agents in asynchronous message-passing distributed systems in particular, using (temporal-)epistemic logic. For this purpose, we perform a careful analysis of a canonical distributed computing problem called Firing Rebels with Relay (FRR) within our framework for modeling such systems.

The FRR problem is a problem related to the *consistent broadcasting primitive*, introduced by Srikanth and Toueg in [ST87b]. Srikanth and Toueg use this communication primitive to simulate signed communication in order to be able to convert an authenticated fault-tolerant algorithm into an equivalent non-authenticated fault-tolerant algorithm. This approach has been applied to byzantine fault-tolerant clock synchronization [DFP+14, FS12, RS11, ST87a, WS09]. Moreover, FRR is instrumental for the generic reduction of task solution algorithms for byzantine fault-tolerant systems to algorithms for crash-resilient systems introduced in [MTH14].

The formulation of FRR first appeared in [Fim18] together with the formulation of Firing Rebels without Relay, albeit with a different correctness requirement (see below). The FRR problem assumes that every agent i may observe an event START and may generate an action FIRE according to the following specification:

Correctness: If at least $2f + 1$ agents learn that START occurred at a correct agent, all correct agents perform FIRE eventually.

Unforgeability: If a correct agent performs FIRE, then START occurred at a correct agent.

Relay: If a correct agent performs FIRE, all correct agents perform FIRE eventually.

In [Fim18], the correctness requirement states: if at least $f + 1$ correct agents observe START, then all correct agents perform FIRE eventually. Note also that only the weaker version of Firing Rebels, that is, Firing Rebels without Relay, has been analyzed in [Fim18].

1.3 Contributions and outline of the thesis

In order to understand and formally reason about the decision-making process of correct agents in presence of byzantine faulty agents in asynchronous message-passing distributed systems, we first establish some general results concerning agents' knowledge limitations using our framework for modeling such systems. In doing so, we introduce several epistemic modalities closely related to knowledge, in particular, the *hope* modality. We propose an axiomatic system for hope, which we show to be strongly sound and strongly complete with respect to the $\mathsf{K45}_n$ class of models satisfying some additional properties. We also present an alternative axiomatic system for hope that is sound and complete with respect to the $\mathsf{KB4}_n$ class of models. Based on it, we then propose a joint system for (common) hope and (common) knowledge. Using (temporal-)epistemic logic based on some group notions of hope, namely, *mutual eventual hope* and *common eventual hope*, we perform an in-depth analysis of Firing Rebels with Relay, which represents a canonical distributed computing problem. We prove that common eventual hope plays a crucial

role for meeting its *relay* requirement, instructing agents to act in a non-simultaneous all-or-none fashion. Moreover, assuming there are sufficiently many agents in the system who will always stay correct (i.e., never become byzantine faulty), at least $2f + 1$ to be precise, we show how common eventual hope can collapse to one level of mutual eventual hope. Finally, we also identify conditions that are sufficient for solving FRR.

We now provide a more detailed description of our results contained in individual chapters of this thesis:

- **Chapter 2**: We introduce the basic concepts of multi-agent epistemic logic used throughout the thesis. We also introduce the cornerstones of our framework for modeling byzantine fault-tolerant asynchronous message-passing distributed systems, based on

  – [KPS+19] R. Kuznets, L. Prosperi, U. Schmid, K. Fruzsa, L. Gréaux. *Knowledge in Byzantine Message-Passing Systems I: Framework and the Causal Cone*, Technical Report TUW-260549, TU Wien, 2019.

- **Chapter 3**: We derive generic results about what asynchronous agents can(not) know in byzantine fault-tolerant message-passing distributed systems. In our central result, the Brain-in-a-Vat lemma, we show that no matter what it observed, an asynchronous agent in a byzantine setting can never rule out the possibility of those observations being imaginary results of its malfunction. Using this result, we conclude that the Knowledge of Preconditions principle (according to which any precondition for action must be known by the acting agent) severely restricts the kinds of preconditions for actions agents can rely on in such a setting. Consequently, we investigate how the corresponding adequate preconditions for actions look like, which gives us insight into the epistemic state of an acting agent in systems with byzantine faults.

  This chapter is based on

  – [KPSF19] R. Kuznets, L. Prosperi, U. Schmid, K. Fruzsa. *Epistemic Reasoning with Byzantine-faulty Agents*, in: A. Herzig and A. Popescu, editors, Frontiers of Combining Systems - 12th International Symposium, FroCoS 2019, London, UK, September 4-6, 2019, Proceedings, volume 11715 of Lecture Notes in Computer Science, pages 259–276. Springer, 2019.

- **Chapter 4**: We study the hope modality, introduced in the previous chapter, from a purely logical point of view. Essentially, we aim to get a better understanding of individual hope in order to be able to formally introduce relevant group notions of hope. So, we propose a separate (from knowledge) axiomatization for the individual hope modality while relying on so-called *correctness atoms* in the language. We then provide a detailed proof of strong soundness and strong completess for the proposed axiom system with respect to a newly designed class of Kripke models that precisely captures the properties of hope. The resulting logic turns out to violate

the uniform substitution rule, however. In addition, we also provide a proof of soundness and completeness with respect to the standard S5 models for knowledge via a suitable translation function.

This part of the chapter is based on

– [Fru21] K. Fruzsa, *Hope for Epistemic Reasoning with Faulty Agents!*, in: A. Pavlova, M. Young Pedersen, and R. Bernardi, editors, Selected Reflections in Language, Logic, and Information - ESSLLI 2019, ESSLLI 2020 and ESSLLI 2021 Student Sessions, Selected Papers, volume 14354 of Lecture Notes in Computer Science, pages 93–108. Springer, 2021.

Finally, in an entirely new section 4.3, we prove that the proposed logic of hope has the finite model property as well as that it is decidable.

- **Chapter 5**: We propose an alternative axiomatization for the hope modality, which successfully avoids the use of correctness atoms. The resulting new logic of hope turns out to be a normal multi-agent epistemic logic. We also propose a joint logic of hope and knowledge as well as a logic extended with notions of common hope and common knowledge. The proposed systems enable us to logically characterize the byzantine fault-tolerant model considered throughout the thesis. We provide a thorough soundness and completeness proof of the axiom system for the joint logic of common hope and common knowledge. This part of the chapter is based on

  – [vDFK22] H. van Ditmarsch, K. Fruzsa, R. Kuznets, *A New Hope*, in: D. Fernández-Duque, A. Palmigiano, and S. Pinchinat, editors, Advances in Modal Logic, AiML 2022, Rennes, France, August 22-25, 2022, pages 349–370. College Publications, 2022,

  with the exception of Section 5.4, which is entirely new.

  In an also new section 5.5, we prove that all of the logics presented in the chapter have the finite model property as well as that they are decidable.

  Finally, in the entirely new Section 5.6 as well, we describe a way to define a particular temporal-epistemic group notion of hope called common eventual hope. We end the chapter by deriving useful properties of the common eventual hope modality needed for the epistemic analysis of the Firing Rebels with Relay problem that is performed in the next chapter.

- **Chapter 6**: Using epistemic reasoning, we analyze the Firing Rebels with Relay (FRR) problem using our framework for modeling byzantine fault-tolerant asynchronous message-passing distributed systems. Informally, FRR requires that every correct agent performs an action called FIRE, in an all-or-none fashion (though not necessarily simultaneously), and only if at least one correct agent locally observed a trigger event called START. Through a detailed epistemic analysis, we establish the necessary epistemic state that needs to be acquired by correct agents in order to FIRE in every correct solution of the problem. The respective epistemic state

turns out to involve common eventual hope, which we show to be attained already by achieving one level of mutual eventual hope in case there are at least $3f + 1$ agents in the system in total. Finally, we also identify conditions that are sufficient for solving FRR.

This chapter is based on

- [FKS21] K. Fruzsa, R. Kuznets, U. Schmid. *Fire!*, in: J. Y. Halpern and A. Perea, editors, Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2021, Beijing, China, June 25-27, 2021, volume 335 of EPTCS, pages 139–153, 2021.

- **Chapter 7**: Finally, we summarize our main accomplishments from all the previous chapters and present a brief description of the already existing follow-up work as well as directions for future research.

Each chapter starts with a concise outline and ends usually with an overview of specific related work. Moreover, each chapter is written in a self-contained manner (at least, for the most part), with the exception of Chapter 3 which heavily relies on Chapter 2.

CHAPTER 2

Preliminaries

In this chapter, we introduce the basic concepts of (multi-agent epistemic) modal logic as well as the cornerstones of our framework for modeling asynchronous message-passing distributed systems allowing byzantine faults, that are going to be used throughout the thesis.

2.1 Multi-agent epistemic logic – basic concepts

Syntax. We start with a nonempty countably infinite set of atomic propositions P and continue by forming formulas by closing under the Boolean connectives $\neg$ and $\wedge$ and under n (unary) modal operators $K_1, \ldots, K_n$ to obtain the modal language $\mathcal{L}_K$, i.e., the language $\mathcal{L}_K$ is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi,$$

where $p \in P$ and $i \in \{1, \ldots, n\}$. We take $\top$ to be an abbreviation for some fixed propositional tautology, and take $\bot$ to be an abbreviation for $\neg\top$. Also, we use the following standard abbreviations from propositional logic: $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ for $\neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$.

We will work with the following definition of a *normal (multi-agent epistemic) modal logic* (it follows closely the one given in [BdRV01]).

Definition 2.1 (Normal multi-agent epistemic logic)**.** *A set of formulas $L \subseteq \mathcal{L}_K$ forms a system of a normal multi-agent epistemic logic iff it contains all propositional tautologies and is closed under the modus ponens inference rule*

$$MP: \quad \frac{\varphi \quad \varphi \to \psi}{\psi},$$

the K axiom scheme

$$K: \qquad K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi),$$

the necessitation inference rule

$$Nec: \qquad \frac{\varphi}{K_i\varphi},$$

and the uniform substitution rule

$$US: \qquad \frac{\varphi}{\varphi[p/\psi]}.$$

Remark 2.2. *We note that in many axiom systems, the US rule is built in indirectly (via axiom schemes).*[1]

Remark 2.3. *We will also sometimes call the modality K normal if the underlying axiom system includes the K axiom scheme and the necessitation inference rule.*

Definition 2.4. *Let $\mathscr{L}$ be an axiomatization in the language $\mathcal{L}_K$ given by axioms and inference rules. A $\mathscr{L}$-derivation is a sequence of formulas $\varphi_1, \ldots, \varphi_n \in \mathcal{L}_K$ such that for each $i \in \{1, \ldots, n\}$:*

- *φ_i is an axiom instance of $\mathscr{L}$, or*
- *φ_i follows from $\varphi_{j_1}, \ldots, \varphi_{j_k}$ by a k-ary inference rule of $\mathscr{L}$ for some $j_1, \ldots, j_k < i$.*

A $\mathscr{L}$-derivation $\varphi_1, \ldots, \varphi_n$ is a $\mathscr{L}$-derivation for φ_n. We will write $\vdash_{\mathscr{L}} \varphi$ to denote that there exists a $\mathscr{L}$-derivation for the formula φ, i.e., that φ is a theorem of $\mathscr{L}$.

Throughout the thesis we will often refer to the axiom system $\mathscr{S}5_n$ depicted in Figure 2.1 and the axiom system $\mathscr{K}45_n$ depicted in Figure 2.2.

We define the semantics in terms of possible worlds, which we formalize in terms of *Kripke models*.

Definition 2.5 (Kripke model, Kripke frame)**.** *A* Kripke model *is a structure $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, consisting of:*

- *W, which is a nonempty set of* states *or* possible worlds *representing the domain of M (we will also sometimes write $\mathcal{D}(M)$ to refer to W),*
- *$\pi : P \to \mathcal{P}(W)$, which is a valuation function that associates with each atomic proposition $p \in P$ a set of worlds where p is true,*
- *and $\mathcal{K}_i$, which are binary relations on W called accessibility relations.*

[1] When defining (normal) modal logics, some textbooks such as [Che80] do not mention this rule.

$$
\begin{aligned}
P: &\quad \textit{all propositional tautologies}\\
K: &\quad K_i(\varphi \to \psi) \wedge K_i\varphi \to K_i\psi\\
T: &\quad K_i\varphi \to \varphi\\
4: &\quad K_i\varphi \to K_iK_i\varphi\\
5: &\quad \neg K_i\varphi \to K_i\neg K_i\varphi\\
MP: &\quad \frac{\varphi \quad \varphi \to \psi}{\psi}\\
Nec: &\quad \frac{\varphi}{K_i\varphi}
\end{aligned}
$$

Figure 2.1: Axiom system $\mathscr{S}5_n$

$$
\begin{aligned}
P: &\quad \textit{all propositional tautologies}\\
K: &\quad K_i(\varphi \to \psi) \wedge K_i\varphi \to K_i\psi\\
4: &\quad K_i\varphi \to K_iK_i\varphi\\
5: &\quad \neg K_i\varphi \to K_i\neg K_i\varphi\\
MP: &\quad \frac{\varphi \quad \varphi \to \psi}{\psi}\\
Nec: &\quad \frac{\varphi}{K_i\varphi}
\end{aligned}
$$

Figure 2.2: Axiom system $\mathscr{K}45_n$

The structure $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ is called a Kripke frame. *Sometimes we will write $M = (F, \pi)$.*

Definition 2.6. *Let $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ and $\varphi \in \mathcal{L}_K$. By $M, w \models \varphi$, we denote the fact that formula φ is satisfied at world w of the model M and define the $\models$ relation inductively as follows:*

- *$M, w \models p$ iff $w \in \pi(p)$ for all $p \in P$,*
- *$M, w \models \neg\varphi$ iff $M, w \models \varphi$ does not hold,*
- *$M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$,*
- *$M, w \models K_i\varphi$ iff $M, v \models \varphi$ for all $v \in \mathcal{K}_i(w)$,*

where $\mathcal{K}_i(w) := \{w' \mid w\mathcal{K}_iw'\}$. We will write $M, w \not\models \varphi$ to denote that $M, w \models \varphi$ does not hold.

We distinguish between the following different levels of truth:

- $M \models \varphi$ means that φ is satisfied at all worlds w of model M, i.e., that φ is *valid in model* M,
- $F \models \varphi$ means that φ is valid in all models (F, π), i.e., that φ is *valid in frame* F,
- $\mathsf{C} \models \varphi$ means that φ is valid in all models M from a class of Kripke models C, i.e., that φ is *valid in class* C.

Definition 2.7. *A binary relation $\mathcal{R}$ on a set S is called*

- reflexive *if $s\mathcal{R}s$ for any $s \in S$,*
- symmetric *if $s\mathcal{R}t$ whenever $t\mathcal{R}s$,*
- transitive *if $s\mathcal{R}u$ whenever $s\mathcal{R}t$ and $t\mathcal{R}u$,*
- euclidean *if $t\mathcal{R}u$ whenever $s\mathcal{R}t$ and $s\mathcal{R}u$, and*
- shift serial *if $\mathcal{R}(t) \neq \varnothing$ for any $t \in \mathcal{R}(s)$, $s \in S$.*

In addition, we call $\mathcal{R}$ an equivalence relation *if it is reflexive, symmetric, and transitive.*

Throughout the thesis we will often refer to the following class of Kripke models:

Definition 2.8. *The class $\mathsf{S5_n}$ consists of all Kripke models $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ with equivalence accessibility relations for all $i \in \{1, \ldots, n\}$.*

Definition 2.9. *The class $\mathsf{K45_n}$ consists of all Kripke models $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ with transitive and euclidean accessibility relations for all $i \in \{1, \ldots, n\}$.*

Definition 2.10 ((Weak) soundness)**.** *Let $\mathscr{L}$ be an axiomatization in the language $\mathcal{L}_K$. The axiom system $\mathscr{L}$ is (weakly) sound with respect to a class of Kripke models C if, for any formula $\varphi \in \mathcal{L}_K$,*

$$\vdash_{\mathscr{L}} \varphi \quad \Longrightarrow \quad \models_{\mathsf{C}} \varphi.$$

Definition 2.11 ((Weak) completeness)**.** *Let $\mathscr{L}$ be an axiomatization in the language $\mathcal{L}_K$. The axiom system $\mathscr{L}$ is (weakly) complete with respect to a class of Kripke models C if, for any formula $\varphi \in \mathcal{L}_K$,*

$$\models_{\mathsf{C}} \varphi \quad \Longrightarrow \quad \vdash_{\mathscr{L}} \varphi.$$

The proof of the following theorem can be found in [FHMV95] (Theorem 3.1.5, p. 61).

Theorem 2.12. *The axiom system $\mathscr{S}5_n$ is sound and complete with respect to the $\mathsf{S5_n}$ class of models.*

Similarly, using the well-known fact that axiom 4 characterizes the transitive property of frames and that axiom 5 characterizes the euclidean property of frames [FHMV95], it is easy to show that:

Theorem 2.13. *The axiom system $\mathscr{K}45_n$ is sound and complete with respect to the* $\mathsf{K45_n}$ *class of models.*

Definition 2.14. *A formula $\varphi \in \mathcal{L}_K$ is derivable from a set of premises $\Gamma \subseteq \mathcal{L}_K$ in an axiom system $\mathscr{L}$ — written $\Gamma \vdash_{\mathscr{L}} \varphi$ — iff*

$$\vdash_{\mathscr{L}} \varphi_1 \wedge \cdots \wedge \varphi_n \to \varphi,$$

for some $\varphi_1, \ldots, \varphi_n \in \Gamma$.

Definition 2.15. *Let C be a class of Kripke models and let $\Gamma \cup \{\varphi\} \subseteq \mathcal{L}_K$ be a set of formulas. Then $\Gamma \models_{\mathsf{C}} \varphi$ means that for every model $M \in \mathsf{C}$, and for every world $w \in \mathcal{D}(M)$,*

$$M, w \models \psi \quad \textit{for all} \quad \psi \in \Gamma \quad \Longrightarrow \quad M, w \models \varphi.$$

Definition 2.16 (Strong soundness)**.** *Let $\mathscr{L}$ be an axiomatization in the language $\mathcal{L}_K$. The axiom system $\mathscr{L}$ is strongly sound with respect to a class of Kripke models C if, for any set of formulas $\Gamma \cup \{\varphi\} \subseteq \mathcal{L}_K$,*

$$\Gamma \vdash_{\mathscr{L}} \varphi \quad \Longrightarrow \quad \Gamma \models_{\mathsf{C}} \varphi.$$

Remark 2.17. *Note that weak soundness is the special case of strong soundness when Γ is the empty set.*

Definition 2.18 (Strong completeness)**.** *Let $\mathscr{L}$ be an axiomatization in the language $\mathcal{L}_K$. The axiom system $\mathscr{L}$ is strongly complete with respect to a class of Kripke models C if, for any set of formulas $\Gamma \cup \{\varphi\} \subseteq \mathcal{L}_K$,*

$$\Gamma \models_{\mathsf{C}} \varphi \quad \Longrightarrow \quad \Gamma \vdash_{\mathscr{L}} \varphi.$$

Remark 2.19. *Note that weak completeness is the special case of strong completeness when Γ is the empty set.*

Definition 2.20. *Let C be a class of Kripke models. A set of formulas $\Gamma \subseteq \mathcal{L}_K$ is called C-satisfiable if there exists a model $M \in \mathsf{C}$ and a world $w \in \mathcal{D}(M)$ such that*

$$M, w \models \varphi \quad \textit{for all} \quad \varphi \in \Gamma.$$

Definition 2.21 (Compact logic)**.** *A logic $L \subseteq \mathcal{L}_K$ defined via validity in a class C of Kripke models is called compact if for any set of formulas $\Gamma \subseteq L$:*

$$\Gamma \textit{ is } \mathsf{C}\textit{-satisfiable}$$

$$\Longleftrightarrow$$

$$\textit{every finite subset of } \Gamma \textit{ is } \mathsf{C}\textit{-satisfiable.}$$

Theorem 2.22. *Let $\mathscr{L}$ be an axiomatization in the language $\mathcal{L}_K$. If $\mathscr{L}$ is strongly sound and strongly complete with respect to a class of Kripke models C, then the logic $L \subseteq \mathcal{L}_K$ defined via validity in C is compact.*

Proof. Assume the opposite towards contradiction: there exists a set of formulas $\Gamma \subseteq L$ which is not C-satisfiable, while every finite subset of Γ is C-satisfiable. Since Γ is assumed to be not C-satisfiable, we have (vacuously)

$$\Gamma \models_{\mathsf{C}} \bot.$$

Therefore, by strong completeness, we also have

$$\Gamma \vdash_{\mathscr{L}} \bot.$$

According to Definition 2.14, it follows that there exist some $\varphi_1, \ldots, \varphi_n \in \Gamma$ such that $\vdash_{\mathscr{L}} \varphi_1 \wedge \cdots \wedge \varphi_n \to \bot$. However, for $\Delta := \{\varphi_1, \ldots, \varphi_n\} \subseteq \Gamma$, $\Delta \vdash_{\mathscr{L}} \bot$ then holds too. Using strong soundness, further results in $\Delta \models_{\mathsf{C}} \bot$, contradicting our assumption that every finite subset of Γ is C-satisfiable. □

A set X is called *recursively enumerable* if there exists an algorithm which lists all the elements of X. A set X is called *decidable* if there exists an algorithm which, given an element, recognizes whether it belongs to X or not (the algorithm must halt on all inputs).

The proof of the following proposition can be found in [CZ97] (Proposition 16.1, p. 492).

Proposition 2.23. *Suppose Y is a decidable set and $X \subseteq Y$. Then X is decidable iff both X and $Y \setminus X$ are recursively enumerable.*

The proof of the following lemma can be found in [CZ97] (Lemma 16.8, p. 495).

Lemma 2.24. *Every logic L defined via a recursively enumerable set of axioms is recursively enumerable.*

Definition 2.25 (Finite model property)**.** *A logic L defined via an axiom system $\mathscr{L}$ has the finite model property (FMP) if, for every formula φ that is not a theorem of $\mathscr{L}$, there is a finite model M of $\mathscr{L}$ where φ is not valid.*

The proof of the following theorem can be found in [BdRV01] (Theorem 6.15, p. 344).

Theorem 2.26. *If L is a finitely axiomatizable normal modal logic with the FMP, then L is decidable.*

2.2 Modeling byzantine fault-tolerant asynchronous message-passing systems – basic concepts

We use $\mathcal{A}$ to denote a (finite) set of agents. Without loss of generality, we assume that the agents are numbered

$$\mathcal{A} := \{1, \ldots, n\},$$

for some integer $n > 1$.

Definition 2.27. *Local timestamps, or simply nodes, are identified by pairs*

$$(i, t) \in \mathcal{A} \times \mathbb{N}_0$$

of an agent i and a timestamp t.

We group all actions and events taking place after timestamp t and no later than $t + 1$ into a round, denoted $t.5$, and treat all actions and events of the round as happening simultaneously.

Each agent begins in one of its *initial states*:

Definition 2.28 (Initial states)**.** *Σ_i denotes the set of local initial states of agent $i \in \mathcal{A}$. A joint initial state is a tuple of local initial states from*

$$\mathscr{G}(0) := \prod_{i \in \mathcal{A}} \Sigma_i.$$

Actions and Events

An agent's state can be modified due to *internal actions* of the agent itself and/or *external events* triggered by the *environment*, represented as a designated agent ϵ, that is *not* considered a member of $\mathcal{A}$.

Definition 2.29 (Local internal actions and local external events)**.** *Int_i denotes the set of all local internal actions of agent $i \in \mathcal{A}$. Ext_i denotes the set of all local external events of agent $i \in \mathcal{A}$. We use*

- *$a, a', a'', \ldots, a_1, a_2, \ldots$, for local internal actions,*
- *$e, e', e'', \ldots, e_1, e_2, \ldots$, for local external events, and*
- *$o, o', o'', \ldots, o_1, o_2, \ldots$, as a generic notation for both.*

We consider *message-passing systems* whereby agents communicate exclusively via messages.

Definition 2.30 (Messages)**.** *We denote by Msgs the (possibly infinite) set of messages that agents can send to each other. For any two agents $i, j \in \mathcal{A}$, a message $\mu \in Msgs$ can be:*

- *sent by agent i to agent j (possibly in multiple copies during the round), which constitutes an internal action of i and is recorded[2] in i's history as $send(j, \mu_k)$ for the kth copy of message μ[3]; we consider $send(j, \mu_0)$ to be the master copy and denote it simply by $send(j, \mu)$ when multiple copies are not necessary;*
- *received by agent j from agent i, which constitutes an external event for j and is recorded in j's history as $recv(i, \mu)$ (j is not aware whether multiple copies of μ have been sent by i to it during the round and cannot tell which copy it has received).*

Remark 2.31. *It is possible that two copies of the same message, possibly sent in different rounds, arrive simultaneously (the receiving agent j would only know that the message μ from agent i is received without being aware of its copies).*

The environment, which also plays the role of the delivery system, is able to distinguish among the multiple copies of the same message. This is modelled by a *global message identifier* $id \in \mathbb{N}_0$, or simply GMI, which can be compared to a tracking number used by the environment to uniquely identify each copy of a message. Agents never observe the GMI.

Definition 2.32 (Global message identifier function)**.** *We fix a function for computing GMIs to be any computable one-to-one total function $id\colon \mathcal{A} \times \mathcal{A} \times Msgs \times \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0$.*[4]

Further, while an agent only observes the messages sent by itself and its own performed actions and observed events, the environment distinguishes between a message μ sent to j from i and the same message μ sent to j by another agent i', between action a performed by i and the same action a performed by j, between event e observed by i and the same event e observed by j.

Definition 2.33 (Global view)**.** *The environment represents*

- *copy $k \in \mathbb{N}_0$ of a message $\mu \in Msgs$ when sent from $i \in \mathcal{A}$ to $j \in \mathcal{A}$ in the format $gsend(i, j, \mu, id)$ with the copy number k transferred to the GMI $id \in \mathbb{N}_0$;*
- *copy $k \in \mathbb{N}_0$ of a message $\mu \in Msgs$ when received by $j \in \mathcal{A}$ from $i \in \mathcal{A}$ in the format $grecv(j, i, \mu, id)$ with the copy number k transferred to the GMI $id \in \mathbb{N}_0$;*

[2] The exception is the case when sending of the message was a byzantine action. Then the record of sending can be missing or corrupted, in particular, it can look like another action (see Definition 2.36).

[3] There is no obligation to use consecutive numbers for copies nor to always use a fresh copy number in following rounds. However, within one round each new copy requires a fresh number. Otherwise, it will be conflated with the same-numbered copy because messages form a set.

[4] A simple though not necessarily the most efficient possibility is to use $2^i \cdot 3^j \cdot 5^{\lceil \mu \rceil} \cdot 7^k \cdot 11^t$, where $\lceil \mu \rceil$ represents the numerical code of the message μ according to some arbitrary but fixed coding scheme.

- *action* $a \in Int_i$ *performed by* $i \in \mathcal{A}$ *in the format* $A = internal\,(i, a)$*;*
- *event* $e \in Ext_i$ *observed by* $i \in \mathcal{A}$ *in the format* $E = external\,(i, e)$*.*

We use

- $A, A', A'', \ldots, A_1, A_2, \ldots,$ *for globally presented local internal actions,*
- $E, E', E'', \ldots, E_1, E_2, \ldots,$ *for globally presented local external events, and*
- $O, O', O'', \ldots, O_1, O_2, \ldots,$ *as a generic notation for both.*

Definition 2.34 (Internal actions)**.** *From the point of view of agent* $i \in \mathcal{A}$*, its correct internal actions consist of the send actions from Definition 2.30 and local internal actions* $a \in Int_i$ *from Definition 2.29:*

$$\overline{Actions_i} := \{send(j, \mu_k) \mid j \in \mathcal{A}, \mu \in Msgs, k \in \mathbb{N}_0\} \sqcup Int_i.$$

The same actions from the point of view of the environment look like:

$$\overline{GActions_i} := \{gsend(i, j, \mu, id) \mid j \in \mathcal{A}, \mu \in Msgs, id \in \mathbb{N}_0\} \sqcup \{internal\,(i, a) \mid a \in Int_i\}.$$

In addition, we abbreviate:

$$\overline{Actions} := \bigcup_{i \in \mathcal{A}} \overline{Actions_i},$$
$$\overline{GActions} := \bigsqcup_{i \in \mathcal{A}} \overline{GActions_i}.$$

The overline in $\overline{Actions_i}$, $\overline{GActions_i}$, $\overline{Actions}$, and $\overline{GActions}$ is used to indicate that the set of considered actions is correct (non-byzantine).

Definition 2.35 (External events)**.** *From the point of view of agent* $i \in \mathcal{A}$*, the correct external events that it can observe consist of the recv events from Definition 2.30 and local external events* $e \in Ext_i$ *from Definition 2.29:*

$$\overline{Events_i} := \{recv(j, \mu) \mid j \in \mathcal{A}, \mu \in Msgs\} \sqcup Ext_i.$$

The same events from the point of view of the environment look like:

$$\overline{GEvents_i} := \{grecv(i, j, \mu, id) \mid j \in \mathcal{A}, \mu \in Msgs, id \in \mathbb{N}_0\} \sqcup \{external\,(i, e) \mid e \in Ext_i\}.$$

In addition, we abbreviate:

$$\overline{Events} := \bigcup_{i \in \mathcal{A}} \overline{Events_i},$$
$$\overline{GEvents} := \bigsqcup_{i \in \mathcal{A}} \overline{GEvents_i}.$$

Just like for actions, the overline in $\overline{Events_i}$, $\overline{GEvents_i}$, $\overline{Events}$, and $\overline{GEvents}$ is used to indicate that the set of considered events is correct (non-byzantine).

Definition 2.36 (Modeling byzantine behaviour)**.** *For each correct external event $E \in \overline{GEvents_i}$ of agent $i \in \mathcal{A}$, we define a matching byzantine external event*

$$fake\,(i, E)$$

representing agent i being mistaken about observing the (local version of the) event E.

*For $A, A' \in \{\boldsymbol{noop}\} \sqcup \overline{GActions_i}$, each of which is either a correct global action of agent i or a **no-op** operation **noop**, we define a matching byzantine external event*

$$fake\,(i, A \mapsto A')$$

representing the situation when agent $i \in \mathcal{A}$ performs (the local version of the) action A but "thinks" (and therefore records) that it performed (the local version of the) action A'.

When agent i faithfully records the performed byzantine action, we abbreviate

$$fake\,(i, A) := fake\,(i, A \mapsto A).$$

Note that $fake\,(i, \boldsymbol{noop})$ acts as a malfunction without any action or any trace in the local history. Hence, we abbreviate

$$fail\,(i) := fake\,(i, \boldsymbol{noop}).$$

In addition, we define the following set:

$$BEvents_i := \{fake\,(i, E) \mid E \in \overline{GEvents_i}\} \sqcup \{fake\,(i, A \mapsto A') \mid A, A' \in \{\boldsymbol{noop}\} \sqcup \overline{GActions_i}\}.$$

Remark 2.37. *We do not impose any* a priori *restrictions on E, A, or A', i.e., a byzantine faulty agent can mistakenly observe any event, mistakenly perform any action, and mistake any performed action or inaction for any other action or for inaction.*

Remark 2.38 (GMIs for byzantine messages)**.** *For byzantine received messages, the environment creates the message "out of thin air". Such a message will be supplied with a GMI for uniformity's sake but it is not assumed to carry any information. Similarly, a byzantine sent message is created already with a well-formed GMI (unlike the correctly sent messages, which are supplied with a GMI in a separate step after creation). At the same time, if the agent is mistaken about having sent a message, its GMI is immaterial as the environment will never deliver this "message".*

Definition 2.39 (System events)**.** *For agent $i \in \mathcal{A}$, we define the set of system events*

$$SysEvents_i := \{go(i), sleep\,(i), hibernate\,(i)\},$$

where:

- *$go(i)$ wakes i up and prompts it to act according to its protocol;*

- *sleep* (i) *wakes* i *up but prevents it from acting and makes it byzantine faulty;*
- *hibernate* (i) *prevents* i *from acting or from waking up and makes it byzantine faulty.*

For each round, the environment determines whether an agent is to be awoken to follow its protocol and/or observe some external events or is to skip the round.

Remark 2.40. *None of the system events are recorded by the agents.*

We abbreviate

$$FEvents_i := BEvents_i \sqcup \{sleep\,(i), hibernate\,(i)\}.$$

The complete set of events affecting agent $i \in \mathcal{A}$ is:

$$GEvents_i := \overline{GEvents}_i \sqcup BEvents_i \sqcup SysEvents_i.$$

In addition, we abbreviate:

$$GEvents := \bigsqcup_{i \in \mathcal{A}} GEvents_i.$$

Protocols

Definition 2.41 (Local states)**.** *A local history* h_i *of agent* $i \in \mathcal{A}$*, or its local state, is a nonempty sequence*

$$h_i = [\lambda_m, \ldots, \lambda_1, \lambda_0],$$

for some $m \geq 0$*, such that* $\lambda_0 \in \Sigma_i$ *and* $\forall j \in \{1, \ldots, m\}$ *we have* $\lambda_j \subset \overline{Actions}_i \sqcup \overline{Events}_i$*. In this case,* m *is called the length of the history* h_i *and is denoted* $|h_i|$*.*

In addition, we use $\mathscr{L}_i$ *to denote the set of local states of agent* $i \in \mathcal{A}$*, i.e., the set of all local histories of agent* $i \in \mathcal{A}$*.*

Remark 2.42. *Note that agents' local states consist of correct internal actions and correct external events — whether or not those actions and events have been actually performed/observed in a byzantine fashion is visible only in the history of the environment.*

Remark 2.43 (Perfect recall)**.** *We consider agents capable of perfect recall. Perfect recall is the assumption that at all times an agent remember everything, or in other words, that at any time the local state of an agent contains a record of all its previous local states.*

Definition 2.44 (Global state)**.** *A global history* h *of the system, or the global state, is a tuple*

$$h := (h_\epsilon, h_1, \ldots, h_n),$$

where the history of the environment ϵ *is a sequence*

$$h_\epsilon = [\Lambda_m, \ldots, \Lambda_1, \Lambda_0],$$

for some $m \geq 0$, such that $\forall j \in \{1, \ldots, m\}$ we have $\Lambda_j \subseteq \overline{GActions} \sqcup GEvents$ and h_i is a local state of each agent $i \in \mathcal{A}$. In this case m is called the length of the history h and is denoted $|h| := |h_\epsilon|$.

In addition, we use $\mathscr{L}_\epsilon$ to denote the set of all histories of the environment, and $\mathscr{G}$ to denote the set of global states, i.e., the set of all global histories of the system.

Definition 2.45. *A (non-deterministic) protocol for agent $i \in \mathcal{A}$ is any function*

$$P_i \colon \mathscr{L}_i \to 2^{2^{\overline{Actions_i}}} \setminus \{\varnothing\}.$$

By $P = (P_1, \ldots, P_n)$ we denote agents' joint protocol.

Remark 2.46. *Agent i's protocol P_i can only rely on i's local state at any given moment. In particular, it is crucial for modeling asynchronous agents that agent i's protocol P_i does not use a timestamp t as a parameter.*

Note that, for a local state $h_i \in \mathscr{L}_i$ of agent $i \in \mathcal{A}$, each member $S \in P_i(h_i)$ is a subset of $\overline{Actions_i}$ and represents one of the non-deterministic choices for i's actions. In case of multiple options, the choice is up to the *adversary* part of the environment. $P_i(h_i) \neq \varnothing$ means that there is always at least one such choice S for i, which might be to perform no actions if $S = \varnothing$.

Definition 2.47 (Coherent sets of events)**.** *Let $t \in \mathbb{N}_0$ be a timestamp. A set $S \subset GEvents$ of events is called t-coherent if it satisfies the following conditions:*

1. *for any fake$(i, gsend(i, j, \mu, id) \mapsto A) \in S$, the GMI $id = id(i, j, \mu, k, t)$ for some $k \in \mathbb{N}_0$[5];*
2. *for any $i \in \mathcal{A}$, at most one of system events $go(i)$, sleep(i), and hibernate(i) is present in S;*
3. *for any $i \in \mathcal{A}$ and any $e \in Ext_i$, at most one of events external(i, e) and fake$(i,$ external$(i, e))$ is present in S;*
4. *for any $grecv(i, j, \mu, id_1) \in S$, no event of the form fake$(i, grecv(i, j, \mu, id_2))$ belongs to S for any $id_2 \in \mathbb{N}_0$;*
5. *for any fake$(i, grecv(i, j, \mu, id_1)) \in S$, no event of the form $grecv(i, j, \mu, id_2)$ belongs to S for any $id_2 \in \mathbb{N}_0$;*

Remark 2.48. *We assume that the environment never attempts simultaneously a correct and fake external event that leave the same trace in agent's local history.*

[5]We assume that all messages actually sent, whether correctly or otherwise, are treated by the environment in the same way, i.e, the environment always assigns a GMI.

Definition 2.49. *A (non-deterministic) protocol for the environment is any function*

$$P_\epsilon \colon \mathbb{N}_0 \longrightarrow 2^{2^{GEvents}} \setminus \{\varnothing\}$$

such that every set $S \in P_\epsilon(t)$ is t-coherent.

Remark 2.50. *The dependence of the environment's protocol P_ϵ on time enables modeling of time-sensitive actions. For instance, we become able to model global prohibition on message delivery during designated quiet time. At the same time, the environment's protocol P_ϵ should not depend on the global state at any given moment to preserve the unbiased representation of the physical laws.*

Note that, for each $t \in \mathbb{N}_0$, each member $S \in P_\epsilon(t)$ is a t-coherent subset of $GEvents$, i.e.,

$$\begin{aligned} S \quad \subset \quad & \{grecv(i,j,\mu,id) \mid i,j \in \mathcal{A}, \mu \in Msgs, id \in \mathbb{N}_0\} \quad \sqcup \quad \{go(i) \mid i \in \mathcal{A}\} \quad \sqcup \\ & \{external(i,e) \mid i \in \mathcal{A}, e \in Ext_i\} \quad \sqcup \quad \{sleep(i) \mid i \in \mathcal{A}\} \quad \sqcup \\ & \{fake(i,E) \mid i \in \mathcal{A}, E \in \overline{GEvents_i}\} \quad \sqcup \quad \{hibernate(i) \mid i \in \mathcal{A}\} \quad \sqcup \\ & \{fake(i, A \mapsto A') \mid i \in \mathcal{A}, A, A' \in \{\mathbf{noop}\} \quad \sqcup \quad \overline{GActions_i}\}, \end{aligned}$$

and represents one of the non-deterministic choices for the events to be imposed by the environment. In case of multiple options, the choice is up to the *adversary* part of the environment. $P_\epsilon(t) \neq \varnothing$ means that there is always at least one such choice S for the environment, which might be to impose no events if $S = \varnothing$.

Different types of agents

In our framework, we distinguish between a number of types of agents. We list below the ones that appear in this thesis:

Definition 2.51. *Environment's protocol P_ϵ makes an agent $i \in \mathcal{A}$:*

1. *delayable if for any $X \in P_\epsilon(t)$,*

$$X \setminus GEvents_i \in P_\epsilon(t).$$

In other words, all activities of a delayable agent can be correctly postponed at any time.

2. *fallible if for any $X \in P_\epsilon(t)$,*

$$X \cup \{fail(i)\} \in P_\epsilon(t).$$

In other words, a fallible agent can fail at any time (which implies that it can become byzantine faulty at any time).

3. *gullible if for any* $Y \subseteq FEvents_i$, *and any* $X \in P_\epsilon(t)$,

$$Y \sqcup (X \setminus GEvents_i) \in P_\epsilon(t), \quad \text{whenever } Y \sqcup (X \setminus GEvents_i) \text{ is } t\text{-coherent.}$$

In other words, all activities of a gullible agent can be replaced with an arbitrary set of faulty events at any time (which implies that it can become byzantine faulty at any time).

Labeling functions

Definition 2.52. *For an agent* $i \in \mathcal{A}$, *we define a labeling function*

$$label_i \colon \overline{Actions_i} \times \mathbb{N} \longrightarrow \overline{GActions_i}$$

converting the local format of i*'s actions into the global format as follows:*

$$label_i(a,t) := \begin{cases} gsend(i,j,\mu,id(i,j,\mu,k,t)) & \text{if } a = send(j,\mu_k) \\ internal(i,a) & \text{if } a \in Int_i \end{cases}$$

We collect all these functions into one tuple $label := (label_1, \ldots, label_n)$.

We also define the *"reverse" labeling function* $label^{-1}$:

Definition 2.53. *The "reverse" labeling function* $label^{-1}$

$$label^{-1} \colon \overline{GActions} \sqcup \overline{GEvents} \longrightarrow \overline{Actions} \sqcup \overline{Events}$$

is defined as follows:

$$label^{-1}(U) := \begin{cases} send(j,\mu_k) & \text{if } U = gsend(i,j,\mu,id(i,j,\mu,k,t)) \\ send(j,\mu_0) & \text{if } U = gsend(i,j,\mu,M) \text{ and } M \neq id(i,j,\mu,k,t)\ k,t \in \mathbb{N}_0 \\ recv(j,\mu) & \text{if } U = grecv(i,j,\mu,id) \\ a & \text{if } U = internal(i,a) \\ e & \text{if } U = external(i,e) \end{cases}$$

The function $label^{-1}$ *extends to sets in the standard way:*

$$label^{-1}(X) := \{label^{-1}(U) \mid U \in X\}.$$

Remark 2.54. *Note that the labeling functions deal with correct internal actions and correct external events only. This is because byzantine behaviour was defined only using the global format.*

Remark 2.55. *The injectivity of the function id used in* $label_i$ *ensures that each message is unique from the point of view of the environment.*

Remark 2.56. *The second clause in the definition of* $label^{-1}(U)$ *is mostly cosmetic: we make GMIs id unforgeable, and, hence, this clause will never be used. It is added solely to make the function* $label^{-1}(U)$ *total, thus, avoiding irrelevant complications stemming from the use of potentially partial functions.*

Given a timestamp $t \in \mathbb{N}_0$, a global history $h = (h_\epsilon, h_1, \ldots, h_n) \in \mathscr{G}$ and protocols P_ϵ for the environment and $P_1, \ldots, P_n$ for the agents, the sets of actions and events to be attempted in round $t.5$ are obtained as described below:

1. Events to be imposed by the environment form a t-coherent set
$$\alpha_\epsilon^t := X_\epsilon \tag{2.1}$$
for some set $X_\epsilon \in P_\epsilon(t)$ non-deterministically chosen by the adversary.
2. Actions to be performed by agent $i \in \mathcal{A}$ form a set
$$\alpha_i^{h,t} := label_i(X_i, t) \tag{2.2}$$
for some set $X_i \in P_i(h_i)$ non-deterministically chosen by the adversary.
3. The choices from 1. and 2. are combined in the *joint attempted actions/events*
$$\alpha^{h,t} := (\alpha_\epsilon^t, \alpha_1^{h,t}, \ldots, \alpha_n^{h,t}).$$

Among the events α_ϵ^t we distinguish between the following subsets:

- Correct external events (to be imposed by the environment) for agent $i \in \mathcal{A}$
$$\overline{\alpha}_{\epsilon_i}^t := \alpha_\epsilon^t \cap \overline{GEvents_i} = \{grecv(i, j, \mu, id) \in \alpha_\epsilon^t \mid j \in \mathcal{A}, \mu \in Msgs, id \in \mathbb{N}_0\} \sqcup \{external(i, e) \in \alpha_\epsilon^t \mid e \in Ext_i\}, \tag{2.3}$$
- Instructions regarding waking up agent $i \in \mathcal{A}$
$$\alpha_{g_i}^t := \alpha_\epsilon^t \cap SysEvents_i, \tag{2.4}$$
- Byzantine external events (to be imposed by the environment) for agent $i \in \mathcal{A}$
$$\alpha_{b_i}^t := \alpha_\epsilon^t \cap BEvents_i = \left\{fake(i, A \mapsto A') \in \alpha_\epsilon^t \mid A, A' \in \{\mathbf{noop}\} \sqcup \overline{GActions_i}\right\} \sqcup \{fake(i, E) \in \alpha_\epsilon^t \mid E \in \overline{GEvents_i}\}, \tag{2.5}$$
- Instructions making agent $i \in \mathcal{A}$ byzantine faulty
$$\alpha_{f_i}^t := \alpha_{b_i}^t \sqcup \left(\alpha_\epsilon^t \cap \{sleep(i), hibernate(i)\}\right). \tag{2.6}$$

Remark 2.57. *Note that sleep*(i) *and hibernate*(i) *may be present in both* $\alpha^t_{g_i}$ *and* $\alpha^t_{f_i}$.

In addition, we abbreviate:

$$\overline{\alpha}^t_\epsilon := \bigsqcup_{i\in\mathcal{A}} \overline{\alpha}^t_{\epsilon_i},$$
$$\alpha^t_g := \bigsqcup_{i\in\mathcal{A}} \alpha^t_{g_i},$$
$$\alpha^t_b := \bigsqcup_{i\in\mathcal{A}} \alpha^t_{b_i},$$
$$\alpha^t_f := \bigsqcup_{i\in\mathcal{A}} \alpha^t_{f_i}.$$

Filter functions

We assume that the environment does not create impossible situations. Most of them are implicitly prohibited by the definition of the environment's protocol (via t-coherent sets). There is, however, one common type of causal impossibility that is not excluded by definition: a message *cannot* be delivered correctly without being previously[6] sent. Since the environment's protocol is independent of the global history, the environment cannot check whether the message in question was actually sent. Therefore, we create a special *filter function* that weeds out such situations.

Firstly, to simplify notation, we introduce the following abbreviations:

Definition 2.58 (Active/passive, aware/unaware)**.** *For a set* $X \subseteq GEvents$*, we define*

$$active(i,X) := \begin{cases} t & \text{if } X \cap SysEvents_i = \{go(i)\}, \\ f & \text{otherwise.} \end{cases}$$
$$aware(i,X) := \begin{cases} t & \text{if } \varnothing \neq X \cap SysEvents_i \in \{\{go(i)\}, \{sleep\,(i)\}\}, \\ f & \text{otherwise.} \end{cases}$$

For readability's sake we write $active(i,X)$ *instead of* $active(i,X) = t$ *and* $passive(i,X)$ *instead of* $active(i,X) = f$*, as well as* $aware(i,X)$ *instead of* $aware(i,X) = t$ *and* $unaware(i,X)$ *instead of* $aware(i,X) = f$.

Definition 2.59 (Byzantine filter functions)**.** *We define the byzantine event filter function*

$$filter^{B}_\epsilon{}^{7} \colon \mathscr{G} \times 2^{GEvents} \times 2^{\overline{GActions}_1} \times \cdots \times 2^{\overline{GActions}_n} \longrightarrow 2^{GEvents}$$

[6] Here *previously* sent means sent in one of the preceding rounds or in the same round (based in part on the actions chosen by the adversary for the sending agent, the presence of the *go* command for it, and other events imposed on this agent), whether correctly or in a byzantine fashion.

[7] We use 'B' to indicate that the event filter function in question is byzantine.

as follows: for a a global history $h = (h_\epsilon, h_1, \ldots, h_n) \in \mathscr{G}$*, a set* $X_\epsilon \subset GEvents$*, and sets* $X_i \subset \overline{GActions_i}$ *for each agent* $i \in \mathcal{A}$*, we define*

$$\begin{aligned} filter_\epsilon^B\,(h, X_\epsilon, X_1, \ldots, X_n) \quad := \\ X_\epsilon \quad \setminus \quad \Big\{ grecv(j,i,\mu,id) \quad | \quad gsend(i,j,\mu,id) \notin h_\epsilon \quad \wedge \\ (\forall A \in \{\textbf{\textit{noop}}\} \sqcup \overline{GActions_i})\, fake\,(i, gsend(i,j,\mu,id) \mapsto A) \notin h_\epsilon \quad \wedge \\ (gsend(i,j,\mu,id) \notin X_i \vee passive(i, X_\epsilon)) \quad \wedge \\ (\forall A \in \{\textbf{\textit{noop}}\} \sqcup \overline{GActions_i})\, fake\,(i, gsend(i,j,\mu,id) \mapsto A) \notin X_\epsilon \Big\}. \end{aligned} \tag{2.7}$$

In addition, we define the byzantine action filter function for agent $i \in \mathcal{A}$

$$filter_i^B{}^{8} : 2^{\overline{GActions_1}} \times \cdots \times 2^{\overline{GActions_n}} \times 2^{GEvents} \longrightarrow 2^{\overline{GActions_i}}$$

as follows: for sets $X_j \subset \overline{GActions_j}$ *for each agent* $j \in \mathcal{A}$ *and a set* $X_\epsilon \subset GEvents$*, we define*

$$filter_i^B\,(X_1, \ldots, X_n, X_\epsilon) = \begin{cases} X_i & \text{if } active(i, X_\epsilon) \\ \varnothing & \text{otherwise} \end{cases}. \tag{2.8}$$

Thus, after the adversary chose the collection α_ϵ^t (2.1) of events to be imposed by the environment and collections $\alpha_i^{h,t}$ (2.2) of actions to be performed by each agent $i \in \mathcal{A}$, the filter functions determine which of these actions and events are to actually happen during the round $t.5$. For this second stage, the resulting sets are called β-sets by analogy with α-sets.

Definition 2.60. *For a global history* $h \in \mathscr{G}$*, a timestamp* $t \in \mathbb{N}_0$*, a tuple of joint attempted actions/events* $\alpha^{h,t} = (\alpha_\epsilon^t, \alpha_1^{h,t}, \ldots, \alpha_n^{h,t})$*, and agent* $i \in \mathcal{A}$*:*

1. $\beta_\epsilon^{h,\alpha^{h,t}} := filter_\epsilon^B \left(h, \alpha_\epsilon^t, \alpha_1^{h,t}, \ldots, \alpha_n^{h,t}\right)$;
2. $\beta_i^{h,\alpha^{h,t}} := filter_i^B \left(\alpha_1^{h,t}, \ldots, \alpha_n^{h,t}, \beta_\epsilon^{h,\alpha^{h,t}}\right)$;
3. $\beta^{h,\alpha^{h,t}} := (\beta_\epsilon^{h,\alpha^{h,t}}, \beta_1^{h,\alpha^{h,t}}, \ldots, \beta_n^{h,\alpha^{h,t}})$.

As for α_ϵ^t, we also distinguish between the following subsets of $\beta_\epsilon^{h,\alpha^{h,t}}$:

1. Correct external events for agent $i \in \mathcal{A}$

$$\begin{aligned} \overline{\beta}_{\epsilon_i}^{h,\alpha^{h,t}} := \beta_\epsilon^{h,\alpha^{h,t}} \cap \overline{GEvents_i} = \\ \{grecv(i,j,\mu,id) \in \beta_\epsilon^{h,\alpha^{h,t}} \mid j \in \mathcal{A}, \mu \in Msgs, id \in \mathbb{N}_0\} \sqcup \\ \{external\,(i,e) \in \beta_\epsilon^{h,\alpha^{h,t}} \mid e \in Ext_i\} \subset \overline{\alpha}_{\epsilon_i}^t, \end{aligned} \tag{2.9}$$

[8] We use 'B' to indicate that the action filter function in question is byzantine.

2. Instructions regarding waking up agent $i \in \mathcal{A}$

$$\beta_{g_i}^{h,\alpha^{h,t}} := \beta_{\epsilon}^{h,\alpha^{h,t}} \cap \mathit{SysEvents}_i \subset \alpha_{g_i}^t, \tag{2.10}$$

3. Byzantine external events for agent $i \in \mathcal{A}$

$$\begin{aligned}\beta_{b_i}^{h,\alpha^{h,t}} := \beta_{\epsilon}^{h,\alpha^{h,t}} \cap \mathit{BEvents}_i = \\ \{\mathit{fake}\,(i, A \mapsto A') \in \beta_{\epsilon}^{h,\alpha^{h,t}} \mid A, A' \in \{\mathbf{noop}\} \sqcup \overline{\mathit{GActions}_i}\} \sqcup \\ \{\mathit{fake}\,(i, E) \in \beta_{\epsilon}^{h,\alpha^{h,t}} \mid E \in \overline{\mathit{GEvents}_i}\} \subset \alpha_{b_i}^t,\end{aligned} \tag{2.11}$$

4. Instructions making agent $i \in \mathcal{A}$ byzantine faulty

$$\beta_{f_i}^{h,\alpha^{h,t}} := \beta_{b_i}^{h,\alpha^{h,t}} \sqcup \left(\beta_{\epsilon}^{h,\alpha^{h,t}} \cap \{\mathit{sleep}\,(i), \mathit{hibernate}\,(i)\}\right) \subset \alpha_{f_i}^t. \tag{2.12}$$

Remark 2.61. *Note that* $\mathit{sleep}\,(i)$ *and* $\mathit{hibernate}\,(i)$ *may be present in both* $\beta_{g_i}^{h,\alpha^{h,t}}$ *and* $\beta_{f_i}^{h,\alpha^{h,t}}$.

In addition, we abbreviate:

$$\begin{aligned}\overline{\beta}_{\epsilon}^{h,\alpha^{h,t}} &:= \bigsqcup_{i \in A} \overline{\beta}_{\epsilon_i}^{h,\alpha^{h,t}} \subset \overline{\alpha}_{\epsilon}^t, \\ \beta_{g}^{h,\alpha^{h,t}} &:= \bigsqcup_{i \in A} \beta_{g_i}^{h,\alpha^{h,t}} \subset \alpha_{g}^t, \\ \beta_{b}^{h,\alpha^{h,t}} &:= \bigsqcup_{i \in A} \beta_{b_i}^{h,\alpha^{h,t}} \subset \alpha_{b}^t, \\ \beta_{f}^{h,\alpha^{h,t}} &:= \bigsqcup_{i \in A} \beta_{f_i}^{h,\alpha^{h,t}} \subset \alpha_{f}^t.\end{aligned}$$

Update functions

One of our central assumptions is that the agents are not able to tell the difference between correct and byzantine actions and events. For example, they are not able to tell the difference between observing an external event E and (mistakenly) thinking they have observed E, as represented by $\mathit{fake}\,(i, E)$, nor between performing an internal action A' and thinking they have performed A' when A was the internal action they actually performed, as represented by $\mathit{fake}\,(i, A \mapsto A')$.

Formally, agents' local states are purged of

(1) *fake* modifiers,

(2) GMIs,

(3) system commands $\mathit{go}(i)$, $\mathit{sleep}\,(i)$, and $\mathit{hibernate}\,(i)$.

This is performed by the *localization function* σ:

Definition 2.62 (Localization function)**.** *The function*

$$\sigma\colon 2^{\overline{GActions} \sqcup GEvents} \longrightarrow 2^{\overline{Actions} \sqcup \overline{Events}}$$

is defined as follows:

$$\sigma(X) := label^{-1}\Big(\big(X \cap (\overline{GActions} \sqcup \overline{GEvents})\big) \quad \cup \\ \{E \mid (\exists i)\, fake\,(i, E) \in X\} \quad \cup \\ \{A' \neq \textbf{\textit{noop}} \mid (\exists i)(\exists A)\, fake\,(i, A \mapsto A') \in X\}\Big),$$

where $label^{-1}$ is the "reverse" labeling function defined in Definition 2.53.

Remark 2.63. *Note that byzantine external events of the form $fake\,(i, A \mapsto \textbf{\textit{noop}})$ leave no trace in i's local state.*

Finally, we define state update functions that record the performed actions and events of a round into all the histories:

Definition 2.64 (State update functions)**.** *Given a global history $h = (h_\epsilon, h_1, \ldots, h_n) \in \mathscr{G}$, and a tuple of performed actions and events $X := (X_\epsilon, X_1, \ldots, X_n) \in 2^{GEvents} \times 2^{\overline{GActions_1}} \times \ldots \times 2^{\overline{GActions_n}}$, agent i's state update function*

$$update_i\colon \mathscr{L}_i \times 2^{\overline{GActions_i}} \times 2^{GEvents} \to \mathscr{L}_i$$

outputs a new local history from $\mathscr{L}_i$ based on i's performed actions X_i and events X_ϵ as follows:

$$update_i\,(h_i, X_i, X_\epsilon) := \begin{cases} h_i & \text{if } \sigma(X_{\epsilon_i}) = \varnothing \text{ and } unaware(i, X_\epsilon) \\ \Big[\sigma\big(X_i \sqcup X_{\epsilon_i}\big)\Big] : h_i & \text{otherwise} \end{cases},$$

where $X_{\epsilon_i} := X_\epsilon \cap GEvents_i$, and $:$ represents sequence concatenation.

Similarly, the environment's state update function

$$update_\epsilon\colon \mathscr{L}_\epsilon \times 2^{GEvents} \times 2^{\overline{GActions_1}} \times \ldots \times 2^{\overline{GActions_n}} \to \mathscr{L}_\epsilon$$

outputs a new history of the environment from $\mathscr{L}_\epsilon$ based on all performed actions and events $X = (X_\epsilon, X_1, \ldots, X_n)$:

$$update_\epsilon\,(h_\epsilon, X) := (X_\epsilon \sqcup X_1 \sqcup \ldots \sqcup X_n)\colon h_\epsilon.$$

Thus, the global state is modified as follows:

$$update\,(h, X) := \Big(update_\epsilon\,(h_\epsilon, X)\,, update_1\,(h_1, X_1, X_\epsilon)\,, \ldots, update_n\,(h_n, X_n, X_\epsilon)\Big).$$

Transition function

Definition 2.65 (Byzantine transition function)**.** *For an agents' joint protocol* $P = (P_1, \ldots, P_n)$ *and a protocol* P_ϵ *of the environment, we define a byzantine transition function*

$$\tau^B_{P_\epsilon,P}{}^{9} \colon 2^{GEvents} \times 2^{\overline{GActions}_1} \times \ldots \times 2^{\overline{GActions}_n} \to (\mathscr{G} \to \mathscr{G})$$

as a function that outputs a global state transformer function

$$\tau^B_{P_\epsilon,P}(Y) \colon \mathscr{G} \to \mathscr{G}$$

from global states to global states given joint attempted actions/events

$$Y \in 2^{GEvents} \times 2^{\overline{GActions}_1} \times \ldots \times 2^{\overline{GActions}_n}$$

defined as follows: for a global state $h = (h_\epsilon, h_1, \ldots, h_n) \in \mathscr{G}$*, we consider two possibilities*

- *if* $Y = \alpha^{h,|h|} = \left(\alpha_\epsilon^{|h|}, \alpha_1^{h,|h|}, \ldots, \alpha_n^{h,|h|}\right)$ *for some* $\alpha_\epsilon^{|h|} \in P_\epsilon(|h|)$ *and some* $X_i \in P_i(h_i)$ *for each* $i \in \mathcal{A}$ *such that* $\alpha_i^{h,|h|} = label_i(X_i, |h|)$*, we define*

$$\tau^B_{P_\epsilon,P}(Y)(h) := update\left(h, \beta^{h,\alpha^{h,|h|}}\right), \tag{2.13}$$

  where the β*-sets are computed from* $\alpha^{h,|h|}$ *according to Definition 2.60;*

- *otherwise, we define* $\tau^B_{P_\epsilon,P}(Y)(h) = h$.[10]

Remark 2.66. *By a slight abuse of notation, we write* $h' \in \tau^B_{P_\epsilon,P}(h)$ *to mean that there is a protocol-conformant set of joint attempted actions/events* $\alpha^{h,|h|}$ *satisfying the first clause of the above definition such that* $\tau^B_{P_\epsilon,P}(\alpha^{h,|h|})(h) = h'$*.*

More generally:

Definition 2.67 (Transition template and transition function)**.** *Let* $\mathscr{C}_\epsilon$ *be the set of all environment protocols and* $\mathscr{C}$ *be the set of all agents' joint protocols. A transition template*

$$\tau \colon \mathscr{C}_\epsilon \times \mathscr{C} \to \left(2^{GEvents} \times 2^{\overline{GActions}_1} \times \cdots \times 2^{\overline{GActions}_n} \to (\mathscr{G} \to \mathscr{G})\right)$$

is a two-place function that takes a protocol $P_\epsilon \in \mathscr{C}_\epsilon$ *of the environment and an agents' joint protocol* $P \in \mathscr{C}$ *and outputs a transition function* $\tau(P_\epsilon, P)$*, which we denote by* $\tau_{P_\epsilon,P}$

$$\tau_{P_\epsilon,P} \colon 2^{GEvents} \times 2^{\overline{GActions}_1} \times \cdots \times 2^{\overline{GActions}_n} \to (\mathscr{G} \to \mathscr{G}).$$

Thus, $\tau^B_{P_\epsilon,P}$ is only an instance of a transition function.

[9] We use 'B' to indicate that the transition function in question is byzantine.

[10] The latter case will never be used and is only provided to make the transition function total.

Transitional runs

Definition 2.68 (Run)**.** *A run is a function that assigns a global state to each timestamp*

$$r : \mathbb{N}_0 \longrightarrow \mathscr{G}.$$

We denote the set of all possible runs by R.

The part of a run $r \in R$ that an agent $i \in \mathcal{A}$ can see is called i's local view. It is a function that assigns i's local state to each timestamp

$$r_i \colon \mathbb{N}_0 \longrightarrow \mathscr{L}_i.$$

Similarly, we define the environment's view to be a function that assigns the environment's history to each timestamp

$$r_\epsilon \colon \mathbb{N}_0 \longrightarrow \mathscr{L}_\epsilon.$$

Given an agents' joint protocol P and an environment's protocol P_ϵ, we are usually interested in runs $r \in R$ that are built according to these protocols by a transition function $\tau_{P_\epsilon,P}$ defined in Definition 2.67.

Definition 2.69 (Transitional run)**.** *A run $r \in R$ is called $\tau_{P_\epsilon,P}$-transitional, or simply transitional if, for each timestamp $t \in \mathbb{N}_0$:*

$$r\,(t+1) \in \tau_{P_\epsilon,P}\,(r\,(t))\,.$$

For a transitional run r, we denote its initial state by $r\,(0)$ and the global state after the round $(t-1).5$ by $r\,(t)$.

One step of a $\tau^B_{P_\epsilon,P}$-transition for runs

In the interest of generality and modularity of concepts, we defined the byzantine transition function (see Definition 2.65) in terms of arbitrary histories. For the case of histories comprising a transitional run, the notation can be simplified, as we will see below.

Figure 2.3 represents one step of a transition according to $\tau^B_{P_\epsilon,P}$, which consists of the following five consecutive phases:

1. **Protocol phase** (note that the protocols are explicit arguments to the transition template τ): First, the protocol P_i for each agent $i \in \mathcal{A}$ lays out a range $P_i\,(r_i\,(t))$ of possible sets of i's actions for the round $t.5$ based on i's local state $r_i\,(t)$. Similarly, the protocol P_ϵ of the environment lays out a range $P_\epsilon\,(t)$ of possible (t-coherent) sets of events for the round $t.5$ based on time t.

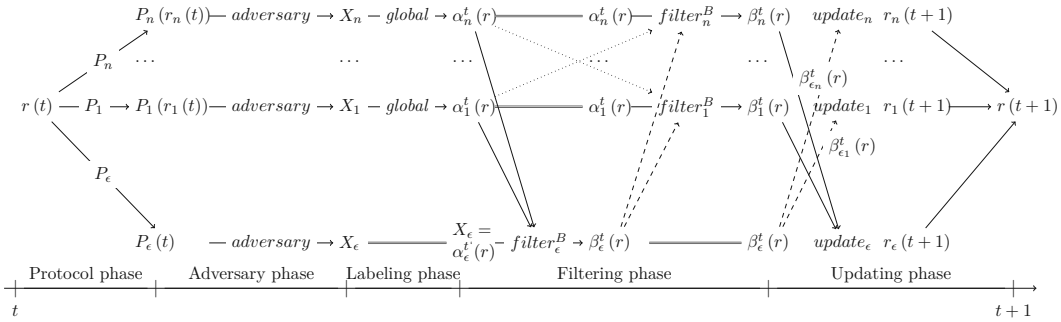


Figure 2.3: Details of round $t.5$ of a $\tau^B_{P_\epsilon,P}$-transitional run r.

2. **Adversary phase** (note that this phase is stable: it does not change from template to template or from protocols to protocols): The adversary non-deterministically picks one set

$$X_i \in P_i\left(r_i\left(t\right)\right) \tag{2.14}$$

of actions for each agent $i \in \mathcal{A}$ and a set

$$X_\epsilon \in P_\epsilon\left(t\right) \tag{2.15}$$

of events for the environment. These are the actions the agents intend to perform and the events the environment intends to impose in the round $t.5$.

Note that $X_i \subset \overline{Actions_i}$ and $X_\epsilon \subset GEvents$.

3. **Labeling phase** (note that this phase is stable: it does not change from template to template or from protocols to protocols): The environment processes the intended actions X_i of each agent $i \in \mathcal{A}$ by converting them into their corresponding global formats, in particular, it assigns GMIs to message send requests.

We denote the resulting sets in the following way:

$$\alpha_i^t\left(r\right) := label_i\left(X_i, t\right), \tag{2.16}$$

where $label_i$ is the labeling function defined in Definition 2.52. The set of intended events X_ϵ of the environment is already in the global format and requires no converting:

$$\alpha_\epsilon^t\left(r\right) := X_\epsilon. \tag{2.17}$$

Note that $\alpha_i^t\left(r\right) \subset \overline{GActions_i}$ and $\alpha_\epsilon^t\left(r\right) \subset GEvents$.

4. **Filtering phase** (note that this phase depends on the filtering functions $filter_\epsilon^B$ and $filter_i^B$, which are considered to be part of the template): In this phase, the intended actions of the agents and events of the environment that are deemed "causally impossible" are filtered out.

The filtering phase is divided into two subphases:

a) first, impossible intended events of the environment are filtered out by the byzantine event filter function $filter_{\epsilon}^{B}$ defined in Definition 2.59 (based on $\alpha_{\epsilon}^{t}(r)$ and $\alpha_{i}^{t}(r)$ for each $i \in \mathcal{A}$), resulting in set $\beta_{\epsilon}^{t}(r)$:

$$\beta_{\epsilon}^{t}(r) := filter_{\epsilon}^{B}\left(r(t), \alpha_{\epsilon}^{t}(r), \alpha_{1}^{t}(r), \ldots, \alpha_{n}^{t}(r)\right), \tag{2.18}$$

b) then, for each agent $i \in \mathcal{A}$, the byzantine action filter function $filter_{i}^{B}$ defined in Definition 2.59 performs the same task on agents' intended actions by taking into account the already filtered out events, resulting in set $\beta_{i}^{t}(r)$:

$$\beta_{i}^{t}(r) := filter_{i}^{B}\left(\alpha_{1}^{t}(r), \ldots, \alpha_{n}^{t}(r), \beta_{\epsilon}^{t}(r)\right). \tag{2.19}$$

Note that $\beta_{i}^{t}(r) \subset \alpha_{i}^{t}(r) \subset \overline{GActions_{i}}$ and $\beta_{\epsilon}^{t}(r) \subset \alpha_{\epsilon}^{t}(r) \subset GEvents$.

In compliance with (2.9), (2.10), (2.11), and (2.12), we denote by:

- $\overline{\beta}_{\epsilon_i}^{t}(r)$ the correct external events observed by agent $i \in \mathcal{A}$, i.e.,

$$\overline{\beta}_{\epsilon_i}^{t}(r) := \beta_{\epsilon}^{t}(r) \cap \overline{GEvents_{i}}; \tag{2.20}$$

- $\beta_{g_i}^{t}(r)$ the system events imposed on agent $i \in \mathcal{A}$, i.e.,

$$\beta_{g_i}^{t}(r) := \beta_{\epsilon}^{t}(r) \cap SysEvents_{i}; \tag{2.21}$$

- $\beta_{b_i}^{t}(r)$ the byzantine external events for agent $i \in \mathcal{A}$, i.e.,

$$\beta_{b_i}^{t}(r) := \beta_{\epsilon}^{t}(r) \cap BEvents_{i}; \tag{2.22}$$

- $\beta_{f_i}^{t}(r)$ the faulty events observed/imposed specifically by/on agent $i \in \mathcal{A}$, i.e.,

$$\beta_{f_i}^{t}(r) := \beta_{\epsilon}^{t}(r) \cap FEvents_{i}. \tag{2.23}$$

5. **Updating phase** (note that this phase is stable: it does not change from template to template or from protocols to protocols): The actions $\beta_{i}^{t}(r)$ and events $\beta_{\epsilon}^{t}(r)$ actually happening in the round $t.5$ are faithfully recorded into the environment's history and are translated into their local format for being recorded into agents' local histories using state update functions defined in Definition 2.64.

   Once again, the local state of each agent $i \in \mathcal{A}$ is only affected by the actions $\beta_{i}^{t}(r)$ it performs and the events $\beta_{\epsilon_i}^{t}(r)$ it observes, whereas the global state is modified

based on the complete information about all actions and events performed in the round t.5.

Therefore:

$$r_i(t+1) := update_i\left(r_i(t), \beta_i^t(r), \beta_\epsilon^t(r)\right), \tag{2.24}$$

$$\beta^t(r) := \left(\beta_\epsilon^t(r), \beta_1^t(r), \ldots, \beta_n^t(r)\right), \tag{2.25}$$

$$r_\epsilon(t+1) := update_\epsilon\left(r_\epsilon(t), \beta^t(r)\right). \tag{2.26}$$

Agent-context

While it is preferrable to directly build the desired properties of runs into the transition functions, in a manner of speech, to *hardwire* them, there are properties that cannot be implemented on a round-by-round basis. For example, *liveness conditions*, which require that something happens *eventually* in a run, cannot be translated into local terms because they are properties of the whole (infinite) run. Therefore, to enforce such properties, we restrict the set of runs being considered using *admissibility conditions*:

Definition 2.70 (Admissibility condition)**.** *An admissibility condition Ψ is any subset of the set of all runs R.*

Definition 2.71 (Context)**.** *A context $\gamma = (P_e, \mathscr{G}(0), \tau, \Psi)$ consists of*

- *an environment protocol P_e,*
- *a set of joint initial states $\mathscr{G}(0)$,*
- *a transition template τ, and*
- *an admissibility condition Ψ.*

Definition 2.72 (Agent-context)**.** *Given a context γ and an agents' joint protocol P, we combine them in an agent-context $\chi = (\gamma, P)$.*

Definition 2.73 (Consistency)**.** *For a context $\gamma = (P_e, \mathscr{G}(0), \tau, \Psi)$ and an agents' joint protocol P, we define the set of runs weakly consistent with P in γ (or weakly consistent with $\chi = (\gamma, P)$), denoted $R^{w\chi} = R^{w(\gamma,P)}$, to be the set of $\tau_{P_e,P}$-transitional runs that start at some joint initial state from $\mathscr{G}(0)$:*

$$R^{w(\gamma,P)} := \{r \in R \mid r(0) \in \mathscr{G}(0) \quad \text{and} \quad (\forall t \in \mathbb{N}_0)\, r(t+1) \in \tau_{P_e,P}(r(t))\}.$$

A run r is called strongly consistent, or simply consistent, with P in γ (or with $\chi = (\gamma, P)$) if it is weakly consistent with P in γ and, additionally, satisfies the admissibility condition, i.e., $r \in \Psi$.

We denote the system of all runs consistent with P in γ by

$$R^\chi = R^{(\gamma,P)} := R^{w(\gamma,P)} \cap \Psi.$$

Furthermore, we say that an agent-context $\chi = (\gamma, P)$ is *non-excluding*, if any prefix of a run that is weakly consistent with P in γ can be extended to a run that is strongly consistent with P in γ:

Definition 2.74 (Non-excluding agent-context)**.** *For any agent-context χ, χ is non-excluding iff*

$$R^{\chi} \neq \varnothing \quad \textit{and} \quad (\forall r \in R^{w\chi})(\forall t \in \mathbb{N}_0)(\exists r' \in R^{\chi})(\forall t' \leq t)\, r'\left(t'\right) = r\left(t'\right).$$

CHAPTER 3

Epistemic analysis

In this chapter, we demonstrate how, using the framework for modeling asynchronous byzantine fault-tolerant distributed systems, introduced in Chapter 2, one can obtain insights into the properties of such systems by performing *epistemic* analysis. By associating a Kripke model with a given set of runs, various tools from (temporal)-epistemic logic can be used to study the system corresponding to those runs. Such an analysis proved to be particularly useful for obtaining impossibility results, for example: if the necessary epistemic states for performing actions of interest cannot be reached by agents in the given system, the underlying problem is not solvable in that system. Our central result in this chapter, the *Brain-in-a-Vat lemma*, enables us to precisely capture epistemic dilemmas agents often find themselves in when performing actions in asynchronous byzantine fault-tolerant systems. As we shall see, one of the main consequences of the Brain-in-a-Vat lemma is that an agent in such a setting cannot *know* if an event really happened. In other words, even if an agent is correct, it cannot rule out the possibility of an alternative reality — alternative execution of the system — in which the event in question has not happened. This leads us to investigate how agents can make decisions in these systems since, as preconditions for actions, usually occurrences of events are used.

Chapter organization

We start by defining interpreted systems and showing how one can associate a Kripke model with an arbitrary interpreted system in Section 3.1. In addition, we introduce a language enriched with special atomic propositions enabling us to reason about correctness of agents, occurrences of events in the system, as well as about agents' actions. In Section 3.2, by introducing a method for run modifications, we model the above mentioned brain-in-a-vat-like scenarios occurring in asynchronous byzantine systems. This allows us to expose various limitations of agents' knowledge in such systems elegantly, i.e., as

theorems. Finally, we discuss the consequences of these findings by figuring out how the preconditions of agents' actions are affected in Section 3.3.

3.1 Obtaining Kripke models

First of all, we fix the following set of atomic propositions (atoms):

$$\begin{aligned} P \quad := \quad \mathsf{Prop} \quad \cup \quad & \{correct_i \mid i \in \mathcal{A}\} \quad \cup \\ & \{fake_{(i,t)}(o) \mid (i,t) \in \mathcal{A} \times \mathbb{N}_0, o \in \overline{Actions} \sqcup \overline{Events}\} \quad \cup \\ & \{\overline{occurred}_{(i,t)}(o) \mid (i,t) \in \mathcal{A} \times \mathbb{N}_0, o \in \overline{Actions} \sqcup \overline{Events}\} \quad \cup \\ & \{\overline{occurred}_i(o) \mid i \in \mathcal{A}, o \in \overline{Actions} \sqcup \overline{Events}\} \quad \cup \\ & \{occurred_i(o) \mid i \in \mathcal{A}, o \in \overline{Actions} \sqcup \overline{Events}\}, \end{aligned}$$

where Prop is a nonempty countably infinite set of atoms.

Syntax. We start with P and continue by forming formulas by closing under the Boolean connectives $\neg$ and $\wedge$ and under the (unary) modal operators $K_1, \ldots, K_n$ to obtain the language $\mathcal{L}$, i.e., the language $\mathcal{L}$ is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi,$$

where $p \in P$ and $i \in \mathcal{A}$. We take $\top$ to be an abbreviation for some fixed propositional tautology, and take $\bot$ to be an abbreviation for $\neg\top$. Also, we use the following standard abbreviations from propositional logic: $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ for $\neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$.

Remark 3.1. *Just like before, the overline in $\overline{occurred}_{(i,t)}(o)$ and $\overline{occurred}_i(o)$ is used to indicate that the occurrence of o is correct (non-byzantine).*

Definition 3.2 (Interpreted system)**.** *An interpreted system is a pair (R, π), consisting of a set of runs R and a valuation function $\pi\colon P \to \mathcal{P}(R \times \mathbb{N}_0)$.*

To reason epistemically via interpreted systems, we associate with a given interpreted system $\mathcal{I} = (R, \pi)$ a special Kripke model

$$M_{\mathcal{I}} := (R \times \mathbb{N}_0, \pi, \sim_1, \ldots, \sim_n) \in \mathsf{S5_n},$$

where

$$(r,t) \sim_i (r',t') \quad \textit{iff} \quad r_i(t) = r'_i(t').$$

A pair $(r,t) \in R \times \mathbb{N}_0$ will be called a *point*.

Definition 3.3. *By $(\mathcal{I}, r, t) \models \varphi$ we denote the fact that formula φ is satisfied at the point (r,t) and define the $\models$ relation as follows:*

$$(\mathcal{I}, r, t) \models \varphi \quad \textit{iff} \quad M_{\mathcal{I}}, (r,t) \models \varphi.$$

Therefore, we obtain:

1. For an atom p, $(\mathcal{I}, r, t) \models p$ iff $(r, t) \in \pi(p)$,
2. $(\mathcal{I}, r, t) \models \neg\varphi$ iff $(\mathcal{I}, r, t) \models \varphi$ does not hold,
3. $(\mathcal{I}, r, t) \models \varphi \wedge \psi$ iff $(\mathcal{I}, r, t) \models \varphi$ and $(\mathcal{I}, r, t) \models \psi$,
4. $(\mathcal{I}, r, t) \models K_i\varphi$ iff $(\mathcal{I}, r', t') \models \varphi$ for all (r', t') such that $r_i(t) = r'_i(t')$.

We will write $(\mathcal{I}, r, t) \not\models \varphi$ to denote that $(\mathcal{I}, r, t) \models \varphi$ does not hold. By $\mathcal{I} \models \varphi$ we denote the fact that φ is satisfied at all the points (r, t), i.e., that φ is *valid* in $\mathcal{I}$.

Definition 3.4. *Let $i \in \mathcal{A}$, $t, t' \in \mathbb{N}_0$ such that $t \leq t'$, and $o \in \overline{Actions} \sqcup \overline{Events}$. For the atoms from $P \setminus \mathsf{Prop}$, the truth value is determined in the following way:*

- *$correct_i$ is true at (r, t') iff no faulty event happened to i yet, i.e., no event from $FEvents_i$ appears in $r_\epsilon(t')$:*

$$(r, t') \in \pi(correct_i) \quad \textit{iff} \quad (\forall t'' \leq t')(\Lambda_{t''} \cap FEvents_i = \varnothing),$$

  where $r_\epsilon(t') = [\Lambda_{t'}, \ldots, \Lambda_1, \Lambda_0]$ is the history of the environment at t' defined in Definition 2.44 and $FEvents_i = \{fake\,(i, E) \mid E \in \overline{GEvents_i}\} \sqcup \{fake\,(i, A \mapsto A') \mid A, A' \in \{\textbf{\textit{noop}}\} \sqcup \overline{GActions_i}\} \sqcup \{sleep\,(i), hibernate\,(i)\}$.

- *$fake_{(i,t)}\,(o)$ is true at (r, t') iff i has a faulty reason to believe that $o \in \overline{Actions} \sqcup \overline{Events}$ occurred in round $(t-1).5$, i.e., $o \in r_i(t)$ because (at least in part) of some $O \in \beta_{b_i}^{t-1}(r)$, i.e.,*

$$(r, t') \in \pi(fake_{(i,t)}\,(o)) \quad \textit{iff} \quad t \geq 1 \text{ and } o \in \sigma\left(\beta_{b_i}^{t-1}(r)\right),$$

  where σ is the localization function defined in Definition 2.62 and $\beta_{b_i}^{t-1}(r)$ is the set of all byzantine external events of agent i in run r until $t-1$ defined in (2.22).

- *$\overline{occurred}_{(i,t)}(o)$ is true at (r, t') iff i has a correct reason to believe that $o \in \overline{Actions} \sqcup \overline{Events}$ occurred in round $(t-1).5$, i.e., $o \in r_i(t)$ because (at least in part) of some $O \in \beta_i^{t-1}(r) \sqcup \overline{\beta}_{\epsilon_i}^{t-1}(r)$. i.e.,*

$$(r, t') \in \pi(\overline{occurred}_{(i,t)}(o)) \quad \textit{iff} \quad t \geq 1 \text{ and } o \in label^{-1}\left(\beta_i^{t-1}(r) \sqcup \overline{\beta}_{\epsilon_i}^{t-1}(r)\right),$$

  where $label^{-1}$ is the "reverse" labeling function defined in Definition 2.53 and $\overline{\beta}_{\epsilon_i}^{t-1}(r)$ is the set of all correct external events observed by agent i in run r until $t-1$ defined in (2.20).

- $\overline{occurred}_i(o)$ *is true at* (r,t') *iff at least one of* $\overline{occurred}_{(i,m)}(o)$ *for* $1 \leq m \leq t'$ *is; also* $\overline{occurred}(o) := \bigvee_{i\in\mathcal{A}} \overline{occurred}_i(o)$*, i.e.,*

$$(r,t') \in \pi(\overline{occurred}_i(o)) \quad \textit{iff} \quad (\exists t < t') \ \textit{such that} \ o \in label^{-1}\left(\beta_i^t(r) \sqcup \overline{\beta}^t_{\epsilon_i}(r)\right),$$

where $label^{-1}$ *is the "reverse" labeling function defined in Definition 2.53 and* $\overline{\beta}^t_{\epsilon_i}(r)$ *is the set of all correct external events observed by agent* i *in run* r *until* t *defined in* (2.20)*.*

- $occurred_i(o)$ *is true at* (r,t') *iff either* $\overline{occurred}_i(o)$ *is or at least one of* $fake_{(i,m)}(o)$ *for* $1 \leq m \leq t'$ *is, i.e.,*

$$(r,t') \in \pi(occurred_i(o)) \quad \textit{iff} \quad o \in r_i(t').$$

Definition 3.5. *A formula* $\varphi \in \mathcal{L}$ *is called localized for an agent* $i \in \mathcal{A}$ *within an agent-context* χ *iff*

$$r_i(t) = r'_i(t') \quad \textit{implies} \quad (\mathcal{I},r,t) \models \varphi \Longleftrightarrow (\mathcal{I},r',t') \models \varphi,$$

for any interpreted system $\mathcal{I} = (R^\chi, \pi)$*, runs* $r, r' \in R^\chi$*, and timestamps* $t, t' \in \mathbb{N}_0$*.*

The proof of the following lemma follows immediately:

Proposition 3.6. *The following statements are valid for any formula* $\varphi \in \mathcal{L}$ *localized for an agent* $i \in \mathcal{A}$ *within an agent-context* χ *and any interpreted system* $\mathcal{I} = (R^\chi, \pi)$*:*

$$\mathcal{I} \models \varphi \leftrightarrow K_i\varphi \qquad \textit{and} \qquad \mathcal{I} \models \neg\varphi \leftrightarrow K_i\neg\varphi.$$

The Knowledge of Preconditions principle [Mos15] postulates that, in order to be able to act on a precondition φ, an agent must know φ. Thus, the preceding lemma shows that formulas localized for i can *always* be used as preconditions for actions. Our first observation is that agent's *perceptions* of a run are one example of such epistemically acceptable (though not necessarily reliable) preconditions:

Proposition 3.7. *For any agent-context* χ*, agent* $i \in \mathcal{A}$*, and* $o \in \overline{Actions} \sqcup \overline{Events}$*, the formula* $occurred_i(o)$ *is localized for* i *within* χ*.*

Proof. Let χ be an arbitrary agent-context, $i \in \mathcal{A}$ and $o \in \overline{Actions} \sqcup \overline{Events}$. Assume further that $(\mathcal{I},r,t) \models occurred_i(o)$ for some $\mathcal{I} = (R^\chi, \pi)$, $r \in R^\chi$ and $t \in \mathbb{N}_0$. According to Definition 3.4, this is equivalent to $o \in r_i(t)$. Take now an arbitrary $r' \in R^\chi$ and $t' \in \mathbb{N}_0$ such that $r_i(t) = r'_i(t')$. We immediately get that $o \in r'_i(t')$ must hold as well, which in turn is equivalent to $(\mathcal{I},r',t') \models occurred_i(o)$ (again according to Definition 3.4). Therefore, by Definition 3.5, we can conclude that the formula $occurred_i(o)$ is indeed localized for i within χ since we have obtained $(\mathcal{I},r,t) \models occurred_i(o) \Longleftrightarrow (\mathcal{I},r',t') \models occurred_i(o)$ by assuming $r_i(t) = r'_i(t')$. □

3.2 Modeling the brain-in-a-vat scenario

By contrast, as we will demonstrate, correctness of agent's perceptions is not localized. In fact, such correctness can never be established by an agent. We prove such impossibility results by means of controlled *run modifications*.

Definition 3.8. *A function*

$$\rho\colon R^{\chi} \longrightarrow \mathcal{P}(\overline{GActions_i}) \times \mathcal{P}(GEvents_i)$$

is called an i-intervention for an agent-context χ and agent $i \in \mathcal{A}$. A joint intervention

$$B = (\rho_1, \ldots, \rho_n)$$

consists of i-interventions ρ_i for each agent $i \in \mathcal{A}$. An adjustment

$$[B_t; \ldots; B_0]$$

is a sequence of joint interventions $B_0 \ldots, B_t$ to be performed in rounds from 0.5 to $t.5$ for some timestamp $t \in \mathbb{N}_0$.

An i-intervention $\rho(r) = (X, X_\epsilon)$ applied to a round $t.5$ of a given run r can be seen as a meta-action modifying the results of this round for i in such a way that

$$\beta_i^t\,(r') = X$$

and

$$\beta_{\epsilon_i}^t\,(r') = \beta_\epsilon^t\,(r') \cap GEvents_i = X_\epsilon$$

in the artificially constructed new run r'.

Given an i-intervention $\rho(r) = (X, X_\epsilon)$, we denote $\mathfrak{a}\rho(r) := X$ and $\mathfrak{e}\rho(r) := X_\epsilon$. Accordingly, a joint intervention $(\rho_1, \ldots, \rho_n)$ prescribes actions $\beta_i^t\,(r') = \mathfrak{a}\rho_i(r)$ for each agent i and events $\beta_\epsilon^t\,(r') = \bigsqcup_{i \in \mathcal{A}} \mathfrak{e}\rho_i(r)$ for the round $t.5$. Thus, an adjustment $[B_t; \ldots; B_0]$ fully determines actions and events in the initial $t+1$ rounds of run r'.

Definition 3.9. *Let*

$$adj = [B_t; \ldots; B_0]$$

be an adjustment where

$$B_m = (\rho_1^m, \ldots, \rho_n^m)$$

for each $0 \leq m \leq t$ and each ρ_i^m is an i-intervention for an agent-context $\chi = ((P_\epsilon, \mathscr{G}(0), \tau_B, \Psi), P)$. A run r' is obtained from $r \in R^{\chi}$ by adjustment adj iff for all $t' \leq t$, all $T > t$, and all $i \in \mathcal{A}$:

1. $r'\,(0) := r\,(0)$,

2. $r'_i(t'+1) := update_i(r'_i(t'), \mathfrak{a}\rho_i^{t'}(r), \bigsqcup_{i\in\mathcal{A}} \mathfrak{e}\rho_i^{t'}(r))$,

3. $r'_\epsilon(t'+1) := update_\epsilon(r'_\epsilon(t'), \bigsqcup_{i\in\mathcal{A}} \mathfrak{e}\rho_i^{t'}(r), \mathfrak{a}\rho_1^{t'}(r), \ldots, \mathfrak{a}\rho_n^{t'}(r))$,

4. $r'(T+1) \in \tau^B_{P_\epsilon,P}(r'(T))$.

We denote by $R\left(\tau^B_{P_\epsilon,P}, r, adj\right)$ *the set of all runs obtained from* r *by* adj.

Remark 3.10. *The last property ensures that beyond its adjusted segment, run* r' *extends in a* $\tau^B_{P_\epsilon,P}$*-transitional manner.*

Remark 3.11. *Note that, even though run* r *is assumed to be* $\tau^B_{P_\epsilon,P}$*-transitional (since* $r \in R^\chi$*), the adjusted runs obtained from it need not be, i.e., they need not obey the last property also for* $t' \leq t$.

To demonstrate the impossibility of establishing knowledge about perception correctness, we use several adjustment types to formalize the infamous *brain in a vat* scenario[1], where one agent, the "brain", is to experience a fabricated, i.e., faulty, version of its local history, whereas all other agents are to remain in their initial states (and made byzantine faulty or not at will). This is achieved by using interventions

(a) $Fake_i$ for brain i,

(b) *CFreeze* for other agents j that are to be correct, and

(c) $BFreeze_j$ for other agents j that are to be byzantine faulty.

Definition 3.12. *For an agent* $i \in \mathcal{A}$*, an agent-context* χ*, and a run* $r \in R^\chi$*, we define the following* i*-interventions:*

$$\begin{aligned} CFreeze(r) &:= (\varnothing, \varnothing),\\ BFreeze_i(r) &:= (\varnothing, \{fail(i)\}),\\ Fake_i^t(r) &:= (\varnothing, \{fail(i)\} \cup \beta^t_{b_i}(r) \cup \{fake(i,E) \mid E \in \overline{\beta}^t_{\epsilon_i}(r)\} \cup\\ &\qquad \{fake(i, \textbf{\textit{noop}} \mapsto A) \mid A \in \beta^t_i(r)\} \cup \{sleep(i) \mid aware(i, \beta^t_{\epsilon_i}(r))\}). \end{aligned}$$

Remark 3.13. *Note that interventions CFreeze and* $BFreeze_i$ *are constant, i.e., the modifications they impose are run-independent.*

The following lemma will prove to be particularly useful when proving Lemma 3.15.

Lemma 3.14. *Let* $i \in \mathcal{A}$*,* $t \in \mathbb{N}_0$*, and* $r \in R^\chi$*, where* χ *is an agent-context. Then*

[1] For connections to the semantic externalism and a survey of philosophical literature on the subject, see [PG96].

1. $\mathfrak{a}Fake_i^t(r) = \varnothing$, *i.e.,* $Fake_i^t$ *removes all actions.*
2. $go(i) \notin \mathfrak{e}Fake_i^t(r)$, *i.e.,* $Fake_i^t$ *never lets agent* i *act.*
3. $\sigma(\mathfrak{a}Fake_i^t(r) \sqcup \mathfrak{e}Fake_i^t(r)) = \sigma(\mathfrak{e}Fake_i^t(r)) = \sigma(\beta_i^t(r) \sqcup \beta_{\epsilon_i}^t(r))$, *i.e., actions and events to be appended to the local history of agent* i *as a result of round* $t.5$ *after the* i*-intervention* $Fake_i^t(r)$ *are the same as before it in the same round of the original run* r*.*
4. $aware(i, \mathfrak{e}Fake_i^t(r)) = t$ *iff* $aware(i, \beta_{\epsilon_i}^t(r)) = t$, *i.e., agent* i*'s awareness of the passing of round* $t.5$ *is not changed by* $Fake_i^t(r)$*.*

Proof. 1. From Definition 3.12, we immediately obtain

$$\begin{aligned}\mathfrak{a}Fake_i^t(r) = \mathfrak{a}(\varnothing, \{fail(i)\} \cup \beta_{b_i}^t(r) \cup \{fake(i,E) \mid E \in \overline{\beta}_{\epsilon_i}^t(r)\} \cup \\ \{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta_i^t(r)\} \cup \{sleep(i) \mid aware(i, \beta_{\epsilon_i}^t(r))\}) \\ = \varnothing.\end{aligned}$$

2. From Definition 3.12,

$$\begin{aligned}\mathfrak{e}Fake_i^t(r) = \mathfrak{e}(\varnothing, \{fail(i)\} \cup \beta_{b_i}^t(r) \cup \{fake(i,E) \mid E \in \overline{\beta}_{\epsilon_i}^t(r)\} \cup \\ \{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta_i^t(r)\} \cup \{sleep(i) \mid aware(i, \beta_{\epsilon_i}^t(r))\}) \\ = \{fail(i)\} \cup \beta_{b_i}^t(r) \cup \{fake(i,E) \mid E \in \overline{\beta}_{\epsilon_i}^t(r)\} \cup \\ \{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta_i^t(r)\} \cup \{sleep(i) \mid aware(i, \beta_{\epsilon_i}^t(r))\},\end{aligned}$$

follows. Now, it is obvious that $go(i) \notin \mathfrak{e}Fake_i^t(r)$.

3. It follows from Definition 2.62 and Definition 3.12, since

$$\sigma(\{fake(i,E) \mid E \in \overline{\beta}_{\epsilon_i}^t(r)\}) = \sigma(\{E \mid E \in \overline{\beta}_{\epsilon_i}^t(r)\}) = \sigma(\overline{\beta}_{\epsilon_i}^t(r))$$

and

$$\sigma(\{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta_i^t(r)\}) = \sigma(\{A \mid A \in \beta_i^t(r)\}) = \sigma(\beta_i^t(r)).$$

4. From Definition 2.58 and Definition 3.12, we get

$$\begin{aligned}aware(i, \mathfrak{e}Fake_i^t(r)) = t \quad &\text{iff} \quad \mathfrak{e}Fake_i^t(r) \cap \{go(i), sleep(i)\} \neq \varnothing \\ &\text{iff} \quad \mathfrak{e}Fake_i^t(r) \cap \{sleep(i)\} \neq \varnothing \\ &\text{iff} \quad aware(i, \beta_{\epsilon_i}^t(r)) = t.\end{aligned}$$

□

Recall that (see Definition 2.51): the environment's protocol P_ϵ makes an agent $i \in \mathcal{A}$ delayable if for any $X \in P_\epsilon(t)$, $X \setminus GEvents_i \in P_\epsilon(t)$; the environment's protocol P_ϵ makes an agent $i \in \mathcal{A}$ fallible if for any $X \in P_\epsilon(t)$, $X \cup \{fail(i)\} \in P_\epsilon(t)$; the environment's protocol P_ϵ makes an agent $i \in \mathcal{A}$ gullible if for any $Y \subseteq FEvents_i$, and any $X \in P_\epsilon(t)$, $Y \sqcup (X \setminus GEvents_i) \in P_\epsilon(t)$, whenever $Y \sqcup (X \setminus GEvents_i)$ is t-coherent.

Lemma 3.15 (Brain in a Vat)**.** *For an agent $i \in \mathcal{A}$, for an agent-context $\chi = ((P_\epsilon, \mathscr{G}(0), \tau_B, \Psi), P)$ such that P_ϵ makes i gullible and every $j \neq i$ delayable and fallible, for a set $Byz \subseteq \mathcal{A} \setminus \{i\}$ such that $f \geq 1 + |Byz|$, for a run $r \in R^\chi$, and for a timestamp $t > 0$, we consider an adjustment*

$$adj = [B_{t-1}; \ldots; B_0] \quad \textit{such that} \quad B_m = (\rho_1^m, \ldots, \rho_n^m)$$

with

$$\rho_i^m = Fake_i^m, \quad \rho_j^m = BFreeze_j \text{ for } j \in Byz, \quad \textit{and} \quad \rho_j^m = CFreeze \text{ for } j \notin \{i\} \sqcup Byz$$

for all $0 \leq m \leq t-1$. Then each run $r' \in R\left(\tau^B_{P_\epsilon,P}, r, adj\right)$ satisfies the following properties:

1. $r' \in R^{w\chi}$;
2. $(\forall m \leq t)\, r'_i(m) = r_i(m)$;
3. $(\forall m \leq t)\,(\forall j \neq i)\; r'_j(m) = r'_j(0)$;
4. *agents from $\mathcal{A} \setminus (\{i\} \sqcup Byz)$ remain correct until t;*
5. *agent i and all agents from Byz become byzantine faulty already in round 0.5;*
6. $(\forall m < t)\,(\forall j \neq i)\; \beta^m_{\epsilon_j}(r') \subseteq \{fail(j)\}$. *More precisely,*

   $$\beta^m_{\epsilon_j}(r') = \varnothing \quad \textit{iff} \quad \rho_j^m = CFreeze \quad \textit{and} \quad \beta^m_{\epsilon_j}(r') = \{fail(j)\} \quad \textit{iff} \quad \rho_j^m = BFreeze_j;$$

7. $(\forall m < t)\, \beta^m_{\epsilon_i}(r') \setminus FEvents_i = \varnothing$;
8. $(\forall m < t)(\forall j \in \mathcal{A})\, \beta^m_j(r') = \varnothing$.

Proof. Let $r' \in R\left(\tau^B_{P_\epsilon,P}, r, adj\right)$.

For property 5, using Definition 3.4, we immediately obtain $(r', 1) \notin \pi(correct_i)$ since the intervention $Fake_i^0$ is performed in round 0.5, i.e.,

$$\Lambda'_1 \cap FEvents_i = FEvents_i \neq \varnothing,$$

where $r'_\epsilon(1) = [\Lambda'_1, \Lambda'_0]$.

Similarly, for $j \in Byz$, $(r', 1) \notin \pi(correct_j)$ since the intervention $BFreeze_j$ is already performed in round 0.5.

For property 6, let $m < t$. According to Definition 3.12, for $j \notin \{i\} \sqcup Byz$, we have

$$\beta^m_{\epsilon_j}(r') = \mathfrak{e}\rho_j^m = \mathfrak{e}\,CFreeze = \mathfrak{e}(\varnothing, \varnothing) = \varnothing,$$

and, for $j \in Byz$, we have

$$\beta^m_{\epsilon_j}(r') = \mathfrak{e}\rho^m_j = \mathfrak{e}BFreeze_j = \mathfrak{e}(\varnothing, \{fail(j)\}) = \{fail(j)\}.$$

For property 7, let $m < t$. According to Definition 3.12, we have

$$\begin{aligned}
\beta^m_{\epsilon_i}(r') &= \mathfrak{e}\rho^m_i \\
&= \mathfrak{e}Fake^m_i(r) \\
&= \mathfrak{e}(\varnothing, \{fail(i)\} \cup \beta^m_{b_i}(r) \cup \{fake(i,E) \mid E \in \overline{\beta}^m_{\epsilon_i}(r)\} \cup \\
&\{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta^m_i(r)\} \cup \{sleep(i) \mid aware(i, \beta^m_{\epsilon_i}(r))\}) \\
&= \{fail(i)\} \cup \beta^m_{b_i}(r) \cup \{fake(i,E) \mid E \in \overline{\beta}^m_{\epsilon_i}(r)\} \cup \\
&\{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta^m_i(r)\} \cup \{sleep(i) \mid aware(i, \beta^m_{\epsilon_i}(r))\} \\
&\subseteq FEvents_i,
\end{aligned}$$

since $FEvents_i = \{fake(i,E) \mid E \in \overline{GEvents_i}\} \sqcup \{fake(i, A \mapsto A') \mid A, A' \in \{\mathbf{noop}\} \sqcup \overline{GActions_i}\} \sqcup \{sleep(i), hibernate(i)\}$. Thus, $\beta^m_{\epsilon_i}(r') \setminus FEvents_i = \varnothing$ indeed holds.

For property 8, let $m < t$. According to Definition 3.12, we have

- $\beta^m_j(r') = \mathfrak{a}CFreeze = \mathfrak{a}(\varnothing, \varnothing) = \varnothing$, for $j \notin \{i\} \sqcup Byz$,
- $\beta^m_j(r') = \mathfrak{a}BFreeze_j = \mathfrak{a}(\varnothing, \{fail(j)\}) = \varnothing$, for $j \in Byz$,
- $\beta^m_i(r') = \mathfrak{a}Fake^m_i(r) = \varnothing$, for agent i, by Lemma 3.14.

According to Definition 2.73, in order to prove property 1, we need to show

$$r'(0) \in \mathscr{G}(0)$$

and

$$r'(m+1) \in \tau^B_{P_\epsilon, P}(r'(m)). \tag{3.1}$$

According to Definition 3.9, $r'(0) = r(0)$. Therefore, $r'(0) \in \mathscr{G}(0)$ indeed holds.

Property (3.1) for $m > t$ directly follows from Definition 3.9. We prove that it holds for $m \leq t$ as well based on the gullibility of i and delayability and fallibility of all other $j \neq i$:

Let $m \leq t$. Consider $\alpha^m_\epsilon(r) \in P_\epsilon(m)$ from the original run r. The set $\alpha^m_\epsilon(r)$ is m-coherent by Definition 2.49. Note that $\alpha^m_\epsilon(r) \subset GEvents = \bigsqcup_{i \in \mathcal{A}} GEvents_i$. Thus, by the delayability of all $j \neq i$,

$$\alpha^m_{\epsilon_i}(r) := \alpha^m_\epsilon(r) \cap GEvents_i = \alpha^m_\epsilon(r) \setminus \bigsqcup_{j \neq i} GEvents_j \in P_\epsilon(m),$$

since $\alpha_{\epsilon}^{m}(r) \in P_{\epsilon}(m)$. Note also that for any $Z \subseteq FEvents_i$, $Z \sqcup \left(\alpha_{\epsilon_i}^{m}(r) \setminus GEvents_i\right) = Z \sqcup \varnothing = Z$ because $\alpha_{\epsilon_i}^{m}(r) \subseteq GEvents_i$. Thus, by the gullibility of i,

$$\alpha_{\epsilon_i}^{m}(r') := \{fail(i)\} \cup \beta_{b_i}^{m}(r) \cup \{fake(i,E) \mid E \in \overline{\beta}_{\epsilon_i}^{m}(r)\} \cup \{fake(i, \mathbf{noop} \mapsto A) \mid A \in \beta_i^m(r)\} \sqcup \{sleep(i) \mid aware(i, \beta_{\epsilon_i}^{t}(r))\} \in P_{\epsilon}(m),$$

since $\alpha_{\epsilon}^{m}(r) \in P_{\epsilon}(m)$. The set $\alpha_{\epsilon_i}^{m}(r')$ is m-coherent because it contains no correct events and neither $go(i)$ nor $hibernate(i)$. Finally, by the fallibility of all agents $j \in Byz$,

$$\alpha_{\epsilon}^{m}(r') := \alpha_{\epsilon_i}^{m}(r') \sqcup \{fail(j) \mid j \in Byz\} \in P_{\epsilon}(m),$$

since $\alpha_{\epsilon_i}^{m}(r') \in P_{\epsilon}(m)$. The set $\alpha_{\epsilon}^{m}(r')$ is m-coherent because $\alpha_{\epsilon_i}^{m}(r')$ is.

It remains to show that filtering turns the sets $\alpha_{\epsilon}^{m}(r')$, $\alpha_1^m(r'), \ldots, \alpha_n^m(r')$ into the exact β-sets prescribed by the adjustment adj. Let us abbreviate:

$$\Upsilon := filter_{\epsilon}^{B}(r'(m), \alpha_{\epsilon}^{m}(r'), \alpha_1^m(r'), \ldots, \alpha_n^m(r')),$$
$$\Xi_j := filter_j^B(\alpha_1^m(r'), \ldots, \alpha_n^m(r'), \Upsilon).$$

Our goal is to show that

$$\Upsilon_j := \Upsilon \cap GEvents_j = \beta_{\epsilon_j}^{m}(r'),$$

and

$$\Xi_j = \beta_j^m(r'),$$

for each $j \in \mathcal{A}$.

The set $\alpha_{\epsilon}^{m}(r')$ is unaffected by $filter_{\epsilon}^{B}$ (there are no correct receives in $\alpha_{\epsilon}^{m}(r')$ to be filtered out). So, according to Definition 2.59, after the filtering phase, we have the following:

$$\Upsilon_i = \alpha_{\epsilon}^{m}(r') \cap GEvents_i = \alpha_{\epsilon_i}^{m}(r'),$$

the latter being exactly $\beta_{\epsilon_i}^{m}(r')$ as shown above for property 7, and

$$\Upsilon_j = \alpha_{\epsilon}^{m}(r') \cap GEvents_j = \begin{cases} \varnothing & \text{if } j \notin Byz \\ \{fail(j)\} & \text{if } j \in Byz \end{cases} = \begin{cases} \varnothing & \text{if } \rho_j^m = CFreeze \\ \{fail(j)\} & \text{if } \rho_j^m = BFreeze_j \end{cases}$$

the latter being exactly $\beta_{\epsilon_j}^{m}(r')$ by property 6.

Since $go(j) \notin \Upsilon$ for any $j \in \mathcal{A}$, we also have that

$$\Xi_j = \varnothing = \beta_j^m(r'),$$

according to Definition 2.59.

Properties 2–4 depend solely on rounds from 0.5 to $(t-1).5$ of r'. We prove them for $m \leq t$ by induction on m.

Base case: $m = 0$**.** Property 2 follows from Definition 3.9. Properties 3–4 follow immediately.

Step from m **to** $m+1$**.**

For the induction step for property 2, we have the following cases:

1. If $\sigma\left(\beta_{\epsilon_i}^m(r)\right) \neq \varnothing$, then using (2.24), Definition 2.64, IH, Lemma 3.14 (3), and Definition 3.9, we obtain

$$\begin{aligned} r_i(m+1) &= \mathit{update}_i\left(r_i(m), \beta_i^m(r), \beta_\epsilon^m(r)\right) \\ &= \sigma\left(\beta_i^m(r) \sqcup \beta_{\epsilon_i}^m(r)\right) : r_i(m) \\ &= \sigma\left(\beta_i^m(r) \sqcup \beta_{\epsilon_i}^m(r)\right) : r_i'(m) \\ &= \sigma\left(\beta_i^m(r') \sqcup \beta_{\epsilon_i}^m(r')\right) : r_i'(m) \\ &= \mathit{update}_i\left(r_i'(m), \beta_i^m(r'), \beta_\epsilon^m(r')\right) \\ &= r_i'(m+1). \end{aligned}$$

2. If $\sigma\left(\beta_{\epsilon_i}^m(r)\right) = \varnothing$, but $\mathit{aware}(i, \beta_{\epsilon_i}^m(r)) = t$, then using (2.24), Definition 2.64, IH, Lemma 3.14 (3–4), and Definition 3.9, we obtain

$$\begin{aligned} r_i(m+1) &= \mathit{update}_i\left(r_i(m), \beta_i^m(r), \beta_\epsilon^m(r)\right) \\ &= \sigma\left(\beta_i^m(r) \sqcup \beta_{\epsilon_i}^m(r)\right) : r_i(m) \\ &= \sigma\left(\beta_i^m(r)\right) : r_i(m) \\ &= \sigma\left(\beta_i^m(r)\right) : r_i'(m) \\ &= \sigma\left(\beta_i^m(r') \sqcup \beta_{\epsilon_i}^m(r')\right) : r_i'(m) \\ &= \mathit{update}_i\left(r_i'(m), \beta_i^m(r'), \beta_\epsilon^m(r')\right) \\ &= r_i'(m+1). \end{aligned}$$

3. If $\sigma\left(\beta_{\epsilon_i}^m(r)\right) = \varnothing$ and $\mathit{unaware}(i, \beta_{\epsilon_i}^m(r)) = t$, then using (2.24), Definition 2.64, IH, Lemma 3.14 (4), and Definition 3.9, we obtain

$$\begin{aligned} r_i(m+1) &= \mathit{update}_i\left(r_i(m), \beta_i^m(r), \beta_\epsilon^m(r)\right) \\ &= r_i(m) \\ &= r_i'(m) \\ &= \mathit{update}_i\left(r_i'(m), \beta_i^m(r'), \beta_\epsilon^m(r')\right) \\ &= r_i'(m+1). \end{aligned}$$

This completes the proof of the induction step for property 2.

For the induction step for property 3, using Definition 3.9, Definition 2.64, property 6, Definition 2.62, and IH, we obtain

$$\begin{aligned} r'_j\,(m+1) &= update_j\left(r'_j\,(m)\,, \beta_j^m\,(r'), \beta_\epsilon^m\,(r')\right) \\ &= r'_j\,(m) \\ &= r'_j\,(0)\,. \end{aligned}$$

For the induction step for property 4, using the fact that, for $j \notin \{i\} \sqcup Byz$, $\beta_{\epsilon_j}^m\,(r') = \varnothing$ by property 6, and IH, we obtain that such agents j remain correct after the round $m.5$, i.e,

$$\Lambda'_{m+1} \cap FEvents_j = \varnothing,$$

where $r'_\epsilon(m+1) = [\Lambda'_{m+1}, \ldots, \Lambda'_1, \Lambda'_0]$. □

Remark 3.16. *The previous lemma states that for a designated agent $i \in A$ in an arbitrary local state $r_i(t)$ in run r there is always an i-indistinguishable local state $r'_i(t)$ in an alternative (transitional) run r' such that all other agents are yet to leave their initial local states, with i definitely byzantine faulty while other agents can be made byzantine faulty or correct at will. We call this the Brain-in-the-Vat lemma because agent i attains this indistinguishable local state by "imagining" that all actions and events from the original run r happened to it without any participation of other agents.*

Corollary 3.17. *If χ is non-excluding, then for any timestamp $t > 0$ there is a run $r' \in R^\chi$ constructed according to Lemma 3.15, such that for any $\mathcal{I} = (R^\chi, \pi)$, $o \in \overline{Actions} \sqcup \overline{Events}$, $j \in \{i\} \sqcup Byz$, and $k \notin \{i\} \sqcup Byz$:*

$$(\mathcal{I}, r', t) \not\models \overline{occurred}(o), \qquad (\mathcal{I}, r', t) \not\models correct_j, \qquad (\mathcal{I}, r', t) \models correct_k. \tag{3.2}$$

The ability to construct a Brain-in-a-Vat run r' in Lemma 3.15 and its properties in Corollary 3.17 enable us to prove that asynchronous agents in byzantine settings are not able to learn that a particular event actually happened, nor that they are not byzantine faulty:

Theorem 3.18. *Let $i \in \mathcal{A}$, let $\chi = ((P_\epsilon, \mathscr{G}(0), \tau_B, \Psi), P)$ be a non-excluding agent-context such that $f \geq 1$ and P_ϵ makes agent i gullible and every other agent $k \neq i$ delayable and fallible, and let $\mathcal{I} = (R^\chi, \pi)$ be an arbitrary interpreted system. Then for any $o \in \overline{Actions} \sqcup \overline{Events}$, $r \in R^\chi$, and $t > 0$:*

$$(\mathcal{I}, r, t) \models \neg K_i \overline{occurred}(o), \qquad (\mathcal{I}, r, t) \models \neg K_i correct_i, \qquad (\mathcal{I}, r, t) \models \neg K_i \neg correct_k.$$

Proof. For arbitrary $r \in R^\chi$ and $t > 0$, by Lemma 3.15 with $Byz = \varnothing$ and non-excludingness of χ, there exists $r' \in R^\chi$ such that (3.2) holds for $j = i$ and $k \neq i$ by Corollary 3.17. Therefore,

$$(\mathcal{I}, r, t) \models \neg K_i \overline{occurred}(o) \wedge \neg K_i correct_i \wedge \neg K_i \neg correct_k$$

follows according to Definition 3.3 since $r_i(t) = r'_i(t)$ by Lemma 3.15 (2). □

Remark 3.19. *While agent i can never learn that it is correct or that another agent k is byzantine faulty, agent i might be able to detect its own faults, for instance, by comparing actions prescribed by its protocol against actions recorded in its local history.*

The case of $f = 0$ corresponds to a system without byzantine faults, where correctness of all agents, actions, and events is common knowledge among agents. When $f = 1$, in view of Remark 3.19, the byzantine faulty agent may be able to conclude that all other agents are correct. However, for $f \geq 2$ this is not possible either:

Theorem 3.20. *Let $i \in \mathcal{A}$, let $\chi = ((P_\epsilon, \mathscr{G}(0), \tau_B, \Psi), P)$ be a non-excluding agent-context such that $f \geq 2$ and P_ϵ makes agent i gullible and every other agent $k \neq i$ delayable and fallible, and let $\mathcal{I} = (R^\chi, \pi)$ be an interpreted system. Then for any $r \in R^\chi$ and $t > 0$:*

$$(\mathcal{I}, r, t) \models \neg K_i \mathit{correct}_k.$$

Proof. For arbitrary $r \in R^\chi$ and $t > 0$, by Lemma 3.15 with $Byz = \{k\}$ and non-excludingness of χ, there exists $r' \in R^\chi$ such that $(\mathcal{I}, r', t) \not\models \mathit{correct}_j$ holds for $j \in \{i, k\}$ by Corollary 3.17. Therefore,

$$(\mathcal{I}, r, t) \models \neg K_i \mathit{correct}_k$$

follows according to Definition 3.3 since $r_i(t) = r'_i(t)$ by Lemma 3.15 (2). □

3.3 Relativizing the preconditions of actions

The results of the previous section clearly show that occurrences of trigger events alone cannot be used as preconditions for actions in asynchronous byzantine settings. The knowledge of a precondition requirement stated in [Mos15], i.e., that an agent acts on φ only when the agent knows φ, would typically lead (for such simple preconditions) to no actions being taken at all — even when an asynchronous agent is correct (at the current time in the current run), it can never discount the scenario of being a "brain in a vat". This first led us to consider the following notion of *belief* (belief as defeasible knowledge [MS93])

$$B_i\varphi := K_i(\mathit{correct}_i \to \varphi)^2 \tag{3.3}$$

as the adequate epistemic state of an acting agent in this settings. According to it, while φ is the desired property, agent i acts on the precondition $\mathit{correct}_i \to \varphi$.

We believe that (3.3) can be further improved in at least two directions:

i.) Firstly, a typical problem specification for a byzantine fault-tolerant system does not impose any restrictions on the actions the byzantine faulty agents perform (for instance,

[2]Note that, if instead $K_i(\underline{\mathit{correct}_i} \to \varphi)$ we require $\mathit{correct}_i \to K_i\varphi$, it would not be helpful since, as we saw, in Theorem 3.18, $K_i\overline{\mathit{occurred}}(o)$ fails to hold even when agent i is correct at the current time in the current run — it is the fact that we can construct (at any given time) an indistinguishable to i brain-in-a-vat run that renders it impossible for i to gain such knowledge.

in case of distributed consensus, all correct agents must agree on a common value, whereas the byzantine faulty agents are completely exempted from this). Consequently, what is of actual interest here is that $B_i\varphi$ is satisfied for the correct agents. By this reasoning, we arrived at the notion of, what we call *hope* (see Remark 3.22),

$$H_i\varphi := correct_i \to K_i(correct_i \to \varphi). \tag{3.4}$$

Moreover, $H_i\varphi$ neatly encapsulates the epistemic state of an acting agent in a uniform way — be it a correct agent acting or a faulty one. This is because (3.4) is automatically satisfied for faulty agents, so it captures the fact that the (malicious) byzantine faulty agents can act irrespective of any preconditions, while for correct agents it simply collapses to belief.

ii.) Secondly, per Remark 3.19, the possibility that an agent can learn that it is byzantine faulty is not excluded completely. Assuming the byzantine faulty agent in question is malfunctioning rather than malicious, its knowledge about its faults could be used to minimize the effects of them on the system as a whole and it would be of crucial importance in scenarios where self-correction is possible. Therefore, for such cases, we propose the following notion of, what we call *credence*,

$$Cr_i\varphi := \neg K_i \neg correct_i \wedge K_i(correct_i \to \varphi), \tag{3.5}$$

as the adequate necessary epistemic state for acting.

Remark 3.21. *We note that generalized versions such that instead of $correct_i$ there can be (almost) any formula α of both belief $B_i\varphi = K_i(correct_i \to \varphi)$ and credence $Cr_i\varphi = \neg K_i \neg correct_i \wedge K_i(correct_i \to \varphi)$ have been studied in [MS93] (in the single-agent case, however).*

Remark 3.22. *As we saw in Theorem 3.18, an agent itself can never ascertain its own correctness. So, $H_i\varphi$ can be read as the following from the point of view of agent i:* "Unless I am incorrect, I believe φ to be the case." *Moreover, in Theorem 3.20, we saw that an agent itself can also never ascertain the correctness of another agent (assuming $f \geq 2$). This observation becomes of crucial importance in situations where agent i has to reason about the epistemic state of some other agent j. As we will see in the epistemic analysis of the FRR problem in Chapter 6, instead of $K_i(correct_i \to K_j(correct_j \to \varphi))$, i.e., $B_iB_j\varphi$, what we often have is actually $K_i(correct_i \to (correct_j \to K_j(correct_j \to \varphi)))$, i.e., $B_iH_j\varphi$, which can be read as the following from the point of view of agent i:* "I believe that, unless j is incorrect, j believes φ to be the case." *It is these specific readings that inspired the name "hope" of the H_i modality since hope can be thought of as allowing more room for uncertainty than belief.*[3]

In the following proposition we establish some basic connections between the proposed modalities.

[3]While we do not attempt to study the true meaning of hope here, hope has been subject of rigorous analysis (especially in relation to belief) by many philosophers, see [BS22] for an overview.

Proposition 3.23. *For any formula $\varphi \in \mathcal{L}$, any agent $i \in \mathcal{A}$, the following formulas are valid in every interpreted system:*

$$\models K_i\varphi \to B_i\varphi \qquad \models correct_i \to (Cr_i\varphi \to \varphi)$$
$$\models Cr_i\varphi \to B_i\varphi \qquad \models correct_i \to (B_i\varphi \to \varphi)$$
$$\models B_i\varphi \to H_i\varphi \qquad \models correct_i \to (H_i\varphi \to \varphi)$$
$$\models correct_i \to (H_i\varphi \to Cr_i\varphi) \qquad \models B_i\varphi \to K_iB_i\varphi$$
$$\models \neg correct_i \to H_i\varphi \qquad \models Cr_i\varphi \to K_iCr_i\varphi$$
$$\models K_i\varphi \to \varphi \qquad \models K_i correct_i \to (H_i\varphi \to K_i\varphi)$$

Proof. Let $\varphi \in \mathcal{L}$ be an arbitrary formula and $i \in \mathcal{A}$ be an arbitrary agent.

The proofs of $\models K_i\varphi \to B_i\varphi$, $\models Cr_i\varphi \to B_i\varphi$ and $\models B_i\varphi \to H_i\varphi$ follow immediately from the definitions of $B_i\varphi$, $Cr_i\varphi$ and $H_i\varphi$.

Let us show $\models correct_i \to (H_i\varphi \to Cr_i\varphi)$. Take an arbitrary interpreted system $\mathcal{I} = (R, \pi)$ and let $r \in R$ and $t \in \mathbb{N}_0$. Assume that $(\mathcal{I}, r, t) \models correct_i \wedge H_i\varphi$. We need to show $(\mathcal{I}, r, t) \models Cr_i\varphi$, i.e., $(\mathcal{I}, r, t) \models \neg K\neg correct_i \wedge K_i(correct_i \to \varphi)$. Using $(\mathcal{I}, r, t) \models correct_i \wedge H_i\varphi$, that is, $(\mathcal{I}, r, t) \models correct_i \wedge (correct_i \to K_i(correct_i \to \varphi))$, we immediately get $(\mathcal{I}, r, t) \models K_i(correct_i \to \varphi)$. Using $(\mathcal{I}, r, t) \models correct_i$, we obtain $(\mathcal{I}, r, t) \models \neg K\neg correct_i$. Thus, $(\mathcal{I}, r, t) \models Cr_i\varphi$ indeed holds.

The proof of $\models \neg correct_i \to H_i\varphi$ follows using propositional reasoning.

The proof of $\models K_i\varphi \to \varphi$ is well-known. The proofs of $\models correct_i \to (Cr_i\varphi \to \varphi)$, $\models correct_i \to (B_i\varphi \to \varphi)$ and $\models correct_i \to (H_i\varphi \to \varphi)$ are straightforward.

The proofs of $\models B_i\varphi \to K_iB_i\varphi$ and $\models Cr_i\varphi \to K_iCr_i\varphi$ follow using the positive and negative introspection properties of knowledge.

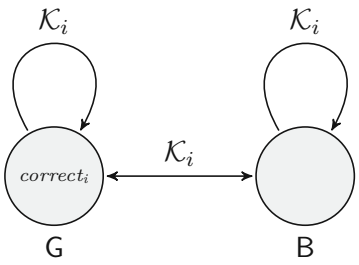
Finally, we show $\models K_i correct_i \to (H_i\varphi \to K_i\varphi)$. Take an arbitrary interpreted system $\mathcal{I} = (R, \pi)$ and let $r \in R$ and $t \in \mathbb{N}_0$. Assume that $(\mathcal{I}, r, t) \models K_i correct_i \wedge H_i\varphi$. We need to show $(\mathcal{I}, r, t) \models K_i\varphi$. Take $r' \in R$ and $t' \in \mathbb{N}_0$ such that $r_i(t) = r'_i(t')$. We need to show $(\mathcal{I}, r', t') \models \varphi$. First of all, from $(\mathcal{I}, r, t) \models K_i correct_i \wedge H_i\varphi$, that is, $(\mathcal{I}, r, t) \models K_i correct_i \wedge (correct_i \to K_i(correct_i \to \varphi))$, follows $(\mathcal{I}, r, t) \models K_i(correct_i \to \varphi)$. Therefore, we have $(\mathcal{I}, r', t') \models correct_i \to \varphi$. Finally, using $(\mathcal{I}, r, t) \models K_i correct_i$, we obtain $(\mathcal{I}, r', t') \models correct_i$ and hence $(\mathcal{I}, r', t') \models \varphi$. □

As follows from the preceding lemma, credence is stronger than belief, which is stronger than hope, while knowledge is stronger than belief. However, for a correct agent, credence, belief, and hope all become equivalent, while knowledge generally remains stronger (note that for fault-free systems they all collapse to the standard notion of knowledge). At the same time, all four modalities are factive for correct agents (knowledge is factive for all agents). Finally, belief and credence satisfy the "self-awareness" condition. On the other hand, hope does not generally satisfy $H_i\varphi \to K_iH_i\varphi$, as we show below.

Let $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$ be such that:

- $W = \{\mathsf{G}, \mathsf{B}\}$,
- $\mathcal{K}_j = W \times W$ for all $j \in \mathcal{A}$,
- $\pi(correct_i) = \{G\}$, and
- $\pi(p) = W$, for all atoms $p \neq correct_i$.

Thus, we have the following situation for agent i, in particular:



It is easy to see now from the picture above that

$$M, \mathsf{B} \not\models (correct_i \to K_i(correct_i \to \bot)) \to K_i(correct_i \to K_i(correct_i \to \bot)),$$

that is,

$$M, \mathsf{B} \not\models H_i\bot \to K_iH_i\bot,$$

since $M, \mathsf{B} \not\models correct_i$ and $M, \mathsf{G} \not\models K_i(correct_i \to \bot)$.

3.4 Related work

Interestingly, the possibility to model the brain-in-a-vat scenario later turned out to be a widespread phenomenon in systems with byzantine faults. Not only that it can be modeled in systems with synchronous agents as well, but even the perfectly synchronized clocks available in lockstep synchronous systems cannot be used to avoid it [SSK20].

CHAPTER 4

A modal logic of hope

In this chapter, we aim to obtain a better understanding of the hope modality, introduced in Chapter 3. For this purpose, we try to capture all its essential properties via an independent logical system. Thus, we propose a separate axiomatization for it, which we prove to be sound and complete with respect to a class of Kripke models designed for hope. The corresponding completeness proof relies on the standard method of canonical model construction, albeit the canonical model itself is non-standard in that its accessibility relations depend on the valuations of correctness atoms. In other words, some of the axioms for hope are not purely frame characterizable. For example, whether or not a certain world (in the Kripke model) has any accessible worlds heavily depends on the correctness status of agents in it. The proposed axiom system also turns out to be strongly sound and strongly complete with respect to the same class of Kripke models. This in turn allows us to conclude that the logic of hope is compact. We also provide a proof of soundness and completeness with respect to the standard S5 models for knowledge via a suitable translation function. Finally, we show that the proposed logic of hope has the finite model property and is decidable.

Chapter organization

We start by introducing the relevant syntax as well as the axiom system for the hope modality in Section 4.1. The soundness and completeness results are all grouped together in Section 4.2: we prove both weak and strong completeness with respect to special models for hope, as well as completeness with respect to the standard S5 models for knowledge via a suitable translation function. Finally, in Section 4.3, we prove (using the same translation) that the proposed logic of hope has the finite model property as well as that it is decidable.

4.1 Towards a logic of hope

First of all, we fix:

- a finite set $\mathcal{A} := \{1, \ldots, n\}$ of *agents*,
- a nonempty countably infinite set Prop of *atomic propositions* (*atoms*),
- a finite set $\mathsf{Co} := \{correct_i \mid i \in \mathcal{A}\}$ of designated *correctness atoms*.

Syntax. We start with $\mathsf{Prop} \cup \mathsf{Co}$ and continue by forming formulas by closing under the Boolean connectives $\neg$ and $\wedge$ and under the unary modal operators (one for each agent) $H_1, \ldots, H_n$ to obtain the language $\mathcal{L}_H^{\mathrm{co}}$, i.e., the language $\mathcal{L}_H^{\mathrm{co}}$ is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid H_i\varphi,$$

where $p \in \mathsf{Prop} \cup \mathsf{Co}$ and $i \in \mathcal{A}$. We take $\top$ to be an abbreviation for some fixed propositional tautology, and take $\bot$ to be an abbreviation for $\neg\top$. Also, we use the following standard abbreviations from propositional logic: $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ for $\neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$. In addition, for each $i \in \mathcal{A}$, we abbreviate $faulty_i := \neg correct_i$.

Remark 4.1. *We use $K_1, \ldots, K_n$ modalities instead of $H_1, \ldots, H_n$ modalities in the above BNF to obtain the language $\mathcal{L}_K^{\mathrm{co}}$.*

The axiom system $\mathscr{H}_{\mathrm{co}}$ is depicted in Figure 4.1.

$$\begin{aligned}
P: &\quad \textit{all propositional tautologies}\\
K^H: &\quad H_i(\varphi \to \psi) \wedge H_i\varphi \to H_i\psi\\
4^H: &\quad H_i\varphi \to H_iH_i\varphi\\
5^H: &\quad \neg H_i\varphi \to H_i\neg H_i\varphi\\
T'^H: &\quad correct_i \to (H_i\varphi \to \varphi)\\
H: &\quad H_i\, correct_i\\
F: &\quad faulty_i \to H_i\varphi\\
MP: &\quad \frac{\varphi \quad \varphi \to \psi}{\psi}\\
Nec^H: &\quad \frac{\varphi}{H_i\varphi}
\end{aligned}$$

Figure 4.1: Axiom system $\mathscr{H}_{\mathrm{co}}$

Remark 4.2. *Axioms P, K^H, 4^H, and 5^H, along with inference rules MP and Nec^H represent the standard $\mathscr{K}45_n$ axiom system (see Figure 2.2). Axiom T'^H is axiom T restricted to correct agents; axiom H states that agents always hope to be correct; axiom F means that the hopes of byzantine faulty agents are unrestricted and all encompassing, in particular, alongside tautologies they also hope for contradictions, making their hopes inconsistent.*

Recall that (see Definition 2.1): a set of formulas forms a system of a normal multi-agent epistemic logic if and only if it contains all propositional tautologies and is closed under the modus ponens inference rule, the K axiom scheme, the necessitation inference rule, and the uniform substitution rule.

Proposition 4.3. *The logic of $\mathscr{H}_{\text{co}}$ is not a normal multi-agent epistemic logic.*

Proof. The uniform substitution rule is violated because of axiom H. □

4.2 Soundness and completeness results

Soundness with respect to S5 models via translation

Definition 4.4. *For any $\varphi, \psi \in \mathcal{L}_H^{\text{co}}$, any $p \in \mathsf{Prop} \cup \mathsf{Co}$, and any $i \in \mathcal{A}$, the translation function $t : \mathcal{L}_H^{\text{co}} \to \mathcal{L}_K^{\text{co}}$ is defined recursively in the following way:*

- $t(p) := p$;
- $t(\neg\varphi) := \neg t(\varphi)$;
- $t(\varphi \wedge \psi) := t(\varphi) \wedge t(\psi)$;
- $t(H_i\varphi) := correct_i \to K_i(correct_i \to t(\varphi))$.

Lemma 4.5. *For all $M^K \in \mathsf{S5_n}$, all $w \in \mathcal{D}(M^K)$, all $\varphi, \psi \in \mathcal{L}_H^{\text{co}}$, and all $i \in \mathcal{A}$ holds the following:*

1. *If φ is an instance of a propositional tautology, then $M^K \models t(\varphi)$;*
2. $\mathsf{S5_n} \models t(H_i(\varphi \to \psi) \wedge H_i\varphi \to H_i\psi)$;
3. $\mathsf{S5_n} \models t(H_i\varphi \to H_iH_i\varphi)$;
4. $\mathsf{S5_n} \models t(\neg H_i\varphi \to H_i\neg H_i\varphi)$;
5. $\mathsf{S5_n} \models t(correct_i \to (H_i\varphi \to \varphi))$;
6. $\mathsf{S5_n} \models t(faulty_i \to H_i\varphi)$;

7. $\mathsf{S5_n} \models t(H_i correct_i)$;

8. *If* $M^K, w \models t(\varphi)$ *and* $M^K, w \models t(\varphi \to \psi)$, *then* $M^K, w \models t(\psi)$;

9. *if* $M^K \models t(\varphi)$, *then* $M^K \models t(H_i\varphi)$.

Proof. 1. This follows immediately from the fact that if φ is an instance of a propositional tautology, then $t(\varphi)$ is also an instance of a propositional tautology.

2. By applying the translation function from Definition 4.4, we obtain

$$t(H_i(\varphi \to \psi) \land H_i\varphi \to H_i\psi) = (correct_i \to K_i(correct_i \to t(\varphi \to \psi))) \land (correct_i \to K_i(correct_i \to t(\varphi))) \to (correct_i \to K_i(correct_i \to t(\psi))).$$

Take an arbitrary model $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5_n}$ and an arbitrary state $w \in W$. We need to show that

$$M^K, w \not\models (correct_i \to K_i(correct_i \to t(\varphi \to \psi))) \land (correct_i \to K_i(correct_i \to t(\varphi))) \text{ or } M^K, w \models (correct_i \to K_i(correct_i \to t(\psi)))$$

holds. Assume

$$M^K, w \models (correct_i \to K_i(correct_i \to t(\varphi \to \psi))) \land (correct_i \to K_i(correct_i \to t(\varphi))).$$

This means that either $M^K, w \not\models correct_i$ or

$$M^K, w \models K_i(correct_i \to t(\varphi \to \psi)) \text{ and } M^K, w \models K_i(correct_i \to t(\varphi)).$$

If $M^K, w \not\models correct_i$, then $M^K, w \models correct_i \to K_i(correct_i \to t(\psi))$ immediately follows. So, assume $M^K, w \models K_i(correct_i \to t(\varphi \to \psi))$ and $M^K, w \models K_i(correct_i \to t(\varphi))$. Take an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{K}_i$. Now, by assumption, $M^K, w' \models correct_i \to t(\varphi \to \psi)$ and $M^K, w' \models correct_i \to t(\varphi)$. By propositional reasoning, from this we get that $M^K, w' \models correct_i \to t(\psi)$ must also hold. Since $w' \in W$ such that $(w, w') \in \mathcal{K}_i$ was chosen arbitrarily, $M^K, w \models K_i(correct_i \to t(\psi))$ follows. Therefore, $M^K, w \models correct_i \to K_i(correct_i \to t(\psi))$ follows in this case too.

3. By applying the translation function from Definition 4.4, we obtain

$$t(H_i\varphi \to H_iH_i\varphi) = (correct_i \to K_i(correct_i \to t(\varphi))) \to (correct_i \to K_i(correct_i \to (correct_i \to K_i(correct_i \to t(\varphi))))).$$

Take an arbitrary model $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5_n}$ and an arbitrary state $w \in W$. We need to show that

$$M^K, w \not\models correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi)) \text{ or } M^K, w \models correct_i \rightarrow K_i(correct_i \rightarrow (correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))))$$

holds.

Let us assume $M^K, w \models correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))$. This means that either $M^K, w \not\models correct_i$ or $M^K, w \models K_i(correct_i \rightarrow t(\varphi))$. If $M^K, w \not\models correct_i$, then

$$M^K, w \models correct_i \rightarrow K_i(correct_i \rightarrow (correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))))$$

immediately follows. So, assume that $M^K, w \models K_i(correct_i \rightarrow t(\varphi))$ holds. Using the positive introspection property of knowledge, we obtain that $M^K, w \models K_iK_i(correct_i \rightarrow t(\varphi))$ must also hold. Take an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{K}_i$. Now, $M^K, w' \models K_i(correct_i \rightarrow t(\varphi))$ follows. Consequently,

$$M^K, w' \models correct_i \rightarrow (correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi)))$$

also follows. Since $w' \in W$ such that $(w, w') \in \mathcal{K}_i$ was chosen arbitrarily, we obtain $M^K, w \models K_i(correct_i \rightarrow (correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))))$. Finally, this implies

$$M^K, w \models correct_i \rightarrow K_i(correct_i \rightarrow (correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi)))).$$

4. By applying the translation function from Definition 4.4, we obtain

$$t(\neg H_i\varphi \rightarrow H_i\neg H_i\varphi) = \neg(correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))) \rightarrow (correct_i \rightarrow K_i(correct_i \rightarrow \neg(correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))))).$$

Take an arbitrary model $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5_n}$ and an arbitrary state $w \in W$. We need to show that

$$M^K, w \not\models \neg(correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))) \text{ or } M^K, w \models correct_i \rightarrow K_i(correct_i \rightarrow \neg(correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi))))$$

holds.

Let us assume $M^K, w \models \neg(correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi)))$. This means that $M^K, w \models correct_i$ and $M^K, w \not\models K_i(correct_i \rightarrow t(\varphi))$. It is enough to show

$$M^K, w \models K_i(correct_i \rightarrow \neg(correct_i \rightarrow K_i(correct_i \rightarrow t(\varphi)))),$$

that is, $M^K, w \models K_i(correct_i \rightarrow \neg K_i(correct_i \rightarrow t(\varphi)))$, since we have $M^K, w \models correct_i$.

From $M^K, w \not\models K_i(correct_i \rightarrow t(\varphi))$, using the negative introspection property of knowledge , we obtain $M^K, w \models K_i\neg K_i(correct_i \rightarrow t(\varphi))$. Take an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{K}_i$. Now, by assumption, $M^K, w' \models \neg K_i(correct_i \rightarrow t(\varphi))$. Consequently,

$$M^K, w' \models correct_i \rightarrow \neg K_i(correct_i \rightarrow t(\varphi))$$

also follows. Since $w' \in W$ such that $(w, w') \in \mathcal{K}_i$ was chosen arbitrarily, we obtain $M^K, w \models K_i(correct_i \rightarrow \neg K_i(correct_i \rightarrow t(\varphi)))$.

5. By applying the translation function from Definition 4.4, we obtain

$$t(correct_i \to (H_i\varphi \to \varphi)) = correct_i \to ((correct_i \to K_i(correct_i \to t(\varphi))) \to t(\varphi)).$$

Take an arbitrary model $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5_n}$ and an arbitrary state $w \in W$. We need to show that

$$M^K, w \not\models correct_i \text{ or } M^K, w \models (correct_i \to K_i(correct_i \to t(\varphi))) \to t(\varphi)$$

holds. Let us assume $M^K, w \models correct_i$. We now need to show

$$M^K, w \models (correct_i \to K_i(correct_i \to t(\varphi))) \to t(\varphi),$$

that is, $M^K, w \not\models correct_i \to K_i(correct_i \to t(\varphi))$ or $M^K, w \models t(\varphi)$. Let us assume $M^K, w \models correct_i \to K_i(correct_i \to t(\varphi))$. By combining this with $M^K, w \models correct_i$, we get $M^K, w \models K_i(correct_i \to t(\varphi))$. Using the factivity property of knowledge, we get that $M^K, w \models correct_i \to t(\varphi)$ must hold. Finally, using the assumption $M^K, w \models correct_i$ one more time, we obtain $M^K, w \models t(\varphi)$.

6. By applying the translation function from Definition 4.4, we obtain

$$t(faulty_i \to H_i\varphi) = faulty_i \to (correct_i \to K_i(correct_i \to t(\varphi))).$$

Further, take an arbitrary $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5_n}$ and an arbitrary $w \in W$. Since $faulty_i = \neg correct_i$, we immediately obtain

$$M^K, w \models faulty_i \to (correct_i \to K_i(correct_i \to t(\varphi)))$$

because the translated formula is an instance of a propositional tautology.

7. By applying the translation function from Definition 4.4, we obtain

$$t(H_i correct_i) = correct_i \to K_i(correct_i \to correct_i).$$

Further, take an arbitrary $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5_n}$ and an arbitrary $w \in W$. Since $correct_i \to correct_i$ is an instance of a propositional tautology, $M^K, w \models K_i(correct_i \to correct_i)$ holds (agents know all propositional tautologies). Thus, $M^K, w \models correct_i \to K_i(correct_i \to correct_i)$ also holds.

8. Assume $M^K, w \models t(\varphi)$ and $M^K, w \models t(\varphi \to \psi)$. The latter implies $M^K, w \models t(\varphi) \to t(\psi)$, which combined with $M^K, w \models t(\varphi)$ implies that $M^K, w \models t(\psi)$ indeed holds.

9. Assume $M^K \models t(\varphi)$. in order to show $M^K \models t(H_i\varphi)$, take an arbitrary $w \in \mathcal{D}(M^K)$ and an arbitrary $w' \in \mathcal{D}(M^K)$ such that $(w, w') \in \mathcal{K}_i$. Since $M^K \models t(\varphi)$, it follows that $M^K, w' \models t(\varphi)$ holds. Consequently, $M^K, w' \models correct_i \to t(\varphi)$ also holds. Since $w' \in \mathcal{D}(M^K)$ was chosen arbitrarily, $M^K, w \models K_i(correct_i \to t(\varphi))$ follows. Therefore, we get that $M^K, w \models correct_i \to K_i(correct_i \to t(\varphi))$, i.e., $M^K \models t(H_i\varphi)$ also follows. □

Theorem 4.6 (Soundness via translation)**.** *For any formula $\varphi \in \mathcal{L}_H^{\mathrm{co}}$ holds the following:*

$$\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi \quad \Longrightarrow \quad \mathsf{S5_n} \models t(\varphi).$$

Proof. We proceed by induction on the length of the derivation of φ in $\mathscr{H}_{\mathrm{co}}$. Let $\varphi_1, \varphi_2, \ldots, \varphi_k$ be a proof of the formula φ in $\mathscr{H}_{\mathrm{co}}$. If $k = 1$, then φ is an instance of an axiom in $\mathscr{H}_{\mathrm{co}}$. Therefore, according to Lemma 4.5 (1-7), it immediately follows that φ is valid with respect to $\mathsf{S5_n}$ models. Assume now that the desired statement holds for every formula which has a proof of length shorter than k and consider the formula φ which has a proof of length k. This means that the last formula in the proof is φ. There are two possibilities: either φ is an instance of some axiom in $\mathscr{H}_{\mathrm{co}}$, or φ follows by an application of an inference rule. We already dealt with the first possibility in the base case, so we consider the second possibility:

- φ follows by an application of modus ponens to some formulas, for example φ_i and $\varphi_i \to \varphi$. Since these two formulas are earlier in the proof, they have proofs whose lengths are shorter than k. After applying the induction hypothesis on them we get $\mathsf{S5_n} \models t(\varphi_i)$ and $\mathsf{S5_n} \models t(\varphi_i \to \varphi)$. Using Lemma 4.5 (8), we obtain that $\mathsf{S5_n} \models t(\varphi)$ indeed holds.

- φ follows by an application of necessitation to some formula, for example φ_i, meaning φ is of the form $H_j\varphi_i$, for some $j \in \mathcal{A}$. Since φ_i is earlier in the proof it has a proof whose length is shorter than k, which means that we can apply the induction hypothesis on it and get $\mathsf{S5_n} \models t(\varphi_i)$. Using Lemma 4.5 (9), we obtain that $\mathsf{S5_n} \models t(H_j\varphi_i)$ indeed holds. □

Soundness and completeness with respect to special models

Just like before, after establishing Lemma 4.9, soundness follows by an easy induction (see Theorem 4.10). The not-so-easy part is proving completeness, as usual. Luckily, the standard method of *canonical model* construction [FHMV95, BdRV01] can be carried out quite smoothly even though some of the axioms of $\mathscr{H}_{\mathrm{co}}$ are not purely frame characterizable (see Definition 4.7). The idea is as follows: we construct one large model for $\mathscr{H}_{\mathrm{co}}$ by taking all maximal consistent (with respect to $\mathscr{H}_{\mathrm{co}}$) sets of formulas (see Definition 4.11 and *Correctness lemma* 4.17) to be the worlds of the model and by defining the valuation function and the accessibility relations in terms of membership of formulas to such sets (see Definition 4.15). This way, using the properties of maximal consistent sets and the *Lindenbaum lemma* 4.14, we obtain the key result: a formula $\varphi \in \mathcal{L}_H^{\mathrm{co}}$ belongs to a maximal consistent set Γ if and only if it is satisfied in the world w_Γ corresponding to it (see *Truth lemma* 4.16). Finally, using all these results, we obtain completeness by contraposition (see Theorem 4.18).

Recall that $\mathsf{K45_n}$ is the class of models with transitive and euclidean accessibility relations (see Definition 2.9).

Definition 4.7. *The class* $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ *of models consists of all Kripke models of the form*

$$M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_{\mathsf{n}},$$

where for every $i \in \{1, \ldots, n\}$*, and every* $w, w' \in W$*:*

- *if* $M^H, w \models \mathit{correct}_i$*, then* $w \in \mathcal{H}_i(w)$*,*
- *if* $M^H, w \models \mathit{faulty}_i$*, then* $\mathcal{H}_i(w) = \varnothing$*, and*
- *for all* $w' \in \mathcal{H}_i(w)$*, it is the case that* $M^H, w' \models \mathit{correct}_i$*,*

where $\mathcal{H}_i(w) := \{w' : (w, w') \in \mathcal{H}_i\}$*.*

Remark 4.8. *Note that the class* $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ *is not based on any class of Kripke frames.*

We first show that the axioms and inference rules of $\mathscr{H}_{\mathrm{co}}$ from Figure 4.1 are satisfied on this class of Kripke models.

Lemma 4.9. *For all* $M^H \in \mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$*, all* $w \in \mathcal{D}(M^H)$*, all* $\varphi, \psi \in \mathcal{L}_H^{\mathrm{co}}$*, and all* $i \in \mathcal{A}$ *holds the following:*

1. *if* φ *is an instance of a propositional tautology, then* $M^H \models \varphi$*;*
2. $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models H_i\varphi \wedge H_i(\varphi \to \psi) \to H_i\psi$*;*
3. $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models H_i\varphi \to H_iH_i\varphi$*;*
4. $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models \neg H_i\varphi \to H_i\neg H_i\varphi$*;*
5. $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models \mathit{correct}_i \to (H_i\varphi \to \varphi)$*;*
6. $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models \mathit{faulty}_i \to H_i\varphi$*;*
7. $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models H_i\mathit{correct}_i$*;*
8. *If* $M^H, w \models \varphi$ *and* $M^H, w \models \varphi \to \psi$*, then* $M^H, w \models \psi$*;*
9. *if* $M^H \models \varphi$*, then* $M^H \models H_i\varphi$*.*

Proof. 1. This follows immediately from the fact that the interpretation of $\wedge$ and $\neg$ in the definition of the $\models$ relation is the same as in propositional logic.

2. Take an arbitrary model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ and an arbitrary state $w \in W$. Assume $M^H, w \models H_i\varphi \wedge H_i(\varphi \to \psi)$. This means that for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$, we have both $M^H, w' \models \varphi$ and $M^H, w' \models \varphi \to \psi$. By the definition of the $\models$ relation, we have that $M^H, w' \models \psi$ must then also hold. Therefore, $M^H, w \models H_i\psi$ indeed holds.

3. Take an arbitrary model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_\mathsf{n}^\mathsf{co}$ an an arbitrary state $w \in W$. Assume $M^H, w \models H_i\varphi$. Consider an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{H}_i$ and an arbitrary $w'' \in W$ such that $(w', w'') \in \mathcal{H}_i$. Since $\mathcal{H}_i$ is transitive according to Definition 4.7, we also have $(w, w'') \in \mathcal{H}_i$. Therefore, $M^H, w'' \models \varphi$ must hold, by assumption. Since $w'' \in W$ was chosen arbitrarily, now $M^H, w' \models H_i\varphi$ follows. Finally, since $w' \in W$ was also chosen arbitrarily, we obtain $M^H, w \models H_iH_i\varphi$.

4. Take an arbitrary model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_\mathsf{n}^\mathsf{co}$ and an arbitrary state $w \in W$. Assume $M^H, w \models \neg H_i\varphi$. This means that there exists some $w' \in W$ such that $(w, w') \in \mathcal{H}_i$ and $M^H, w' \not\models \varphi$. Consider now an arbitrary $w'' \in W$ such that $(w, w'') \in \mathcal{H}_i$. Since $\mathcal{H}_i$ is euclidean according to Definition 4.7, we have that $(w'', w') \in \mathcal{H}_i$ holds too in this case. Thus, $M^H, w'' \models \neg H_i\varphi$. Since this is true for all $w'' \in W$ such that $(w, w'') \in \mathcal{H}_i$, $M^H, w \models H_i\neg H_i\varphi$ follows.

5. Take an arbitrary model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_\mathsf{n}^\mathsf{co}$ and an arbitrary state $w \in W$. Assume $M^H, w \models \mathit{correct}_i$. According to Definition 4.7, it follows that $w \in \mathcal{H}_i(w)$, i.e., $(w, w) \in \mathcal{H}_i$. Now suppose that $M^H, w \models H_i\varphi$ holds as well. This means that, for any $w' \in W$ such that $(w, w') \in \mathcal{H}_i$, $M^H, w' \models \varphi$ holds. Since we do have that $(w, w) \in \mathcal{H}_i$, it follows that $M^H, w \models \varphi$ holds, as desired.

6. Take an arbitrary model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_\mathsf{n}^\mathsf{co}$ and an arbitrary state $w \in W$. Assume $M^H, w \models \mathit{faulty}_i$. According to Definition 4.7, it follows that $\mathcal{H}_i(w) = \varnothing$, i.e., there exists no $w' \in W$ such that $(w, w') \in \mathcal{H}_i$. To prove $M^H, w \models H_i\varphi$, we have to show that $M^H, w' \models \varphi$ holds for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$. Since there are no such $w' \in W$, we have that $M^H, w \models H_i\varphi$ vacuously holds.

7. Take an arbitrary model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_\mathsf{n}^\mathsf{co}$ and an arbitrary state $w \in W$. To prove $M^H, w \models H_i\mathit{correct}_i$, we need to show that $M^H, w' \models \mathit{correct}_i$ holds for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$. But this follows immediately from Definition 4.7.

8. This follows immediately from the fact that the interpretation of $\wedge$ and $\neg$ in the definition of the $\models$ relation is the same as in propositional logic.

9. If $M^H \models \varphi$, then $M^H, w' \models \varphi$ holds for all states $w' \in W$. In particular, for any state $w \in W$, it follows that $M^H, w' \models \varphi$ for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$. Thus, we have $M^H, w \models H_i\varphi$ for all $w \in W$. Hence, $M^H \models H_i\varphi$. □

Theorem 4.10 (Soundness)**.** *The axiom system $\mathscr{H}_{\mathrm{co}}$ is sound with respect to the $\mathsf{K45}_\mathsf{n}^\mathsf{co}$ class of models.*

Proof. For an arbitrary $\varphi \in \mathcal{L}_H^{\mathrm{co}}$, using the close correspondence of Lemma 4.9 and Figure 4.1, it is straightforward to prove, by induction on the length of the derivation of φ in $\mathscr{H}_{\mathrm{co}}$, that if φ is $\mathscr{H}_{\mathrm{co}}$-provable, then it is also valid with respect to class $\mathsf{K45}_\mathsf{n}^\mathsf{co}$. □

Definition 4.11. *A formula $\varphi \in \mathcal{L}_H^{\text{co}}$ is consistent with respect to $\mathscr{H}_{\text{co}}$ if $\neg\varphi$ is not $\mathscr{H}_{\text{co}}$-provable, i.e., there does not exist a proof of $\neg\varphi$ in $\mathscr{H}_{\text{co}}$. A finite set $\{\varphi_1, \ldots, \varphi_k\} \subseteq \mathcal{L}_H^{\text{co}}$ of formulas is consistent with respect to $\mathscr{H}_{\text{co}}$ exactly if $\varphi_1 \wedge \cdots \wedge \varphi_k$ is consistent with respect to $\mathscr{H}_{\text{co}}$, and an infinite set of formulas is consistent with respect to $\mathscr{H}_{\text{co}}$ exactly if all of its finite subsets are consistent with respect to $\mathscr{H}_{\text{co}}$. Furthermore, a set $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ of formulas is maximal consistent with respect to $\mathscr{H}_{\text{co}}$ if*

- *Γ is consistent with respect to $\mathscr{H}_{\text{co}}$, and*
- *for all φ in $\mathcal{L}_H^{\text{co}}$ but not in Γ, the set $\Gamma \cup \{\varphi\}$ is not consistent with respect to $\mathscr{H}_{\text{co}}$.*

The following technical lemma will prove to be useful when proving various results throughout the chapter:

Lemma 4.12. *Let $\varphi, \psi \in \mathcal{L}_H^{\text{co}}$ and $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ be a maximal consistent set with respect to $\mathscr{H}_{\text{co}}$. Then the following holds:*

1. $\vdash_{\mathscr{H}_{\text{co}}} \varphi \quad \Longrightarrow \quad \varphi \in \Gamma$;
2. $\varphi \in \Gamma \quad \Longleftrightarrow \quad \neg\varphi \notin \Gamma$;
3. $\varphi \wedge \psi \in \Gamma \quad \Longleftrightarrow \quad \varphi \in \Gamma$ *and* $\psi \in \Gamma$.

Proof. 1. Let $\vdash_{\mathscr{H}_{\text{co}}} \varphi$. Assume towards a contradiction that $\varphi \notin \Gamma$. Given the assumption that Γ is maximal consistent with respect to $\mathscr{H}_{\text{co}}$, we get that $\Gamma \cup \{\varphi\}$ must be inconsistent with respect to $\mathscr{H}_{\text{co}}$. This means that there exists a finite set of formulas

$$\{\theta_1, \ldots, \theta_k\} \subseteq \Gamma \cup \{\varphi\}$$

such that

$$\vdash_{\mathscr{H}_{\text{co}}} \neg(\theta_1 \wedge \cdots \wedge \theta_k).$$

Since Γ is assumed to be consistent with respect to $\mathscr{H}_{\text{co}}$, it follows that $\varphi \equiv \theta_i$ must hold for some $i \in \{1, \ldots, k\}$. Using propositional reasoning, we, thus, further obtain $\vdash_{\mathscr{H}_{\text{co}}} \varphi \to \neg(\theta_1 \wedge \ldots \theta_{i-1} \wedge \theta_{i+1} \wedge \cdots \wedge \theta_k)$, which, combined with $\vdash_{\mathscr{H}_{\text{co}}} \varphi$, results in $\vdash_{\mathscr{H}_{\text{co}}} \neg(\theta_1 \wedge \ldots \theta_{i-1} \wedge \theta_{i+1} \wedge \cdots \wedge \theta_k)$, contradicting the consistency of Γ with respect to $\mathscr{H}_{\text{co}}$.

2. ($\Longrightarrow$): Let $\varphi \in \Gamma$. Assume towards a contradiction that $\neg\varphi \in \Gamma$. Given that $\{\varphi, \neg\varphi\} \subseteq \Gamma$ is inconsistent with respect to $\mathscr{H}_{\text{co}}$, we obtain a contradiction with the assumption that Γ is consistent with respect to $\mathscr{H}_{\text{co}}$.

   ($\Longleftarrow$): Let $\neg\varphi \notin \Gamma$. Assume towards a contradiction that $\varphi \notin \Gamma$. Given the assumption that Γ is maximal consistent with respect to $\mathscr{H}_{\text{co}}$, we get that $\Gamma \cup \{\neg\varphi\}$ and $\Gamma \cup \{\varphi\}$ must be inconsistent with respect to $\mathscr{H}_{\text{co}}$. This means that there exist finite sets of formulas

$$\{\theta_1, \ldots, \theta_k\} \subseteq \Gamma \cup \{\neg\varphi\}$$

and

$$\{\theta'_1, \ldots, \theta'_l\} \subseteq \Gamma \cup \{\varphi\}$$

such that

$$\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta_1 \wedge \cdots \wedge \theta_k)$$

and

$$\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta'_1 \wedge \cdots \wedge \theta'_l).$$

Since Γ is assumed to be consistent with respect to $\mathscr{H}_{\mathrm{co}}$, it follows that $\neg\varphi \equiv \theta_i$ and $\varphi \equiv \theta_j$ must hold for some $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, l\}$. Thus, we have $\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta_1 \wedge \cdots \wedge \theta_{i-1} \wedge \neg\varphi \wedge \theta_{i+1} \wedge \cdots \wedge \theta_k)$ and $\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta'_1 \wedge \cdots \wedge \theta'_{j-1} \wedge \varphi \wedge \theta'_{j+1} \wedge \cdots \wedge \theta'_j)$. Let $\xi := \theta_1 \wedge \cdots \wedge \theta_{i-1} \wedge \theta_{i+1} \wedge \cdots \wedge \theta_k$ and $\xi' := \theta'_1 \wedge \cdots \wedge \theta'_{j-1} \wedge \theta'_{j+1} \wedge \cdots \wedge \theta'_j$. So, now $\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\xi \wedge \neg\varphi)$ and $\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\xi' \wedge \varphi)$. From this, using the propositional tautology $\neg(\xi \wedge \neg\varphi) \wedge \neg(\xi' \wedge \varphi) \rightarrow \neg(\xi \wedge \xi')$, we obtain $\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\xi \wedge \xi')$, i.e, $\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta_1 \wedge \cdots \wedge \theta_{i-1} \wedge \theta_{i+1} \wedge \cdots \wedge \theta_k \wedge \theta'_1 \wedge \cdots \wedge \theta'_{j-1} \wedge \theta'_{j+1} \wedge \cdots \wedge \theta'_j)$ contradicting the consistency of Γ with respect to $\mathscr{H}_{\mathrm{co}}$.

3. ($\Longrightarrow$): Let $\varphi \wedge \psi \in \Gamma$. Assume towards a contradiction that $\varphi \notin \Gamma$. From 2., $\neg\varphi \in \Gamma$ follows. Given that the set $\{\varphi \wedge \psi, \neg\varphi\} \subseteq \Gamma$ is inconsistent with respect to $\mathscr{H}_{\mathrm{co}}$, we obtain a contradiction with the assumption that Γ is consistent with respect to $\mathscr{H}_{\mathrm{co}}$. We prove that $\psi \in \Gamma$ holds analogously.

   ($\Longleftarrow$): Let $\varphi \in \Gamma$ and $\psi \in \Gamma$. Assume towards a contradiction that $\varphi \wedge \psi \notin \Gamma$. From 2., $\neg(\varphi \wedge \psi) \in \Gamma$ follows. Given that the set $\{\varphi, \psi, \neg(\varphi \wedge \psi)\} \subseteq \Gamma$ is inconsistent with respect to $\mathscr{H}_{\mathrm{co}}$, we obtain a contradiction with the assumption that Γ is consistent with respect to $\mathscr{H}_{\mathrm{co}}$. □

Lemma 4.13. *Let $\varphi, \psi \in \mathcal{L}_H^{\mathrm{co}}$ and $\Gamma \subseteq \mathcal{L}_H^{\mathrm{co}}$ be a maximal consistent set with respect to $\mathscr{H}_{\mathrm{co}}$. If $\varphi \in \Gamma$ and $\varphi \rightarrow \psi \in \Gamma$, then $\psi \in \Gamma$.*

Proof. Let $\varphi \in \Gamma$ and $\varphi \rightarrow \psi \in \Gamma$. Assume towards a contradiction that $\psi \notin \Gamma$. From Lemma 4.12 (2), $\neg\psi \in \Gamma$ follows. Given that the set $\{\varphi, \varphi \rightarrow \psi, \neg\psi\} \subseteq \Gamma$ is inconsistent with respect to $\mathscr{H}_{\mathrm{co}}$, we obtain a contradiction with the assumption that Γ is consistent with respect to $\mathscr{H}_{\mathrm{co}}$. □

Lemma 4.14 (Lindenbaum lemma)**.** *Let $\Gamma \subseteq \mathcal{L}_H^{\mathrm{co}}$. If Γ is consistent with respect to $\mathscr{H}_{\mathrm{co}}$, then there exists a set $\Gamma^* \supseteq \Gamma$ such that Γ^* is maximal consistent with respect to $\mathscr{H}_{\mathrm{co}}$.*

Proof. Assume that Γ is consistent with respect to $\mathscr{H}_{\mathrm{co}}$. First, let us enumerate all formulas from $\mathcal{L}_H^{\mathrm{co}}$ (without repetitions): $\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots$ Next, we define recursively the following infinite sequence of sets Δ_i:

$$\Delta_0 := \Gamma,$$

$$\Delta_{i+1} := \begin{cases} \Delta_i \cup \{\varphi_i\}, & \text{if } \Delta_i \cup \{\varphi_i\} \text{ is consistent with respect to } \mathscr{H}_{\mathrm{co}}, \\ \Delta_i, & \text{otherwise.} \end{cases}$$

Let $\Gamma^* := \bigcup_{i=0}^{\infty} \Delta_i$. Obviously, $\Gamma^* \supseteq \Gamma$. Each set Δ_i is consistent with respect to $\mathscr{H}_{\text{co}}$, by construction. To show that Γ^* too is consistent with respect to $\mathscr{H}_{\text{co}}$, assume the opposite towards a contradiction: there exists a finite set $\Gamma' \subseteq \Gamma^*$ such that Γ' is inconsistent with respect to $\mathscr{H}_{\text{co}}$. Therefore, $\Gamma' \subseteq \Delta_j$, for some $j \in \mathbb{N}_0$, must hold since $\Delta_0 \subseteq \Delta_1 \subseteq \cdots \subseteq \Delta_n \subseteq \ldots$. However, this now contradicts the consistency of Δ_j. To show that Γ^* is maximal as well, assume the opposite towards a contradiction: there exists $\varphi_j \in \mathcal{L}_H^{\text{co}}$ such that $\varphi_j \notin \Gamma^*$ and $\Gamma^* \cup \{\varphi_j\}$ is consistent with respect to $\mathscr{H}_{\text{co}}$. It follows that $\Delta_j \cup \{\varphi_j\}$ must be inconsistent with respect to $\mathscr{H}_{\text{co}}$ (by the definition of sets Δ_i). However, this now contradicts the consistency of $\Gamma^* \cup \{\varphi_j\}$ since $\Delta_j \cup \{\varphi_j\} \subseteq \Gamma^* \cup \{\varphi_j\}$. □

Definition 4.15 (Canonical model)**.** *We define the so-called canonical Kripke model*

$$M^c = (W^c, \pi^c, \mathcal{H}_1^c, \ldots, \mathcal{H}_n^c)$$

of $\mathscr{H}_{\text{co}}$ *in the following way:*

$$W^c = \{w_\Gamma : \Gamma \textit{ is some maximal consistent set with respect to } \mathscr{H}_{\text{co}}\},$$
$$\pi^c(p) = \{w_\Gamma \in W^c \mid p \in \Gamma\},$$
$$\mathcal{H}_i^c = \{(w_\Gamma, w_\Delta) : \Gamma/H_i \subseteq \Delta\},$$

where $\Gamma/H_i = \{\varphi : H_i\varphi \in \Gamma\}$.

Lemma 4.16 (Truth lemma)**.** *For any formula* $\varphi \in \mathcal{L}_H^{\text{co}}$*, and any maximal consistent set* Γ *with respect to* $\mathscr{H}_{\text{co}}$*,*

$$\varphi \in \Gamma \quad \Longleftrightarrow \quad M^c, w_\Gamma \models \varphi.$$

Proof. We proceed by induction on the structure of φ.

Base case: If φ is $p \in \mathsf{Prop} \cup \mathsf{Co}$, then the statement of the lemma follows immediately from the definition of π^c.

Induction step:

1. If φ is of the form $\neg\psi$, then $\varphi \in \Gamma$ is equivalent to $\neg\psi \in \Gamma$, which is further equivalent to $\psi \notin \Gamma$ by Lemma 4.12 (2). By the induction hypothesis, this is now equivalent to $M^c, w_\Gamma \not\models \psi$, i.e., $M^c, w_\Gamma \models \neg\psi$, i.e., $M^c, w_\Gamma \models \varphi$.

2. If φ is of the form $\psi_1 \wedge \psi_2$, then $\varphi \in \Gamma$ is equivalent to $\psi_1 \wedge \psi_2 \in \Gamma$ which is further equivalent to $\psi_1 \in \Gamma$ and $\psi_2 \in \Gamma$ by Lemma 4.12 (3). By the induction hypothesis, this is now equivalent to $M^c, w_\Gamma \models \psi_1$ and $M^c, w_\Gamma \models \psi_2$, i.e., $M^c, w_\Gamma \models \psi_1 \wedge \psi_2$, i.e., $M^c, w_\Gamma \models \varphi$.

3. Assume that φ is of the form $H_i\psi$.

($\Longrightarrow$): Let $H_i\psi \in \Gamma$. Take an arbitrary $w_\Delta \in W^c$ such that $(w_\Gamma, w_\Delta) \in \mathcal{H}_i^c$. We now have that $\psi \in \Delta$ must hold according to the definition of $\mathcal{H}_i^c$. By applying the induction hypothesis, we obtain $M^c, w_\Delta \models \psi$. Therefore, $M^c, w_\Gamma \models H_i\psi$ indeed holds.

($\Longleftarrow$): Let $M^c, w_\Gamma \models H_i\psi$. It follows that the set $(\Gamma/H_i) \cup \{\neg\psi\}$ must be inconsistent with respect to $\mathscr{H}_{\text{co}}$, because otherwise there would exist a maximal consistent set Δ with respect to $\mathscr{H}_{\text{co}}$ extending it (according to the Lindenbaum lemma 4.14), and, by construction, we would have $(w_\Gamma, w_\Delta) \in \mathcal{H}_i^c$. Using the induction hypothesis, we would then obtain $M^c, w_\Delta \not\models \psi$, and so $M^c, w_\Gamma \not\models H_i\psi$, contradicting our original assumption. Since $(\Gamma/H_i) \cup \{\neg\psi\}$ is hence indeed inconsistent with respect to $\mathscr{H}_{\text{co}}$, some finite subset, say $\{\varphi_1, \dots \varphi_k, \neg\psi\}$, must be inconsistent with respect to $\mathscr{H}_{\text{co}}$. This means that $\neg(\varphi_1 \wedge \dots \wedge \varphi_k \wedge \neg\psi)$ is $\mathscr{H}_{\text{co}}$-provable. Thus, by propositional reasoning, we have

$$\vdash_{\mathscr{H}_{\text{co}}} \varphi_1 \to (\varphi_2 \to (\dots \to (\varphi_k \to \psi) \dots)),$$

and by necessitation,

$$\vdash_{\mathscr{H}_{\text{co}}} H_i(\varphi_1 \to (\varphi_2 \to (\dots \to (\varphi_k \to \psi) \dots))).$$

By induction on k, using axiom K^H and propositional reasoning, it is straightforward to prove

$$\vdash_{\mathscr{H}_{\text{co}}} H_i(\varphi_1 \to (\varphi_2 \to (\dots \to (\varphi_k \to \psi) \dots))) \to (H_i\varphi_1 \to (H_i\varphi_2 \to (\dots \to (H_i\varphi_k \to H_i\psi) \dots))).$$

Now, by modus ponens, we finally get

$$\vdash_{\mathscr{H}_{\text{co}}} H_i\varphi_1 \to (H_i\varphi_2 \to (\dots \to (H_i\varphi_k \to H_i\psi) \dots)).$$

It follows that the set $\{H_i\varphi_1, \dots, H_i\varphi_k, \neg H_i\psi\}$ is inconsistent with respect to $\mathscr{H}_{\text{co}}$ as well. By the definition of Γ/H_i, since $\varphi_1, \dots, \varphi_k \in \Gamma/H_i$, it must be $H_i\varphi_1, \dots, H_i\varphi_k \in \Gamma$. As either $H_i\psi$ or $\neg H_i\psi$ is in Γ according to Lemma 4.12 (2), we must in fact have $H_i\psi \in \Gamma$ or else Γ would not be consistent with respect to $\mathscr{H}_{\text{co}}$. □

Lemma 4.17 (Correctness lemma)**.** $M^c \in \mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$.

Proof. We need to show that the four conditions from Definition 4.7 are satisfied for $M^c = (W^c, \pi^c, \mathcal{H}_1^c, \dots, \mathcal{H}_n^c)$[1].

[1]Note that $W^c \neq \varnothing$ since any consistent set (for example, the set $\{\top\}$, which we know is consistent with respect to $\mathscr{H}_{\text{co}}$ because $\nvdash_{\mathscr{H}_{\text{co}}} \neg\top$ by Theorem 4.10) with respect to $\mathscr{H}_{\text{co}}$ can be extended to a maximal consistent with respect to $\mathscr{H}_{\text{co}}$, according to Lindenbaum lemma 4.14.

1. Assume $(w_\Gamma, w_\Delta) \in \mathcal{H}_i^c$ and $\varphi \in \Gamma/H_i$. Then $H_i\varphi \in \Gamma$, according to the definition of Γ/H_i. Axiom 4^H, Lemma 4.12 (1), and Lemma 4.13 imply $H_iH_i\varphi \in \Gamma$, which further implies $H_i\varphi \in \Gamma/H_i$. Using the definition of $\mathcal{H}_i^c$, we also obtain $H_i\varphi \in \Delta$. Thus, $\varphi \in \Delta/H_i$ follows, which implies that $\Gamma/H_i \subseteq \Delta/H_i$ holds. Assume now $\varphi \notin \Gamma/H_i$. Then $H_i\varphi \notin \Gamma$, according to the definition of Γ/H_i. Axiom 5^H, Lemma 4.12 (1–2), and Lemma 4.13 imply $H_i\neg H_i\varphi \in \Gamma$, which further implies $\neg H_i\varphi \in \Gamma/H_i$. Using the definition of $\mathcal{H}_i^c$, we also get $\neg H_i\varphi \in \Delta$. Since Δ is a maximal consistent set with respect to $\mathscr{H}_{\text{co}}$, we now have $H_i\varphi \notin \Delta$ according to Lemma 4.12 (2). Thus, $\varphi \notin \Delta/H_i$ follows, which implies that $\Delta/H_i \subseteq \Gamma/H_i$ holds as well. Transitivity and euclideanity now easily follow using the fact that $(w_\Gamma, w_\Delta) \in \mathcal{H}_i^c$ implies $\Gamma/H_i = \Delta/H_i$: For transitivity, assume $(w_\Gamma, w_\Delta) \in \mathcal{H}_i^c$ and $(w_\Delta, w_U) \in \mathcal{H}_i^c$. By definition, $\Delta/H_i \subseteq U$ follows ($\Gamma/H_i \subseteq \Delta$ follows as well). From $\Delta/H_i \subseteq U$ and $\Gamma/H_i = \Delta/H_i$ we obtain $\Gamma/H_i \subseteq U$, which implies $(w_\Gamma, w_U) \in \mathcal{H}_i^c$, as desired. The proof of euclideanity is similar.

2. Let $w_\Gamma \in W^c$ be a state satisfying $M^c, w_\Gamma \models \mathit{correct}_i$ and let $\varphi \in \Gamma/H_i$, that is, $H_i\varphi \in \Gamma$. Using Lemma 4.16, we obtain $\mathit{correct}_i \in \Gamma$. From axiom T'^H, Lemma 4.12 (1), and Lemma 4.13, $\varphi \in \Gamma$ follows. Since φ was chosen arbitrarily, this means that $\Gamma/H_i \subseteq \Gamma$ holds. By the definition of $\mathcal{H}_i^c$, this implies $w_\Gamma \in \mathcal{H}_i^c(w_\Gamma)$.

3. Let $w_\Gamma \in W^c$ be a state satisfying $M^c, w_\Gamma \models \mathit{faulty}_i$. Using Lemma 4.16, we obtain $\mathit{faulty}_i \in \Gamma$. From axiom F, Lemma 4.12 (1), and Lemma 4.13, $H_i\varphi \in \Gamma$ follows for any φ. This means that the set Γ/H_i contains all formulas. Since there does not exist a maximal consistent set Δ with respect to $\mathscr{H}_{\text{co}}$ such that $\Gamma/H_i \subseteq \Delta$ holds, we obtain $\mathcal{H}_i^c(w_\Gamma) = \varnothing$.

4. Let $(w_\Gamma, w_\Delta) \in \mathcal{H}_i^c$. Since $H_i\mathit{correct}_i$ is axiom H, from Lemma 4.12 (1) we get $H_i\mathit{correct}_i \in \Gamma$ and, consequently, $\mathit{correct}_i \in \Gamma/H_i$. Using the definition of $\mathcal{H}_i^c$, we conclude that $\mathit{correct}_i \in \Delta$ holds as well. Using Lemma 4.16, we thus obtain $M^H, w_\Delta \models \mathit{correct}_i$. □

Theorem 4.18 (Completeness)**.** *The axiom system $\mathscr{H}_{\text{co}}$ is complete with respect to the $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ class of models.*

Proof. We prove the contrapositive. Assume $\nvdash_{\mathscr{H}_{\text{co}}} \varphi$. Therefore, $\{\neg\varphi\}$ must be consistent with respect to $\mathscr{H}_{\text{co}}$. Using the Lindenbaum lemma 4.14, we obtain that $\{\neg\varphi\}$ is contained in some maximal consistent set Γ with respect to $\mathscr{H}_{\text{co}}$. According to Lemma 4.16, it thus follows $M^c, w_\Gamma \models \neg\varphi$, i.e., $M^c, w_\Gamma \not\models \varphi$, where M^c is the canonical Kripke model for $\mathscr{H}_{\text{co}}$ defined in Figure 4.1. Therefore, $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \not\models \varphi$ since $M^c \in \mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ as shown in Lemma 4.17. □

Corollary 4.19. *The axiom system $\mathscr{H}_{\text{co}}$ is sound and complete with respect to the $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ class of models.*

Completeness with respect to S5 models via translation

Lemma 4.20. *For any model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}^{\mathsf{co}}_{\mathsf{n}}$, there exists a corresponding model $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5}_{\mathsf{n}}$, such that the following holds for any formula $\varphi \in \mathcal{L}^{\mathrm{co}}_H$ and any state $w \in W$:*

$$M^H, w \models \varphi \quad \textit{iff} \quad M^K, w \models t(\varphi).$$

Proof. Let $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}^{\mathsf{co}}_{\mathsf{n}}$. The corresponding model is constructed by taking

$$\mathcal{K}_i := \mathcal{H}_i \cup \{(w, w) \mid M^K, w \models \mathit{faulty}_i\} \tag{4.1}$$

for all $i \in \{1, \ldots, n\}$. We proceed by induction on the structure of φ.

Base case: If φ is $p \in \mathsf{Prop} \cup \mathsf{Co}$, then $t(p) = p$ according to Definition 4.4. Now, the statement of the lemma immediately follows as π is same in M^H and M^K.

Induction step:

1. If φ is of the form $\neg\psi$, then $t(\neg\psi) = \neg t(\psi)$ according to Definition 4.4. Using the induction hypothesis, we obtain $M^H, w \models \varphi$ iff $M^H, w \models \neg\psi$ iff $M^H, w \not\models \psi$ iff $M^K, w \not\models t(\psi)$ iff $M^K, w \models \neg t(\psi)$ iff $M^K, w \models t(\neg\psi)$ iff $M^K, w \models t(\varphi)$.

2. If φ is of the form $\psi_1 \wedge \psi_2$, then $t(\psi_1 \wedge \psi_2) = t(\psi_1) \wedge t(\psi_2)$ according to Definition 4.4. Using the induction hypothesis, we obtain $M^H, w \models \varphi$ iff $M^H, w \models \psi_1 \wedge \psi_2$ iff $M^H, w \models \psi_1$ and $M^H, w \models \psi_2$ iff $M^K, w \models t(\psi_1)$ and $M^K, w \models t(\psi_2)$ iff $M^K, w \models t(\psi_1) \wedge t(\psi_2)$ iff $M^K, w \models t(\psi_1 \wedge \psi_2)$ iff $M^K, w \models t(\varphi)$.

3. Assume that φ is of the form $H_i\psi$. We need to show

$$M^H, w \models H_i\psi \quad \textit{iff} \quad M^K, w \models \mathit{correct}_i \to K_i(\mathit{correct}_i \to t(\psi)).$$

   ($\Rightarrow$): Assume $M^H, w \models H_i\psi$. This means that for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$, $M^H, w' \models \psi$ holds. Assume now $M^K, w \models \mathit{correct}_i$. Take an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{K}_i$. By (4.1), this means that either $(w, w') \in \mathcal{H}_i$ or $w' \equiv w$ and $M^K, w \models \mathit{faulty}_i$. Since, by assumption, $M^K, w \models \mathit{correct}_i$ (i.e., $M^K, w \not\models \mathit{faulty}_i$), it follows that $(w, w') \in \mathcal{H}_i$ must hold. Therefore, $M^H, w' \models \psi$ follows by assumption. From this, by applying the induction hypothesis, we obtain $M^K, w' \models t(\psi)$. Consequently, $M^K, w' \models \mathit{correct}_i \to t(\psi)$ also holds. Since $w' \in W$ was chosen arbitrarily, we get $M^K, w \models K_i(\mathit{correct}_i \to t(\psi))$. Thus, $M^K, w \models \mathit{correct}_i \to K_i(\mathit{correct}_i \to t(\psi))$, i.e., $M^K, w \models t(H_i\psi)$ indeed holds.

   ($\Leftarrow$): Assume $M^K, w \models t(H_i\psi) = \mathit{correct}_i \to K_i(\mathit{correct}_i \to t(\psi))$. Therefore, either $M^K, w \not\models \mathit{correct}_i$ or $M^K, w \models K_i(\mathit{correct}_i \to t(\psi))$. If $M^K, w \not\models \mathit{correct}_i$, then $M^H, w \not\models \mathit{correct}_i$, i.e., $M^H, w \models \mathit{faulty}_i$ follows as shown in the base case. Thus, we get $M^H, w \models H_i\psi$ using Lemma 4.9. Assume now that $M^K, w \models K_i(\mathit{correct}_i \to t(\psi))$. This means that for all $w' \in W$ such that

$(w, w') \in \mathcal{K}_i$, we have that $M^K, w' \models \mathit{correct}_i \to t(\psi)$ holds, that is, either $M^K, w' \not\models \mathit{correct}_i$ or $M^K, w' \models t(\psi)$. In particular, for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$, either $M^K, w' \not\models \mathit{correct}_i$ or $M^K, w' \models t(\psi)$. If $M^K, w' \not\models \mathit{correct}_i$, then $M^H, w' \not\models \mathit{correct}_i$, i.e., $M^H, w' \models \mathit{faulty}_i$ holds too as shown in the base case. Since there exists no $w' \in W$ such that $(w, w') \in \mathcal{H}_i$ and $M^H, w' \models \mathit{faulty}_i$ according to Definition 4.7, it follows that for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$, $M^K, w' \models t(\psi)$ holds. By applying the induction hypothesis now, we obtain that $M^H, w' \models \psi$ holds for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$. Thus, $M^H, w \models H_i\psi$ indeed holds.

It remains to show that $M^K \in \mathsf{S5_n}$, i.e., it remains to show that the relations $\mathcal{K}_i$ are reflexive, transitive and euclidean. Transitivity and euclideanity follow based on the fact that $\mathcal{H}_i$ satisfy these properties. To show reflexivity, let $w \in W$. If $M^K, w \models \mathit{correct}_i$, then $M^H, w \models \mathit{correct}_i$ holds too, thus $(w, w) \in \mathcal{H}_i$ follows by Definition 4.7. Therefore, $(w, w) \in \mathcal{K}_i$ holds too according to the definition of $\mathcal{K}_i$. If $M^K, w \models \mathit{faulty}_i$, then $(w, w) \in \mathcal{K}_i$ immediately follows from the definition of $\mathcal{K}_i$. □

Hence, we can show:

Theorem 4.21 (Completeness via translation)**.** *For any formula $\varphi \in \mathcal{L}_H^{\mathrm{co}}$,*

$$\mathsf{S5_n} \models t(\varphi) \implies \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi.$$

Proof. We prove the contrapositive. Let $\nvdash_{\mathscr{H}_{\mathrm{co}}} \varphi$. Corollary 4.19 now implies that $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \not\models \varphi$ must also hold. This means that there exists a model $M^H \in \mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ such that $M^H \not\models \varphi$. We now have that there exists a state $w \in \mathcal{D}(M^H)$ such that $M^H, w \not\models \varphi$. According to Lemma 4.20, there exists a corresponding model $M^K \in \mathsf{S5_n}$ such that $M^K, w \not\models t(\varphi)$. Therefore, $M^K \not\models t(\varphi)$ holds as well. Finally, we obtain $\mathsf{S5_n} \not\models t(\varphi)$, as desired. □

We gather the results stated in Theorem 4.6 and Theorem 4.21 in the following corollary.

Corollary 4.22. *For any formula $\varphi \in \mathcal{L}_H^{\mathrm{co}}$,*

$$\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi \iff \mathsf{S5_n} \models t(\varphi).$$

which finally allows us to conclude the following:

$$\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \models \varphi \iff \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi \iff \mathsf{S5_n} \models t(\varphi) \iff \vdash_{\mathscr{S}5_n} t(\varphi).$$

Strong soundness and strong completeness

We will now show that the axiom system $\mathscr{H}_{\mathrm{co}}$ is also strongly sound and strongly complete with respect to the $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}}$ class of models.

First, we need to know how to derive a formula from a set of premises.

Definition 4.23. *A $\mathscr{H}_{\text{co}}$-derivation from premises $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ is a sequence of formulas $\varphi_1, \ldots, \varphi_k \in \mathcal{L}_H^{\text{co}}$ such that for each $i = 1, \ldots, k$:*

- *$\varphi_i = H_{a_m} \ldots H_{a_1} \xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\text{co}}$, or*
- *φ_i follows from $\varphi_{j_1} = \varphi_{j_2} \to \varphi_i$ and φ_{j_2} by modus ponens for some $j_1, j_2 < i$, or*
- *$\varphi_i \in \Gamma$.*

A $\mathscr{H}_{\text{co}}$-derivation (from $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$) $\varphi_1, \ldots, \varphi_k$ is a $\mathscr{H}_{\text{co}}$-derivation (from $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$) for φ_k.

We will write $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \varphi$ to denote the fact that φ can be derived from Γ in $\mathscr{H}_{\text{co}}$.

Remark 4.24. *Note that necessation is omitted in the second clause of the Definition 4.23. Therefore, we do not allow derivations of the form*

$$\varphi \vdash_{\mathscr{H}_{\text{co}}} H_i \varphi. \tag{4.2}$$

This is because (4.2) combined with the Deduction theorem 4.29 would lead to $\vdash_{\mathscr{H}_{\text{co}}} \varphi \to H_i\varphi$ (and we know that $\nvdash_{\mathscr{H}_{\text{co}}} \varphi \to H_i\varphi$ holds since $\mathsf{K45}_{\mathsf{n}}^{\mathsf{co}} \not\models \varphi \to H_i\varphi$).

The following lemma states that $\vdash_{\mathscr{H}_{\text{co}}}$ is closed with respect to modus ponens.

Lemma 4.25. *Let $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ and $\varphi, \psi \in \mathcal{L}_H^{\text{co}}$. If $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \varphi$ and $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \varphi \to \psi$, then $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \psi$.*

Proof. Follows immediately according to Definition 4.23. □

Theorem 4.26. *For any $\varphi \in \mathcal{L}_H^{\text{co}}$,*

$$\vdash_{\mathscr{H}_{\text{co}}} \varphi \quad \Longleftrightarrow \quad \varnothing \vdash_{\mathscr{H}_{\text{co}}} \varphi.$$

Proof. Let $\varphi \in \mathcal{L}_H^{\text{co}}$.

($\Longrightarrow$): Assume $\vdash_{\mathscr{H}_{\text{co}}} \varphi$. We proceed by induction on the length of the derivation of φ in $\mathscr{H}_{\text{co}}$. Let $\varphi_1, \varphi_2, \ldots, \varphi_k$ be a proof of the formula φ in $\mathscr{H}_{\text{co}}$. If $k = 1$, then φ is an instance of an axiom in $\mathscr{H}_{\text{co}}$. Therefore, according to Definition 4.23, it immediately follows that $\varnothing \vdash_{\mathscr{H}_{\text{co}}} \varphi$. Assume now that the desired statement holds for every formula which has a proof of length shorter than k and consider the formula φ which has a proof of length k. There are two possibilities: either φ is an instance of some axiom in $\mathscr{H}_{\text{co}}$, or φ follows by an application of an inference rule. We already dealt with the first possibility in the base case, so we consider the second possibility.

– Let us assume that φ follows by an application of modus ponens to some formulas, for example φ_i and $\varphi_i \to \varphi$. Since these two formulas are earlier in the proof, they have proofs whose lengths are shorter than k. After applying the induction hypothesis on them we get $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi_i$ and $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi_i \to \varphi$. Using Lemma 4.25, we obtain that $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi$ indeed holds.

– Let us assume that φ follows by an application of necessitation to some formula, for example φ_i, meaning φ is of the form $H_{a_j}\varphi_i$, for some $a_j \in \mathcal{A}$. Since φ_i is earlier in the proof it has a proof whose length is shorter than k, which means that we can apply the induction hypothesis on it and get $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi_i$. We proceed by induction on the length of the derivation of φ_i in $\mathscr{H}_{\mathrm{co}}$ from the empty set of premises. Let $\psi_1, \psi_2, \ldots, \psi_l$ be a proof of φ_i in $\mathscr{H}_{\mathrm{co}}$ from the empty set of premises. If $l = 1$, it means that $\varphi_i = H_{a_m} \ldots H_{a_1}\xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\mathrm{co}}$. In this case, according to Definition 4.23, $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}H_{a_m} \ldots H_{a_1}\xi$, that is, $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}\varphi_i$, immediately follows. Assume now that the desired statement holds for every formula which has a proof of length shorter than l and consider the formula φ_i which has a proof of length l. There are two possibilities: either $\varphi_i = H_{a_m} \ldots H_{a_1}\xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\mathrm{co}}$, or φ_i follows by an application of modus ponens. We already dealt with the first possibility in the base of the induction, so we consider the second possibility. Let us assume that φ_i follows by an application of modus ponens to some formulas ψ_h and $\psi_h \to \varphi_i$. Since these two formulas are earlier in the proof, they have proofs whose lengths are shorter than l. After applying the induction hypothesis on them we get $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}\psi_h$ and $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}(\psi_h \to \varphi_i)$. We know that $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}(\psi_h \to \varphi_i) \to (H_{a_j}\psi_h \to H_j\varphi_i)$ holds because it is an instance of the K axiom. Thus, from $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}(\psi_h \to \varphi_i)$ and $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}(\psi_h \to \varphi_i) \to (H_{a_j}\psi_h \to H_j\varphi_i)$, using Lemma 4.25, we obtain $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}\psi_h \to H_j\varphi_i$. Finally, from $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}\psi_h$ and $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}\psi_h \to H_j\varphi_i$, by applying Lemma 4.25 again, we obtain $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_j}\varphi_i$, as desired.

($\Longleftarrow$): Assume $\varnothing \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi$. We proceed by induction on the length of the derivation of φ in $\mathscr{H}_{\mathrm{co}}$ from the empty set of premises. Let $\varphi_1, \varphi_2, \ldots, \varphi_k$ be a proof of the formula φ in $\mathscr{H}_{\mathrm{co}}$ from the empty set of premises. If $k = 1$, then $\varphi = H_{a_m} \ldots H_{a_1}\xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\mathrm{co}}$:

– If $m = 0$, that is $\varphi = \xi$, then we immediately obtain $\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi$,

– If $m > 0$, then, by applying necessitation m times on $\vdash_{\mathscr{H}_{\mathrm{co}}} \xi$, we obtain $\vdash_{\mathscr{H}_{\mathrm{co}}} H_{a_m} \ldots H_{a_1}\xi$, that is, $\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi$, as desired.

Assume now that the desired statement holds for every formula which has a proof of length shorter than k and consider the formula φ which has a proof of length k. There are two possibilities: either $\varphi = H_{a_m} \ldots H_{a_1}\xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\mathrm{co}}$, or φ follows by an application of modus ponens. We already dealt with the first possibility in the base of the induction, so we consider

the second possibility. Let us assume that φ follows by an application of modus ponens to some formulas φ_i and $\varphi_i \to \varphi$. Since these two formulas are earlier in the proof, they have proofs whose lengths are shorter than k. After applying the induction hypothesis on them we get $\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi_i$ and $\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi_i \to \varphi$. By applying modus ponens, we obtain $\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi$, as desired. □

Corollary 4.27. *Let $\Gamma \subseteq \mathcal{L}_H^{\mathrm{co}}$ and $\varphi \in \mathcal{L}_H^{\mathrm{co}}$. Then,*

$$\vdash_{\mathscr{H}_{\mathrm{co}}} \varphi \quad \Longrightarrow \quad \Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi.$$

Using this result, we can now prove:

Lemma 4.28. *Let $\Gamma \subseteq \mathcal{L}_H^{\mathrm{co}}$. If Γ is inconsistent with respect to $\mathscr{H}_{\mathrm{co}}$, then $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \bot$.*

Proof. Assume that Γ is inconsistent with respect to $\mathscr{H}_{\mathrm{co}}$. According to Definition 4.11, this means that there exists a finite set of formulas

$$\{\theta_1, \ldots, \theta_k\} \subseteq \Gamma$$

such that

$$\vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta_1 \wedge \cdots \wedge \theta_k)$$

Using the previous corollary, we obtain that

$$\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \neg(\theta_1 \wedge \cdots \wedge \theta_k)$$

holds as well. However, since $\{\theta_1, \ldots, \theta_k\} \subseteq \Gamma$, we have $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \theta_1 \wedge \cdots \wedge \theta_k$ according to Definition 4.23. Thus, $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \bot$ indeed follows. □

Theorem 4.29 (Deduction theorem)**.** *Let $\Gamma \subseteq \mathcal{L}_H^{\mathrm{co}}$ and $\varphi, \psi \in \mathcal{L}_H^{\mathrm{co}}$. Then,*

$$\Gamma \cup \{\psi\} \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi \quad \Longrightarrow \quad \Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \psi \to \varphi.$$

Proof. We proceed by induction on the length of the derivation of φ from $\Gamma \cup \{\psi\}$ in $\mathscr{H}_{\mathrm{co}}$. Let $\varphi_1, \ldots, \varphi_k$ be a proof of the formula φ from $\Gamma \cup \{\psi\}$ in $\mathscr{H}_{\mathrm{co}}$. If $k = 1$, then we have the following three possibilities:

- $\varphi = H_{a_m} \ldots H_{a_1} \xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\mathrm{co}}$. Using the fact that $\varphi \to (\psi \to \varphi)$ is an instance of a propositional tautology and Definition 4.23, we obtain $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \varphi \to (\psi \to \varphi)$. By applying Lemma 4.25, $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \psi \to \varphi$ follows.
- $\varphi \in \Gamma$. Analogously to the previous case, we obtain $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \psi \to \varphi$.
- $\varphi = \psi$. Using the fact that $\psi \to \psi$ is an instance of a propositional tautology and Definition 4.23, we obtain $\Gamma \vdash_{\mathscr{H}_{\mathrm{co}}} \psi \to \varphi$.

Assume now that the desired statement holds for every formula which has a proof of length shorter than k and consider a formula φ which has a proof of length k. This means that the last formula in the proof is φ. There are four possibilities: $\varphi = H_{a_m} \dots H_{a_1} \xi$ for some $m \geq 0$ and some (instance of an) axiom ξ of $\mathscr{H}_{\text{co}}$, or $\varphi \in \Gamma$, or $\varphi = \psi$, or φ follows by an application of modus ponens. We already dealt with the first three possibilities in the base case, so let us consider the remaining possibility: φ follows by an application of modus ponens to some formulas, for example φ_i and $\varphi_i \to \varphi$. Since these two formulas are earlier in the proof, they have proofs whose lengths are shorter than k. By applying the induction hypothesis on them, we get $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \psi \to \varphi_i$ and $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \psi \to (\varphi_i \to \varphi)$. Using Lemma 4.25 and propositional reasoning, we obtain $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \psi \to \varphi$. □

Next, we need to know when a formula is a logical consequence of a set of formulas.

Definition 4.30. *Let* C *be a collection of Kripke models. Let* $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ *and* $\varphi \in \mathcal{L}_H^{\text{co}}$. *We say that* φ *is a local logical consequence of* Γ *and write* $\Gamma \models_{\mathsf{C}} \varphi$ *if, for any* $M \in \mathsf{C}$ *and any* $w \in \mathcal{D}(M)$, *we have*

$$M, w \models \psi \quad \text{for all} \quad \psi \in \Gamma \quad \Longrightarrow \quad M, w \models \varphi.$$

We will now prove that $\models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}}$ is closed too with respect to modus ponens:

Lemma 4.31. *Let* $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ *and* $\varphi, \psi \in \mathcal{L}_H^{\text{co}}$. *If* $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi$ *and* $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi \to \psi$, *then* $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \psi$.

Proof. Assume that $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi$ and $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi \to \psi$ hold. Take an arbitrary $M^H = (W, \pi, \mathcal{H}_1, \dots, \mathcal{H}_n) \in \mathsf{K45}_{\mathsf{n}}^{\text{co}}$ and an arbitrary $w \in W$. Let $M^H, w \models \xi$, for all $\xi \in \Gamma$. Since $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi$, $M^H, w \models \varphi$ follows. Since $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi \to \psi$, $M^H, w \models \varphi \to \psi$ follows too. Consequently, we obtain $M^H, w \models \psi$, as desired. Since $M^H = (W, \pi, \mathcal{H}_1, \dots, \mathcal{H}_n) \in \mathsf{K45}_{\mathsf{n}}^{\text{co}}$ and $w \in W$ were chosen arbitrarily, $\Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \psi$ follows. □

Finally, we have all the necessary pieces to prove the following theorem.

Theorem 4.32 (Strong soundness and strong completeness)**.** *Let* $\Gamma \subseteq \mathcal{L}_H^{\text{co}}$ *and* $\varphi \in \mathcal{L}_H^{\text{co}}$. *Then*

$$\Gamma \vdash_{\mathscr{H}_{\text{co}}} \varphi \quad \Longleftrightarrow \quad \Gamma \models_{\mathsf{K45}_{\mathsf{n}}^{\text{co}}} \varphi.$$

Proof. Strong soundness: Follows by induction on the length of the derivation of φ from Γ in $\mathscr{H}_{\text{co}}$.

Strong completeness: We prove the contrapositive. Assume $\Gamma \nvdash_{\mathscr{H}_{\text{co}}} \varphi$. Then it is easy to prove that $\Gamma \cup \{\neg\varphi\}$ must be consistent with respect to $\mathscr{H}_{\text{co}}$: Assume the opposite, i.e., $\Gamma \cup \{\neg\varphi\}$ is inconsistent with respect to $\mathscr{H}_{\text{co}}$. Therefore, according to Lemma 4.28, $\Gamma \cup \{\neg\varphi\} \vdash \bot$. Deduction theorem 4.29 now implies $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \neg\varphi \to \bot$. Using propositional reasoning, from this we further obtain $\Gamma \vdash_{\mathscr{H}_{\text{co}}} \varphi$, contradicting our original assumption.

Since $\Gamma \cup \{\neg\varphi\}$ is indeed consistent with respect to $\mathscr{H}_{\text{co}}$, it is contained in some maximal consistent set Δ with respect to $\mathscr{H}_{\text{co}}$ according to the Lindenbaum lemma 4.14. From Lemma 4.16, we thus obtain

$$M^c, w_\Delta \models \neg\varphi \quad \text{and} \quad M^c, w_\Delta \models \psi \quad \text{for all} \quad \psi \in \Gamma,$$

where M^c is the canonical Kripke model for $\mathscr{H}_{\text{co}}$ defined in Definition 4.15. However, $M^c, w_\Delta \models \neg\varphi$ means that $M^c, w_\Delta \models \varphi$ does not hold. Therefore, $\Gamma \not\models_{\mathsf{K45}_\mathsf{n}^{\text{co}}} \varphi$. □

Therefore, using Theorem 2.22, we immediately get:

Corollary 4.33. *The logic of hope is compact.*

4.3 Finite model property and decidability

For the translation $t : \mathcal{L}_H^{\text{co}} \to \mathcal{L}_K^{\text{co}}$ stated in Definition 4.4, we can prove the converse of Lemma 4.20:

Lemma 4.34. *For any model $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5}_\mathsf{n}$, there exists a corresponding model $M^H = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_\mathsf{n}^{\text{co}}$, such that the following holds for any formula $\varphi \in \mathcal{L}_H^{\text{co}}$ and any state $w \in W$:*

$$M^H, w \models \varphi \quad \textit{iff} \quad M^K, w \models t(\varphi).$$

Proof. Let $M^K = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathsf{S5}_\mathsf{n}$. The corresponding model is constructed by taking

$$\mathcal{H}_i := \{(w, w') \in \mathcal{K}_i \mid M^H, w \models \mathit{correct}_i \text{ and } M^H, w' \models \mathit{correct}_i\}. \tag{4.3}$$

for all $i \in \{1, \ldots, n\}$. We proceed by induction on the structure of φ.

Base case: If φ is $p \in \mathsf{Prop} \cup \mathsf{Co}$, then $t(p) = p$ according to Definition 4.4. Now, the statement of the lemma immediately follows as π is same in M^H and M^K.

Induction step:

1. If φ is of the form $\neg\psi$, then $t(\neg\psi) = \neg t(\psi)$ according to Definition 4.4. Using the induction hypothesis, we obtain $M^H, w \models \varphi$ iff $M^H, w \models \neg\psi$ iff $M^H, w \not\models \psi$ iff $M^K, w \not\models t(\psi)$ iff $M^K, w \models \neg t(\psi)$ iff $M^K, w \models t(\neg\psi)$ iff $M^K, w \models t(\varphi)$.

2. If φ is of the form $\psi_1 \wedge \psi_2$, then $t(\psi_1 \wedge \psi_2) = t(\psi_1) \wedge t(\psi_2)$ according to Definition 4.4. Using the induction hypothesis, we obtain $M^H, w \models \varphi$ iff $M^H, w \models \psi_1 \wedge \psi_2$ iff $M^H, w \models \psi_1$ and $M^H, w \models \psi_2$ iff $M^K, w \models t(\psi_1)$ and $M^K, w \models t(\psi_2)$ iff $M^K, w \models t(\psi_1) \wedge t(\psi_2)$ iff $M^K, w \models t(\psi_1 \wedge \psi_2)$ iff $M^K, w \models t(\varphi)$.

3. Assume that φ is of the form $H_i\psi$. We need to show

$$M^H, w \models H_i\psi \quad \textit{iff} \quad M^K, w \models \mathit{correct}_i \to K_i(\mathit{correct}_i \to t(\psi)).$$

($\Rightarrow$): Assume $M^H, w \models H_i\psi$. This means that for all $w' \in W$ such that $(w, w') \in \mathcal{H}_i$, $M^H, w' \models \psi$ holds. That is, according to (4.3), for all $w' \in W$ such that $(w, w') \in \mathcal{K}_i$ and $M^H, w \models \mathit{correct}_i$ and $M^H, w' \models \mathit{correct}_i$, $M^H, w' \models \psi$ holds. Assume now $M^K, w \models \mathit{correct}_i$. Therefore, $M^H, w \models \mathit{correct}_i$ as shown in the base case. Take an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{K}_i$. Assume further $M^K, w' \models \mathit{correct}_i$. Therefore, $M^H, w' \models \mathit{correct}_i$ holds too as shown in the base case. Consequently, $M^H, w' \models \psi$, by assumption. From this, by applying the induction hypothesis, we obtain $M^K, w' \models t(\psi)$, as desired.

($\Leftarrow$): Assume $M^K, w \models t(H_i\psi) = \mathit{correct}_i \rightarrow K_i(\mathit{correct}_i \rightarrow t(\psi))$. Therefore, either $M^K, w \not\models \mathit{correct}_i$ or $M^K, w \models K_i(\mathit{correct}_i \rightarrow t(\psi))$. If $M^K, w \not\models \mathit{correct}_i$, then $M^H, w \not\models \mathit{correct}_i$ as shown in the base case. Thus, according to (4.3), $\mathcal{H}(w) = \emptyset$, so $M^H, w \models H_i\psi$ vacuously holds. Assume now that $M^K, w \models K_i(\mathit{correct}_i \rightarrow t(\psi))$. This means that for all $w' \in W$ such that $(w, w') \in \mathcal{K}_i$, we have that $M^K, w' \models \mathit{correct}_i \rightarrow t(\psi)$ holds. Take an arbitrary $w' \in W$ such that $(w, w') \in \mathcal{H}_i$. According to (4.3), this means that $(w, w') \in \mathcal{K}_i$ and $M^H, w \models \mathit{correct}_i$ and $M^H, w' \models \mathit{correct}_i$. Therefore, $M^K, w' \models \mathit{correct}_i$ holds too, in particular, as shown in the base case. Consequently, $M^K, w' \models t(\psi)$, by assumption. From this, by applying the induction hypothesis, we obtain $M^H, w' \models \psi$, as desired.

It remains to note that $M^H \in \mathsf{K45}^{\mathsf{co}}_{\mathsf{n}}$:

- Transitivity and euclideanity of $\mathcal{H}_i$ follow easily since $\mathcal{K}_i$ satisfies these properties.
- If $M^H, w \models \mathit{correct}_i$, then $w \in \mathcal{H}_i(w)$ follows easily since $\mathcal{K}_i$ is reflexive,
- If $M^H, w \models \mathit{faulty}_i$, that is, $M^H, w \not\models \mathit{correct}_i$, then $\mathcal{H}_i(w) = \varnothing$ follows immediately,
- For all $w' \in \mathcal{H}_i(w)$, according to (4.3), it is the case that $M^H, w' \models \mathit{correct}_i$. □

It is well-known [FHMV95] (Theorem 3.2.4, p. 69) that:

Theorem 4.35. *The logic of $\mathscr{S}5_n$ has the FMP.*

Therefore, we can easily show that:

Theorem 4.36. *The logic of $\mathscr{H}_{\mathrm{co}}$ has the FMP.*

Proof. Take an arbitrary $\varphi \in \mathcal{L}^{\mathrm{co}}_H$. If $\nvdash_{\mathscr{H}_{\mathrm{co}}} \varphi$, then $\nvdash_{\mathscr{S}5_n} t(\varphi)$ by Corollary 4.22. According to Theorem 4.35, there exists a finite model $M^K \in \mathsf{S5}_{\mathsf{n}}$ of $\mathscr{S}5_n$ such that $M^K \not\models t(\varphi)$. Therefore, there exists a world $w \in \mathcal{D}(M^K)$ such that $M^K, w \not\models t(\varphi)$. According to Lemma 4.34, there exists a model $M^H \in \mathsf{K45}^{\mathsf{co}}_{\mathsf{n}}$ such that $M^H, w \not\models \varphi$ holds. Therefore, $M^H \not\models \varphi$ also holds. It remains to note that M^H is finite since it has the same exact domain as M^K. □

Similarly, it is well-known [FHMV95] (Corollary 3.2.5, p. 70) that:

Theorem 4.37. *The logic of $\mathscr{S}5_n$ is decidable.*

Using this result, we can obtain:

Theorem 4.38. *The logic of $\mathscr{H}_{\text{co}}$ is decidable.*

Proof. Since the logic of $\mathscr{H}_{\text{co}}$ is finitely axiomatizable, it follows that it is recursively enumerable according to Lemma 2.24. It remains to show that the complement of the logic of $\mathscr{H}_{\text{co}}$ is also recursively enumerable (see Proposition 2.23). By Theorem 4.37, the logic of $\mathscr{S}5_n$ is decidable, so we know that the set of all $\mathscr{S}5_n$-refutable formulas is recursively enumerable, in particular. Combining this with Corollary 4.22 and the fact that for each formula $\varphi \in \mathcal{L}_K^{\text{co}}$ such that $\mathscr{S}5_n \nvdash \varphi$ it is easy to write an algorithm that checks (in finite time) whether there exists a formula $\varphi^* \in \mathcal{L}_H^{\text{co}}$ such that $t(\varphi^*) = \varphi$ for $t : \mathcal{L}_H^{\text{co}} \to \mathcal{L}_K^{\text{co}}$ as defined in Definition 4.4, allows us to conclude that the set of all $\mathscr{H}_{\text{co}}$-refutable formulas is recursively enumerable as well. Therefore, decidability of the logic of $\mathscr{H}_{\text{co}}$ follows.[2] □

4.4 Related work

Moses and Shoham [MS93] introduce three binary modal operators describing single agent's beliefs as a form of knowledge relativized to an assumption (without committing to any type of knowledge). The most relevant of the three for us is the first one $B_1^\alpha\varphi := K(\alpha \to \varphi)$[3], where α is any formula in the bi-modal language restricted to formulas that don't contain any belief operators, i.e., to formulas that can contain only knowledge operators. Thus, dropping the agent subscript for a single agent, our notion of belief $B\varphi = K(\mathit{correct} \to \varphi)$, introduced in Section 3.3, coincides with their $B_1^{\mathit{correct}}\varphi$. The authors also provide independent (from knowledge) sound and complete axiomatizations for all three belief operators. In particular, they show that B_1^α is a $\mathsf{K45}$-type of modality (satisfying two extra properties), assuming K is of type $\mathsf{S5}$.

In [BvDH+16], Bolander et al. consider a version of public announcement logic, called *attention-based announcement logic*, where agents need not pay attention to a public announcement. Not being attentive (which could be viewed as a special type of fault) is modeled by designated atoms h_i for each agent i (thus, much like the knowledge of our agents depends on whether they are correct, i.e., whether $\mathit{correct}_i$ is true, the knowledge of their agents after a public announcement depends on whether h_i is true). With respect to introspective properties, the authors consider systems for both non-fault-introspective and fault-introspective agents, the latter stipulating the *attention introspection property*: an attentive agent believes to be attentive, $h_i \to B_i h_i$, and an inattentive agent believes to be inattentive, $\neg h_i \to B_i \neg h_i$. This results in logic K_n for non-fault-introspective agents

[2] Alternatively, a direct proof can be obtained using techniques from [Kuz08].

[3] Here subscript 1 means "first operator out of three" rather than agent 1.

and a specific extension of logic $\mathsf{K45}_n$ for fault-introspective ones. Note that, by the very nature of their work, [BvDH$^+$16] deals with dynamic epistemic notions. The authors also introduce an adaptation of relativized common belief [bBvEK06] called *attentive relativized common belief* defined as the greatest fixpoint of the equation $x = E^{\chi}_{\mathcal{A}}(\varphi \wedge x)$, where $E^{\chi}_{\mathcal{A}} := \bigwedge_{i \in \mathcal{A}} (h_i \to B_i(\chi \to \varphi))$ is called *attentive relativized shared belief* and χ is the relativizing formula. This closely resembles a group notion of hope called *mutual hope* $E^{H}_{\mathcal{A}} := \bigwedge_{i \in \mathcal{A}} (\mathit{correct}_i \to K_i(\mathit{correct}_i \to \varphi))$.

CHAPTER 5

A new hope

In this chapter, we introduce an alternative axiomatization for the hope modality by removing the reliance on designated atoms denoting correctness of individual agents and show that hope can be viewed as a KB4 type of a modality. We also combine hope modalities with knowledge modalities in a joint logic and present a logic enriched with both common knowledge and common hope. In these logics we formalize as frame-characterizable axioms some of the main properties of byzantine fault-tolerant distributed systems: bounds on the number of byzantine faulty agents and the epistemic limitations due to agents' inability to rule out brain-in-a-vat scenarios. All of the logics presented in the chapter have the finite model property and are decidable. In addition, we describe a way to define the notion of common eventual hope, which is needed for the epistemic analysis of the Firing Rebels with Relay problem in the next chapter.

Chapter organization

In Section 5.1 we propose a new axiomatization of the hope modality. One of the advantages of this new axiomatization is that the resulting logic is a normal multi-agent epistemic logic. We then proceed by providing a joint system for hope and knowledge in Section 5.2. In this joint system, we characterize the brain-in-a-vat-like properties of agents, discussed in Chapter 3, by purely modal logical means. We then enrich the joint logic with both common hope modality and comon knowledge modality in Section 5.3. The resulting axiom system turns out to be sound and complete with respect to the $\mathsf{KB4}_n$ class of models. A thorough soundness and completeness proof is also included in this section. In Section 5.5, we prove that all of the logics presented in the chapter have the finite model property as well as that they are decidable. Finally, in Section 5.6, we describe a way to introduce common eventual hope and prove some of its basic properties.

5.1 Axiomatizing hope (again)

Our first result in this chapter is an alternative axiomatization for hope that deals away with the designated atoms $correct_i$.

This is achieved by adopting the definition

$$correct_i := \neg H_i \bot \tag{5.1}$$

in the language

$$\mathcal{L}_H := \mathcal{L}_H^{\mathrm{co}} | \mathsf{Prop},$$

where the language $\mathcal{L}_H^{\mathrm{co}}$ has been introduced in Section 4.1. It turns out that the logic of hope in this language is the logic of the class $\mathsf{KB4}_n$ of all transitive and symmetric Kripke frames and is axiomatized by the axiom system $\mathscr{H} = \mathscr{KB}4_n$ (depicted in Figure 5.1).

$$\begin{aligned}
P: &\quad \text{all propositional tautologies}\\
K^H: &\quad H_i(\varphi \to \psi) \wedge H_i\varphi \to H_i\psi\\
B^H: &\quad \varphi \to H_i \neg H_i \neg \varphi\\
4^H: &\quad H_i\varphi \to H_i H_i \varphi\\
MP: &\quad \frac{\varphi \quad \varphi \to \psi}{\psi}\\
Nec^H: &\quad \frac{\varphi}{H_i\varphi}
\end{aligned}$$

Figure 5.1: Axiom system $\mathscr{H}$

Remark 5.1. *We will write $\mathscr{H} + Ax_1 + \cdots + Ax_n$ to represent the axiom system obtained by adding the axioms $Ax_1, \ldots, Ax_n$ to the axioms of $\mathscr{H}$ without changing the inference rules of $\mathscr{H}$.*

Recall that (see Definition 2.1): a set of formulas forms a system of a normal multi-agent epistemic logic if and only if it contains all propositional tautologies and is closed under the modus ponens inference rule, the K axiom scheme, the necessitation inference rule, and the uniform substitution rule.

Therefore, we immediately obtain:

Proposition 5.2. *The logic of $\mathscr{H}$ is a normal multi-agent epistemic logic.*

Remark 5.3. *The new axiomatization makes it easier to see how hope is different from the usual notion of belief. Indeed, belief is quite often assumed to be consistent, i.e., satisfying axiom $\neg B_i \bot$ (called axiom D in the literature), which fails for hope due to inconsistent hopes of byzantine faulty agents. On the other hand, axiom B^H is typically*

invalid for belief because, together with 4^H*, it would preclude agents from having consistent but false beliefs.*

Using the well-known fact that axiom B characterizes the symmetric property of frames [BdRV01], it is easy to show:

Theorem 5.4. *The axiom system* $\mathscr{H}$ *is sound and complete with respect to the* $\mathsf{KB4}_n$ *class of models.*

We now show that $\mathscr{H}$ is equivalent to $\mathscr{H}_{\text{co}}$ (see Figure 4.1) modulo abbreviation (5.1):

Theorem 5.5. *i.)* $\mathscr{H} \vdash \varphi$ *implies* $\mathscr{H}_{\text{co}} \vdash \varphi$ *for all* $\varphi \in \mathcal{L}_H$*.*

ii.) $\mathscr{H}_{\text{co}} \vdash \varphi$ *implies* $\mathscr{H} \vdash \varphi^\dagger$*, where* $\varphi^\dagger \in \mathcal{L}_H$ *is the result of replacing each* $correct_i$ *in* $\varphi \in \mathcal{L}_H^{\text{co}}$ *with* $\neg H_i\bot$*, according to* (5.1)*.*

Proof. i.) It is sufficient to show $\mathscr{H}_{\text{co}} \vdash B^H$. Using $faulty_i \to H_i\neg H_i\neg\varphi$, which is an instance of axiom F, we get $\mathscr{H}_{\text{co}} \vdash faulty_i \to (\varphi \to H_i\neg H_i\neg\varphi)$ by propositional reasoning. Similarly, using $correct_i \to (H_i\neg\varphi \to \neg\varphi)$, which is an instance of axiom T'^H, we get $\mathscr{H}_{\text{co}} \vdash correct_i \to (\varphi \to \neg H_i\neg\varphi)$. Combining this further with $\neg H_i\neg\varphi \to H_i\neg H_i\neg\varphi$, which is an instance of axiom 5^H, results now in $\mathscr{H}_{\text{co}} \vdash correct_i \to (\varphi \to H_i\neg H_i\neg\varphi)$. Finally, given that $correct_i \vee faulty_i$ is an instance of a propositional tautology, we obtain $\mathscr{H}_{\text{co}} \vdash \varphi \to H_i\neg H_i\neg\varphi$.

ii.) It is sufficient to show that axiom 5^H, as well as the $\dagger$-translations of axioms T'^H, F, and H are derivable in $\mathscr{H}$.

- That 5^H can be derived from 4^H and B^H is a well-known fact (any transitive and symmetric relation is euclidean).
- The $\dagger$-translation of T'^H is $\neg H_i\bot \to (H_i\psi \to \psi)$ for $\psi = \varphi^\dagger$. It is sufficient to show that $\mathscr{H} \vdash \neg(H_i\psi \to \psi) \to H_i\bot$ holds (the contrapositive). Firstly, $\neg(H_i\psi \to \psi) \to H_i\psi \wedge \neg\psi$ is an instance of a propositional tautology. Further, $H_i\psi \to H_iH_i\psi$ is an instance of axiom 4^H and $\mathscr{H} \vdash \neg\psi \to H_i\neg H_i\psi$ follows by axiom B^H and propositional reasoning. Therefore, $\mathscr{H} \vdash \neg(H_i\psi \to \psi) \to H_iH_i\psi \wedge H_i\neg H_i\psi$ follows. It remains to use the normality of H_i and propositional reasoning to replace $H_iH_i\psi \wedge H_i\neg H_i\psi$ first with $H_i(H_i\psi \wedge \neg H_i\psi)$ and finally with $H_i\bot$.
- The $\dagger$-translation of F is (modulo a double negation) $H_i\bot \to H_i\psi$ for $\psi = \varphi^\dagger$, which follows by the normality of H_i and propositional reasoning from $\bot \to \psi$.
- The $\dagger$-translation of axiom H is $H_i\neg H_i\bot$, which is straightforward to obtain from $\neg\bot \to H_i\neg H_i\neg\neg\bot$, which is an instance of axiom B^H, by propositional reasoning. □

Recall that we use $0 \leq f < n$ to denote the maximal number of byzantine faulty agents occurring in a single execution of the system. We demonstrate the utility of the reformulation of the logic of hope, presented in this chapter, by encoding this assumption in $\mathcal{L}_H$ as a *frame-characterizable* property:

$$Byz_f \qquad := \qquad \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G| = n-f}} \bigwedge_{i \in G} \neg H_i \bot.$$

Remark 5.6. *$Byz_0 = \bigwedge_{i \in \mathcal{A}} \neg H_i \bot$ simply states that all n agents are correct.*

Proposition 5.7 (Characterizing $\leq f$ byzantine faulty agents)**.** *Byz_f is characterized by the* all-but-f-seriality *property of Kripke frames $F = (W, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ requiring each world to have outgoing arrows for all but f agents:*

$$(\forall w \in W)(\exists G \subseteq \mathcal{A})\Big(|G| = n - f \wedge (\forall i \in G)\mathcal{H}_i(w) \neq \varnothing\Big).$$

Proof. Take an arbitrary Kripke frame $F = (W, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ for the language $\mathcal{L}_H$. We need to show

$$F \models Byz_f \qquad \Longleftrightarrow \qquad F \text{ is all-but-}f\text{-serial.}$$

($\Longrightarrow$): We prove the contrapositive. If F is not all-but-f-serial, then there is some world $w \in W$ such that any group $G \subseteq \mathcal{A}$ of $n - f$ agents has some agent $i_G \in G$ such that $\mathcal{H}_{i_G}(w) = \varnothing$. Independent of a valuation π, we can conclude that $(F, \pi), w \not\models \neg H_{i_G} \bot$ holds for all these agents. Hence, we get $(F, \pi), w \not\models Byz_f$ (for any π). Consequently, $F \not\models Byz_f$ follows.

($\Longleftarrow$): Let F be all-but-f-serial. Take an arbitrary world $w \in W$. It follows that there is a group $G \subseteq \mathcal{A}$ of $n - f$ agents such that $\mathcal{H}_i(w) \neq \varnothing$ for all $i \in G$. Independent of a valuation π, we can conclude $(F, \pi), w \models \bigwedge_{i \in G} \neg H_i \bot$. Hence, we get $(F, \pi), w \models Byz_f$ (for any π). This completes the proof that Byz_f is valid in F. □

Definition 5.8. *The class $\mathsf{KB4}_n^{n-f}$ of models consists of all Kripke models from $\mathsf{KB4}_n$ with all-but-f-serial Kripke frames.*

Using Proposition 5.7, we immediately obtain:

Corollary 5.9. *The axiom system $\mathscr{H} + Byz_f$ is sound and complete with respect to the $\mathsf{KB4}_n^{n-f}$ class of models.*

5.2 Individual hope and individual knowledge

Syntax. We start with Prop and continue by forming formulas by closing under the Boolean connectives $\neg$ and $\wedge$ and under the unary modal operators (one for each agent) $H_1, \ldots, H_n$ and $K_1, \ldots, K_n$ to obtain the language $\mathcal{L}_{KH}$, i.e., the language $\mathcal{L}_{KH}$ is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid H_i\varphi,$$

where $p \in \mathsf{Prop}$ and $i \in \mathcal{A}$. We take $\top$ to be an abbreviation for some fixed propositional tautology, and take $\bot$ to be an abbreviation for $\neg\top$. Also, we use the following standard abbreviations from propositional logic: $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ for $\neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$.

Recall that agent i's hope of φ was initially defined in the following way:

$$H_i\varphi \quad \leftrightarrow \quad (\mathit{correct}_i \to K_i(\mathit{correct}_i \to \varphi)),$$

in Chapter 3. Using (5.1), we now get:

$$H_i\varphi \quad \leftrightarrow \quad (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi)).$$

We denote this formula by KH.

Recall also that a relation R on a set S is called shift serial if $R(t) \neq \varnothing$ for any $t \in R(s)$, $s \in S$ (see Definition 2.7).

Our new language enables us to (almost) characterize formula KH by two frame properties for the two directions of the equivalence in the following way:

Proposition 5.10 (Characterizing knowledge-to-hope connection)**.** *Formula*

$$KH^{\leftarrow} := (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi)) \to H_i\varphi$$

is characterized by the $\mathcal{H}\mathrm{in}\mathcal{K}$ property of Kripke frames $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ with shift serial $\mathcal{H}_i$:

$$\mathcal{H}\mathrm{in}\mathcal{K}: \qquad \mathcal{H}_i \subseteq \mathcal{K}_i.$$

Proof. Take an arbitrary Kripke frame $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ for the language $\mathcal{L}_{KH}$ with shift serial $\mathcal{H}_i$. We need to show

$$F \models KH^{\leftarrow} \qquad \Longleftrightarrow \qquad F \text{ satisfies } \mathcal{H}\mathrm{in}\mathcal{K}.$$

($\Longrightarrow$): We prove the contrapositive. If F violates $\mathcal{H}\mathrm{in}\mathcal{K}$, then there are worlds $w, v \in W$ with $w\mathcal{H}_i v$ but not $w\mathcal{K}_i v$. Consider a valuation π such that $\pi(p) = W \setminus \{v\}$ for some atom p. We now have $(F, \pi), w \models K_i(\neg H_i\bot \to p)$ because $\mathcal{K}_i(w) \subseteq W \setminus \{v\} = \pi(p)$. Therefore, we also have that $(F, \pi), w \models \neg H_i\bot \to K_i(\neg H_i\bot \to p)$. However, clearly $(F, \pi), w \not\models H_i p$ because of v. Thus, we have shown $(F, \pi), w \not\models KH^{\leftarrow}$ for $\varphi = p$. Consequently, $F \not\models KH^{\leftarrow}$ follows. Note that this direction does not rely on the shift seriality of $\mathcal{H}_i$.

($\Longleftarrow$): Let us assume that F satisfies $\mathcal{H}\text{in}\mathcal{K}$. Let the antecedent in $KH^{\leftarrow}$ hold at an arbitrary world $w \in W$ for an arbitrary valuation π. In order to show that $(F,\pi), w \models H_i\varphi$ holds, it is sufficient to show $(F,\pi), v \models \varphi$ for all $v \in \mathcal{H}_i(w)$. It is vacuously true if $\mathcal{H}_i(w) = \varnothing$. Otherwise, take any such world v. We have $(F,\pi), w \models \neg H_i\bot$ because $\mathcal{H}_i(w) \neq \varnothing$, thus, $(F,\pi), w \models K_i(\neg H_i\bot \to \varphi)$ by assumption. Since $\mathcal{H}_i(w) \subseteq \mathcal{K}_i(w)$ due to $\mathcal{H}\text{in}\mathcal{K}$, we now get $(F,\pi), v \models \neg H_i\bot \to \varphi$. It remains to make use of $\mathcal{H}_i(v) \neq \varnothing$, which we know to be true due to the shift seriality of $\mathcal{H}_i$. Therefore $(F,\pi), v \models \varphi$ indeed holds. This completes the proof that $KH^{\leftarrow}$ is valid in F. □

Proposition 5.11 (Characterizing hope-to-knowledge connection)**.** *Formula*

$$KH^{\rightarrow} := H_i\varphi \to (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi))$$

is characterized by the one$\mathcal{H}$ *property of Kripke frames* $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$*:*

$$\text{one}\mathcal{H}: \qquad (\forall w, v \in W)\big(\mathcal{H}_i(w) \neq \varnothing \wedge \mathcal{H}_i(v) \neq \varnothing \wedge w\mathcal{K}_i v \Longrightarrow w\mathcal{H}_i v\big).$$

Proof. Take an arbitrary Kripke frame $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ for the language $\mathcal{L}_{KH}$. We need to show

$$F \models KH^{\rightarrow} \qquad \Longleftrightarrow \qquad F \text{ satisfies one}\mathcal{H}.$$

($\Longrightarrow$): We prove the contrapositive. If F violates one$\mathcal{H}$, then there are worlds $w, v \in W$ with $\mathcal{H}_i(w) \neq \varnothing$, $\mathcal{H}_i(v) \neq \varnothing$, $w\mathcal{K}_i v$, but not $w\mathcal{H}_i v$. Consider a valuation π such that $\pi(p) = \mathcal{H}_i(w)$ for some atom p. Clearly, $(F,\pi), w \models H_i p$ and $(F,\pi), w \models \neg H_i\bot$. However, $(F,\pi), w \not\models K_i(\neg H_i\bot \to p)$ since $(F,\pi), v \not\models \neg H_i\bot \to p$. Thus, we have shown $(F,\pi), w \not\models KH^{\rightarrow}$ for $\varphi = p$. Consequently, $F \not\models KH^{\rightarrow}$ follows.

($\Longleftarrow$): Let us assume that F satisfies one$\mathcal{H}$. Let $H_i\varphi$ hold at an arbitrary world $w \in W$ for an arbitrary valuation π. The case of $\mathcal{H}_i(w) = \varnothing$ is trivial since $(F,\pi), w \models H_i\bot$ makes the succedent in $KH^{\rightarrow}$ true at w. Otherwise, $\mathcal{H}_i(w) \neq \varnothing$. Similarly, for any $v \in \mathcal{K}_i(w)$ with $\mathcal{H}_i(v) = \varnothing$, we have $(F,\pi), v \models \neg H_i\bot \to \varphi$. Finally, for any $v \in \mathcal{K}_i(w)$ with $\mathcal{H}_i(v) \neq \varnothing$, we have $v \in \mathcal{H}_i(w)$ by one$\mathcal{H}$. Hence, $(F,\pi), v \models \varphi$ follows by assumption. This now further implies $(F,\pi), v \models \neg H_i\bot \to \varphi$. We have shown that $\neg H_i\bot \to \varphi$ is true in all worlds from $\mathcal{K}_i(w)$ and can again conclude that the succedent in $KH^{\rightarrow}$ is true at w. This completes the proof that $KH^{\rightarrow}$ is valid in F. □

Definition 5.12. *The class* KH *of models for knowledge and hope consists of all Kripke models*

$$M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n, \pi)$$

where:

- *every* $\mathcal{K}_i$ *is an equivalence relation,*

$$
\begin{array}{lll}
 & P: & \text{all propositional tautologies} \\
H^{\dagger}: \; H_i\neg H_i\bot & K^K: & K_i(\varphi\to\psi)\wedge K_i\varphi\to K_i\psi \\
 & 4^K: & K_i\varphi\to K_iK_i\varphi \\
 & 5^K: & \neg K_i\varphi\to K_i\neg K_i\varphi \\
 & T^K: & K_i\varphi\to\varphi \\
MP: \; \dfrac{\varphi \quad \varphi\to\psi}{\psi} & Nec^K: & \dfrac{\varphi}{K_i\varphi} \\
KH: \quad H_i\varphi\leftrightarrow(\neg H_i\bot\to K_i(\neg H_i\bot\to\varphi)) & &
\end{array}
$$

Figure 5.2: Axiom system $\mathscr{KH}$

- *every $\mathcal{H}_i$ is shift serial, and*
- *properties $\mathcal{H}\text{in}\mathcal{K}$ and one$\mathcal{H}$ are satisfied.*

Proposition 5.13. *For all models $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n, \pi) \in \mathsf{KH}$, each accessibility relation $\mathcal{H}_i$ is symmetric and transitive.*

Proof. To prove transitivity of $\mathcal{H}_i$, assume $w\mathcal{H}_iv$ and $v\mathcal{H}_iu$. We get $w\mathcal{K}_iv$ and $v\mathcal{K}_iu$ by $\mathcal{H}\text{in}\mathcal{K}$. Therefore, we also have $w\mathcal{K}_iu$ since $\mathcal{K}_i$ is transitive. $\mathcal{H}_i(w) \ni v$ is not empty, and so is $\mathcal{H}_i(u) \neq \varnothing$, by the shift seriality of $\mathcal{H}_i$ because $v\mathcal{H}_iu$. Hence, $w\mathcal{H}_iu$ by one$\mathcal{H}$.

To prove symmetry of $\mathcal{H}_i$, assume $w\mathcal{H}_iv$. We get $w\mathcal{K}_iv$ by $\mathcal{H}\text{in}\mathcal{K}$. Therefore, we also have $v\mathcal{K}_iw$ since $\mathcal{K}_i$ is symmetric. As before, $\mathcal{H}_i(w) \ni v$ is not empty, and $\mathcal{H}_i(v) \neq \varnothing$, by the shift seriality of $\mathcal{H}_i$ because $w\mathcal{H}_iv$. Hence, $v\mathcal{H}_iw$ by one$\mathcal{H}$. □

Remark 5.14. *A partial equivalence relation is any transitive and symmetric binary relation [MM91]. Hence, $\mathcal{H}_i$ are partial equivalence relations, so that* one$\mathcal{H}$ *can be described as "no $\mathcal{K}_i$-equivalence class contains more than one $\mathcal{H}_i$-partial-equivalence class."*

A natural way of obtaining the combined logic of hope and knowledge would be to combine the axioms and rules for hope, axioms and rules for knowledge, and KH as a connection axiom. Proposition 5.13, however, indicates that this would create redundancies. As we now show, in the presence of KH, KB4 properties of hope originate from S5 properties of knowledge, albeit with the help of the translation of axiom $H = H_i correct_i$ from $\mathscr{H}_{\text{co}}$ into language $\mathcal{L}_{KH}$. This translation $H^{\dagger} = H_i\neg H_i\bot$ can be called *necessary consistency* for hope and is known to be characterized by shift seriality. The resulting simplified axiom system is depicted in Figure 5.2.

Remark 5.15. *As before, we will write $\mathscr{KH} + Ax_1 + \cdots + Ax_n$ to represent the axiom system obtained by adding the axioms $Ax_1, \ldots, Ax_n$ to the axioms of $\mathscr{KH}$ without changing the inference rules of $\mathscr{KH}$.*

Proposition 5.16. *For any $i \in \mathcal{A}$ and any $\varphi, \psi \in \mathcal{L}_{KH}$:*

1. *$\mathscr{KH} \vdash K_i\varphi \to H_i\varphi$,*

2. $\mathscr{KH} \vdash H_i(\varphi \to \psi) \wedge H_i\varphi \to H_i\psi$,

3. *if* $\mathscr{KH} \vdash \varphi$, *then* $\mathscr{KH} \vdash H_i\varphi$,

Proof.

1. 1. $\varphi \to (\neg H_i\bot \to \varphi)$ prop. tautology
   2. $K_i(\varphi \to (\neg H_i\bot \to \varphi))$ by Nec^K from 1.
   3. $K_i(\varphi \to (\neg H_i\bot \to \varphi)) \to (K_i\varphi \to K_i(\neg H_i\bot \to \varphi))$ axiom K^K
   4. $K_i\varphi \to K_i(\neg H_i\bot \to \varphi)$ by MP from 2. and 3.
   5. $K_i(\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi))$ prop. tautology
   6. $(\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi)) \to H_i\varphi$ $KH^{\leftarrow}$
   7. $K_i\varphi \to H_i\varphi$ by syllogism from 4–6.
2. 1. $(\neg H_i\bot \to (\varphi \to \psi)) \to ((\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to \psi))$ prop. tautology
   2. $K_i\Big((\neg H_i\bot \to (\varphi \to \psi)) \to ((\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to \psi))\Big)$ by Nec^K from 1.
   3. $K_i\Big((\neg H_i\bot \to (\varphi \to \psi)) \to ((\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to \psi))\Big) \to \Big(K_i(\neg H_i\bot \to (\varphi \to \psi)) \to K_i((\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to \psi))\Big)$ axiom K^K
   4. $K_i(\neg H_i\bot \to (\varphi \to \psi)) \to K_i((\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to \psi))$ by MP from 2. and 3.
   5. $K_i((\neg H_i\bot \to \varphi) \to (\neg H_i\bot \to \psi)) \to (K_i(\neg H_i\bot \to \varphi) \to K_i(\neg H_i\bot \to \psi))$ axiom K^K
   6. $K_i(\neg H_i\bot \to (\varphi \to \psi)) \to (K_i(\neg H_i\bot \to \varphi) \to K_i(\neg H_i\bot \to \psi))$ by syllogism from 4. and 5.
   7. $H_i(\varphi \to \psi) \to \Big(\neg H_i\bot \to K_i(\neg H_i\bot \to (\varphi \to \psi))\Big)$ $KH^{\to}$
   8. $H_i\varphi \to (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi))$ $KH^{\to}$
   9. $H_i(\varphi \to \psi) \to \Big(\neg H_i\bot \to (K_i(\neg H_i\bot \to \varphi) \to K_i(\neg H_i\bot \to \psi))\Big)$ by prop. reasoning from 6. and 7.
   10. $H_i(\varphi \to \psi) \to \Big(H_i\varphi \to ((\neg H_i\bot \to (K_i(\neg H_i\bot \to \varphi) \to K_i(\neg H_i\bot \to \psi))) \wedge (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi)))\Big)$ by prop. reasoning from 8. and 9.
   11. $\Big((\neg H_i\bot \to (K_i(\neg H_i\bot \to \varphi) \to K_i(\neg H_i\bot \to \psi))) \wedge (\neg H_i\bot \to K_i(\neg H_i\bot \to \varphi))\Big) \to \Big(\neg H_i\bot \to (K_i(\neg H_i\bot \to \psi))\Big)$ prop. tautology

12. $H_i(\varphi \to \psi) \to \Big(H_i\varphi \to (\neg H_i\bot \to K_i(\neg H_i\bot \to \psi))\Big)$
    by prop. reasoning from 10. and 11.
13. $(\neg H_i\bot \to K_i(\neg H_i\bot \to \psi)) \to H_i\psi$ $KH^{\leftarrow}$
14. $H_i(\varphi \to \psi) \wedge H_i\varphi \to H_i\psi$ by prop. reasoning from 12. and 13.

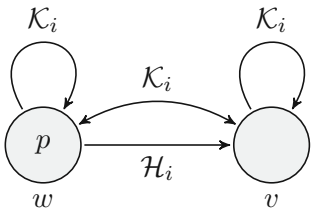
3. 1. φ assumption
   2. $K_i\varphi$ by Nec^K from 1.
   3. $K_i\varphi \to H_i\varphi$ Lemma 5.16 (1).
   4. $H_i\varphi$ by *MP* from 2. and 3.

□

Remark 5.17. *Given that $K_i\varphi \to H_i\varphi$ is also known to characterize frame property $\mathcal{H}\text{in}\mathcal{K}$, one might ask whether $KH^{\leftarrow}$ is equivalent to $K_i\varphi \to H_i\varphi$. The answer is negative, because $KH^{\leftarrow}$ only characterizes $\mathcal{H}\text{in}\mathcal{K}$ under the additional assumption of $\mathcal{H}_i$ being shift serial. For instance, consider a model M with*

- $W = \{w, v\}$,
- $\mathcal{K}_j = W \times W$ *for all* $j \in \mathcal{A}$,
- $\mathcal{H}_j = W \times W$ *for all* $j \neq i$,
- *non-shift-serial* $\mathcal{H}_i = \{(w, v)\}$,
- $\pi(p) = \{w\}$ *for some atom* p, *and*
- $\pi(q) = W$ *for all atoms* $q \neq p$.

Thus, we have the following situation for agent i, in particular:



It is easy to see now from the picture above that

$$M, w \not\models (\neg H_i\bot \to K_i(\neg H_i\bot \to p)) \to H_ip,$$

but

$$M, w \models K_ip \to H_ip.$$

Theorem 5.18. *The axiom system $\mathscr{KH}$ is sound and complete with respect to the* KH *class of models.*

Proof. It is straightforward to prove using the standard canonical model construction. □

Corollary 5.19. *For all $i \in \mathcal{A}$ and all $\varphi \in \mathcal{L}_{KH}$:*

1. $\mathscr{KH} \vdash H_i\varphi \to H_iH_i\varphi$;
2. $\mathscr{KH} \vdash \varphi \to H_i\neg H_i\neg\varphi$.

Proof. The proofs follow immediately from Theorem 5.18 and Proposition 5.13. □

Proposition 5.20. $\mathscr{KH} \vdash \neg H_i\bot \to (H_i\varphi \to \varphi)$, *for all $i \in \mathcal{A}$ and all $\varphi \in \mathcal{L}_{KH}$.*

Proof.

1.	$\neg(H_i\varphi \to \varphi) \to (H_i\varphi \wedge \neg\varphi)$	prop. tautology
2.	$H_i\varphi \to H_iH_i\varphi$	Corollary 5.19 (1)
3.	$(H_i\varphi \to H_iH_i\varphi) \to ((H_i\varphi \wedge \neg\varphi) \to (H_iH_i\varphi \wedge \neg\varphi))$	prop. tautology
4.	$(H_i\varphi \wedge \neg\varphi) \to (H_iH_i\varphi \wedge \neg\varphi)$	by *MP* from 2. and 3.
5.	$\neg\varphi \to H_i\neg H_i\varphi$	Corollary 5.19 (2)
6.	$(\neg\varphi \to H_i\neg H_i\varphi) \to ((H_iH_i\varphi \wedge \neg\varphi) \to (H_iH_i\varphi \wedge H_i\neg H_i\varphi))$	prop. tautology
7.	$(H_iH_i\varphi \wedge \neg\varphi) \to (H_iH_i\varphi \wedge H_i\neg H_i\varphi)$	by *MP* from 5. and 6.
8.	$(H_i\varphi \wedge \neg\varphi) \to (H_iH_i\varphi \wedge H_i\neg H_i\varphi)$	by syllogism from 4. and 7.
9.	$\neg(H_i\varphi \to \varphi) \to H_i(H_i\varphi \wedge \neg H_i\varphi)$	by syllogism from 1. and 8.
10.	$\neg(H_i\varphi \to \varphi) \to H_i\bot$	by prop. reasoning from 9.
11.	$(\neg(H_i\varphi \to \varphi) \to H_i\bot) \to (\neg H_i\bot \to (H_i\varphi \to \varphi))$	prop. tautology
12.	$\neg H_i\bot \to (H_i\varphi \to \varphi)$	by *MP* from 10. and 11.

□

Definition 5.21. Class KH^{n-f} *consists of all Kripke models from* KH *that have all-but-f-serial Kripke frames with respect to $\mathcal{H}_i$ relations.*

Using Proposition 5.7, we immediately obtain the following corollary.

Corollary 5.22. *The axiom system $\mathscr{KH} + Byz_f$ is sound and complete with respect to the KH^{n-f} class of models.*

The following two propositions outline the epistemic attitudes of agents who found out that they are byzantine faulty and agents who know that they are correct.

Proposition 5.23. *$\mathscr{KH} \vdash K_i H_i \bot \to H_i \varphi$ for all $i \in \mathcal{A}$ and all $\varphi \in \mathcal{L}_{KH}$.*

Proof.

1. $\bot \to \varphi$	prop. tautology
2. $H_i(\bot \to \varphi)$	by Lemma 5.16 (3) from 1.
3. $H_i(\bot \to \varphi) \wedge H_i \bot \to H_i \varphi$	Lemma 5.16 (2)
4. $H_i \bot \to H_i \varphi$	by *MP* from 2. and 3.
5. $K_i H_i \bot \to H_i \bot$	axiom T^K
6. $K_i H_i \bot \to H_i \varphi$	by syllogism from 5. and 4.

□

Proposition 5.24. *$\mathscr{KH} \vdash K_i \neg H_i \bot \to (H_i \varphi \leftrightarrow K_i \varphi)$ for all $i \in \mathcal{A}$ and all $\varphi \in \mathcal{L}_{KH}$.*

Proof. $\mathscr{KH} \vdash K_i \neg H_i \bot \to (K_i \varphi \to H_i \varphi)$ is an easy corollary of Lemma 5.16 (1). Let us derive $K_i \neg H_i \bot \to (H_i \varphi \to K_i \varphi)$:

1. $K_i \neg H_i \bot \to \neg H_i \bot$	axiom T^K
2. $H_i \varphi \to (\neg H_i \bot \to K_i(\neg H_i \bot \to \varphi))$	$KH^{\to}$
3. $K_i \neg H_i \bot \to (H_i \varphi \to K_i(\neg H_i \bot \to \varphi))$	by prop. reasoning from 1. and 2.
4. $K_i(\neg H_i \bot \to \varphi) \to (K_i \neg H_i \bot \to K_i \varphi)$	axiom K^K
5. $K_i \neg H_i \bot \to (H_i \varphi \to (K_i \neg H_i \bot \to K_i \varphi))$	by syllogism from 3. and 4.
6. $K_i \neg H_i \bot \to (H_i \varphi \to K_i \varphi)$	by prop. reasoning from 5.

□

Proposition 5.25. *For all $i \in \mathcal{A}$:*

1. *$\mathscr{KH} + Byz_f \vdash K_i Byz_f$;*

2. $\mathsf{KH}^{n-f} \models K_i Byz_f$.

Proof. The proofs follow immediately using Nec^K and Corollary 5.22. □

Corollary 5.26 (In fault-free systems, hope is knowledge)**.** *Recall that axiom Byz_0 rules out the presence of byzantine faulty agents. For any $i \in \mathcal{A}$,*

$$\mathscr{KH} + Byz_0 \vdash H_i\varphi \leftrightarrow K_i\varphi.$$

Proof. Follows from Remark 5.6 and Propositions 5.24 and 5.25. □

Proposition 5.23 can be strengthened because a byzantine faulty agent hopes for anything even without knowing that it is byzantine faulty, i.e., $\mathscr{KH} \vdash H_i\bot \to H_i\varphi$ (as obtained in step 4. in the proof). By contrast, in Proposition 5.24, the knowledge modality cannot be dropped: for a correct agent, hope does not yet mean knowledge, i.e., $\mathscr{KH} \nvdash \neg H_i\bot \to (H_i\varphi \leftrightarrow K_i\varphi)$. Instead, for a correct agent, hope is equivalent to

$$B_i\varphi := K_i(\neg H_i\bot \to \varphi),$$

which is a notion of belief that was introduced in Chapter 3 as $B_i\varphi := K_i(correct_i \to \varphi)$ in a language where $correct_i$ is an atomic proposition.

Proposition 5.27. *For all $i \in \mathcal{A}$ and $\varphi \in \mathcal{L}_{KH}$:*

$$\mathscr{KH} \vdash \neg H_i\bot \to (H_i\varphi \leftrightarrow B_i\varphi). \tag{5.2}$$

Proof. Follows from axiom KH by propositional reasoning. □

Let us introduce the following abbreviations for *mutual belief* E^B_G and *mutual hope* E^H_G among a group $\varnothing \neq G \subseteq \mathcal{A}$ of agents:

$$E^B_G\varphi := \bigwedge_{i\in G} B_i\varphi,$$
$$E^H_G\varphi := \bigwedge_{i\in G} H_i\varphi.$$

It is easy to see that mutual belief among some agents can be "extracted" from mutual hope among all agents in systems with at most $0 \le f < n$ byzantine faulty agents:

Proposition 5.28. $\mathscr{KH} + Byz_f \quad \vdash \quad E^H_{\mathcal{A}}\varphi \to \bigvee_{\substack{G\subseteq\mathcal{A}\\ |G|=n-f}} E^B_G\varphi.$

Proof. Follows from (5.2) and axiom Byz_f. □

Hope, generally, creates neither knowledge of hope nor hope of knowledge, as we show in the following proposition.

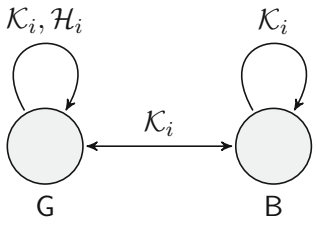
Proposition 5.29 (Knowledge and hope do not mix)**.** *For any $i \in \mathcal{A}$:*

- *it is not the case that* $\mathsf{KH} \models H_i\varphi \to H_iK_i\varphi$ *for all* $\varphi \in \mathcal{L}_{KH}$,
- *it is not the case that* $\mathsf{KH} \models H_i\varphi \to K_iH_i\varphi$ *for all* $\varphi \in \mathcal{L}_{KH}$.

Proof. We use the same countermodel to refute both statements but refute them for different formulas φ. Let $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n, \pi) \in \mathsf{KH}$ be such that:

- $W = \{\mathsf{G}, \mathsf{B}\}$,
- $\mathcal{K}_j = W \times W$ for all $j \in \mathcal{A}$,
- $\mathcal{H}_j = W \times W$ for all $j \neq i$,
- $\mathcal{H}_i = \{(\mathsf{G}, \mathsf{G})\}$, and
- π is arbitrary.

Thus, we have the following situation for agent i, in particular:



Clearly, agent i is correct in world G, i.e., $\mathcal{H}_i(\mathsf{G}) \neq \varnothing$, and byzantine faulty in world B, i.e., $\mathcal{H}_i(\mathsf{B}) = \varnothing$. It is easy to see now from the picture above that

$$M, \mathsf{G} \not\models H_i\neg H_i\bot \to H_iK_i\neg H_i\bot$$

and

$$M, \mathsf{B} \not\models H_i\bot \to K_iH_i\bot.$$ □

Corollary 5.30. *For any $i \in \mathcal{A}$ and any $f > 0$:*

- *it is not the case that* $\mathsf{KH}^{n-f} \models H_i\varphi \to H_iK_i\varphi$ *for all* $\varphi \in \mathcal{L}_{KH}$,
- *it is not the case that* $\mathsf{KH}^{n-f} \models H_i\varphi \to K_iH_i\varphi$ *for all* $\varphi \in \mathcal{L}_{KH}$.

Modal representation of the consequences of the brain-in-a-vat scenario

The goal of this section is to show the utility of the joint axiom system $\mathscr{KH}$ by providing axiomatic descriptions of some of the properties of byzantine fault-tolerant distributed systems that originate from earlier epistemic analyses presented in Chapter 3, such as the following:

- Agents cannot reliably establish their own correctness (Theorem 3.18), as formalized by the formula, for all $i \in \mathcal{A}$,

$$iByz \quad := \quad \neg K_i \neg H_i \bot.$$

- A byzantine faulty agent lacks any reliable information about other agents. In particular, a byzantine faulty agent has no reliable information to decide whether any other agent is correct or byzantine faulty, as formalized by the formula, for all $i \neq j$,

$$BiV \quad := \quad H_i \bot \to \neg K_i H_j \bot \wedge \neg K_i \neg H_j \bot.$$

From these two principles we can derive that no agent knows whether other agents are correct or byzantine faulty (Theorem 3.18 and Theorem 3.20):

Proposition 5.31. *$\mathscr{KH} + iByz + BiV \vdash anyByz_{ij} \wedge anyCor_{ij}$ for all $i \neq j$, where*

$$anyByz_{ij} : \quad \neg K_i \neg H_j \bot; \qquad anyCor_{ij} : \quad \neg K_i H_j \bot.$$

Proof. Let us derive $anyCor_{ij}$. The derivation of $anyByz_{ij}$ is similar.

1. $H_i \bot \to \neg K_i H_j \bot \wedge \neg K_i \neg H_j \bot$ — BiV
2. $H_i \bot \to \neg K_i H_j \bot$ — by prop. reasoning from 1.
3. $K_i H_j \bot \to \neg H_i \bot$ — by prop. reasoning from 2.
4. $K_i(K_i H_j \bot \to \neg H_i \bot)$ — by Nec^K from 3.
5. $K_i(K_i H_j \bot \to \neg H_i \bot) \to (K_i K_i H_j \bot \to K_i \neg H_i \bot)$ — axiom K^K
6. $K_i K_i H_j \bot \to K_i \neg H_i \bot$ — by MP from 4. and 5.
7. $\neg K_i \neg H_i \bot \to \neg K_i K_i H_j \bot$ — by prop. reasoning from 6.
8. $K_i H_j \bot \to K_i K_i H_j \bot$ — axiom 4^K
9. $\neg K_i K_i H_j \bot \to \neg K_i H_j \bot$ — by prop. reasoning from 8.
10. $\neg K_i \neg H_i \bot \to \neg K_i H_j \bot$ — by syllogism from 7. and 9.
11. $\neg K_i \neg H_i \bot$ — $iByz$

12. $\neg K_i H_j \bot$ by *MP* from 10. and 11.

□

Proposition 5.32. *Formula iByz is characterized by the* i-may-aseriality *property of Kripke frames* $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$*:*

$$(\forall w \in W)(\exists w' \in \mathcal{K}_i(w)) \quad \mathcal{H}_i(w') = \varnothing.$$

Proof. Take an arbitrary Kripke frame $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ for the language $\mathcal{L}_{KH}$. We need to show

$$F \models iByz \qquad \Longleftrightarrow \qquad F \text{ is } i\text{-may-aserial.}$$

($\Longrightarrow$): We prove the contrapositive. If F is not i-may-aserial, there is some world $w \in W$ such that $\mathcal{H}_i(w') \neq \varnothing$ for all $w' \in \mathcal{K}_i(w)$. Independent of a valuation π, we can conclude that $(F, \pi), w' \models \neg H_i \bot$ holds for all $w' \in \mathcal{K}_i(w)$. Hence, we get $(F, \pi), w \models K_i \neg H_i \bot$ (for any π). Consequently, $F \not\models iByz$ follows.

($\Longleftarrow$): Let F be i-may-aserial. Take an arbitrary world $w \in W$. It now follows that there is $w' \in \mathcal{K}_i(w)$ such that $\mathcal{H}_i(w') = \varnothing$. Independent of a valuation π, we can conclude that $(F, \pi), w \models \neg K_i \neg H_i \bot$ holds since $(F, \pi), w' \models H_i \bot$ for any π. Hence, we get $(F, \pi), w \models iByz$ (for any π). This completes the proof that $iByz$ is valid in F. □

Proposition 5.33. *Formula BiV is characterized by the* BiValence *property of Kripke frames* $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$*:*

$$(\forall w \in W)\Big(\mathcal{H}_i(w) = \varnothing \Longrightarrow (\exists w', w'' \in \mathcal{K}_i(w))(\mathcal{H}_j(w') \neq \varnothing \wedge \mathcal{H}_j(w'') = \varnothing)\Big).$$

Proof. Take an arbitrary Kripke frame $F = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n)$ for the language $\mathcal{L}_{KH}$. We need to show

$$F \models BiV \qquad \Longleftrightarrow \qquad F \text{ is BiValent.}$$

($\Longrightarrow$): We prove the contrapositive. If F is not BiValent, there is some world $w \in W$ such that $\mathcal{H}_i(w) = \varnothing$ but either $\mathcal{H}_j(w') = \varnothing$ for all $w' \in \mathcal{K}_i(w)$ or $\mathcal{H}_j(w'') \neq \varnothing$ for all $w'' \in \mathcal{K}_i(w)$. Independent of a valuation π, we can conclude that $(F, \pi), w \models K_i \neg H_j \bot \vee K_i H_j \bot$ holds despite $M, w \models H_i \bot$. Hence, we get $(F, \pi), w \not\models BiV$ (for any π). Consequently, $F \not\models BiV$ follows.

($\Longleftarrow$): Let F be BiValent. Take an arbitrary world $w \in W$ such that $\mathcal{H}_i(w) = \varnothing$. It now follows that there are $w' \in \mathcal{K}_i(w)$ such that $\mathcal{H}_j(w') \neq \varnothing$ and $w'' \in \mathcal{K}_i(w)$ such that $\mathcal{H}_j(w'') = \varnothing$. Independent of a valuation π, we can conclude that $(F, \pi), w \models \neg K_i H_j \bot \wedge \neg K_i \neg H_j \bot$ holds whenever $(F, \pi), w \models H_i \bot$. Hence, we get $(F, \pi), w \models BiV$ (for any π). This completes the proof that BiV is valid in F. □

We can also easily derive that brain-in-a-vat scenarios are not compatible with fault-free systems:

Proposition 5.34. *If no agent can become byzantine faulty, then all agents can establish their own correctness: for all $i \in \mathcal{A}$,*

$$\mathscr{KH} + Byz_0 \quad \vdash \quad \neg iByz.$$

Proof. By Proposition 5.25 and the normality of K_i, given that $Byz_0 \to \neg H_i\bot$ is an instance of a propositional tautology, we have $\mathscr{KH} + Byz_0 \vdash K_i\neg H_i\bot$ for $i \in \mathcal{A}$. Deriving $\neg iByz$ is now a matter of propositional reasoning. □

Another interesting special case is $f = 1$ (with $n > 1$):

Proposition 5.35. *If any agent but no more than one can become byzantine faulty, the inability of an agent to establish its own correctness leads to its inability to establish faultiness of somebody else: for all $i \neq j \in \mathcal{A}$,*

$$\mathscr{KH} + Byz_1 + iByz \quad \vdash \quad \neg K_iH_j\bot.$$

Proof. We have $H_j\bot \to \neg H_i\bot$ by Byz_1 for any $j \neq i$. Thus, we can conclude $K_iH_j\bot \to K_i\neg H_i\bot$ by the normality of K_i, which further implies $\neg K_i\neg H_i\bot \to \neg K_iH_j\bot$. Since $\neg K_i\neg H_i\bot$ is axiom $iByz$, we conclude $\neg K_iH_j\bot$ by *MP*. □

Proposition 5.36. *If any agent but no more than one can become byzantine faulty, the inability of a byzantine faulty agent to establish correctness of somebody else leads to its inability to establish its own faultiness: for all $i \neq j \in \mathcal{A}$,*

$$\mathscr{KH} + Byz_1 + (H_i\bot \to \neg K_i\neg H_j\bot) \quad \vdash \quad \neg K_iH_i\bot.$$

Proof. A correct agent i considers its own correctness possible by T^K, i.e., $\neg H_i\bot \to \neg K_iH_i\bot$. Formula $H_i\bot \to \neg K_i\neg H_j\bot$ for at least one $j \neq i$ is an assumption. At the same time, $H_i\bot \to \neg H_j\bot$ by Byz_1. As before, $\neg K_i\neg H_j\bot \to \neg K_iH_i\bot$ follows by the normality of K_i, yielding implication $H_i\bot \to \neg K_iH_i\bot$ by syllogism. Since we have derived $\neg K_iH_i\bot$ from both $\neg H_i\bot$ and $H_i\bot$, we get $\neg K_iH_i\bot$ by propositional reasoning. □

Remark 5.37. *Intuitively, for $f = 1$, if an agent establishes its own faultiness, which does not run afoul of $iByz$, then it will thereby establish the correctness of all other agents. It seems wrong to prohibit this by adopting the respective half of BiV, whereas the other half is derivable anyway. We, therefore, propose using*

$$\mathscr{KH} + Byz_f + BiV + iByz$$

for $f \geq 2$, but

$$\mathscr{KH} + Byz_1 + iByz$$

for $f = 1$. (The case of $f = 0$, which can be axiomatized by $\mathscr{KH} + Byz_0$, is more efficiently dealt with in the standard epistemic language.)

5.3 Common hope and common knowledge

In this section, we introduce the common hope modality by analogy with the common knowledge modality and explore their relationship.

We start by extending the language $\mathcal{L}_{KH}$ with the unary modal operator C_G^H for common hope and the unary modal operator C_G^K for common knowledge, where $\varnothing \neq G \subseteq \mathcal{A}$ is an arbitrary group of agents. We denote this extended language by $\mathcal{L}_{KH}^C$.

Let $\varnothing \neq G \subseteq \mathcal{A}$. By common hope of φ (in G) we intuitively mean mutual hope of φ (in G) and mutual hope of mutual hope of φ (in G), etc.:

$$C_G^H \qquad \leftrightsquigarrow \qquad E_G^H\varphi \wedge E_G^H E_G^H\varphi \wedge E_G^H E_G^H E_G^H\varphi \wedge \ldots,$$

just like by common knowledge of φ (in G) we intuitively mean mutual knowledge of φ (in G) and mutual knowledge of mutual knowledge of φ (in G), etc.

Definition 5.38. *Axiom system $\mathscr{KHC}$ consists of all the axioms and inference rules of $\mathscr{KH}$ (formulated for $\mathcal{L}_{KH}^C$ formulas) plus the following axioms and inference rules for all $\varnothing \neq G \subseteq \mathcal{A}$ and all formulas $\varphi, \psi \in \mathcal{L}_{KH}^C$:*

$$Mix^H: \quad C_G^H\varphi \to E_G^H(\varphi \wedge C_G^H\varphi); \qquad Mix^K: \quad C_G^K\varphi \to E_G^K(\varphi \wedge C_G^K\varphi);$$

$$Ind^H: \quad \frac{\psi \to E_G^H(\varphi \wedge \psi)}{\psi \to C_G^H\varphi}; \qquad Ind^K: \quad \frac{\psi \to E_G^K(\varphi \wedge \psi)}{\psi \to C_G^K\varphi}.$$

In Section 5.4, we will prove that $\mathscr{KHC}$ is sound and complete with respect to KH.

That common S5 knowledge has the properties of individual S5 knowledge is well-known [FHMV95, vDvdHK08]. Still, it may be surprising that common hope has the properties of individual hope. (Recall that common KD45 belief does not have the properties of individual KD45 belief as it lacks negative introspection.)

Proposition 5.39. *For any $\varnothing \neq G \subseteq \mathcal{A}$ and any $\varphi, \psi \in \mathcal{L}_{KH}^C$:*

$\mathscr{KHC} \vdash C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to C_G^H\psi$	$\mathscr{KHC} \vdash C_G^K(\varphi \to \psi) \wedge C_G^K\varphi \to C_G^K\psi$
$\mathscr{KHC} \vdash C_G^H\varphi \to C_G^H C_G^H\varphi$	$\mathscr{KHC} \vdash C_G^K\varphi \to C_G^K C_G^K\varphi$
	$\mathscr{KHC} \vdash \neg C_G^K\varphi \to C_G^K \neg C_G^K\varphi$
$\mathscr{KHC} \vdash \varphi \Longrightarrow \mathscr{KHC} \vdash C_G^H\varphi$	$\mathscr{KHC} \vdash \varphi \Longrightarrow \mathscr{KHC} \vdash C_G^K\varphi$
$\mathscr{KHC} \vdash \varphi \to C_G^H \neg C_G^H \neg\varphi$	$\mathscr{KHC} \vdash C_G^K\varphi \to \varphi$

Proof. We only prove the properties for the common hope operator:

$$\mathscr{KHC} \vdash C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to C_G^H\psi$$

1. $C_G^H\varphi \to E_G^H(\varphi \wedge C_G^H\varphi)$ axiom Mix^H

2. $C_G^H\varphi \to E_G^H\varphi$ by normality of E_G^H and prop. reasoning from 1.
3. $C_G^H(\varphi \to \psi) \to E_G^H((\varphi \to \psi) \wedge C_G^H(\varphi \to \psi))$ axiom Mix^H
4. $C_G^H(\varphi \to \psi) \to E_G^H(\varphi \to \psi)$ by normality of E_G^H and prop. reasoning from 3.
5. $C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to E_G^H(\varphi \to \psi) \wedge E_G^H\varphi$ by prop. reasoning from 2. and 4.
6. $E_G^H(\varphi \to \psi) \wedge E_G^H\varphi \to E_G^H\psi$ theorem of $\mathscr{H}$
7. $C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to E_G^H\psi$ by syllogism from 5. and 6.
8. $C_G^H\varphi \to E_G^HC_G^H\varphi$ by normality of E_G^H and prop. reasoning from 1.
9. $C_G^H(\varphi \to \psi) \to E_G^HC_G^H(\varphi \to \psi)$ by normality of E_G^H and prop. reasoning from 3.
10. $C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to E_G^HC_G^H(\varphi \to \psi) \wedge E_G^HC_G^H\varphi$ by prop. reasoning from 8. and 9.
11. $C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to E_G^H(\psi \wedge C_G^H(\varphi \to \psi) \wedge C_G^H\varphi)$ by normality of E_G^H and prop. reasoning from 7. and 10.
12. $C_G^H(\varphi \to \psi) \wedge C_G^H\varphi \to C_G^H\psi$ by Ind^H from 11.

$$\mathscr{KHC} \vdash C_G^H\varphi \to C_G^HC_G^H\varphi$$

1. $C_G^H\varphi \to E_G^H(\varphi \wedge C_G^H\varphi)$ axiom Mix^H
2. $C_G^H\varphi \to E_G^HC_G^H\varphi$ by normality of E_G^H and prop. reasoning from 1.
3. $C_G^H\varphi \to C_G^HC_G^H\varphi$ by Ind^H from 2.

$$\mathscr{KHC} \vdash \varphi \to C_G^H\neg C_G^H\neg\varphi$$

1. $\varphi \to H_i\neg H_i\neg\varphi$ axiom B^H
2. $\varphi \to E_G^H\neg E_G^H\neg\varphi$ by normality of H_i and prop. reasoning from 1.
3. $C_G^H\neg\varphi \to E_G^H(\neg\varphi \wedge C_G^H\neg\varphi)$ axiom Mix^H
4. $C_G^H\neg\varphi \to E_G^H\neg\varphi$ by normality of E_G^H and prop. reasoning from 3.
5. $(C_G^H\neg\varphi \to E_G^H\neg\varphi) \to (\neg E_G^H\neg\varphi \to \neg C_G^H\neg\varphi)$ prop. tautology
6. $\neg E_G^H\neg\varphi \to \neg C_G^H\neg\varphi$ by MP from 4. and 5.
7. $E_G^H(\neg E_G^H\neg\varphi \to \neg C_G^H\neg\varphi)$ by Nec^H and prop. reasoning from 6.
8. $E_G^H(\neg E_G^H\neg\varphi \to \neg C_G^H\neg\varphi) \to (E_G^H\neg E_G^H\neg\varphi \to E_G^H\neg C_G^H\neg\varphi)$ theorem of $\mathscr{H}$

9. $E_G^H \neg E_G^H \neg\varphi \to E_G^H \neg C_G^H \neg\varphi$ by *MP* from 7. and 8.
10. $E_G^H \neg E_G^H \neg\varphi \to E_G^H E_G^H \neg E_G^H \neg\varphi$ theorem of $\mathscr{H}$
11. $E_G^H \neg E_G^H \neg\varphi \to E_G^H(\neg C_G^H \neg\varphi \wedge E_G^H \neg E_G^H \neg\varphi)$ by normality of E_G^H and prop. reasoning from 9. and 10.
12. $E_G^H \neg E_G^H \neg\varphi \to C_G^H \neg C_G^H \neg\varphi$ by Ind^H from 11.
13. $\varphi \to C_G^H \neg C_G^H \neg\varphi$ by syllogism from 2. and 12.

$$\mathscr{KHC} \vdash \varphi \Longrightarrow \mathscr{KHC} \vdash C_G^H \varphi$$

1. φ assumption
2. $E_G^H \varphi$ by Nec^H and prop. reasoning from 1.
3. $E_G^H \varphi \to (\top \to E_G^H \varphi)$ prop. tautology
4. $\top \to E_G^H \varphi$ by *MP* from 2. and 3.
5. $\top \to C_G^H \varphi$ by Ind^H from 4.
6. $C_G^H \varphi$ by *MP* from $\top$ and 5.

□

Proposition 5.40. *$\mathscr{KHC} \vdash C_G^K \varphi \to C_G^H \varphi$ for all $\varnothing \neq G \subseteq \mathcal{A}$ and all $\varphi \in \mathcal{L}_{KH}^C$.*

Proof.

1. $C_G^K \varphi \to E_G^K(\varphi \wedge C_G^K \varphi)$ axiom Mix^K
2. $E_G^K(\varphi \wedge C_G^K \varphi) \to E_G^H(\varphi \wedge C_G^K \varphi)$ follows from Lemma 5.16 (1)
3. $C_G^K \varphi \to E_G^H(\varphi \wedge C_G^K \varphi)$ by syllogism from 1. and 2.
4. $C_G^K \varphi \to C_G^H \varphi$ by Ind^H from 3.

□

Formulas of $\mathcal{L}_{KH}^C$ are also evaluated on models from KH, with the new clauses for common knowledge and common hope stated in the following definition.

Definition 5.41. *For a model* $(W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n, \pi) \in \mathsf{KH}$*, let*

$$\mathcal{K}_G^C := \left(\bigcup_{i \in G} \mathcal{K}_i\right)^+, \qquad \mathcal{H}_G^C := \left(\bigcup_{i \in G} \mathcal{H}_i\right)^+,$$

where R^+ *is the transitive (but not reflexive) closure of a relation* R*. Then we define*

$$M, w \models C_G^K \varphi \quad \textit{iff} \quad M, v \models \varphi \quad \textit{for all} \quad v \in \mathcal{K}_G^C(w)$$

and

$$M, w \models C_G^H \varphi \quad \textit{iff} \quad M, v \models \varphi \quad \textit{for all} \quad v \in \mathcal{H}_G^C(w).$$

So far, by and large, the relationship between common knowledge and common hope exhibited the same traits as the relationship between their individual variants. But the naive generalization of connection axiom KH is invalid for the common modalities (for $|G| \geq 2$). We recall that $KH^{\rightarrow}$ corresponds to property one$\mathcal{H}$ that each knowledge equivalence class contains at most one hope partial equivalence class. It is easy to see that when lifted to the common modalities, each common knowledge equivalence class may contain more than one common hope partial equivalence class, thus, invalidating the generalization. The proof of the proposition below provides a simple four-world countermodel that demonstrates this fact:

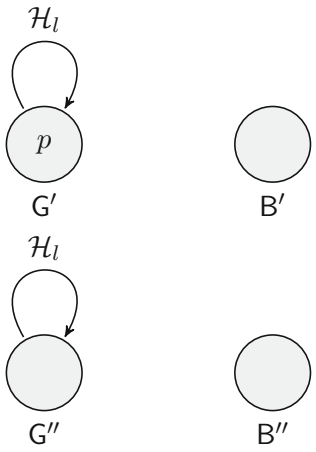
Proposition 5.42. *For any* $\varnothing \neq G \subseteq \mathcal{A}$ *such that* $|G| \geq 2$*, it is not the case that* $\mathsf{KH} \models C_G^H \varphi \leftrightarrow (\neg C_G^H \bot \rightarrow C_G^K(\neg C_G^H \bot \rightarrow \varphi))$ *for all* $\varphi \in \mathcal{L}_{KH}^C$*.*

Proof. To show this, we construct a countermodel from KH. Let $i \neq j \in G$.

Consider a model $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n, \pi) \in \mathsf{KH}$ such that

- $W = \{\mathsf{G}', \mathsf{G}'', \mathsf{B}', \mathsf{B}''\}$;
- $\mathcal{K}_i = \{(\mathsf{G}', \mathsf{G}'), (\mathsf{G}', \mathsf{B}'), (\mathsf{B}', \mathsf{G}'), (\mathsf{B}', \mathsf{B}'), (\mathsf{G}'', \mathsf{G}''), (\mathsf{G}'', \mathsf{B}'), (\mathsf{B}', \mathsf{G}''), (\mathsf{B}', \mathsf{B}')\}$, that is, $\mathcal{K}_i$ splits W into equivalence classes $\{\mathsf{G}', \mathsf{B}'\}$ and $\{\mathsf{G}'', \mathsf{B}''\}$;
- $\mathcal{K}_j = \{(\mathsf{G}', \mathsf{G}'), (\mathsf{G}', \mathsf{B}''), (\mathsf{B}'', \mathsf{G}'), (\mathsf{B}'', \mathsf{B}''), (\mathsf{G}'', \mathsf{G}''), (\mathsf{G}'', \mathsf{B}'), (\mathsf{B}', \mathsf{G}''), (\mathsf{B}', \mathsf{B}')\}$, that is, $\mathcal{K}_j$ splits W into equivalence classes $\{\mathsf{G}', \mathsf{B}''\}$ and $\{\mathsf{G}'', \mathsf{B}'\}$;
- $\mathcal{K}_l = \mathcal{K}_i$ for all $l \in G \setminus \{i, j\}$;
- $\mathcal{H}_l = \{(\mathsf{G}', \mathsf{G}'), (\mathsf{G}'', \mathsf{G}'')\}$, for all $l \in G$;
- $\pi(p) = \{\mathsf{G}'\}$ for some atom p, and
- $\pi(q) = W$ for all atoms $q \neq p$.

Thus, we have the following situation for hope relations of all agents $l \in G$, in particular (we omit the knowledge relations for the sake of clarity):



Clearly all agents from G are correct in G' and G'', i.e., $\mathcal{H}_l(\mathsf{G}') \neq \varnothing$ and $\mathcal{H}_l(\mathsf{G}'') \neq \varnothing$, and byzantine faulty in B' and B'', i.e., $\mathcal{H}_l(\mathsf{B}') = \varnothing$ and $\mathcal{H}_l(\mathsf{B}'') = \varnothing$. On the one hand,

$$M, \mathsf{G}' \models C_G^H p.$$

On the other hand, $M, w \models C_G^H \bot$ iff $w \in \{\mathsf{B}', \mathsf{B}''\}$. In particular, we have $M, \mathsf{G}' \models \neg C_G^H \bot$ and $M, \mathsf{G}'' \models \neg C_G^H \bot$. Now, $M, \mathsf{G}'' \not\models \neg C_G^H \bot \to p$ and, consequently, $M, \mathsf{G}' \not\models C_G^K(\neg C_G^H \bot \to p)$. Overall, we can conclude that

$$M, \mathsf{G}' \not\models \neg C_G^H \bot \to C_G^K(\neg C_G^H \bot \to p).$$

Therefore, $\mathsf{KH} \not\models C_G^H p \leftrightarrow (\neg C_G^H \bot \to C_G^K(\neg C_G^H \bot \to p))$. □

It turns out that even though we can "extract" mutual belief (among some agents) from mutual hope (among all agents), as per Proposition 5.28, we cannot do the same in case of common belief and common hope:

Proposition 5.43. *Let $n > 2$. For any $0 < f < n$, it is not the case that $\mathsf{KH}^{n-f} \models C_{\mathcal{A}}^H \varphi \to \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G| = n-f}} C_G^B \varphi$, for all $\varphi \in \mathcal{L}_{KH}^C$.*

Proof. Let us construct a countermodel from KH^{n-1} for $f = 1$.

Consider a model $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n, \pi) \in \mathsf{KH}^{n-1}$ such that

- $W = \{w, v_1, \ldots, v_n, u_1, \ldots, u_n\}$;

- for all $i \in \mathcal{A}$, the relation $\mathcal{K}_i$ splits W into the following equivalence classes:

$$\{w\} \cup (V \setminus \{v_i\}), \quad \{v_i, u_i\}, \quad \text{and} \quad U \setminus \{u_i\},$$

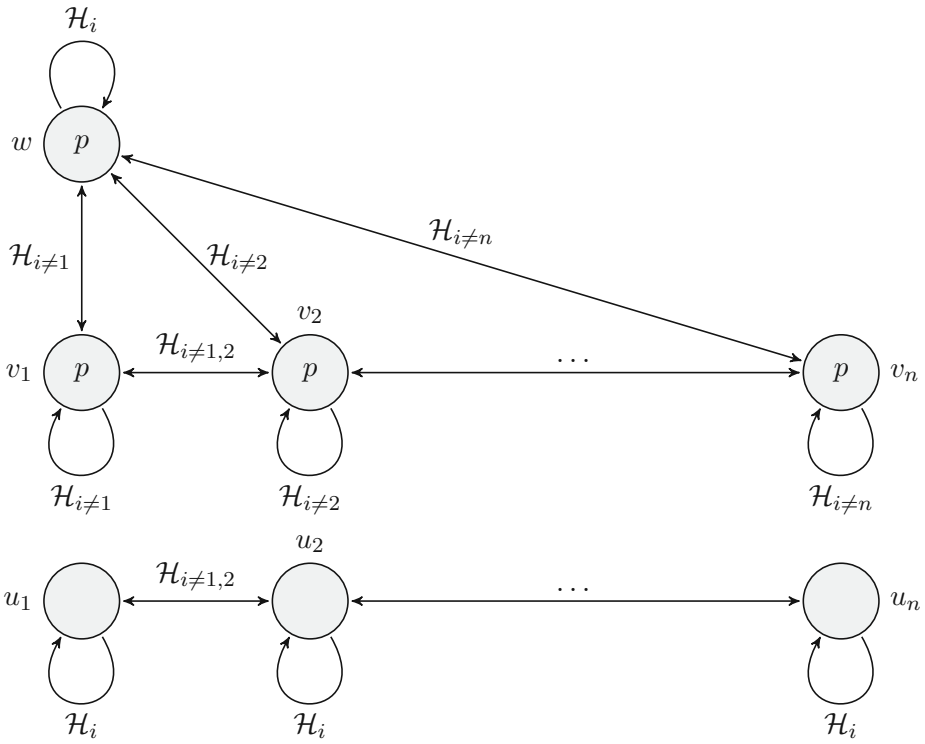
  where $V := \{v_1, \ldots, v_n\}$ and $U := \{u_1, \ldots, u_n\}$;

- $\mathcal{H}_i = \mathcal{K}_i \setminus \{(v_i, v_i), (v_i, u_i), (u_i, v_i)\}$, for all $i \in \mathcal{A}$ (therefore, for any $i \in \mathcal{A}$, agent i is byzantine faulty in world v_i since $\mathcal{H}_i(v_i) = \varnothing$);

- $\pi(p) = \{w, v_1, \ldots, v_n\}$ for some atom p, and

- $\pi(q) = W$ for all atoms $q \neq p$.

Thus, we have the following situation for hope relations, in particular (we omit the knowledge relations for the sake of clarity):



On the one hand,

$$M, w \models C^H_{\mathcal{A}} p$$

because $M, w' \models p$ for all $w' \in \mathcal{H}^C_{\mathcal{A}}(w) = \{w\} \cup V$.

On the other hand, for any $G \subseteq \mathcal{A}$ such that $|G| \geq 2$ we have

$$M, w \not\models C^B_G p$$

since $M, w \not\models B_j B_i p$ for any $j \neq i \in \mathcal{A}$ (as $M, v_i \not\models B_i p$). Therefore, $M, w \not\models C_G^B p$ for all $G \subseteq \mathcal{A}$ such that $|G| = n - 1$, in particular.

Similar countermodels can be constructed for all $1 < f < n$. □

As we will now see, there is also a way to define common hope using knowledge relations:

Definition 5.44. *For any $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{KH}$ and any $i \in \mathcal{A}$, we define*

$$\mathcal{K}_i^{\neg H_i \bot} := \{(w, v) \in \mathcal{K}_i \mid M, w \models \neg H_i \bot \quad \text{and} \quad M, v \models \neg H_i \bot\}.$$

Proposition 5.45. *For $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{KH}$ and any $i \in \mathcal{A}$,*

$$\mathcal{H}_i = \mathcal{K}_i^{\neg H_i \bot}.$$

Proof. ($\subseteq$): Assume $(w, v) \in \mathcal{H}_i$. Therefore, $M, w \models \neg H_i \bot$ holds. Since $\mathcal{H}_i$ is symmetric, we also have $(v, w) \in \mathcal{H}_i$. Therefore, $M, v \models \neg H_i \bot$ holds too. By $\mathcal{H}$in$\mathcal{K}$, $(w, v) \in \mathcal{K}_i$ follows. Thus, $(w, v) \in \mathcal{K}_i^{\neg H_i \bot}$.

($\supseteq$): Assume $(w, v) \in \mathcal{K}_i^{\neg H_i \bot}$. This means $(w, v) \in \mathcal{K}_i$, $M, w \models \neg H_i \bot$ and $M, v \models \neg H_i \bot$. That is, $(w, v) \in \mathcal{K}_i$, $\mathcal{H}_i(w) \neq \varnothing$ and $\mathcal{H}_i(v) \neq \varnothing$. Now, by one$\mathcal{H}$, $(w, v) \in \mathcal{H}_i$ follows. □

Corollary 5.46. *For any $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{KH}$ and any $\varnothing \neq G \subseteq \mathcal{A}$,*

$$M, w \models C_G^H \varphi \quad \textit{iff} \quad M, v \models \varphi \quad \textit{for all} \quad v \in W \quad \textit{such that} \quad (w, v) \in (\bigcup_{i \in G} \mathcal{K}_i^{\neg H_i \bot})^+.$$

5.4 Soundness and completeness of $\mathscr{KHC}$ with respect to KH

It is well-known that constructing one uniform-for-all-formulas canonical model does not work when common knowledge is added to the standard epistemic language [FHMV95, vDvdHK08]. The same is the case for the language $\mathcal{L}_{KH}^C$, where in addition to common knowledge we also have common hope. Therefore, for each formula $\varphi \in \mathcal{L}_{KH}^C$, we construct a canonical φ-model of $\mathscr{KHC}$ based on the *Fischer-Ladner closure* of φ, i.e., $cl(\varphi)$ (see Definition 5.52). The main tools used in this case are *maximal φ-consistent sets*, which are, in contrast to before, finite sets (since they are contained in $cl(\varphi)$, see Definition 5.55). The idea is standard: we take all maximal φ-consistent (with respect to $\mathscr{H}_{\text{co}}$) sets of formulas to be the worlds of the model and define the valuation function and the accessibility relations in terms of membership of formulas to such sets (see Definition 5.58). This way, using the properties of maximal φ-consistent sets (see Lemma 5.56) and the *Lindenbaum lemma* 5.57, we obtain the *Truth lemma* 5.63, according to which a formula $\psi \in cl(\varphi)$ belongs to a maximal φ-consistent set Γ if and only if it is satisfied in the

world Γ. Finally, just like before, we then obtain completeness by contraposition (see Theorem 5.64). The "only" challenge lies in defining the Fischer-Ladner closure of a formula $\varphi \in \mathcal{L}^C_{KH}$ appropriately.

Theorem 5.47. *The logic of common hope and common knowledge is not compact.*

Proof. Follows immediately from the well-known fact that the logic of common knowledge is not compact [vDvdHK08]. □

Therefore, using Theorem 2.22, we immediately get:

Corollary 5.48. *No axiomatization of common hope and common knowledge is strongly sound and strongly complete.*

Definition 5.49. *For all sets of formulas $\Gamma \subseteq \mathcal{L}^C_{KH}$ and all formulas $\varphi \in \mathcal{L}^C_{KH}$ we define*

$$\Gamma \vdash_{\mathcal{KHC}} \varphi \quad \textit{iff} \quad \vdash_{\mathcal{KHC}} \psi_1 \wedge \cdots \wedge \psi_n \to \varphi,$$

for some $\psi_1, \ldots, \psi_n \in \Gamma$.

Theorem 5.50 (Deduction theorem)**.** *For all sets of formulas $\Gamma \subseteq \mathcal{L}^C_{KH}$ and all formulas $\varphi, \psi \in \mathcal{L}^C_{KH}$:*

$$\Gamma \cup \{\psi\} \vdash_{\mathcal{KHC}} \varphi \iff \Gamma \vdash_{\mathcal{KHC}} \psi \to \varphi.$$

Proof. Let $\Gamma \subseteq \mathcal{L}^C_{KH}$ and $\varphi, \psi \in \mathcal{L}^C_{KH}$.

($\Longrightarrow$): Assume $\Gamma \cup \{\psi\} \vdash_{\mathcal{KHC}} \varphi$. According to the previous definition, this means that there exist some $\psi_1, \ldots, \psi_n \in \Gamma \cup \{\psi\}$ such that

$$\vdash_{\mathcal{KHC}} \psi_1 \wedge \cdots \wedge \psi_n \to \varphi.$$

- If $\psi \equiv \psi_i$ for some $i \in \{1, \ldots, n\}$, then, using propositional reasoning, we can rewrite the above in the following way:

  $$\vdash_{\mathcal{KHC}} \psi_1 \wedge \ldots \psi_{i-1} \wedge \psi_{i+1} \wedge \cdots \wedge \psi_n \to (\psi \to \varphi).$$

  So, we can conclude that indeed $\Gamma \vdash_{\mathcal{KHC}} \psi \to \varphi$.

- If $\psi \notin \{\psi_1, \ldots, \psi_n\}$, using $\vdash_{\mathcal{KHC}} \varphi \to (\psi \to \varphi)$ and propositional reasoning we obtain:

  $$\vdash_{\mathcal{KHC}} \psi_1 \wedge \cdots \wedge \psi_n \to (\psi \to \varphi).$$

  Again, we can conclude that indeed $\Gamma \vdash_{\mathcal{KHC}} \psi \to \varphi$.

($\Longleftarrow$): Assume $\Gamma \vdash_{\mathscr{KHC}} \psi \to \varphi$. According to the previous definition, this means that there exist $\psi_1, \ldots, \psi_n \in \Gamma$ such that

$$\vdash_{\mathscr{KHC}} \psi_1 \wedge \cdots \wedge \psi_n \to (\psi \to \varphi).$$

Using propositional reasoning, we can rewrite the above in the following way:

$$\vdash_{\mathscr{KHC}} \psi_1 \wedge \cdots \wedge \psi_n \wedge \psi \to \varphi.$$

This now means that $\Gamma \cup \{\psi\} \vdash_{\mathscr{KHC}} \varphi$ indeed holds. □

Recall that the language $\mathcal{L}^C_{KH}$ is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid H_i\varphi \mid C^K_G\varphi \mid C^H_G\varphi,$$

where $p \in \mathsf{Prop}$, $i \in \mathcal{A}$ and $G \subseteq \mathcal{A}$.

Definition 5.51. *Let $\varphi \in \mathcal{L}^C_{KH}$. The set $Sub(\varphi)$ of subformulas of φ is defined by induction on the construction of φ in the following way:*

$$\begin{aligned}
Sub(p) &:= \{p\};\\
Sub(\neg\psi) &:= \{\neg\psi\} \cup Sub(\psi);\\
Sub(\psi_1 \wedge \psi_2) &:= \{\psi_1 \wedge \psi_2\} \cup Sub(\psi_1) \cup Sub(\psi_2);\\
Sub(K_i\psi) &:= \{K_i\psi\} \cup Sub(\psi);\\
Sub(H_i\psi) &:= \{H_i\psi\} \cup Sub(\psi);\\
Sub(C^K_G\psi) &:= \{C^K_G\psi\} \cup Sub(\psi);\\
Sub(C^H_G\psi) &:= \{C^H_G\psi\} \cup Sub(\psi).
\end{aligned}$$

The Fischer-Ladner closure of formula $\varphi \in \mathcal{L}^C_{KH}$, i.e., $cl(\varphi)$, is defined in the following way:

Definition 5.52. *We define the following six sets for $\varphi \in \mathcal{L}^C_{KH}$:*

1. *$cl_0(\varphi)$ is the smallest set closed with respect to following rules*
   - *$\varphi \in cl_0(\varphi)$;*
   - *$H_i\neg H_i\bot \in cl_0(\varphi)$ for all $i \in \mathcal{A}$;*
   - *if $\psi \in cl_0(\varphi)$ and $\theta \in Sub(\psi)$, then $\theta \in cl_0(\varphi)$;*
   - *if $C^K_G\psi \in cl_0(\varphi)$, then $E^K_G(\psi \wedge C^K_G\psi) \in cl_0(\varphi)$;*
   - *if $C^H_G\psi \in cl_0(\varphi)$, then $E^H_G(\psi \wedge C^H_G\psi) \in cl_0(\varphi)$;*
2. *$cl_1(\varphi) := cl_0(\varphi) \cup \{\neg\psi \mid \psi \in cl_0(\varphi)$ and ψ is not a negation[1]$\}$;*

[1] That is, ψ does not start with the $\neg$ operator.

3. $cl_2(\varphi) := cl_1(\varphi) \cup \{H_i\psi, K_i(\neg H_i\bot \to \psi), H_i(\neg H_i\bot \to \psi), \neg H_i\bot \to \psi \mid K_i\psi \in cl_1(\varphi)\}$;
4. $cl_3(\varphi) := cl_2(\varphi) \cup \{\neg\psi \mid \psi \in cl_2(\varphi)$ *and* ψ *is not a negation*$\}$;
5. $cl_4(\varphi) := cl_3(\varphi) \cup \{K_iK_i\psi, H_iK_i\psi \mid K_i\psi \in cl_3(\varphi)\} \cup \{K_i\neg K_i\psi, H_i\neg K_i\psi \mid \neg K_i\psi \in cl_3(\varphi)\}$;
6. $cl(\varphi) := cl_4(\varphi) \cup \{\neg\psi \mid \psi \in cl_4(\varphi)$ *and* ψ *is not a negation*$\}$.

Lemma 5.53. *For any formula* $\varphi \in \mathcal{L}^C_{KH}$, $cl(\varphi)$ *is finite.*

Proof. First of all, $cl(H_i\neg H_i\bot) = \{H_i\neg H_i\bot, \neg H_i\bot, H_i\bot, \bot, \neg H_i\neg H_i\bot, \neg\bot\}$ is evidently finite. We proceed by induction on the structure of φ.

Base case: If $\varphi = p$, then $cl(\varphi) = \{p, \neg p\} \cup cl(H_i\neg H_i\bot)$, which is finite.

Induction step:

1. φ is of the form $\neg\psi$. Then
$$cl(\varphi) = \{\neg\psi\} \cup cl(\psi) \cup cl(H_i\neg H_i\bot).$$
Since $cl(\psi)$ is finite according to the induction hypothesis, it follows that $cl(\varphi)$ is finite as well.
2. φ is of the form $\psi_1 \wedge \psi_2$. Then
$$cl(\varphi) = \{\psi_1 \wedge \psi_2, \neg(\psi_1 \wedge \psi_2)\} \cup cl(\psi_1) \cup cl(\psi_2) \cup cl(H_i\neg H_i\bot).$$
Since $cl(\psi_1)$ and $cl(\psi_2)$ are finite according to the induction hypothesis, it follows that $cl(\varphi)$ is finite as well.
3. φ is of the form $K_i\psi$. Then
$$\begin{aligned} cl(\varphi) = \{&K_i\psi, \neg K_i\psi, K_iK_i\psi, \neg K_iK_i\psi, K_i\neg K_i\psi, \neg K_i\neg K_i\psi, H_iK_i\psi, \neg H_iK_i\psi, \\ &H_i\neg K_i\psi, \neg H_i\neg K_i\psi, H_i\psi, \neg H_i\psi, K_i(\neg H_i\bot \to \psi), \neg K_i(\neg H_i\bot \to \psi), \\ &H_i(\neg H_i\bot \to \psi), \neg H_i(\neg H_i\bot \to \psi), \neg H_i\bot \to \psi, \neg(\neg H_i\bot \to \psi)\} \cup cl(\psi) \cup \\ &cl(H_i\neg H_i\bot). \end{aligned}$$
Since $cl(\psi)$ is finite according to the induction hypothesis, it follows that $cl(\varphi)$ is finite as well.
4. φ is of the form $H_i\psi$. Then
$$cl(\varphi) = \{H_i\psi, \neg H_i\psi\} \cup cl(\psi) \cup cl(H_i\neg H_i\bot).$$
Since $cl(\psi)$ is finite according to the induction hypothesis, it follows that $cl(\varphi)$ is finite as well.

5. φ is of the form $C_G^K\psi$. Then

$$\begin{aligned} cl(\varphi) =&\{C_G^K\psi, E_G^K(\psi \wedge C_G^K\psi), \psi \wedge C_G^K\psi, \neg C_G^K\psi, \neg E_G^K(\psi \wedge C_G^K\psi),\\ &\neg(\psi \wedge C_G^K\psi)\} \cup cl(\psi) \cup cl(H_i\neg H_i\bot). \end{aligned}$$

Since $cl(\psi)$ is finite according to the induction hypothesis, it follows that $cl(\varphi)$ is finite as well.

6. φ is of the form $C_G^H\psi$. Then

$$\begin{aligned} cl(\varphi) =&\{C_G^H\psi, E_G^H(\psi \wedge C_G^H\psi), \psi \wedge C_G^H\psi, \neg C_G^H\psi, \neg E_G^H(\psi \wedge C_G^H\psi),\\ &\neg(\psi \wedge C_G^H\psi)\} \cup cl(\psi) \cup cl(H_i\neg H_i\bot). \end{aligned}$$

Since $cl(\psi)$ is finite according to the induction hypothesis, it follows that $cl(\varphi)$ is finite as well. □

Lemma 5.54. *For any formula $\varphi, \psi \in \mathcal{L}_{KH}^C$:*

1. *if $\psi \in cl(\varphi)$ is not a negation, then $\neg\psi \in cl(\varphi)$,*
2. *if $\psi \in cl(\varphi)$, then $Sub(\psi) \subseteq cl(\varphi)$.*

Proof. 1. Let $\psi \in cl(\varphi)$. If ψ is not a negation, then $\psi \in cl_4(\varphi)$ according to Definition 5.52. Now it immediately follows that $\neg\psi \in cl(\varphi)$ according to Definition 5.52.

2. Let $\psi \in cl(\varphi)$. We proceed by induction on the structure of ψ.

Base case: If $\psi = p$, then $Sub(\psi) = \{p\} \subseteq cl(\varphi)$ since $p \in cl(\varphi)$ by assumption.

Induction step:

1. ψ is of the form $\neg\theta$. According to Definition 5.52, $\neg\theta \in cl(\varphi)$ means that either $\neg\theta \in cl_4(\varphi)$ or $\theta \in cl_4(\varphi)$. Once we unfold this, we get that either $\neg\theta \in cl_0(\varphi)$ or $\theta \in cl_0(\varphi)$ or $\theta \in cl_2(\varphi)$ or $\theta \in cl_4(\varphi)$. In all these cases $\theta \in cl(\varphi)$ follows. Thus, using Definition 5.51 and the induction hypothesis, we obtain $Sub(\psi) = \{\neg\theta\} \cup Sub(\theta) \subseteq cl(\varphi)$.
2. ψ is of the form $\theta_1 \wedge \theta_2$. According to Definition 5.52, $\theta_1 \wedge \theta_2 \in cl(\varphi)$ means that $\theta_1 \wedge \theta_2 \in cl_0(\varphi)$, so $\theta_1 \in cl(\varphi)$ and $\theta_2 \in cl(\varphi)$ follow. Thus, using Definition 5.51 and the induction hypothesis, we obtain $Sub(\psi) = \{\theta_1 \wedge \theta_2\} \cup Sub(\theta_1) \cup Sub(\theta_2) \subseteq cl(\varphi)$.
3. ψ is of the form $K_i\theta$. According to Definition 5.52, $K_i\theta \in cl(\varphi)$ means that either:
   - $K_i\theta \in cl_0(\varphi)$, in which case $\theta \in cl_0(\varphi)$ immediately follows.
   - $K_i\theta \in cl_2(\varphi) \setminus cl_0(\varphi)$, in which case $\theta = \neg H_i\bot \to \theta'$, for $K_i\theta' \in cl_1(\varphi)$. Hence $\neg H_i\bot \to \theta' = \theta \in cl_2(\varphi)$ follows.

- $K_i\theta \in cl_4(\varphi) \setminus cl_2(\varphi)$, in which case either:
  - $\theta = K_i\theta'$, for $K_i\theta' \in cl_3(\varphi)$.
  - $\theta = \neg K_i\theta'$, for $\neg K_i\theta' \in cl_3(\varphi)$.

Regardless of the case, $\theta \in cl(\varphi)$ follows. Thus, using Definition 5.51 and the induction hypothesis, we obtain $Sub(\psi) = \{K_i\theta\} \cup Sub(\theta) \subseteq cl(\varphi)$.

4. ψ is of the form $H_i\theta$. According to Definition 5.52, $H_i\theta \in cl(\varphi)$ means that either:

- $H_i\theta \in cl_0(\varphi)$, in which case $\theta \in cl_0(\varphi)$ immediately follows.
- $H_i\theta \in cl_2(\varphi) \setminus cl_0(\varphi)$, in which case either:
  - $K_i\theta \in cl_1(\varphi)$, in which case $\theta \in cl_0(\varphi)$ immediately follows.
  - $\theta = \neg H_i\bot \to \theta'$, for $K_i\theta' \in cl_1(\varphi)$. Hence $\neg H_i\bot \to \theta' = \theta \in cl_2(\varphi)$ follows.
- $H_i\theta \in cl_4(\varphi) \setminus cl_2(\varphi)$, in which case either:
  - $\theta = K_i\theta'$, for $K_i\theta' \in cl_3(\varphi)$.
  - $\theta = \neg K_i\theta'$, for $\neg K_i\theta' \in cl_3(\varphi)$.

Regardless of the case, $\theta \in cl(\varphi)$ follows. Thus, using Definition 5.51 and the induction hypothesis, we obtain $Sub(\psi) = \{H_i\theta\} \cup Sub(\theta) \subseteq cl(\varphi)$.

5. ψ is of the form $C^K_G\theta$. According to Definition 5.52, $C^K_G\theta \in cl(\varphi)$ means that $C^K_G\theta \in cl_0(\varphi)$, so $\theta \in cl(\varphi)$ follows. Thus, using Definition 5.51 and the induction hypothesis, we obtain $Sub(\psi) = \{C^K_G\theta\} \cup Sub(\theta) \subseteq cl(\varphi)$.

6. ψ is of the form $C^H_G\theta$. According to Definition 5.52, $C^H_G\theta \in cl(\varphi)$ means that $C^H_G\theta \in cl_0(\varphi)$, so $\theta \in cl(\varphi)$ follows. Thus, using Definition 5.51 and the induction hypothesis, we obtain $Sub(\psi) = \{C^H_G\theta\} \cup Sub(\theta) \subseteq cl(\varphi)$. □

Definition 5.55. *Let $\Gamma \subseteq \mathcal{L}^C_{KH}$ and $\varphi \in \mathcal{L}^C_{KH}$. We say that Γ is φ-consistent with respect to $\mathscr{KHC}$ if:*

- $\Gamma \subset cl(\varphi)$ *and*
- $\Gamma \nvdash_{\mathscr{KHC}} \bot$.

Furthermore, we say that Γ is maximal φ-consistent with respect to $\mathscr{KHC}$ if:

- *Γ is φ-consistent with respect to $\mathscr{KHC}$ and*
- *for any $\Gamma \subset \Gamma' \subseteq cl(\varphi)$, we have $\Gamma' \vdash_{\mathscr{KHC}} \bot$.*

The following lemma will prove to be particularly useful.

Lemma 5.56. *Let $\varphi, \psi, \theta \in \mathcal{L}^C_{KH}$ and $\Gamma \subseteq \mathcal{L}^C_{KH}$ be a maximal φ-consistent set with respect to $\mathscr{KHC}$. Then the following holds:*

1. *If $\psi \in cl(\varphi)$, then $\Gamma \vdash_{\mathscr{KHC}} \psi \implies \psi \in \Gamma$.*
2. *If $\neg\psi \in cl(\varphi)$, then $\psi \in \Gamma \iff \neg\psi \notin \Gamma$.*
3. *If $\psi \wedge \theta \in cl(\varphi)$, then $\psi \wedge \theta \in \Gamma \iff \psi \in \Gamma$ and $\theta \in \Gamma$.*

Proof. 1. Let $\psi \in cl(\varphi)$ and $\Gamma \vdash_{\mathscr{KHC}} \psi$. Assume towards a contradiction that $\psi \notin \Gamma$. Given the assumption that Γ is a maximal φ-consistent set with respect to $\mathscr{KHC}$, now we get $\Gamma \cup \{\psi\} \vdash_{\mathscr{KHC}} \bot$. According to the Deduction theorem 5.50, $\Gamma \vdash_{\mathscr{KHC}} \psi \to \bot$ follows. By propositional reasoning, we obtain $\Gamma \vdash_{\mathscr{KHC}} \bot$, contradicting the φ-consistency of Γ.

2. Let $\neg\psi \in cl(\varphi)$.

   ($\Longrightarrow$): Let $\psi \in \Gamma$. Assume towards a contradiction that $\neg\psi \in \Gamma$. We now obtain $\Gamma \vdash_{\mathscr{KHC}} \bot$, contradicting the φ-consistency of Γ.

   ($\Longleftarrow$): Let $\neg\psi \notin \Gamma$. Given the assumption that Γ is a maximal φ-consistent set with respect to $\mathscr{KHC}$, we get $\Gamma \cup \{\neg\psi\} \vdash_{\mathscr{KHC}} \bot$. According to the Deduction theorem 5.50, $\Gamma \vdash_{\mathscr{KHC}} \neg\psi \to \bot$ follows. By propositional reasoning, we obtain $\Gamma \vdash_{\mathscr{KHC}} \psi$. From this, using 1., we now obtain $\psi \in \Gamma$ (note that $\psi \in cl(\varphi)$ according to Lemma 5.54).

3. Let $\psi \wedge \theta \in cl(\varphi)$.

   ($\Longrightarrow$): Let $\psi \wedge \theta \in \Gamma$. Assume towards a contradiction that $\psi \notin \Gamma$. From 2. it follows that $\neg\psi \in \Gamma$ (note that $\neg\psi \in cl(\varphi)$ according to Lemma 5.54). However, using $\psi \wedge \theta \in \Gamma$ and $\neg\psi \in \Gamma$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \bot$, contradicting the φ-consistency of Γ. We prove that $\theta \in \Gamma$ holds analogously.

   ($\Longleftarrow$): Let $\psi \in \Gamma$ and $\theta \in \Gamma$. Assume towards a contradiction that $\psi \wedge \theta \notin \Gamma$. From 2. it now follows that $\neg(\psi \wedge \theta) \in \Gamma$ (note that $\neg(\psi \wedge \theta) \in cl(\varphi)$ according to Lemma 5.54). However, from $\psi \in \Gamma$, $\theta \in \Gamma$ and $\neg(\psi \wedge \theta) \in \Gamma$ it follows $\Gamma \vdash_{\mathscr{KHC}} \bot$, contradicting the φ-consistency of Γ. □

Lemma 5.57 (Lindenbaum lemma)**.** *Let $\Gamma \subseteq \mathcal{L}^C_{KH}$ and $\varphi \in \mathcal{L}^C_{KH}$. If Γ is φ-consistent with respect to $\mathscr{KHC}$, then there exists a set $\Gamma^* \supseteq \Gamma$ such that Γ^* is maximal φ-consistent with respect to $\mathscr{KHC}$.*

Proof. Assume that Γ is φ-consistent with respect to $\mathscr{KHC}$. Let $|cl(\varphi)| = n$. First, let us enumerate all formulas from $cl(\varphi)$ (without repetitions): $\theta_0, \theta_1, \ldots, \theta_{n-1}$. Next, we

define recursively the following sequence of sets Δ_i:

$$\Delta_0 := \Gamma,$$
$$\Delta_{i+1} := \begin{cases} \Delta_i \cup \{\theta_i\}, & \text{if } \Delta_i \cup \{\theta_i\} \text{ is } \varphi\text{-consistent with respect to } \mathscr{KHC} \\ \Delta_i, & \text{otherwise} \end{cases}$$

Obviously, $\Gamma \subseteq \Delta_n$. Each set Δ_i is φ-consistent with respect to $\mathscr{KHC}$, by construction. In particular, Δ_n is φ-consistent with respect to $\mathscr{KHC}$. We show that Δ_n is maximal too. Assume the opposite towards a contradiction: let Δ' be a φ-consistent set with respect to $\mathscr{KHC}$ such that $\Delta_n \subset \Delta'$. Therefore, there exists a formula $\theta_l \in \Delta' \setminus \Delta_n$, where $0 \leq l \leq n-1$. In particular, $\theta_l \notin \Delta_{l+1} \subseteq \Delta_n$. Now, we can conclude that $\Delta_l \cup \{\theta_l\}$ is not φ-consistent with respect to $\mathscr{KHC}$. However, this contradicts the φ-consistency of Δ' with respect to $\mathscr{KHC}$ given that $\Delta_l \cup \{\theta_l\} \subseteq \Delta'$ holds. □

Definition 5.58 (Canonical model)**.** *Let $\varphi \in \mathcal{L}^C_{KH}$. The canonical φ-model*

$$M^\varphi = (W^\varphi, \pi^\varphi, \mathcal{K}^\varphi_1, \ldots, \mathcal{K}^\varphi_n, \mathcal{H}^\varphi_1, \ldots, \mathcal{H}^\varphi_n)$$

of $\mathscr{KHC}$ is defined in the following way:

1. $W^\varphi := \{\Gamma \mid \Gamma$ *is maximal φ-consistent set with respect to $\mathscr{KHC}$}*;*
2. $\Gamma \mathcal{K}^\varphi_i \Delta$ *iff for all $\sigma \in cl(\varphi)$: $K_i\sigma \in \Gamma \to \sigma \in \Delta$;*
3. $\Gamma \mathcal{H}^\varphi_i \Delta$ *iff for all $\sigma \in cl(\varphi)$: $H_i\sigma \in \Gamma \to \sigma \in \Delta$;*
4. $\pi^\varphi(p) := \{\Gamma \in W^\varphi \mid p \in \Gamma\}$.

Lemma 5.59 (Correctness lemma)**.** *For any $\varphi \in \mathcal{L}^C_{KH}$, $M^\varphi \in \mathsf{KH}$ holds.*

Proof. We need to show that $\mathcal{K}^\varphi_i$ are equivalence relations, that $\mathcal{H}^\varphi_i$ are shift serial and that the conditions $\mathcal{H}$in$\mathcal{K}$ and one$\mathcal{H}$ are satisfied as well.

- $\mathcal{K}^\varphi_i$ is reflexive: We need to show that $\Gamma \mathcal{K}^\varphi_i \Gamma$ holds. Assume $K_i\theta \in \Gamma$ for some $\theta \in cl(\varphi)$. Using axiom T^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to \theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \theta$ by Definition 5.49. Using Lemma 5.56 (1), we obtain $\theta \in \Gamma$, as desired.
- $\mathcal{K}^\varphi_i$ is transitive: Assume that $\Gamma \mathcal{K}^\varphi_i \Delta$ and $\Delta \mathcal{K}^\varphi_i \Sigma$. Therefore:

$$(\forall \sigma \in cl(\varphi))(K_i\sigma \in \Gamma \to \sigma \in \Delta), \tag{5.3}$$

$$(\forall \sigma \in cl(\varphi))(K_i\sigma \in \Delta \to \sigma \in \Sigma). \tag{5.4}$$

  We need to show that $\Gamma \mathcal{K}^\varphi_i \Sigma$ holds too. Assume $K_i\theta \in \Gamma$ for some $\theta \in cl(\varphi)$. We need to show $\theta \in \Sigma$.

1. Assume $K_i\theta \in cl_3(\varphi) \subset cl(\varphi)$. Using axiom 4^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to K_iK_i\theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} K_iK_i\theta$ according to Definition 5.49. Using Lemma 5.56 (1), we obtain that $K_iK_i\theta \in \Gamma$ holds (note that $K_iK_i\theta \in cl(\varphi)$). According to assumption (5.3), this implies that $K_i\theta \in \Delta$ holds. Finally, using assumption (5.4), we obtain $\theta \in \Sigma$, as desired.
2. Assume $K_i\theta \in cl(\varphi)\backslash cl_3(\varphi)$.
   - $\theta = K_i\theta'$ for $K_i\theta' \in cl_3(\varphi)$. Using axiom T^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to \theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \theta$ by Definition 5.49, that is, $\Gamma \vdash_{\mathscr{KHC}} K_i\theta'$. Using Lemma 5.56 (1), we obtain that $K_i\theta' \in \Gamma$ holds. Using axiom 4^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta' \to K_iK_i\theta'$, we further obtain $\Gamma \vdash_{\mathscr{KHC}} K_iK_i\theta'$ according to Definition 5.49. Using Lemma 5.56 (1), we obtain that $K_iK_i\theta' \in \Gamma$ holds (note that $K_iK_i\theta' \in cl(\varphi)$). According to assumption (5.3), this implies that $K_i\theta' \in \Delta$ holds. Using axiom 4^K again, more precisely $\vdash_{\mathscr{KHC}} K_i\theta' \to K_iK_i\theta'$, we obtain $\Delta \vdash_{\mathscr{KHC}} K_iK_i\theta'$ according to Definition 5.49. Using Lemma 5.56 (1) one more time, we further obtain that $K_iK_i\theta' \in \Delta$ holds (note that $K_iK_i\theta' \in cl(\varphi)$). Finally, using assumption (5.4), we obtain $K_i\theta' = \theta \in \Sigma$, as desired.
   - $\theta = \neg K_i\theta'$ for $\neg K_i\theta' \in cl_3(\varphi)$. Using axiom T^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to \theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \theta$ by Definition 5.49, that is, $\Gamma \vdash_{\mathscr{KHC}} \neg K_i\theta'$. Using Lemma 5.56 (1), we obtain that $\neg K_i\theta' \in \Gamma$ holds. Using axiom 5^K, more precisely $\vdash_{\mathscr{KHC}} \neg K_i\theta' \to K_i\neg K_i\theta'$, we further obtain $\Gamma \vdash_{\mathscr{KHC}} K_i\neg K_i\theta'$ according to Definition 5.49. Using Lemma 5.56 (1), we obtain that $K_i\neg K_i\theta' \in \Gamma$ holds (note that $K_i\neg K_i\theta' \in cl(\varphi)$). According to assumption (5.3), this implies that $\neg K_i\theta' \in \Delta$ holds. Using axiom 5^K again, more precisely $\vdash_{\mathscr{KHC}} \neg K_i\theta' \to K_i\neg K_i\theta'$, we obtain $\Delta \vdash_{\mathscr{KHC}} K_i\neg K_i\theta'$ according to Definition 5.49. Using Lemma 5.56 (1) one more time, we further obtain that $K_i\neg K_i\theta' \in \Delta$ holds (note that $K_i\neg K_i\theta' \in cl(\varphi)$). Finally, using assumption (5.4), we obtain $\neg K_i\theta' = \theta \in \Sigma$, as desired.

- $\mathcal{K}_i^\varphi$ is euclidean: Assume that $\Gamma\mathcal{K}_i^\varphi\Delta$ and $\Gamma\mathcal{K}_i^\varphi\Sigma$. Therefore:

$$(\forall\sigma \in cl(\varphi))(K_i\sigma \in \Gamma \to \sigma \in \Delta), \tag{5.5}$$

$$(\forall\sigma \in cl(\varphi))(K_i\sigma \in \Gamma \to \sigma \in \Sigma). \tag{5.6}$$

  We need to show that $\Delta\mathcal{K}_i^\varphi\Sigma$ holds too. Assume $K_i\theta \in \Delta$ for some $\theta \in cl(\varphi)$. We need to show $\theta \in \Sigma$.

1. Assume $\neg K_i\theta \in cl_3(\varphi) \subset cl(\varphi)$. From $K_i\theta \in \Delta$ it follows that $\neg K_i\theta \notin \Delta$ (otherwise, Δ would not be φ-consistent with respect to $\mathscr{KHC}$). Therefore, according to (5.5), $K_i\neg K_i\theta \notin \Gamma$ follows. Using Lemma 5.56 (1), we obtain $\Gamma \nvdash_{\mathscr{KHC}} K_i\neg K_i\theta$ (note that $K_i\neg K_i\theta \in cl(\varphi)$). Using axiom 5^K, more precisely, $\vdash_{\mathscr{KHC}} \neg K_i\theta \to K_i\neg K_i\theta$, we further obtain $\neg K_i\theta \notin \Gamma$ by Definition 5.49.

Therefore, $K_i\theta \in \Gamma$ (otherwise Γ would not be φ-consistent with respect to $\mathscr{KHC}$). Finally, assumption (5.6) implies that $\theta \in \Sigma$ indeed holds.

2. Assume $\neg K_i\theta \in cl(\varphi)\backslash cl_3(\varphi)$.
   - $\theta = K_i\theta'$ for $K_i\theta' \in cl_3(\varphi)$. Using axiom T^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta \rightarrow \theta$, we obtain $\Delta \vdash_{\mathscr{KHC}} \theta$ by Definition 5.49, that is, $\Delta \vdash_{\mathscr{KHC}} K_i\theta'$. Using Lemma 5.56 (1), we obtain that $K_i\theta' \in \Delta$ holds. Therefore, $\neg K_i\theta' \notin \Delta$ (otherwise, Δ would not be φ-consistent with respect to $\mathscr{KHC}$). Assumption (5.5) implies that $K_i\neg K_i\theta' \notin \Gamma$. Using Lemma 5.56 (1), we obtain $\Gamma \nvdash_{\mathscr{KHC}} K_i\neg K_i\theta'$. Using axiom 5^K, more precisely, $\vdash_{\mathscr{KHC}} \neg K_i\theta' \rightarrow K_i\neg K_i\theta'$, we further obtain $\neg K_i\theta' \notin \Gamma$ by Definition 5.49. Lemma 5.56 (2) implies that then $K_i\theta' \in \Gamma$ holds. Using axiom 4^K, more precisely, $\vdash_{\mathscr{KHC}} K_i\theta' \rightarrow K_iK_i\theta'$ we obtain $\Gamma \vdash_{\mathscr{KHC}} K_iK_i\theta'$ by Definition 5.49. Using Lemma 5.56 (1), we obtain $K_iK_i\theta' \in \Gamma$ (note that $K_iK_i\theta' \in cl(\varphi)$). Assumption (5.6) implies that $K_i\theta' = \theta \in \Sigma$.
   - $\theta = \neg K_i\theta'$ for $\neg K_i\theta' \in cl_3(\varphi)$. Using axiom T^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta \rightarrow \theta$, we obtain $\Delta \vdash_{\mathscr{KHC}} \theta$ by Definition 5.49, that is, $\Delta \vdash_{\mathscr{KHC}} \neg K_i\theta'$. Using Lemma 5.56 (1), we obtain that $\neg K_i\theta' \in \Delta$ holds. Therefore, $K_i\theta' \notin \Delta$ (otherwise, Δ would not be φ-consistent with respect to $\mathscr{KHC}$). Assumption (5.5) implies that $K_iK_i\theta' \notin \Gamma$. Using Lemma 5.56 (1), we obtain $\Gamma \nvdash_{\mathscr{KHC}} K_iK_i\theta$. Using axiom 4^K, more precisely, $\vdash_{\mathscr{KHC}} K_i\theta' \rightarrow K_iK_i\theta'$, we further obtain $K_i\theta' \notin \Gamma$ by Definition 5.49. Lemma 5.56 (2) implies that then $\neg K_i\theta' \in \Gamma$ holds. Using axiom 5^K, more precisely, $\vdash_{\mathscr{KHC}} \neg K_i\theta' \rightarrow K_i\neg K_i\theta'$ we obtain $\Gamma \vdash_{\mathscr{KHC}} K_i\neg K_i\theta'$ by Definition 5.49. Using Lemma 5.56 (1), we obtain $K_i\neg K_i\theta' \in \Gamma$ (note that $K_i\neg K_i\theta' \in cl(\varphi)$). Assumption (5.6) implies that $\neg K_i\theta' = \theta \in \Sigma$.

- $\mathcal{H}_i^\varphi$ is shift serial: Assume that $\Gamma\mathcal{H}_i^\varphi\Delta$. Therefore:

$$(\forall\sigma \in cl(\varphi))(H_i\sigma \in \Gamma \rightarrow \sigma \in \Delta). \tag{5.7}$$

  We need to show that $\mathcal{H}_i^\varphi(\Delta) \neq \varnothing$. Using axiom H^H, more precisely, $\vdash_{\mathscr{KHC}} H_i\neg H_i\bot$, we obtain $\Gamma \vdash_{\mathscr{KHC}} H_i\neg H_i\bot$ by Definition 5.49. Using Lemma 5.56 (1), we further obtain that $H_i\neg H_i\bot \in \Gamma$ holds (note that $H_i\neg H_i\bot \in cl(\varphi)$). Assumption (5.7) now implies that $\neg H_i\bot \in \Delta$ (note that $\neg H_i\bot \in cl(\varphi)$). Assume $H_i\theta \in \Delta$ for some $\theta \in cl(\varphi)$. Using Proposition 5.20, more precisely $\vdash_{\mathscr{KHC}} \neg H_i\bot \rightarrow (H_i\theta \rightarrow \theta)$, we get $\Delta \vdash_{\mathscr{KHC}} \theta$ by Definition 5.49. Finally, according to Lemma 5.56 (1), $\theta \in \Delta$ (note that $\theta \in cl(\varphi)$). Therefore, $\mathcal{H}_i^\varphi(\Delta) \neq \varnothing$ indeed holds since we have just shown that $\Delta\mathcal{H}_i^\varphi\Delta$.

- $\mathcal{H}\text{in}\mathcal{K}$: Assume that $\Gamma\mathcal{H}_i^\varphi\Delta$. Therefore:

$$(\forall\sigma \in cl(\varphi))(H_i\sigma \in \Gamma \rightarrow \sigma \in \Delta). \tag{5.8}$$

  We need to show that $\Gamma\mathcal{K}_i^\varphi\Delta$ holds too. Assume $K_i\theta \in \Gamma$ for some $\theta \in cl(\varphi)$. We need to show $\theta \in \Delta$.

– Assume $K_i\theta \in cl_1(\varphi) \subset cl(\varphi)$. Using Proposition 5.16 (1), more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to H_i\theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} H_i\theta$ by Definition 5.49. Using Lemma 5.56 (1), we now obtain that $H_i\theta \in \Gamma$ holds (note that $H_i\theta \in cl(\varphi)$). According to assumption (5.8), this implies that $\theta \in \Delta$ indeed holds.

– Assume $K_i\theta \in cl(\varphi)\backslash cl_1(\varphi)$.

∗ $\theta = \neg H_i\bot \to \theta'$ for $K_i\theta' \in cl_1(\varphi)$. Using Proposition 5.16 (1), more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to H_i\theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} H_i\theta$ by Definition 5.49, that is, $\Gamma \vdash_{\mathscr{KHC}} H_i(\neg H_i\bot \to \theta')$. Using Lemma 5.56 (1), we now obtain that $H_i(\neg H_i\bot \to \theta') \in \Gamma$ holds (note that $H_i(\neg H_i\bot \to \theta') \in cl(\varphi)$). According to assumption (5.8), this implies that $\neg H_i\bot \to \theta' = \theta \in \Delta$ indeed holds.

∗ $\theta = K_i\theta'$ for $K_i\theta' \in cl_3(\varphi)$. Using Proposition 5.16 (1), more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to H_i\theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} H_i\theta$ by Definition 5.49, that is, $\Gamma \vdash_{\mathscr{KHC}} H_iK_i\theta'$. Using Lemma 5.56 (1), we now obtain that $H_iK_i\theta' \in \Gamma$ holds (note that $H_iK_i\theta' \in cl(\varphi)$). According to assumption (5.8), this implies that $K_i\theta' = \theta \in \Delta$ indeed holds.

∗ $\theta = \neg K_i\theta'$ for $\neg K_i\theta' \in cl_3(\varphi)$. Using Proposition 5.16 (1), more precisely $\vdash_{\mathscr{KHC}} K_i\theta \to H_i\theta$, we obtain $\Gamma \vdash_{\mathscr{KHC}} H_i\theta$ by Definition 5.49, that is, $\Gamma \vdash_{\mathscr{KHC}} H_i\neg K_i\theta'$. Using Lemma 5.56 (1), we now obtain that $H_i\neg K_i\theta' \in \Gamma$ holds (note that $H_i\neg K_i\theta' \in cl(\varphi)$). According to assumption (5.8), this implies that $\neg K_i\theta' = \theta \in \Delta$ indeed holds.

- one$\mathcal{H}$: Let $\Gamma, \Delta \in W^\varphi$. Assume that $\mathcal{H}_i^\varphi(\Gamma) \neq \varnothing$ (i.e., that there exists $\Gamma' \in W^\varphi$ such that $\Gamma\mathcal{H}_i^\varphi\Gamma'$), $\mathcal{H}_i^\varphi(\Delta) \neq \varnothing$ (i.e., that there exists $\Delta' \in W^\varphi$ such that $\Delta\mathcal{H}_i^\varphi\Delta'$) and that $\Gamma\mathcal{K}_i^\varphi\Delta$. Therefore:

$$\neg H_i\bot \in \Gamma^2 \tag{5.9}$$

$$\neg H_i\bot \in \Delta^3, \tag{5.10}$$

$$(\forall\sigma \in cl(\varphi))(K_i\sigma \in \Gamma \to \sigma \in \Delta). \tag{5.11}$$

We need to show that $\Gamma\mathcal{H}_i^\varphi\Delta$ holds too. Assume $H_i\theta \in \Gamma$ for some $\theta \in cl(\varphi)$. We need to show $\theta \in \Delta$.

1. Assume $H_i\theta \in cl_3(\varphi) \subset cl(\varphi)$.

– $\theta = \neg H_i\bot$. Assumption (5.10) immediately implies $\theta \in \Delta$.

– $K_i\theta \in cl_1(\varphi)$. Using axiom KH, more precisely $\vdash_{\mathscr{KHC}} H_i\theta \leftrightarrow (\neg H_i\bot \to K_i(\neg H_i\bot \to \theta))$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \neg H_i\bot \to K_i(\neg H_i\bot \to \theta)$ by Definition 5.49. Using the Deduction theorem 5.50 and assumption (5.9), we further obtain $\Gamma \vdash_{\mathscr{KHC}} K_i(\neg H_i\bot \to \theta)$. Lemma 5.56 (1) now implies that

[2]Otherwise, $H_i\bot \in \Gamma$ (according to Lemma 5.56 (2)) which implies that it must be $\mathcal{H}_i^\varphi(\Gamma) = \varnothing$ since no φ-consistent set can contain $\bot$.

[3]Otherwise, $H_i\bot \in \Delta$ (according to Lemma 5.56 (2)) which implies that it must be $\mathcal{H}_i^\varphi(\Delta) = \varnothing$ since no φ-consistent set can contain $\bot$.

$K_i(\neg H_i\bot \to \theta) \in \Gamma$ (note that $K_i(\neg H_i\bot \to \theta) \in cl(\varphi)$). According to assumption (5.11), this now implies that $\neg H_i\bot \to \theta \in \Delta$ holds. Assumption (5.10) and Definition 5.49 now imply that $\Delta \vdash_{\mathscr{KHC}} \theta$ holds. Finally, according to Lemma 5.56 (1), $\theta \in \Delta$ holds.

- $\theta = \neg H_i\bot \to \theta'$ for $K_i\theta' \in cl_1(\varphi)$. First of all, assumption $H_i(\neg H_i\bot \to \theta') = H_i\theta \in \Gamma$ implies that $H_i\theta' \in \Gamma$ also holds since $H_i\neg H_i\bot \in \Gamma$. Using axiom KH, more precisely $\vdash_{\mathscr{KHC}} H_i\theta' \leftrightarrow (\neg H_i\bot \to K_i(\neg H_i\bot \to \theta'))$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \neg H_i\bot \to K_i(\neg H_i\bot \to \theta')$ by Definition 5.49. Using the Deduction theorem 5.50 and assumption (5.9), we further obtain $\Gamma \vdash_{\mathscr{KHC}} K_i(\neg H_i\bot \to \theta')$. Lemma 5.56 (1) implies that $K_i(\neg H_i\bot \to \theta') \in \Gamma$. According to assumption (5.11), $\neg H_i\bot \to \theta' = \theta \in \Delta$ hence holds.

2. Assume $H_i\theta \in cl(\varphi) \setminus cl_3(\varphi)$.
   - $\theta = K_i\theta'$ for $K_i\theta' \in cl_3(\varphi)$. Using Proposition 5.20, more precisely $\vdash_{\mathscr{KHC}} \neg H_i\bot \to (H_i\theta \to \theta)$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \theta$, that is, $\Gamma \vdash_{\mathscr{KHC}} K_i\theta'$ by Definition 5.49. Lemma 5.56 (1) now implies that $K_i\theta' \in \Gamma$. Using axiom 4^K, more precisely $\vdash_{\mathscr{KHC}} K_i\theta' \to K_iK_i\theta'$, we further obtain $\Gamma \vdash_{\mathscr{KHC}} K_iK_i\theta'$ according to Definition 5.49. Lemma 5.56 (1) now implies that $K_iK_i\theta' \in \Gamma$ (note that $K_iK_i\theta' \in cl(\varphi)$). Finally, assumption (5.11) implies that $K_i\theta' = \theta \in \Delta$ indeed holds.
   - $\theta = \neg K_i\theta'$ for $\neg K_i\theta' \in cl_3(\varphi)$. Using $\vdash_{\mathscr{KHC}} \neg H_i\bot \to (H_i\theta \to \theta)$, we obtain $\Gamma \vdash_{\mathscr{KHC}} \theta$, that is, $\Gamma \vdash_{\mathscr{KHC}} \neg K_i\theta'$ by Definition 5.49. Lemma 5.56 (1) implies that $\neg K_i\theta' \in \Gamma$. Using axiom 5^K, more precisely $\vdash_{\mathscr{KHC}} \neg K_i\theta' \to K_i\neg K_i\theta'$, we further obtain $\Gamma \vdash_{\mathscr{KHC}} K_i\neg K_i\theta'$ according to Definition 5.49. Lemma 5.56 (1) now implies that $K_i\neg K_i\theta' \in \Gamma$ (note that $K_i\neg K_i\theta' \in cl(\varphi)$). Finally, assumption (5.11) implies that $\neg K_i\theta' = \theta \in \Delta$ indeed holds. □

For the purpose of proving the following lemmas we introduce some notations:

- $\underline{\Gamma} := \bigwedge_{\xi\in\Gamma} \xi$, for any finite set of formulas $\Gamma \subset \mathcal{L}^C_{KH}$;
- $\sim\xi := \begin{cases} \tau, & \text{if } \xi = \neg\tau \\ \neg\xi, & \text{otherwise} \end{cases}$, for any $\xi \in \mathcal{L}^C_{KH}$.

It is easy to prove:

Lemma 5.60. *For any $\varphi, \xi \in \mathcal{L}^C_{KH}$:*

1. $\vdash_{\mathscr{KHC}} \sim\xi \leftrightarrow \neg\xi$;
2. *If $\xi \in cl(\varphi)$, then $\sim\xi \in cl(\varphi)$.*

Lemma 5.61. *For any $\varphi \in \mathcal{L}^C_{KH}$:*

1. $\vdash_{\mathscr{KHC}} \bigvee_{\Gamma \in W^\varphi} \underline{\Gamma}$;
2. $\vdash_{\mathscr{KHC}} \neg\underline{\Gamma} \vee \neg\underline{\Delta}$ *for any $\Gamma, \Delta \in W^\varphi$ such that $\Gamma \neq \Delta$.*

Proof. 1. Assume the opposite towards a contradiction. By distributing $\vee$ through $\underline{\Gamma} = \bigwedge_{\xi \in \Gamma} \xi$, we get an equivalent conjuction of disjunctions, wherein each disjunction contains exactly one formula from each $\Gamma \in W^\varphi$. By assumption, that conjuction is not derivable in $\mathscr{KHC}$. Therefore, at least one of its conjuncts is not derivable in $\mathscr{KHC}$, that is, for some $\sigma_\Gamma \in \Gamma$ for each $\Gamma \in W^\varphi$, we have

$$\nvdash_{\mathscr{KHC}} \bigvee_{\Gamma \in W^\varphi} \sigma_\Gamma.$$

From this, by propositional reasoning, we obtain $\nvdash_{\mathscr{KHC}} \bigwedge_{\Gamma \in W^\varphi} {\sim}\sigma_\Gamma \to \bot$ which implies that $\{{\sim}\sigma_\Gamma \mid \Gamma \in W^\varphi\}$ is φ-consistent with respect to $\mathscr{KHC}$ (according to the previous lemma, ${\sim}\sigma_\Gamma \in cl(\varphi)$ for any $\Gamma \in W^\varphi$). According to the Lindenbaum lemma 5.57, $\{{\sim}\sigma_\Gamma \mid \Gamma \in W^\varphi\} \subseteq \Delta$ for some $\Delta \in W^\varphi$. Thus, ${\sim}\sigma_\Delta \in \Delta$, in particular. However, this contradicts the φ-consistency of Δ since $\sigma_\Delta \in \Delta$.

2. Assume $\Gamma \neq \Delta$ for some $\Gamma, \Delta \in W^\varphi$. Given that $\Gamma \cup \Delta \supset \Gamma$ (as well as $\Gamma \cup \Delta \supset \Delta$), the set $\Gamma \cup \Delta$ cannot be φ-consistent with respect to $\mathscr{KHC}$ by the maximality of Γ (as well as by the maximality of Δ). Since we do have $\Gamma \cup \Delta \subset cl(\varphi)$, it must be

$$\Gamma \cup \Delta \vdash_{\mathscr{KHC}} \bot.$$

From this, using the Deduction theorem 5.50, we obtain $\vdash_{\mathscr{KHC}} \underline{\Gamma} \wedge \underline{\Delta} \to \bot$. Finally, by propositional reasoning, we get $\vdash_{\mathscr{KHC}} \neg\underline{\Gamma} \vee \neg\underline{\Delta}$. □

Corollary 5.62. *For any $\varphi \in \mathcal{L}^C_{KH}$ and any $W \subseteq W^\varphi$,*

$$\vdash_{\mathscr{KHC}} \bigvee_{\Gamma \in W} \underline{\Gamma} \leftrightarrow \bigwedge_{\Delta \in W^\varphi \setminus W} \neg\underline{\Delta}.$$

Proof. Let $\varphi \in \mathcal{L}^C_{KH}$ and $W \subseteq W^\varphi$. From Lemma 5.61 (1), we immediately get $\vdash_{\mathscr{KHC}} \bigwedge_{\Delta \in W^\varphi \setminus W} \neg\underline{\Delta} \to \bigvee_{\Gamma \in W} \underline{\Gamma}$ by propositional reasoning. Using Lemma 5.61 (2), it is easy to show that, for any $\Gamma \in W$, we have

$$\vdash_{\mathscr{KHC}} \Gamma \to \bigwedge_{\Delta \in W^\varphi \setminus \{\Gamma\}} \neg\underline{\Delta}.$$

Therefore, for any $\Gamma \in W$, we also have $\vdash_{\mathscr{KHC}} \Gamma \to \bigwedge_{\Delta \in W^\varphi \setminus W} \neg\underline{\Delta}$. Finally, using propositional reasoning, we obtain $\vdash_{\mathscr{KHC}} \bigvee_{\Gamma \in W} \underline{\Gamma} \to \bigwedge_{\Delta \in W^\varphi \setminus W} \neg\underline{\Delta}$. □

Lemma 5.63 (Truth lemma)**.** *For any $\varphi \in \mathcal{L}^C_{KH}$, any maximal φ-consistent set Γ with respect to $\mathscr{KHC}$, and any $\psi \in cl(\varphi)$,*

$$\psi \in \Gamma \quad \Longleftrightarrow \quad M^\varphi, \Gamma \models \psi.$$

Proof. We proceed by induction on the structure of ψ.

Base case: If $\psi = p$, then the statement of the theorem follows from Definition 5.58 (4).

Induction step:

1. If ψ is of the form $\neg\theta$, then $\psi \in \Gamma$ is equivalent to $\theta \notin \Gamma$ by Lemma 5.56 (2). By the induction hypothesis, this is further equivalent to $M^\varphi, \Gamma \not\models \theta$, i.e., $M^\varphi, \Gamma \models \psi$.

2. If ψ is of the form $\theta_1 \wedge \theta_2$, then $\theta_1 \wedge \theta_2 \in \Gamma$ is equivalent to $\theta_1 \in \Gamma$ and $\theta_2 \in \Gamma$ by Lemma 5.56 (3). By the induction hypothesis, this is further equivalent to $M^\varphi, \Gamma \models \theta_1$ and $M^\varphi, \Gamma \models \theta_2$, i.e., $M^\varphi, \Gamma \models \theta_1 \wedge \theta_2$. In other words, $M^\varphi, \Gamma \models \psi$ holds.

3. Assume that ψ is of the form $K_i\theta$.

   ($\Longrightarrow$): Assume that $K_i\theta \in \Gamma$. Take an arbitrary set Δ such that $\Gamma \mathcal{K}_i^\varphi \Delta$. By Definition 5.58 (2), $\theta \in \Delta$ holds. Using the induction hypothesis, we now obtain $M^\varphi, \Delta \models \theta$. Therefore, $M^\varphi, \Gamma \models K_i\theta$ indeed holds.

   ($\Longleftarrow$): Assume that $M^\varphi, \Gamma \models K_i\theta$. We first show that the set

   $$\{\sigma \mid K_i\sigma \in \Gamma\} \cup \{\neg\theta\}$$

   is not φ-consistent with respect to $\mathscr{KHC}$ (note that $\neg\theta \in cl(\varphi)$). Assume the opposite towards a contradiction. Then, according to the Lindenbaum lemma 5.57 there exists a maximal φ-consistent set Δ with respect to $\mathscr{KHC}$ such that $\Delta \supset \{\sigma \mid K_i\sigma \in \Gamma\} \cup \{\neg\theta\}$. According to Definition 5.58 (2), we now obtain that $\Gamma \mathcal{K}_i^\varphi \Delta$ holds. By assumption, $M^\varphi, \Delta \models \theta$ follows. By applying the induction hypothesis, we now obtain $\theta \in \Delta$, contradicting the φ-consistency of Δ. Thus, $\vdash_{\mathscr{KHC}} \bigwedge_{K_i\sigma \in \Gamma} \sigma \wedge \neg\theta \rightarrow \bot$. Using propositional reasoning, we obtain $\vdash_{\mathscr{KHC}} \bigwedge_{K_i\sigma \in \Gamma} \sigma \rightarrow \theta$. Using $\mathscr{KHC}$ reasoning, we now obtain $\vdash_{\mathscr{KHC}} \bigwedge_{K_i\sigma \in \Gamma} K_i\sigma \rightarrow K_i\theta$. Thus, $\Gamma \vdash_{\mathscr{KHC}} K_i\theta$, by Definition 5.49. By applying Lemma 5.56 (1), we finally obtain $K_i\theta \in \Gamma$.

4. Assume that ψ is of the form $H_i\theta$.

   ($\Longrightarrow$): Assume that $H_i\theta \in \Gamma$. Take an arbitrary set Δ such that $\Gamma \mathcal{H}_i^\varphi \Delta$. By Definition 5.58 (3), $\theta \in \Delta$ holds. Using the induction hypothesis, we now obtain $M^\varphi, \Delta \models \theta$. Therefore, $M^\varphi, \Gamma \models H_i\theta$ indeed holds.

($\Longleftarrow$): Assume that $M^\varphi, \Gamma \models H_i\theta$. We first show that the set

$$\{\sigma \mid H_i\sigma \in \Gamma\} \cup \{\neg\theta\}$$

is not φ-consistent with respect to $\mathscr{KHC}$ (note that $\neg\theta \in cl(\varphi)$). Assume the opposite towards a contradiction. Then, according to the Lindenbaum lemma 5.57 there exists a maximal φ-consistent set Δ with respect to $\mathscr{KHC}$ such that $\Delta \supset \{\sigma \mid H_i\sigma \in \Gamma\} \cup \{\neg\theta\}$. According to Definition 5.58 (3), we now obtain that $\Gamma \mathcal{H}_i^\varphi \Delta$ holds. By assumption, $M^\varphi, \Delta \models \theta$ follows. By applying the induction hypothesis, we now obtain $\theta \in \Delta$, contradicting the φ-consistency of Δ. Thus, $\vdash_{\mathscr{KHC}} \bigwedge_{H_i\sigma\in\Gamma} \sigma \wedge \neg\theta \rightarrow \bot$. Using propositional reasoning, we obtain $\vdash_{\mathscr{KHC}} \bigwedge_{H_i\sigma\in\Gamma} \sigma \rightarrow \theta$. Using $\mathscr{KHC}$ reasoning, we now obtain $\vdash_{\mathscr{KHC}} \bigwedge_{H_i\sigma\in\Gamma} H_i\sigma \rightarrow H_i\theta$. Thus, $\Gamma \vdash_{\mathscr{KHC}} H_i\theta$, by Definition 5.49. By applying Lemma 5.56 (1), we finally obtain $H_i\theta \in \Gamma$.

5. Assume that ψ is of the form $C_G^K\theta$.

($\Longrightarrow$): Assume that $C_G^K\theta \in \Gamma$. In order to show $M^\varphi, \Gamma \models C_G^K\theta$, we need to show that $M^\varphi, \Delta_n \models \theta$ holds for any sequence $\Gamma\mathcal{K}_1^\varphi\Delta_1 \ldots \mathcal{K}_n^\varphi\Delta_n$.

Let $\Delta_0 := \Gamma$. We first show by induction on $n \geq 0$ that $C_G^K\theta \in \Delta_n$:

- $C_G^K\theta \in \Delta_0$ holds by assumption.
- Assume that $C_G^K\theta \in \Delta_{n-1}$. Using Mix^K, we obtain $\Delta_{n-1} \vdash_{\mathscr{KHC}} E_G^H(\theta \wedge C_G^K\theta)$. By applying Lemma 5.56 (1), we obtain $E_G^K(\theta \wedge C_G^K\theta) \in \Delta_{n-1}$ (note that $E_G^K(\theta \wedge C_G^K\theta) \in cl(\varphi)$ since $C_G^K\theta \in cl(\varphi)$). By applying Lemma 5.56 (3), we further obtain $K_n(\theta \wedge C_G^K\theta) \in \Delta_{n-1}$, in particular. Thus, $\theta \wedge C_G^K\theta \in \Delta_n$ according to Definition 5.58 (3). Finally, $C_G^K\theta \in \Delta_n$ follows according to Lemma 5.56 (3).

Now that we have $C_G^K\theta \in \Delta_n$, $\Delta_n \vdash_{\mathscr{KHC}} E_G^K(\theta \wedge C_G^K\theta)$ follows according to Mix^K. By applying Lemma 5.56 (1), we now obtain $E_G^K(\theta \wedge C_G^K\theta) \in \Delta_n$ (note that $E_G^K(\theta \wedge C_G^K\theta) \in cl(\varphi)$ since $C_G^K\theta \in cl(\varphi)$). From this, using $\vdash_{\mathscr{KHC}} E_G^K\theta \rightarrow \theta$, we obtain that $\Delta_n \vdash_{\mathscr{KHC}} \theta$ holds too. By applying Lemma 5.56 (1), we further obtain $\theta \in \Delta_n$ (note that $\theta \in cl(\varphi)$). Finally, by applying the induction hypothesis, we obtain $M^\varphi, \Delta_n \models \theta$. Therefore, $M^\varphi, \Gamma \models C_G^K\theta$.

($\Longleftarrow$): Assume that $M^\varphi, \Gamma \models C_G^K\theta$. Let $W := \{\Delta \in W^\varphi \mid M^\varphi, \Delta \models C_G^K\theta\}$. We first show the following two things:

a) $\vdash_{\mathscr{KHC}} \underline{\Delta} \rightarrow K_i\theta$ for all $i \in G$ and $\Delta \in W$ and
b) $\vdash_{\mathscr{KHC}} \underline{\Delta} \rightarrow K_i\neg\underline{\Sigma}$ for all $i \in G$, $\Delta \in W$, and $\Sigma \in W^\varphi \backslash W$.

Proof of (a): Let $i \in G$ and $\Delta \in W$. Therefore, $M^\varphi, \Delta \models C_K^H\theta$. This implies that, in particular, $M^\varphi, \Delta \models K_i\theta$ holds too. As in case 3, it can be shown that $\vdash_{\mathscr{KHC}} \bigwedge_{K_i\sigma\in\Delta} K_i\sigma \to K_i\theta$ holds because the set $\{\sigma \mid K_i\sigma \in \Delta\} \cup \{\neg\theta\}$ is not φ-consistent with respect to $\mathscr{KHC}$. Thus, $\vdash_{\mathscr{KHC}} \underline{\Delta} \to K_i\theta$ indeed holds.
Proof of (b): Let $i \in G$, $\Delta \in W$ and $\Sigma \in W^\varphi \backslash W$. Therefore, $M^\varphi, \Delta \models C_G^K\theta$ and $M^\varphi, \Sigma \not\models C_G^K\theta$. From this we get that, in particular, $\Delta \mathcal{K}_i^\varphi \Sigma$ does not hold. This means that $\tau \notin \Sigma$ for some $K_i\tau \in \Delta$, where $\tau \in cl(\varphi)$.

Therefore:

1. $\neg\tau \in \Sigma$ for some $K_i\tau \in \Delta$ — Lemma 5.56 (2)
2. $\vdash_{\mathscr{KHC}} \underline{\Sigma} \to \neg\tau$ for some $K_i\tau \in \Delta$ — by prop. reasoning from 1.
3. $\vdash_{\mathscr{KHC}} \tau \to \neg\underline{\Sigma}$ for some $K_i\tau \in \Delta$ — by prop. reasoning from 2.
4. $\vdash_{\mathscr{KHC}} K_i(\tau \to \neg\underline{\Sigma})$ for some $K_i\tau \in \Delta$ — by Nec^K from 3.
5. $\vdash_{\mathscr{KHC}} K_i\tau \to K_i\neg\underline{\Sigma}$ for some $K_i\tau \in \Delta$ — by $\mathscr{KHC}$ reasoning from 4.
6. $\vdash_{\mathscr{KHC}} \underline{\Delta} \to K_i\neg\underline{\Sigma}$ — by prop. reasoning from 5.

By combining (a) and (b), we obtain that, for all $i \in G$ and $\Delta \in W$, the following holds:

$$\vdash_{\mathscr{KHC}} \underline{\Delta} \to K_i(\theta \wedge \bigwedge_{\Sigma\in W^\varphi\backslash W} \neg\underline{\Sigma}). \tag{5.12}$$

Using Corollary 5.62, we obtain $\vdash_{\mathscr{KHC}} \underline{\Delta} \to K_i(\theta \wedge \bigvee_{\Delta\in W} \underline{\Delta})$. Since (5.12) holds for all $i \in G$ and $\Delta \in W$, we get $\vdash_{\mathscr{KHC}} \bigvee_{\Delta\in W} \underline{\Delta} \to E_G^K(\theta \wedge \bigvee_{\Delta\in W} \underline{\Delta})$. The application of Ind^K now results in $\vdash_{\mathscr{KHC}} \bigvee_{\Delta\in W} \underline{\Delta} \to C_G^K\theta$. Thus, $\vdash_{\mathscr{KHC}} \underline{\Gamma} \to C_G^K\theta$ holds, by propositional reasoning, because $\Gamma \in W$ by assumption. This means that $\Gamma \vdash_{\mathscr{KHC}} C_G^K\theta$. Finally, by applying Lemma 5.56 (1), we obtain $C_G^K\theta \in \Gamma$.

6. Assume that ψ is of the form $C_G^H\theta$.

($\Longrightarrow$): Assume that $C_G^H\theta \in \Gamma$. In order to show $M^\varphi, \Gamma \models C_G^H\theta$, we need to show that $M^\varphi, \Delta_n \models \theta$ holds for any sequence $\Gamma\mathcal{H}_1^\varphi\Delta_1 \ldots \mathcal{H}_n^\varphi\Delta_n$.

Let $\Delta_0 := \Gamma$. We first show by induction on $n \geq 0$ that $C_G^H\theta \in \Delta_n$:

- $C_G^H\theta \in \Delta_0$ holds by assumption.
- Assume that $C_G^H\theta \in \Delta_{n-1}$. Using Mix^H, we obtain $\Delta_{n-1} \vdash_{\mathscr{KHC}} E_G^H(\theta \wedge C_G^H\theta)$. By applying Lemma 5.56 (1), we obtain $E_G^H(\theta \wedge C_G^H\theta) \in \Delta_{n-1}$ (note that $E_G^H(\theta \wedge C_G^H\theta) \in cl(\varphi)$ since $C_G^H\theta \in cl(\varphi)$). By applying Lemma 5.56 (3), we further obtain $H_n(\theta \wedge C_G^H\theta) \in \Delta_{n-1}$, in particular. Thus, $\theta \wedge C_G^H\theta \in \Delta_n$ according to Definition 5.58 (3). Finally, $C_G^H\theta \in \Delta_n$ follows according to Lemma 5.56 (3).

Now that we have $C_G^H\theta \in \Delta_n$, $\Delta_n \vdash_{\mathscr{KHC}} E_G^H(\theta \wedge C_G^H\theta)$ follows according to Mix^H.

($\Longleftarrow$): Assume that $M^\varphi, \Gamma \models C_G^H\theta$. Let

$$W := \{\Delta \in W^\varphi \mid M^\varphi, \Delta \models C_G^H\theta\}.$$

We first show the following two things:

a) $\vdash_{\mathscr{KHC}} \underline{\Delta} \to H_i\theta$ for all $i \in G$ and $\Delta \in W$ and
b) $\vdash_{\mathscr{KHC}} \underline{\Delta} \to H_i\neg\underline{\Sigma}$ for all $i \in G$, $\Delta \in W$, and $\Sigma \in W^\varphi \setminus W$.

Proof of (a): Let $i \in G$ and $\Delta \in W$. Therefore, $M^\varphi, \Delta \models C_G^H\theta$. This implies that, in particular, $M^\varphi, \Delta \models H_i\theta$ holds too. As in the previous case, it can be shown that $\vdash_{\mathscr{KHC}} \bigwedge_{H_i\sigma\in\Delta} H_i\sigma \to H_i\theta$ holds because the set $\{\sigma \mid H_i\sigma \in \Delta\} \cup \{\neg\theta\}$ is not φ-consistent with respect to $\mathscr{KHC}$. Thus, $\vdash_{\mathscr{KHC}} \underline{\Delta} \to H_i\theta$ indeed holds.
Proof of (b): Let $i \in G$, $\Delta \in W$ and $\Sigma \in W^\varphi \setminus W$. Therefore, $M^\varphi, \Delta \models C_G^H\theta$ and $M^\varphi, \Sigma \not\models C_G^H\theta$. From this we get that, in particular, $\Delta \mathcal{H}_i^\varphi \Sigma$ does not hold. This means that $\tau \notin \Sigma$ for some $H_i\tau \in \Delta$, where $\tau \in cl(\varphi)$.

Therefore:

1. $\neg\tau \in \Sigma$ for some $H_i\tau \in \Delta$	Lemma 5.56 (2)
2. $\vdash_{\mathscr{KHC}} \underline{\Sigma} \to \neg\tau$ for some $H_i\tau \in \Delta$	by prop. reasoning from 1.
3. $\vdash_{\mathscr{KHC}} \tau \to \neg\underline{\Sigma}$ for some $H_i\tau \in \Delta$	by prop. reasoning from 2.
4. $\vdash_{\mathscr{KHC}} H_i(\tau \to \neg\underline{\Sigma})$ for some $H_i\tau \in \Delta$	by Nec^H from 3.
5. $\vdash_{\mathscr{KHC}} H_i\tau \to H_i\neg\underline{\Sigma}$ for some $H_i\tau \in \Delta$	by $\mathscr{KHC}$ reasoning from 4.
6. $\vdash_{\mathscr{KHC}} \underline{\Delta} \to H_i\neg\underline{\Sigma}$	by prop. reasoning from 5.

By combining (a) and (b), we obtain that, for all $i \in G$ and $\Delta \in W$, the following holds:

$$\vdash_{\mathscr{KHC}} \underline{\Delta} \to H_i(\theta \wedge \bigwedge_{\Sigma\in W^\varphi\setminus W} \neg\underline{\Sigma}). \tag{5.13}$$

Using Corollary 5.62, we obtain $\vdash_{\mathscr{KHC}} \underline{\Delta} \to H_i(\theta \wedge \bigvee_{\Delta\in W} \underline{\Delta})$. Since (5.13) holds for all $i \in G$ and $\Delta \in W$, we get $\vdash_{\mathscr{KHC}} \bigvee_{\Delta\in W} \underline{\Delta} \to E_G^H(\theta \wedge \bigvee_{\Delta\in W} \underline{\Delta})$. The application of Ind^H now results in $\vdash_{\mathscr{KHC}} \bigvee_{\Delta\in W} \underline{\Delta} \to C_G^H\theta$. Thus, $\vdash_{\mathscr{KHC}} \underline{\Gamma} \to C_G^H\theta$ holds, by propositional reasoning, because $\Gamma \in W$ by assumption. This means that $\Gamma \vdash_{\mathscr{KHC}} C_G^H\theta$. Finally, by applying Lemma 5.56 (1), we obtain $C_G^H\theta \in \Gamma$. □

Finally, we obtain:

Theorem 5.64 (Soundness and completeness)**.** *The axiom system $\mathscr{KHC}$ is sound and complete with respect to the* KH *class of models.*

Proof. Soundness: For an arbitrary $\varphi \in \mathcal{L}^C_{KH}$, it follows by induction on the length of the derivation of φ that if φ is $\mathscr{KHC}$-provable, then it is also valid with respect to class KH.

Completeness: We prove the contrapositive. Let $\nvdash_{\mathscr{KHC}} \varphi$. It follows that $\{\neg\varphi\}$ is φ-consistent with respect to $\mathscr{KHC}$ and, as such, it is contained in some maximal φ-consistent set with respect to $\mathscr{KHC}$, i.e., $\neg\varphi \in \Gamma$ for some $\Gamma \in W^\varphi$, according to the Lindenbaum lemma 5.57. By applying the previous theorem, we obtain $M^\varphi, \Gamma \models \neg\varphi$, i.e., $M^\varphi, \Gamma \not\models \varphi$. Therefore, $\mathsf{KH} \not\models \varphi$ follows, since $M^\varphi \in \mathsf{KH}$ (as shown in Lemma 5.59). □

Corollary 5.65. *The axiom system $\mathscr{KHC} + Byz_f$ is sound and complete with respect to the KH^{n-f} class of models.*

5.5 Finite model property and decidability

Definition 5.66. *A set of formulas $\Gamma \subseteq \mathcal{L}_H$ is subformula-closed if for all formulas φ and ψ such that $\varphi \in \Gamma$ and ψ is a subformula of φ, it holds that $\psi \in \Gamma$.*

Definition 5.67 (Filtration)**.** *Let $M = (W, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{KB4}_n$ and $\Gamma \subseteq \mathcal{L}_H$ be a subformula-closed set of formulas. Let $\sim_\Gamma$ be the relation on the worlds of M defined by:*

$$w \sim_\Gamma v$$
$$\textit{iff}$$
$$(\forall\varphi \in \Gamma)(M, w \models \varphi \iff M, v \models \varphi)$$

Note that $\sim_\Gamma$ is an equivalence relation. We denote the equivalence class of a world w with respect to Γ by $|w|_\Gamma$. Let

$$W_\Gamma := \{|w|_\Gamma \mid w \in W\}.$$

Suppose M^f_Γ is any model $(W^f, \pi^f, \mathcal{H}^f_1, \ldots, \mathcal{H}^f_n)$ such that:

1. *$W^f = W_\Gamma$.*
2. *If $(w, v) \in \mathcal{H}_i$ then $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$.*
3. *If $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$ then for all $H_i\varphi \in \Gamma$:*
   a) *if $M, w \models H_i\varphi$ then $M, v \models H_i\varphi \wedge \varphi$,*
   b) *if $M, v \models H_i\varphi$ then $M, w \models H_i\varphi \wedge \varphi$,*
4. *$\pi^f(p) = \{|w|_\Gamma \mid M, w \models p\}$, for all atomic propositions $p \in \Gamma$.*

Then M^f_Γ is called a filtration of M through Γ.

Lemma 5.68 (Filtration lemma)**.** *Let M^f_Γ be a filtration of $M \in \mathsf{KB4}_n$ through a subformula-closed set $\Gamma \subseteq \mathcal{L}_H$. Then, for all formulas $\varphi \in \Gamma$ and all worlds $w \in M$, we have*

$$M, w \models \varphi \quad \textit{iff} \quad M^f_\Gamma, |w|_\Gamma \models \varphi.$$

Proof. We proceed by induction on the structure of φ.

Base case: If φ is $p \in \mathsf{Prop}$, then

$$M, w \models p \quad \textit{iff} \quad M^f_\Gamma, |w|_\Gamma \models p$$

follows immediately according to Definition 5.67 (4).

Induction step:

1. If φ is of the form $\neg\psi$, then $M, w \models \neg\psi$ iff $M, w \not\models \psi$ iff $M^f_\Gamma, |w|_\Gamma \not\models \psi$ by the induction hypothesis ($\psi \in \Gamma$ since $\varphi \in \Gamma$ and Γ is subformula-closed) iff $M^f_\Gamma, |w|_\Gamma \models \neg\psi$.

2. If φ is of the form $\psi_1 \wedge \psi_2$, then $M, w \models \psi_1 \wedge \psi_2$ iff $M, w \models \psi_1$ and $M, w \models \psi_2$ iff $M^f_\Gamma, |w|_\Gamma \models \psi_1$ by the induction hypothesis ($\psi_1 \in \Gamma$ since $\varphi \in \Gamma$ and Γ is subformula-closed) and $M^f_\Gamma, |w|_\Gamma \models \psi_2$ by the induction hypothesis ($\psi_2 \in \Gamma$ since $\varphi \in \Gamma$ and Γ is subformula-closed) iff $M^f_\Gamma, |w|_\Gamma \models \psi_1 \wedge \psi_2$.

3. Assume that φ is of the form $H_i\psi$.

   ($\Longrightarrow$): Assume $M, w \models H_i\psi$. Take an arbitrary $|v|_\Gamma \in W^f$ such that $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$. According to Definition 5.67 (3), this means that for all $H_i\varphi \in \Gamma$, if $M, w \models H_i\varphi$ then $M, v \models H_i\varphi \wedge \varphi$ (and if $M, v \models H_i\varphi$ then $M, w \models H_i\varphi \wedge \varphi$). Since $M, w \models H_i\psi$, it follows that $M, v \models \psi$, in particular. From this, by the induction hypothesis ($\psi \in \Gamma$ since $\varphi \in \Gamma$ and Γ is subformula-closed), we get $M^f_\Gamma, |v|_\Gamma \models \psi$. Therefore, $M^f_\Gamma, |w|_\Gamma \models H_i\psi$ indeed holds.

   ($\Longleftarrow$): Assume $M^f_\Gamma, |w|_\Gamma \models H_i\psi$. This means that for all $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$, it holds that $M^f_\Gamma, |v|_\Gamma \models \psi$. From this, by the induction hypothesis ($\psi \in \Gamma$ since $\varphi \in \Gamma$ and Γ is subformula-closed), we get $M, v \models \psi$ for all $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$. Take an arbitrary $v \in W$ such that $(w, v) \in \mathcal{H}_i$. Therefore, according to Definition 5.67 (2), $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$, which further implies $M, v \models \psi$. Consequently, $M, w \models H_i\psi$ follows. □

Lemma 5.69. $M^f_\Gamma \in \mathsf{KB4}_n$.

Proof. We need to show that the relations $\mathcal{H}^f_i$ are symmetric and transitive.

- Assume $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}^f_i$. Then $(|v|_\Gamma, |w|_\Gamma) \in \mathcal{H}^f_i$ follows immediately from Definition 5.67 (3).

- Assume $(|w|_\Gamma, |v|_\Gamma) \in \mathcal{H}_i^f$ and $(|v|_\Gamma, |u|_\Gamma) \in \mathcal{H}_i^f$. Take an arbitrary $H_i\varphi \in \Gamma$ and assume $M, w \models H_i\varphi$. Then, $M, v \models H_i\varphi$ since $M, v \models H_i\varphi \wedge \varphi$. Consequently, $M, u \models H_i\varphi \wedge \varphi$. Conversely, assume $M, u \models H_i\varphi$. Then, $M, v \models H_i\varphi$ since $M, v \models H_i\varphi \wedge \varphi$. Consequently, $M, w \models H_i\varphi \wedge \varphi$. It follows that $(|w|_\Gamma, |u|_\Gamma) \in \mathcal{H}_i^f$ indeed holds. □

Lemma 5.70. *Let $\Gamma \subseteq \mathcal{L}_H$ be a finite subformula-closed set of formulas. For any model $M \in \mathsf{KB4}_n$, if M_Γ^f is a filtration of M through Γ, then M_Γ^f is finite.*

Proof. Take an arbitrary $M \in \mathsf{KB4}_n$ and let $M_\Gamma^f = (W^f, \pi^f, \mathcal{H}_1^f, \ldots, \mathcal{H}_n^f)$ be a filtration of M through Γ. Note that $W^f = W_\Gamma$. Let $g : W_\Gamma \to \mathcal{P}(\Gamma)$ be the function defined by

$$g(|w|_\Gamma) := \{\varphi \in \Gamma \mid M, w \models \varphi\}.$$

It follows from the definition of $\sim_\Gamma$ that g is well-defined and injective. Thus, the size of W_Γ is at most 2^n, where n is the size of Γ. □

Theorem 5.71. *The logic of $\mathscr{H}$ has the FMP.*

Proof. Take an arbitrary $\varphi \in \mathcal{L}_H$. If $\mathscr{H} \nvdash \varphi$, there exists a model $M \in \mathsf{KB4}_n$ of $\mathscr{H}$ such that $M \not\models \varphi$ (according to Theorem 5.4). Let $M_{Sub(\varphi)}^f$ be a filtration of M through $Sub(\varphi)$ (which is a finite set). By the Filtration lemma 5.68, $M_{Sub(\varphi)}^f \not\models \varphi$ also holds. Therefore, $M_{Sub(\varphi)}^f$ is a finite model of $\mathscr{H}$ (according to Lemma 5.69 and Lemma 5.70) that is a countermodel for φ. □

Therefore, according to Theorem 2.26, we obtain:

Corollary 5.72. *(Decidability) The logic of $\mathscr{H}$ is decidable.*

Theorem 5.73. *The logic of $\mathscr{KHC}$ has the FMP.*

Proof. If $\mathscr{KHC} \nvdash \varphi$, then the canonical φ-model M^φ is a finite countermodel for φ (as already seen in the proof of Theorem 5.64). □

Therefore, according to Theorem 2.26, we obtain:

Corollary 5.74 (Decidability)**.** *The logic of $\mathscr{KHC}$ is decidable.*

Definition 5.75. *An axiom system $\mathscr{L}_1$ in a language $\mathcal{L}_1$ is conservative over another axiom system $\mathscr{L}_2$ in a language $\mathcal{L}_2 \subseteq \mathcal{L}_1$, if:*

- *$\mathscr{L}_1 \vdash \varphi$ whenever $\mathscr{L}_2 \vdash \varphi$ for any $\varphi \in \mathcal{L}_1$ and*
- *$\mathscr{L}_2 \vdash \varphi$ whenever $\mathscr{L}_1 \vdash \varphi$ for any $\varphi \in \mathcal{L}_2$.*

Lemma 5.76. *$\mathscr{KHC}$ is conservative over $\mathscr{KH}$.*

Proof. If $\mathscr{KH} \vdash \varphi$ for $\varphi \in \mathcal{L}^C_{KH}$, then $\mathscr{KHC} \vdash \varphi$ follows immediately according to Definition 5.38. If $\mathscr{KH} \nvdash \varphi$ for $\varphi \in \mathcal{L}_{KH}$, then $\mathsf{KH} \not\models \varphi$ by Theorem 5.18. Therefore, $\mathscr{KHC} \nvdash \varphi$ by Theorem 5.64. □

Corollary 5.77. *The logic of $\mathscr{KH}$ has the FMP.*

Therefore, according to Theorem 2.26, we obtain:

Corollary 5.78 (Decidability)**.** *The logic of $\mathscr{KH}$ is decidable.*

5.6 Defining common eventual hope

Inspired by a related approach described in [HM90, FHMV95], we will show how to introduce *common eventual hope*, albeit without using a greatest fixpoint operator [Tar55].

As a first step, we extend the language $\mathcal{L}_H$ by adding the temporal modality *eventually* $\Diamond$ and denote the extended language with $\mathcal{L}_\Diamond$.

Given an interpreted system $\mathcal{I} = (R, \pi)$, we define

$$(\mathcal{I}, r, t) \models \Diamond\varphi \quad \textit{iff} \quad (\mathcal{I}, r, t') \models \varphi \quad \textit{for some} \quad t' \geq t.$$

We further define *mutual eventual hope* of φ (in the language $\mathcal{L}_\Diamond$)

$$E_G^{\Diamond H}\varphi := \bigwedge_{i \in G} \Diamond H_i \varphi,$$

where $\varnothing \neq G \subseteq \mathcal{A}$. Thus, we obtain

$$(\mathcal{I}, r, t) \models E_G^{\Diamond H}\varphi \quad \textit{iff} \quad (\mathcal{I}, r, t) \models \Diamond H_i\varphi \quad \textit{for all} \quad i \in G.$$

We also associate with a formula $\varphi \in \mathcal{L}_\Diamond$ its *intension* [FHMV95] in the following way:

$$\varphi^{\mathcal{I}} := \{(r, t) \mid (\mathcal{I}, r, t) \models \varphi\}.$$

As a next step, we use $\mathcal{L}^{CH}_\Diamond$ to denote the language obtained by extending $\mathcal{L}_\Diamond$ with a unary modal operator for common eventual hope $C_G^{\Diamond H}$, and we use $\mathcal{L}^{CH}_{\Diamond x}$ to denote the language obtained by extending $\mathcal{L}^{CH}_\Diamond$ with a single propositional variable x.

We call an occurrence of the propositional variable x in a formula φ *free*, if it is outside the scope of the operator $C_G^{\Diamond H}$. We say that a free occurrence of x in a formula φ is *positive (negative)* if it is in the scope of an even (odd) number of negation symbols.

Definition 5.79. *Let $A, B \subseteq R \times \mathbb{N}_0$ and let $\varnothing \neq G \subseteq \mathcal{A}$. We associate a function with the Boolean connectives and modal operators of the language $\mathcal{L}^{CH}_{\Diamond x}$ in the following way:*

1. $f_\neg(A) = (R \times \mathbb{N}_0) \setminus A$ *(the complement of A),*
2. $f_\wedge(A, B) = A \cap B$,
3. $f_{H_i}(A) = \{(r, t) \mid (r', t') \in A$ *whenever* $(r, t)\mathcal{H}_i(r', t')\}$,
4. $f_\Diamond(A) = \{(r, t) \mid (r, t') \in A$ *for some* $t' \geq t\}$,

Definition 5.80. *Let $f : R \times \mathbb{N}_0 \to R \times \mathbb{N}_0$. The function f is monotonically increasing (resp. decreasing) iff for all $A, B \subseteq R \times \mathbb{N}_0$, $A \subseteq B$ implies $f(A) \subseteq f(B)$ (resp. $f(A) \supseteq f(B)$).*

Definition 5.81. *Let $f : (R \times \mathbb{N}_0) \times (R \times \mathbb{N}_0) \to R \times \mathbb{N}_0$. The function f is monotonically increasing (resp. decreasing) iff for all $A, A', B, B' \subseteq R \times \mathbb{N}_0$, $A \subseteq A'$ and $B \subseteq B'$ imply $f(A, B) \subseteq f(A', B')$ (resp. $f(A, B) \supseteq f(A', B')$).*

Lemma 5.82. *The functions defined in Definition 5.79 are monotonically increasing, except the function associated with negation which is monotonically decreasing.*

Proof. Assume that $A \subseteq A' \subseteq R \times \mathbb{N}_0$ and $B \subseteq B' \subseteq R \times \mathbb{N}_0$ and $\varnothing \neq G \subseteq \mathcal{A}$. Then, we obtain the following:

1. $f_\neg(A) = \overline{A} \supseteq \overline{A'} = f_\neg(A')$,
2. $f_\wedge(A, B) = A \cap B \subseteq A' \cap B' = f_\wedge(A', B')$,
3. $f_{H_i}(A) = \{(r, t) \mid (r', t') \in A$ whenever $(r, t)\mathcal{H}_i(r', t')\} \subseteq \{(r, t) \mid (r', t') \in A'$ whenever $(r, t)\mathcal{H}_i(r', t')\} = f_{H_i}(A')$,
4. $f_\Diamond(A) = \{(r, t) \mid (r, t') \in A$ for some $t' \geq t\} \subseteq \{(r, t) \mid (r, t') \in A'$ for some $t' \geq t\} = f_\Diamond(A')$. □

Definition 5.83. *We define $f_\varphi(A)$ for every set $A \subseteq R \times \mathbb{N}_0$ by induction on the structure of $\varphi \in \mathcal{L}^{CH}_{\Diamond x}$ as follows:*

1. $f_p(A) = p^{\mathcal{I}}$, *where* $p \in \mathsf{Prop}$,
2. $f_x(A) = A$,
3. $f_{\neg\varphi}(A) = f_\neg(f_\varphi(A))$,
4. $f_{\varphi\wedge\psi}(A) = f_\wedge(f_\varphi(A), f_\psi(A))$,
5. $f_{H_i\varphi}(A) = f_{H_i}(f_\varphi(A))$,

6. $f_{\Diamond\varphi}(A) = f_{\Diamond}(f_{\varphi}(A))$,
7. $f_{C_G^{\Diamond H}\varphi}(A) = \bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E_G^{\Diamond H}(\varphi \wedge x)}(B)\}$.

Notice that the function $f_{C_G^{\Diamond H}\varphi}$ does not depend on the set A in any way, i.e., it is a constant function. As we will show in Lemma 5.88, functions associated with formulas that have no occurrences of x are constant as well.

Lemma 5.84. *If every free occurrence of x in $\varphi \in \mathcal{L}_{\Diamond x}^{CH}$ is positive (resp. negative), then the function f_φ is monotonically increasing (resp. monotonically decreasing).*

Proof. We proceed by induction on the structure of φ.

Base case: If φ is p, then there are no occurrences of x in φ, so, the statement of the lemma vacuously holds. If φ is x, we immediately obtain that f_x is monotonically increasing because it is an identity function. If φ is $\neg x$, we obtain that $f_{\neg x}$ is monotonically decreasing because the composition of a monotonically decreasing function and a monotonically increasing function is a monotonically decreasing function.

Induction step:

1. Assume that φ is of the form $\neg\psi$.
   a) If every free occurrence of x in φ is positive, then every free occurrence of x in ψ is negative. Therefore, using the induction hypothesis, we obtain that the function f_ψ is monotonically decreasing. Finally, $f_{\neg\psi}$ is monotonically increasing as a composition of two monotonically decreasing functions.
   b) If every free occurrence of x in φ is negative, then every free occurrence of x in ψ is positive. Similarly to the first case, we conclude that $f_{\neg\psi}$ is monotonically decreasing.
2. Assume that φ is of the form $\varphi_1 \wedge \varphi_2$.
   a) If every free occurrence of x in φ is positive, then every free occurrence of x in φ_1 and φ_2 is positive. Using the induction hypothesis and the fact that the composition of monotonically increasing functions is a monotonically increasing function, we obtain that $f_{\varphi_1\wedge\varphi_2}$ is indeed monotonically increasing.
   b) If every free occurrence of x in φ is negative, then every free occurrence of x in φ_1 and φ_2 is negative. Similarly to the first case, we conclude that $f_{\varphi_1\wedge\varphi_2}$ is monotonically decreasing.
3. Assume that φ is of the form $H_i\psi$.
   a) If every free occurrence of x in φ is positive, then every free occurrence of x in ψ is positive. Using the induction hypothesis and the fact that the composition of monotonically increasing functions is a monotonically increasing function, we obtain that $f_{H_i\psi}$ is indeed monotonically increasing.

b) If every free occurrence of x in φ is negative, then every free occurrence of x in ψ is negative. Similarly to the first case, we conclude that $f_{H_i\psi}$ is monotonically decreasing.

4. Assume that φ is of the form $\Diamond\psi$. Analogously to 3., we obtain the desired monotonicity.

5. Assume that φ is of the form $C_G^{\Diamond H}\psi$. Since no occurrence of x is free in φ in this case, the statement of the lemma vacuously holds. □

Corollary 5.85. *For all formulas $\varphi \in \mathcal{L}_{\Diamond}^{CH}$ and groups $\varnothing \neq G \subseteq \mathcal{A}$ of agents, the function $f_{E_G^{\Diamond H}(\varphi\wedge x)}$ is monotonically increasing.*

Proof. By definition, $E_G^{\Diamond H}(\varphi \wedge x) = \bigwedge_{i\in G} \Diamond H_i(\varphi \wedge x)$. Using Lemma 5.84, we immediately obtain that the function $f_{E_G^{\Diamond H}(\varphi\wedge x)}$ is indeed monotonically increasing. □

Definition 5.86. *Given an interpreted system $\mathcal{I} = (R, \pi)$, a point $(r, t) \in R \times \mathbb{N}_0$, and a formula $\varphi \in \mathcal{L}_{\Diamond}^{CH}$ we define*

$$(\mathcal{I}, r, t) \models C_G^{\Diamond H}\varphi \quad \textit{iff} \quad (r,t) \in \bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}(B)\}.$$

In other words,

$$(C_G^{\Diamond H}\varphi)^{\mathcal{I}} := \bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}(B)\}.$$

Lemma 5.87. *Given an interpreted system $\mathcal{I} = (R, \pi)$, for any formula $\varphi \in \mathcal{L}_{\Diamond}^{CH}$,*

$$f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}}) = (C_G^{\Diamond H}\varphi)^{\mathcal{I}}.$$

Proof. Let us first show that $(C_G^{\Diamond H}\varphi)^{\mathcal{I}} \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}})$, using Definition 5.86:

$$\begin{aligned}
(C_G^{\Diamond H}\varphi)^{\mathcal{I}} &= \bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}(B)\} \\
&\subseteq \bigcup\{f_{E_G^{\Diamond H}(\varphi\wedge x)}(B) \mid B \subseteq R \times \mathbb{N}_0 \text{ and } B \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}(B)\} \\
&= f_{E_G^{\Diamond H}(\varphi\wedge x)}(\bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}(B)\}) \\
&= f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}}).
\end{aligned}$$

Since $f_{E_G^{\Diamond H}(\varphi\wedge x)}$ is monotonically increasing, from $(C_G^{\Diamond H}\varphi)^{\mathcal{I}} \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}})$ it follows that

$$f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}}) \subseteq f_{E_G^{\Diamond H}(\varphi\wedge x)}(f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}})).$$

Therefore, according to Definition 5.86, $f_{E_G^{\Diamond H}(\varphi\wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}}) \subseteq (C_G^{\Diamond H}\varphi)^{\mathcal{I}}$ holds as well. □

Lemma 5.88. *Given an interpreted system $\mathcal{I} = (R, \pi)$ and a set $A \subseteq R \times \mathbb{N}_0$, for all formulas $\varphi \in \mathcal{L}_{\Diamond}^{CH}$,*

$$f_\varphi(A) = \varphi^{\mathcal{I}}.$$

Proof. We proceed by induction on the structure of φ.

Base case: If φ is $p \in \mathsf{Prop}$, then Definition 5.83 (1) implies

$$f_p(A) = p^{\mathcal{I}}.$$

Induction step:

1. Assume that φ is of the form $\neg\psi$. Using Definition 5.83 (3) and the induction hypothesis, we obtain

$$f_{\neg\psi}(A) = f_\neg(f_\psi(A)) = (R \times \mathbb{N}_0) \setminus f_\psi(A) = (R \times \mathbb{N}_0) \setminus \psi^{\mathcal{I}} = (\neg\psi)^{\mathcal{I}}.$$

2. Assume that φ is of the form $\psi_1 \wedge \psi_2$. Using Definition 5.83 (4) and the induction hypothesis, we obtain

$$f_{\psi_1\wedge\psi_2}(A) = f_\wedge(f_{\psi_1}(A), f_{\psi_2}(A)) = f_{\psi_1}(A) \cap f_{\psi_2}(A) = \psi_1^{\mathcal{I}} \cap \psi_2^{\mathcal{I}} = (\psi_1 \wedge \psi_2)^{\mathcal{I}}.$$

3. Assume that φ is of the form $H_i\psi$. Using Definition 5.83 (5) and the induction hypothesis, we obtain

$$\begin{aligned} f_{H_i\psi}(A) &= f_{H_i}(f_\psi(A)) \\ &= \{(r,t) \mid (r',t') \in f_\psi(A) \text{ for all } (r',t') \text{ s.t. } (r,t)\mathcal{H}_i(r',t')\} \\ &= \{(r,t) \mid (r',t') \in \psi^{\mathcal{I}} \text{ for all } (r',t') \text{ s.t. } (r,t)\mathcal{H}_i(r',t')\} \\ &= (H_i\psi)^{\mathcal{I}}. \end{aligned}$$

4. Assume that φ is of the form $\Diamond\psi$. Similarly to the previous case, we obtain the desired equality.

5. Assume that φ is of the form $C_G^{\Diamond H}\psi$. Then, using Definition 5.83 (7) and Definition 5.86, we immediately get

$$f_{C_G^{\Diamond H}\psi}(A) = (C_G^{\Diamond H}\psi)^{\mathcal{I}}$$

regardless of the induction hypothesis. □

Now we turn to showing that the well-known *Fixpoint Axiom* and *Induction Rule* [FHMV95] hold for common eventual hope modality.

First, we need the following lemma.

Lemma 5.89. *Given an interpreted system $\mathcal{I} = (R, \pi)$, for any formula $\varphi \in \mathcal{L}^{CH}_{\Diamond x}$ and any formula $\psi \in \mathcal{L}^{CH}_{\Diamond}$,*

$$f_\varphi(\psi^{\mathcal{I}}) = (\varphi[x/\psi])^{\mathcal{I}}.$$

Proof. We proceed by induction on the structure of φ. Let $\psi \in \mathcal{L}^{CH}_{\Diamond}$.

Base case: If φ is $p \in \mathsf{Prop}$, then $p[x/\psi] = p$. The result $f_p(\psi^{\mathcal{I}}) = p^{\mathcal{I}}$ follows immediately by Definition 5.83 (1). If φ is the propositional variable x, then $x[x/\psi] = \psi$. The result $f_x(\psi^{\mathcal{I}}) = \psi^{\mathcal{I}}$ follows immediately by Definition 5.83 (2).

Induction step:

1. Assume that φ is of the form $\neg\chi$. Using Definition 5.83 (3) and the induction hypothesis, we obtain

$$f_{\neg\chi}(\psi^{\mathcal{I}}) = f_\neg(f_\chi(\psi^{\mathcal{I}})) = (R \times \mathbb{N}_0) \setminus f_\chi(\psi^{\mathcal{I}}) = (R \times \mathbb{N}_0) \setminus (\chi[x/\psi])^{\mathcal{I}} = (\neg\chi[x/\psi])^{\mathcal{I}}.$$

2. Assume that φ is of the form $\chi_1 \wedge \chi_2$. Using Definition 5.83 (4) and the induction hypothesis, we obtain

$$\begin{aligned} f_{\chi_1\wedge\chi_2}(\psi^{\mathcal{I}}) &= f_\wedge(f_{\chi_1}(\psi^{\mathcal{I}}), f_{\chi_2}(\psi^{\mathcal{I}})) \\ &= f_{\chi_1}(\psi^{\mathcal{I}}) \cap f_{\chi_2}(\psi^{\mathcal{I}}) \\ &= (\chi_1[x/\psi])^{\mathcal{I}} \cap (\chi_2[x/\psi])^{\mathcal{I}} \\ &= (\chi_1[x/\psi] \wedge \chi_2[x/\psi])^{\mathcal{I}} \\ &= ((\chi_1 \wedge \chi_2)[x/\psi])^{\mathcal{I}}. \end{aligned}$$

3. Assume that φ is of the form $H_i\chi$. Using Definition 5.83 (5) and the induction hypothesis, we obtain

$$\begin{aligned} f_{H_i\chi}(\psi^{\mathcal{I}}) &= f_{H_i}(f_\chi(\psi^{\mathcal{I}})) \\ &= \{(r,t) \mid (r',t') \in f_\chi(\psi^{\mathcal{I}}) \text{ for all } (r',t') \text{ s.t. } (r,t)\mathcal{H}_i(r',t')\} \\ &= \{(r,t) \mid (r',t') \in (\chi[x/\psi])^{\mathcal{I}} \text{ for all } (r',t') \text{ s.t. } (r,t)\mathcal{H}_i(r',t')\} \\ &= (H_i(\chi[x/\psi]))^{\mathcal{I}} = ((H_i\chi)[x/\psi])^{\mathcal{I}}. \end{aligned}$$

4. Assume that φ is of the form $\Diamond\chi$. Similarly to the previous case, we obtain the desired equality.

5. Assume that φ is of the form $C^{\Diamond H}_G\chi$. Then, using Definition 5.83 (7) and Definition 5.86, we immediately get

$$\begin{aligned} f_{C^{\Diamond H}_G\chi}(\psi^{\mathcal{I}}) &= \bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E^{\Diamond H}_G(\varphi\wedge x)}(B)\} \\ &= (C^{\Diamond H}_G\chi[x/\psi])^{\mathcal{I}} \end{aligned}$$

regardless of the induction hypothesis. □

Theorem 5.90. *Given an interpreted system $\mathcal{I} = (R, \pi)$, for all formulas $\varphi, \psi \in \mathcal{L}_{\Diamond}^{CH}$ and groups $\varnothing \neq G \subseteq \mathcal{A}$ of agents:*

$$\mathcal{I} \models E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi) \leftrightarrow C_G^{\Diamond H}\varphi \qquad \textit{Fixpoint Axiom} \qquad (5.14)$$

$$\textit{If } \mathcal{I} \models \psi \rightarrow E_G^{\Diamond H}(\varphi \wedge \psi), \textit{ then } \mathcal{I} \models \psi \rightarrow C_G^{\Diamond H}\varphi \qquad \textit{Induction Rule} \qquad (5.15)$$

Proof. Using Lemma 5.89, we obtain $(E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi))^{\mathcal{I}} = f_{E_G^{\Diamond H}(\varphi \wedge x)}((C_G^{\Diamond H}\varphi)^{\mathcal{I}})$. Therefore, according to Lemma 5.87, $(E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi))^{\mathcal{I}} = (C_G^{\Diamond H}\varphi)^{\mathcal{I}}$ follows, that is,

$$(I, r, t) \models C_G^{\Diamond H}\varphi \leftrightarrow E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi),$$

for all $(r, t) \in R \times \mathbb{N}_0$.

Let now $\mathcal{I} \models \psi \rightarrow E_G^{\Diamond H}(\varphi \wedge \psi)$. This means that $\psi^{\mathcal{I}} \subseteq (E_G^{\Diamond H}(\varphi \wedge \psi))^{\mathcal{I}}$. Using Lemma 5.89, we, thus, obtain

$$\psi^{\mathcal{I}} \subseteq f_{E_G^{\Diamond H}(\varphi \wedge x)}(\psi^{\mathcal{I}}).$$

Finally, from Definition 5.86, follows that $\psi^{\mathcal{I}} \subseteq (C_G^{\Diamond H}\varphi)^{\mathcal{I}}$, that is, $\mathcal{I} \models \psi \rightarrow C_G^{\Diamond H}\varphi$. □

Lemma 5.91. *Let $\mathcal{I} = (R, \pi)$ be an interpreted system, $\varphi, \psi \in \mathcal{L}_{\Diamond}^{CH}$, and $\varnothing \neq G \subseteq \mathcal{A}$. Then*

$$\mathcal{I} \models E_G^{\Diamond H}(\varphi \wedge \psi) \rightarrow E_G^{\Diamond H}\varphi \wedge E_G^{\Diamond H}\psi.$$

Proof. Let $(r, t) \in R \times \mathbb{N}_0$. Assume $(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\varphi \wedge \psi)$. This means that for all $i \in G$ it holds that $(\mathcal{I}, r, t) \models \Diamond H_i(\varphi \wedge \psi)$. Therefore, for all $i \in G$ there exists $t_i \in \mathbb{N}_0$ such that $t \leq t_i$ and $(\mathcal{I}, r, t_i) \models H_i(\varphi \wedge \psi)$. This is further equivalent to: for all $i \in G$ there exists $t_i \in \mathbb{N}_0$ such that $t \leq t_i$ and $(\mathcal{I}, r, t_i) \models H_i\varphi \wedge H_i\psi$. It now follows that for all $i \in G$ it holds that $(\mathcal{I}, r, t) \models \Diamond H_i\varphi$, and for all $i \in G$ it holds that $(\mathcal{I}, r, t) \models \Diamond H_i\psi$. Thus, $(\mathcal{I}, r, t) \models E_G^{\Diamond H}\varphi \wedge E_G^{\Diamond H}\psi$ indeed holds. □

Therefore, we can prove the following:

Theorem 5.92. *Let $\mathcal{I} = (R, \pi)$ be an interpreted system, $\varphi \in \mathcal{L}_{\Diamond}^{CH}$, and $\varnothing \neq G \subseteq \mathcal{A}$. Then*

$$\mathcal{I} \models C_G^{\Diamond H}\varphi \rightarrow (E_G^{\Diamond H})^k\varphi \quad \textit{for all} \quad k > 0. \qquad (5.16)$$

Proof. Let $k > 0$ and $(r, t) \in R \times \mathbb{N}_0$. Assume that $(\mathcal{I}, r, t) \models C_G^{\Diamond H}\varphi$. Therefore, $(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi)$ according to the Fixpoint Axiom. Using the Fixpoint Axiom again, we obtain $(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\varphi \wedge E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi))$. Similarly (by applying the Fixpoint Axiom $k-2$ more times), we obtain:

$$(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\varphi \wedge E_G^{\Diamond H}(\varphi \wedge \cdots \wedge E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi)\ldots),$$

in which the modal operator $E_G^{\Diamond H}$ appears k times. Now, according to Lemma 5.91, the latter implies

$$(\mathcal{I}, r, t) \models E_G^{\Diamond H}\varphi \wedge (E_G^{\Diamond H})^2\varphi \wedge \cdots \wedge (E_G^{\Diamond H})^k\varphi \wedge (E_G^{\Diamond H})^k C_G^{\Diamond H}\varphi,$$

which further implies that $(\mathcal{I}, r, t) \models (E_G^{\Diamond H})^k\varphi$ indeed holds. Since $k > 0$ and $(r, t) \in R \times \mathbb{N}_0$ were chosen arbitrarily, we can conclude that (5.16) holds. □

Theorem 5.93. *Let $\mathcal{I} = (R, \pi)$ be an interpreted system, $\varphi, \psi \in \mathcal{L}_{\Diamond}^{CH}$, and $\varnothing \neq G \subseteq \mathcal{A}$. Then*

$$\textit{If} \quad \mathcal{I} \models \varphi \to \psi, \quad \textit{then} \quad \mathcal{I} \models C_G^{\Diamond H}\varphi \to C_G^{\Diamond H}\psi.$$

Proof. Assume $\mathcal{I} \models \varphi \to \psi$. Let $(r, t) \in R \times \mathbb{N}_0$. Assume further $(\mathcal{I}, r, t) \models C_G^{\Diamond H}\varphi$. Using the Fixpoint Axiom, we get

$$(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi). \tag{5.17}$$

Since $(\mathcal{I}, r, t) \models \varphi \to \psi$ holds by assumption, $(\mathcal{I}, r, t) \models (\varphi \wedge C_G^{\Diamond H}\varphi) \to (\psi \wedge C_G^{\Diamond H}\varphi)$ holds as well. Now $(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\varphi \wedge C_G^{\Diamond H}\varphi) \to E_G^{\Diamond H}(\psi \wedge C_G^{\Diamond H}\varphi)$ easily follows as both H_i (for all $i \in G$) and $\Diamond$ are monotone. Using (5.17), we obtain $(\mathcal{I}, r, t) \models E_G^{\Diamond H}(\psi \wedge C_G^{\Diamond H}\varphi)$. To conclude, we have shown that

$$\mathcal{I} \models C_G^{\Diamond H}\varphi \to E_G^{\Diamond H}(\psi \wedge C_G^{\Diamond H}\varphi),$$

since $(r, t) \in R \times \mathbb{N}_0$ was chosen arbitrarily. Thus, according to the Induction Rule, $\mathcal{I} \models C_G^{\Diamond H}\varphi \to C_G^{\Diamond H}\psi$ follows. □

5.7 Related work

It is notable that, independently, based on algebraic topological modeling, Goubault et al. [GLR22] proposed a KB4-type of a modality to model the epistemic attitudes of agents in synchronous systems restricted to crash failures. They call their (KB4) modalities 'knowledge,' use K_i for them, and define a dead agent as $K_i\bot$. We call our (KB4) modalities 'hope,' use H_i for them, and define a byzantine faulty agent as $H_i\bot$, whereas 'knowledge' for us is a separate modality of type S5. In addition, recall that we model the epistemic attitudes of agents in asynchronous systems with byzantine faults. All this suggests KB4 to be a good choice for epistemically studying a wide range of fault-tolerant systems.

CHAPTER 6

The Firing Rebels with Relay

In this chapter, we perform a knowledge-based analysis of a canonical distributed computing problem called *Firing Rebels with Relay* (FRR) within the byzantine fault-tolerant asynchronous model of distributed systems introduced in Chapter 2. Through a detailed epistemic analysis, we establish the necessary level of knowledge that needs to be acquired by correct agents in every correct solution of the FRR problem (that is, in any protocol that is supposed to solve it). Even though identifying such epistemic conditions does not immediately lead to practical protocols for FRR, it is an important first step towards this goal. Indeed, we expect it to lead to necessary *communication structures*, which must be present in every run of any protocol that is supposed to solve FRR. Knowing the latter would not only enable us to decide right away whether the communication guarantees provided by a given model of distributed systems allow to solve FRR, but would also facilitate the design of efficient protocols for it.

Chapter organization

The exact formulation of the FRR problem is introduced in Section 6.1. Using an appropriately chosen language, we model FRR in Section 6.2 by "translating" its specification into epistemic formulas. Then, in Section 6.3, we perform a thorough epistemic analysis by studying the corresponding interpreted systems. In particular, we establish the necessary epistemic state that needs to be achieved in every correct solution of the FRR problem. Interestingly, the required epistemic state turns out to include a variant of common hope, namely, common eventual hope. We also explore the relationship between mutual eventual hope and common eventual hope specifically in case there are at least $3f + 1$ agents present in the system, where f is the maximal number of agents that can turn byzantine. Finally, we also identify sufficient conditions for solving FRR.

6.1 Formulating the problem

The FRR problem assumes that every agent $i \in \mathcal{A}$ may observe an event START and may generate an action FIRE according to the following specification:

Definition 6.1 (Firing Rebels with Relay)**.** *A system is consistent with* Firing Rebels with Relay (FRR) *for $f > 0$, iff all runs satisfy:*

(C) Correctness*: If at least $2f+1$ agents learn that* START *occurred at a correct agent, then all correct agents perform* FIRE *eventually.*

(U) Unforgeability*: If a correct agent performs* FIRE*, then* START *occurred at a correct agent.*

(R) Relay*: If a correct agent performs* FIRE*, then all correct agents perform* FIRE *eventually.*

Remark 6.2. *A different specification for Correctness can be found in literature: "*If at least $f+1$ reliable agents locally observed START, then some reliable agent fires eventually*" (see, e.g., [BL87]). Here, a reliable agent is one that* will always *follow its protocol, which corresponds to a forever correct agent in our terminology. In the case of FRR, by invoking* (R)*, this specification implies "*If at least $f+1$ reliable agents locally observed START, then all reliable agents fire eventually*." We require $2f+1$ arbitrary (correct or byzantine faulty) agents instead. Of course, given the limit of f byzantine faulty agents per run, at least $f+1$ (not necessarily the same) of these agents will remain forever correct in every run. Moreover, we relax the condition of the $2f+1$ agents locally observing* START *to each of them learning that* START *occurred at a correct agent. This is preferable, because direct observation is only one possible way of ascertaining that* START *occurred. For instance, if an agent has already determined*[1] *who the f byzantine faulty agents are, e.g., due to their erratic behaviour in the past, then a confirmation of* START *from just one other agent would be sufficient.*

Note that in crash-prone systems, FRR is trivial to solve, even for large f: Indeed, every agent who observes START or receives a notification message (for the first time) just invokes FIRE and sends a notification message to everyone. This guarantees that if a single correct agent observes START, every correct agent will invoke FIRE (agents that crash during the run may or may not issue FIRE here). Observe that this solution involves a trivial silent choir [GM18], namely, when no agent observes START. In the presence of byzantine faulty agents, however, this solution does not work, as byzantine faulty agents may send a notification without having observed anything. A correct solution for FRR must, hence, prevent the byzantine faulty agents from triggering FIRE at any correct agent.

[1] Strictly speaking, the agent in this situation *does not know* that the f agents are byzantine faulty, but rather that they are byzantine faulty if it itself is not. By the same token, whenever we say "learned," "determined," or "ascertained" above, what we mean is reasoning under the assumption of its own correctness, i.e., the belief modality B_i rather than the knowledge modality K_i.

6.2 Modeling via interpreted systems

First of all, we fix:

- a finite set $\mathcal{A} := \{1, \ldots, n\}$ of *agents*,
- a nonempty countably infinite set Prop of *atomic propositions* (*atoms*),
- a finite set $\mathsf{Co} := \{correct_i \mid i \in \mathcal{A}\}$ of designated *correctness atoms*,
- a finite set $\mathsf{Start} := \{\overline{occurred}_i(\text{START}) \mid i \in \mathcal{A}\}$ of designated *start event atoms*,
- a finite set $\mathsf{Fire} := \{\overline{occurred}_i(\text{FIRE}) \mid i \in \mathcal{A}\}$ of designated *firing atoms*.

Syntax. We start with $\mathsf{Prop} \cup \mathsf{Co} \cup \mathsf{Start} \cup \mathsf{Fire}$ and continue by forming formulas by closing under the Boolean connectives $\neg$ and $\wedge$ and under the following (unary) modal operators: $K_1, \ldots, K_n$, $H_1, \ldots, H_n$, $\Diamond$, $C^{\Diamond H}\varphi$, and Y to obtain the language $\mathcal{L}_{FRR}$, i.e., the language $\mathcal{L}_{FRR}$ is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid H_i\varphi \mid \Diamond\varphi \mid C^{\Diamond H}\varphi \mid Y\varphi,$$

where $p \in \mathsf{Prop} \cup \mathsf{Co} \cup \mathsf{Start} \cup \mathsf{Fire}$ and $i \in \mathcal{A}$. We take $\top$ to be an abbreviation for some fixed propositional tautology, and take $\bot$ to be an abbreviation for $\neg\top$. Also, we use the following standard abbreviations from propositional logic: $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi$ for $\neg\varphi \vee \psi$, and $\varphi \leftrightarrow \psi$ for $(\varphi \to \psi) \wedge (\psi \to \varphi)$. In addition, we also write:

$$\begin{aligned} \Box\varphi &:= \neg\Diamond\neg\varphi, \\ B_i\varphi &:= K_i(correct_i \to \varphi), \\ E^{\Diamond B}\varphi &:= \bigwedge_{j\in\mathcal{A}} \Diamond B_j\varphi, \\ E^{\Diamond H}\varphi &:= \bigwedge_{j\in\mathcal{A}} \Diamond H_j\varphi. \end{aligned}$$

Remark 6.3. *Just like before, the overline in* $\overline{occurred}_i(\text{START})$ *and* $\overline{occurred}_i(\text{FIRE})$ *is used to indicate that the occurrences of* START *and* FIRE *are correct (non-byzantine).*

Semantics. Truth of formulas from $\mathcal{L}_{FRR}$ is defined in the following way:

1. For an atom $p \in \mathsf{Prop} \cup \mathsf{Co} \cup \mathsf{Start} \cup \mathsf{Fire}$, $(\mathcal{I}, r, t) \models p$ iff $(r, t) \in \pi(p)$,
2. $(\mathcal{I}, r, t) \models \neg\varphi$ iff $(\mathcal{I}, r, t) \models \varphi$ does not hold,
3. $(\mathcal{I}, r, t) \models \varphi \wedge \psi$ iff $(\mathcal{I}, r, t) \models \varphi$ and $(\mathcal{I}, r, t) \models \psi$,
4. $(\mathcal{I}, r, t) \models K_i\varphi$ iff $(\mathcal{I}, r', t') \models \varphi$ for all (r', t') such that $r_i(t) = r'_i(t')$,

5. $(\mathcal{I}, r, t) \models H_i\varphi$ iff $(\mathcal{I}, r', t') \models \varphi$ for all (r', t') such that $(\mathcal{I}, r, t) \models \mathit{correct}_i$ and $(\mathcal{I}, r', t') \models \mathit{correct}_i$ and $r_i(t) = r'_i(t')$,

6. $(\mathcal{I}, r, t) \models \Diamond\varphi$ iff $(\mathcal{I}, r, t') \models \varphi$ for some $t' \geq t$,

7. $(\mathcal{I}, r, t) \models C_G^{\Diamond H}\varphi$ iff $(r, t) \in \bigcup\{B \subseteq R \times \mathbb{N}_0 \mid B \subseteq f_{E_G^{\Diamond H}(\varphi \wedge x)}(B)\}$,

8. $(\mathcal{I}, r, t) \models Y\varphi$ iff $t > 0$ and $(\mathcal{I}, r, t-1) \models \varphi$,

where the function $f_{E_G^{\Diamond H}(\varphi \wedge x)}$ is defined as in Definition 5.83.

Note that

$$M_{\mathcal{I}} = (R \times \mathbb{N}_0, \pi, \mathcal{H}_1, \ldots, \mathcal{H}_n) \in \mathsf{K45}_{\mathsf{n}}^{\mathsf{co}},$$

whenever, for $i \in \mathcal{A}$:

$$(r, t)\mathcal{H}_i(r', t') \;\; \textit{iff} \;\; (\mathcal{I}, r, t) \models \mathit{correct}_i \;\; \textit{and} \;\; (\mathcal{I}, r', t') \models \mathit{correct}_i \;\; \textit{and} \;\; r_i(t) = r'_i(t').$$

We will write $(\mathcal{I}, r, t) \not\models \varphi$ to denote that $(\mathcal{I}, r, t) \models \varphi$ does not hold. By $\mathcal{I} \models \varphi$, we denote the fact that φ is satisfied at all the points (r, t), i.e., that φ is *valid* in $\mathcal{I}$.

Lemma 6.4. *For any agent $i \in \mathcal{A}$, formula $\varphi \in \mathcal{L}_{FRR}$, and interpreted system $\mathcal{I}$, it holds that $\mathcal{I} \models H_i\varphi \leftrightarrow (\mathit{correct}_i \rightarrow B_i\varphi)$.*

Proof. Let $\mathcal{I} = (R, \pi)$. Consider a run $r \in R$ and a node $(i, t) \in \mathcal{A} \times \mathbb{N}_0$. Assume $(\mathcal{I}, r, t) \models H_i\varphi$. This means that $(\mathcal{I}, r', t') \models \varphi$ for all (r', t') such that $(\mathcal{I}, r, t) \models \mathit{correct}_i$ and $(\mathcal{I}, r', t') \models \mathit{correct}_i$ and $r_i(t) = r'_i(t')$. Assume now $(\mathcal{I}, r, t) \models \mathit{correct}_i$. In order to show $(\mathcal{I}, r, t) \models B_i\varphi$, i.e., $(\mathcal{I}, r, t) \models K_i(\mathit{correct}_i \rightarrow \varphi)$, take an arbitrary (r^*, t^*) such that $r_i(t) = r^*_i(t^*)$. If $(\mathcal{I}, r^*, t^*) \models \mathit{correct}_i$, our initial assumption implies that $(\mathcal{I}, r^*, t^*) \models \varphi$ must hold. Therefore, $(\mathcal{I}, r^*, t^*) \models \mathit{correct}_i \rightarrow \varphi$.

For the other direction, assume $(\mathcal{I}, r, t) \models \mathit{correct}_i \rightarrow B_i\varphi$. In order to show $(\mathcal{I}, r, t) \models H_i\varphi$, take an arbitrary (r^*, t^*) such that $(\mathcal{I}, r, t) \models \mathit{correct}_i$ and $(\mathcal{I}, r^*, t^*) \models \mathit{correct}_i$ and $r_i(t) = r^*_i(t^*)$. It thus follows from the initial assumption that $(\mathcal{I}, r, t) \models B_i\varphi$, i.e., $(\mathcal{I}, r, t) \models K_i(\mathit{correct}_i \rightarrow \varphi)$. This means that $(\mathcal{I}, r', t') \models \mathit{correct}_i \rightarrow \varphi$ for all (r', t') such that $r_i(t) = r'_i(t')$. Therefore, $(\mathcal{I}, r^*, t^*) \models \mathit{correct}_i \rightarrow \varphi$, in particular. Finally, since $(\mathcal{I}, r^*, t^*) \models \mathit{correct}_i$, it follows that $(\mathcal{I}, r^*, t^*) \models \varphi$. □

Corollary 6.5. *For any formula $\varphi \in \mathcal{L}_{FRR}$ and interpreted system $\mathcal{I}$, it holds that $\mathcal{I} \models E^{\Diamond H}\varphi \leftrightarrow \bigwedge_{i \in \mathcal{A}} \Diamond(\mathit{correct}_i \rightarrow B_i\varphi)$.*

Definition 6.6. *For an agent $i \in \mathcal{A}$, we define:*

$$\begin{aligned}
\overline{start_i} &:= Y\overline{occurred_i}(\text{START}) \wedge correct_i \\
\overline{fire_i} &:= \overline{occurred_i}(\text{FIRE}) \wedge correct_i \\
\overline{start} &:= \bigvee_{j \in \mathcal{A}} \overline{start_j} \\
\overline{fire} &:= \bigvee_{j \in \mathcal{A}} \overline{fire_j}
\end{aligned}$$

Note that for one of these formulas to be true, it is necessary for (one of) the involved agent(s) to be correct not only at the time the event/action in question occurred but also at the time of the evaluation. Using the yesterday modality Y in $\overline{start_i}$ accounts for the fact that our agents cannot act on a precondition in the same round it is established.

Using Definition 6.6, we can translate the specification of FRR (stated in Definition 6.1) as follows:

Definition 6.7 (Modeling Firing Rebels with Relay)**.** *An interpreted system $\mathcal{I}$ is consistent with Firing Rebels with Relay for $f > 0$, if the conditions Correctness* (C)*, Unforgeability* (U)*, and Relay* (R) *hold:*

$$\begin{aligned}
&\text{(C)} \quad \mathcal{I} \models \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G| = 2f+1}} \bigwedge_{j \in G} B_j \overline{start} \rightarrow \bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \rightarrow \overline{fire_i}) \\
&\text{(U)} \quad \mathcal{I} \models \overline{fire} \rightarrow \overline{start} \\
&\text{(R)} \quad \mathcal{I} \models \overline{fire} \rightarrow \bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \rightarrow \overline{fire_i})
\end{aligned}$$

Remark 6.8 (Variants of eventuality)**.** *The phrase* all correct agents fulfill φ_i eventually *in Definition 6.1 can be formalized in two different ways:*

- $\bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \rightarrow \varphi_i)$*, which states that each agent will either become byzantine faulty at some point in the future or will fulfill its respective φ_i at some point in the future.*
- $\Diamond \bigwedge_{i \in \mathcal{A}} (correct_i \rightarrow \varphi_i)$*, which states that there is one moment in the future by which every agent still correct fulfills its respective φ_i.*

The second statement is a strengthening of the first as it demands the existence of a common moment in time at which for all correct agents formulas φ_i are satisfied. As we show in Corollary 6.11, for $\varphi_i = \overline{fire_i}$, the two formulations are equivalent because, due to our agents having perfect recall (Remark 2.43), $correct_i \rightarrow \overline{fire_i}$ is a stable fact (Lemma 6.10), i.e., once the formula evaluates to true, it stays true forever. In order to prove this, we first show that faultiness of an agent is a stable fact too.

Lemma 6.9. *$\mathcal{I} \models \neg correct_i \rightarrow \Box\neg correct_i$ for any agent $i \in \mathcal{A}$ and any interpreted system $\mathcal{I}$.*

Proof. Let $\mathcal{I} = (R, \pi)$. Consider a run $r \in R$ and a node $(i, t) \in \mathcal{A} \times \mathbb{N}_0$. Assume $(\mathcal{I}, r, t) \models \neg correct_i$. By Definition 3.4, this means that for some $t_i \leq t$,

$$\Lambda_{t_i} \cap FEvents_i \neq \varnothing$$

holds. Consider an arbitrary $t' \geq t$. Now $(\mathcal{I}, r, t') \models \neg correct_i$ immediately follows since $t_i \leq t \leq t'$. Therefore, $(\mathcal{I}, r, t) \models \Box\neg correct_i$ indeed holds. □

Lemma 6.10. *$\mathcal{I} \models (correct_i \rightarrow \overline{fire_i}) \rightarrow \Box(correct_i \rightarrow \overline{fire_i})$ for any agent $i \in \mathcal{A}$ and any interpreted system $\mathcal{I}$.*

Proof. Let $\mathcal{I} = (R, \pi)$. Consider a run $r \in R$ and a node $(i, t) \in \mathcal{A} \times \mathbb{N}_0$. Assume $(\mathcal{I}, r, t) \models correct_i \rightarrow \overline{fire_i}$. Consider further an arbitrary $t' \geq t$. If $(\mathcal{I}, r, t) \models \neg correct_i$, then, according to the previous lemma, $(\mathcal{I}, r, t') \models \neg correct_i$ follows. Thus, $(\mathcal{I}, r, t') \models correct_i \rightarrow \overline{fire_i}$ also follows. Let us assume now $(\mathcal{I}, r, t) \models \overline{fire_i}$. To show that $(\mathcal{I}, r, t') \models correct_i \rightarrow \overline{fire_i}$ holds in this case as well, assume further $(\mathcal{I}, r, t') \models correct_i$. According to Definition 6.6, $(\mathcal{I}, r, t) \models \overline{fire_i}$ means $(\mathcal{I}, r, t) \models \overline{occurred_i}(\text{FIRE}) \wedge correct_i$. Consequently, according to Definition 3.4, there exists some $t^* < t$ such that $\text{FIRE} \in label^{-1}\left(\beta_i^{t^*}(r) \sqcup \overline{\beta}_{\epsilon_i}^{t^*}(r)\right)$. Therefore, $(\mathcal{I}, r, t') \models \overline{occurred_i}(\text{FIRE})$ must also hold since $t^* < t \leq t'$. Finally, using $(\mathcal{I}, r, t') \models correct_i$, we obtain $(\mathcal{I}, r, t') \models correct_i \rightarrow \overline{fire_i}$. □

Corollary 6.11. *For any interpreted system $\mathcal{I}$:*

$$\mathcal{I} \models \bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \rightarrow \overline{fire_i}) \leftrightarrow \Diamond \bigwedge_{i \in \mathcal{A}} (correct_i \rightarrow \overline{fire_i}).$$

Proof. Let $\mathcal{I} = (R, \pi)$. Consider a run $r \in R$ and a timestamp $t \in \mathbb{N}_0$. Assume $(\mathcal{I}, r, t) \models \bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \rightarrow \overline{fire_i})$. This means that for every $i \in \mathcal{A}$ there exists some $t_i \geq t$ such that $(\mathcal{I}, r, t_i) \models correct_i \rightarrow \overline{fire_i}$. Now, let $t_{max} := max\{t_i \mid i \in \mathcal{A}\}$. By applying the previous lemma, we obtain $(\mathcal{I}, r, t_{max}) \models correct_i \rightarrow \overline{fire_i}$ for all $i \in \mathcal{A}$, i.e., $(\mathcal{I}, r, t_{max}) \models \bigwedge_{i \in \mathcal{A}} correct_i \rightarrow \overline{fire_i}$. Since $t_{max} \geq t$, $(\mathcal{I}, r, t) \models \Diamond \bigwedge_{i \in \mathcal{A}} (correct_i \rightarrow \overline{fire_i})$ follows.

For the other direction, let us assume $(\mathcal{I}, r, t) \models \Diamond \bigwedge_{i \in \mathcal{A}} (correct_i \rightarrow \overline{fire_i})$. This means that there exists some $t' \geq t$ such that $(\mathcal{I}, r, t') \models \bigwedge_{i \in \mathcal{A}} (correct_i \rightarrow \overline{fire_i})$, i.e., $(\mathcal{I}, r, t') \models (correct_i \rightarrow \overline{fire_i})$ for all $i \in \mathcal{A}$. Thus, indeed, there exists some $t_i \geq t$ for all $i \in \mathcal{A}$ (namely, $t_i = t'$ for all $i \in \mathcal{A}$) such that $(\mathcal{I}, r, t_i) \models correct_i \rightarrow \overline{fire_i}$. In other words, $(\mathcal{I}, r, t) \models \bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \rightarrow \overline{fire_i})$. □

6.3 Necessary and sufficient conditions

The goal of this section is to

1. lift the given necessary conditions on a single correct agent's firing — namely, that $\overline{start}$ must hold by Unforgeability (U) and $\bigwedge_{i \in \mathcal{A}} \Diamond(correct_i \to \overline{fire_i})$ must hold by Relay (R) — to statements that describe the epistemic state that is necessary to be achieved by any correct agent before firing;
2. strengthen the obtained necessary epistemic conditions for firing so that they become sufficient for satisfying the conditions Unforgeability (U) and Relay (R);
3. show how Correctness (C) helps simplifying the obtained strengthened necessary epistemic conditions in the presence of at least $3f + 1$ agents in the system;
4. find conditions that are sufficient for solving FRR.

Note that the case when insufficiently many agents learn that START occurred at a correct agent, trivially satisfies condition (C). In this case, FRR reduces to (U)+(R), a problem with a trivial solution, namely, all correct agents not firing. It is the combination of all three conditions that makes FRR a problem worth the analysis.

The first lemma formalizes the fact that, since our agents have perfect recall, reasoning under the assumption of their own correctness leads them to believe that their perceptions are accurate. For instance, an agent who recalls observing START believes that, unless it is byzantine faulty, a correct agent (namely, itself) observed START.

Lemma 6.12. *For any interpreted system $\mathcal{I}$ and any agent $i \in \mathcal{A}$:*

$$\mathcal{I} \models \overline{fire_i} \to B_i \overline{fire_i} \tag{6.1}$$
$$\mathcal{I} \models \overline{fire_i} \to B_i \overline{fire} \tag{6.2}$$
$$\mathcal{I} \models \overline{start_i} \to B_i \overline{start_i} \tag{6.3}$$
$$\mathcal{I} \models \overline{start_i} \to B_i \overline{start} \tag{6.4}$$

Proof. The argument is the same for FIRE and START. We only provide it for the former. Let $\mathcal{I} = (R, \pi)$. Consider a run $r \in R$ and a node $(i, t) \in \mathcal{A} \times \mathbb{N}_0$. Assume $(\mathcal{I}, r, t) \models \overline{fire_i}$. According to Definition 6.6, this means that $(\mathcal{I}, r, t) \models \overline{occurred_i}(\text{FIRE}) \wedge correct_i$. Consequently, according to Definition 3.4, there exists some $t^* < t$ such that $\text{FIRE} \in label^{-1}\left(\beta_i^{t^*}(r) \sqcup \overline{\beta}_{\epsilon_i}^{t^*}(r)\right)$. Consider any $r' \in R$ and $t' \in \mathbb{N}_0$ such that $r_i(t) = r'_i(t')$. Then, $r'_i(t')$ also contains a record of FIRE since agent i has perfect recall. If $(\mathcal{I}, r', t') \models correct_i$, this record must correspond to a correct action and, consequently, $(\mathcal{I}, r', t') \models \overline{fire_i}$. Since $(\mathcal{I}, r', t') \models correct_i \to \overline{fire_i}$ whenever $r_i(t) = r'_i(t')$, we have $(\mathcal{I}, r, t) \models K_i(correct_i \to \overline{fire_i})$, i.e., $(\mathcal{I}, r, t) \models B_i \overline{fire_i}$. The other statement about FIRE follows from $\models \overline{fire_i} \to \overline{fire}$ and monotonicity of B_i. □

Unforgeability (U) states that $\overline{start}$ is a necessary condition for a correct agent firing. Lifting this condition to the level of agent's knowledge yields that, in order to fire, it must believe in $\overline{start}$.

Lemma 6.13 (Epistemic state necessary for firing in presence of Unforgeability (U))**.** *Let $\mathcal{I}$ be an interpreted system consistent with Unforgeability* (U)*. For any agent $i \in \mathcal{A}$,*

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i \overline{start}. \tag{6.5}$$

Proof. The proof follows immediately from (6.2), (U), and monotonicity of B_i. □

Similarly, lifting the Relay condition (R) to the level of agent's knowledge yields the requirement that, in order to fire, a correct agent must believe that all correct agents eventually will have fired.

Lemma 6.14 (Epistemic state necessary for firing in presence of Relay (R))**.** *Let $\mathcal{I}$ be an interpreted system consistent with Relay* (R)*. For any agent $i \in \mathcal{A}$,*

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i \bigwedge_{j \in \mathcal{A}} \Diamond(correct_j \to \overline{fire}_j). \tag{6.6}$$

Proof. Immediately follows from (6.2), (R), and monotonicity of B_i. □

Combining the conditions necessary for (U) and (R), we establish the following level of knowledge necessary for firing:

Theorem 6.15 (Epistemic state necessary for firing in presence of both (U) and (R))**.** *Let $\mathcal{I}$ be an interpreted system consistent with* (U) *and* (R)*. For any agent $i \in \mathcal{A}$,*

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i(\overline{start} \wedge E^{\Diamond H}\overline{start}).$$

Proof. Since the system is consistent with (U), (6.5) holds according to Lemma 6.13. Thus, it only remains to show that

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i E^{\Diamond H}\overline{start}. \tag{6.7}$$

Using $\mathcal{I} \models \overline{fire}_j \to B_j\overline{start}$, that is (6.5), we easily obtain

$$\mathcal{I} \quad \models \quad B_i \bigwedge_{j \in \mathcal{A}} \Diamond(correct_j \to \overline{fire}_j) \to B_i \bigwedge_{j \in \mathcal{A}} \Diamond(correct_j \to B_j\overline{start}), \tag{6.8}$$

using propositional reasoning and monotonicity of both B_j (for all $j \in G$) and $\Diamond$.

Since the system is consistent with (R) as well, (6.6) holds according to Lemma 6.14. Therefore, using (6.8) we further obtain

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i \bigwedge_{j \in \mathcal{A}} \Diamond(correct_j \to B_j\overline{start})$$

by propositional reasoning. Finally, according to Corollary 6.5, it follows that (6.7) indeed holds. □

Remark 6.16 (Emergence of hope)**.** *Note that requiring* $\mathcal{I} \models \overline{fire}_i \to B_i E^{\Diamond B} \overline{start}$ *would not work. We cannot strengthen the necessary condition in Theorem 6.15 by replacing mutual eventual hope with mutual eventual belief, i.e., by omitting* correct_j *therein. In other words, the use of hope for deeper iterations is crucial for the correct formulation. Indeed, in case of our notion of belief, agent* i *can rarely have unconditional beliefs about another agent* j*'s beliefs. The problematic situation is when agent* j*'s perception is compromised. In that case, agent* i *has no way of ascertaining what* j*'s erroneous input data might be and, hence, cannot determine what a correct agent would have inferred from these incorrect inputs. According to our notion of belief, whether agent* i *itself is correct or not, it reasons assuming that its own perceptions are the objective reality. The* correct_j *assumption is, therefore, necessary to anchor* j *to the same (allegedly) objective reality contemplated by* i*, even though* j*'s access to the facts of this objective reality is generally different from* i*'s.*

Remark 6.17 (Relation to indexical sets)**.** *Another approach to describing beliefs of fault-prone agents is via so-called indexical sets [FHMV95, MT88], which are variable (non-rigid) sets that can be used to represent the set of all correct agents at every point in the system. While our results could be reformulated in terms of indexical sets, there were several reasons for us to choose another language. Besides the ability to reason about all agents, whether correct or byzantine faulty, in a uniform way, we tried to stay as close as possible to the standard language of epistemic modal logic. Perhaps more importantly, however, was the moral lesson of the already mentioned Knowledge of Preconditions Principle [Mos15], which reveals how important it is for an agent to know all ingredients affecting its behaviour, correctness of itself and other agents being one of them. Thus, we believe that the transparent and explicit use of correctness in our language is advantageous. An immediate example is the distinction between belief and hope discussed in Remark 6.16, which would have remained somewhat obscured in the indexical set notation.*

Remark 6.18 (Mutual eventual hope is not sufficient)**.** *While using* $B_i(\overline{start} \wedge E^{\Diamond H} \overline{start})$ *as a trigger for agent* i *firing would ensure Unforgeability* (U)*, it is too weak to guarantee Relay* (R)*. Indeed, consider a system with 3 agents (*$n = 3$*), at most one of which can become byzantine faulty (*$f = 1$*). In such a system, receiving the same information from two independent sources is sufficient to believe in its validity, while information from only one source without observing it first hand is not. Suppose that the protocol forces a correct agent to notify all other agents whenever it observed* START*. Consider a run where agent* b *is byzantine from the beginning, whereas agents* c_1 *and* c_2 *remain correct. Let* c_1 *and* c_2 *each observe* START *and, hence, notify all agents about it. Meanwhile* b *falsely notifies* c_2 *that it too observed* START *but will never duplicate this message to* c_1*. Thus,*

- *correct* c_2 *observed* START *and eventually received 2 confirmations of* START *from* c_1 *and* b*;*

- *correct* c_1 *observed* START *and eventually received 1 confirmation of* START *from* c_2*;*
- *byzantine faulty* b *did not observe* START *but was eventually notified of* START *by both* c_1 *and* c_2*.*

In this situation, all agents eventually believe that START *was correctly observed (*c_1 *and* c_2 *saw it themselves, whereas* b *has 2 independent confirmations). Moreover,* c_2 *has a reason to believe in the mutual eventual hope of* START*. Indeed, hope would be trivially satisfied for a byzantine faulty agent, whereas any correct agent would eventually receive at least 2 confirmations out of 3 that* c_2 *itself possesses. Thus, according to the proposed knowledge threshold,* c_2 *should fire. On the other hand,* c_1 *will never fire because it cannot be sure that* b *will eventually hope that* START *occurred. In* c_1*'s mind, if* b *were correct and* c_2 *were byzantine faulty and did not send a confirmation to* b*, then* b *would only ever receive 1 confirmation, which is not sufficient to make it trust* START *truly occurred. Hence,* c_1 *would never fire, and Relay* (R) *would be violated. The issue here is that* $B_i E^{\Diamond H}\overline{start}$ *for one correct agent* i *does not generally imply that eventually* $B_j E^{\Diamond H}\overline{start}$ *for all other correct agents* j*.*

Thus, although $B_i E^{\Diamond H}\overline{start}$ is necessary before i can fire, acting on it may be premature. The necessary level of knowledge must be further strengthened. Since FRR involves an agreement property (if one correct agent fires, all other correct agents also fire eventually), it is not very surprising that, in fact, some form of common level of knowledge, namely *common eventual hope*, plays a role. We show that Unforgeability (U) and Relay (R) together imply that, in order to fire, an agent must ascertain (modulo its own correctness) both that START was observed by some correct agent and the common eventual hope of the same fact:

Theorem 6.19 (Strengthened epistemic state necessary for firing in presence of both (U) and (R))**.** *Let* $\mathcal{I}$ *be an interpreted system consistent with* (U) *and* (R)*. For any agent* $i \in \mathcal{A}$*,*

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i\Big(\overline{start} \wedge C^{\Diamond H}\overline{start}\Big). \tag{6.9}$$

Proof. Since (6.5) holds by Lemma 6.13, it is sufficient to demonstrate

$$\mathcal{I} \models \overline{fire}_i \to B_i C^{\Diamond H}\overline{start}.$$

Combining (R) with (6.2) by applying the replacement property for positive subformulas, we obtain $\mathcal{I} \models \overline{fire} \to E^{\Diamond H}\overline{fire}$. Thus, using the Induction Rule (5.15) with $\varphi = \psi = \overline{fire}$, we conclude

$$\mathcal{I} \models \overline{fire} \to C^{\Diamond H}\overline{fire}.$$

According to Theorem 5.93, it follows from (U) that $\mathcal{I} \models \overline{fire} \to C^{\Diamond H}\overline{start}$. It remains to use (6.2), monotonicity of B_i and propositional reasoning to obtain $\mathcal{I} \models \overline{fire}_i \to B_i C^{\Diamond H}\overline{start}$. □

Corollary 6.20. *For any interpreted system consistent with FRR,* (6.9) *is satisfied for all agents.*

We now show that, unlike belief in mutual eventual hope (see Remark 6.18), belief in common eventual hope is sufficient to fulfill Unforgeability (U) and Relay (R), i.e., that firing as soon as the necessary level of knowledge from Theorem 6.19 is achieved does guarantee that both (U) and (R) are fulfilled:

Theorem 6.21 (Sufficient conditions for (U) and (R))**.** *For any interpreted system $\mathcal{I}$:*

1. (U) *is fulfilled if* $\quad \mathcal{I} \models \bigwedge_{i\in\mathcal{A}} (\neg B_i \overline{start} \to \neg \overline{fire_i})$.

2. *Both* (U) *and* (R) *are fulfilled if*

$$\begin{aligned} \mathcal{I} \models \bigwedge_{i\in\mathcal{A}} \Big(\Big(\neg B_i \Big(\overline{start} \wedge C^{\Diamond H} \overline{start} \Big) \to \neg \overline{fire_i} \Big) \\ \wedge \Big(B_i \Big(\overline{start} \wedge C^{\Diamond H} \overline{start} \Big) \to \Diamond (correct_i \to \overline{fire_i}) \Big) \Big). \end{aligned} \tag{6.10}$$

Proof. 1. Assume $\mathcal{I} \models \bigwedge_{i\in\mathcal{A}} (\neg B_i \overline{start} \to \neg \overline{fire_i})$. Fix $i \in \mathcal{A}$. Therefore, we have $\mathcal{I} \models \overline{fire_i} \to B_i \overline{start}$, by assumption. Combining this with $\mathcal{I} \models \overline{fire_i} \to correct_i$ (holds according to Definition 6.6) and $\mathcal{I} \models correct_i \to (B_i\varphi \to \varphi)$ results in $\mathcal{I} \models \overline{fire_i} \to \overline{start}$. Now, (U) follows by propositional reasoning since $\overline{fire} = \bigvee_{j\in\mathcal{A}} \overline{fire_j}$.

2. Assume (6.10). Analogously to the previous case, we can show that (U) holds. Fix $i \in \mathcal{A}$. From the first conjunct of (6.10), it follows that $I \models \overline{fire_i} \to B_i C^{\Diamond H} \overline{start}$. Now, just like before, combining this with $\mathcal{I} \models \overline{fire_i} \to correct_i$ and $\mathcal{I} \models correct_i \to (B_i\varphi \to \varphi)$ results in $\mathcal{I} \models \overline{fire_i} \to C^{\Diamond H} \overline{start}$. Since $\mathcal{I} \models C^{\Diamond H}\varphi \to \bigwedge_{j\in\mathcal{A}} \Diamond H_j(\varphi \wedge C^{\Diamond H}\varphi)$ for any formula φ according to the Fixpoint Axiom (5.14),

$$\mathcal{I} \models \overline{fire_i} \to \bigwedge_{j\in\mathcal{A}} \Diamond \Big(correct_j \to B_j (\overline{start} \wedge C^{\Diamond H} \overline{start}) \Big). \tag{6.11}$$

Using the second conjunct of (6.10) and monotonicity of $\Diamond$ in (6.11), we further obtain

$$\mathcal{I} \models \overline{fire_i} \to \bigwedge_{j\in\mathcal{A}} \Diamond \Big(correct_j \to \Diamond (correct_j \to \overline{fire_j}) \Big).$$

In order to show (R), just like before, it is sufficient to demonstrate that

$$\mathcal{I} \models \overline{fire_i} \to \bigwedge_{j\in\mathcal{A}} \Diamond (correct_j \to \overline{fire_j}).$$

It remains to note that $\mathcal{I} \models \Diamond \Big(\varphi \to \Diamond(\varphi \to \psi) \Big) \to \Diamond(\varphi \to \psi)$ for all formulas φ and ψ. □

The following "Lifting lemma" shows that Correctness (C) lifts mutual eventual hope to common eventual hope. This way, the arbitrarily deep nested hope implied by the latter effectively collapses, a phenomenon that has also been reported for other problems [BZM10].

Lemma 6.22 (Lifting lemma)**.** *Let $\mathcal{I}$ be an interpreted system consistent with* (C) *and let $|\mathcal{A}| \geq 3f+1$, where $f > 0$. Furthermore, assume that*

$$\mathcal{I} \quad \models \quad \overline{fire}_i \to B_i\left(\overline{start} \wedge E^{\Diamond H}\overline{start}\right) \tag{6.12}$$

holds. Then,

$$\mathcal{I} \quad \models \quad E^{\Diamond H}\overline{start} \to C^{\Diamond H}\overline{start}. \tag{6.13}$$

Proof. Let $\mathcal{I} = (R, \pi)$. Assume $(\mathcal{I}, r, t) \models E^{\Diamond H}\overline{start}$ for some $r \in R$ and $t \in \mathbb{N}_0$. This means that, for every agent $j \in \mathcal{A}$, there exists some $t'_j \geq t$ such that $(\mathcal{I}, r, t_j) \models H_j\overline{start}$. Since $|\mathcal{A}| \geq 3f+1$, it follows that there exists a group G of $2f+1$ correct agents such that $(\mathcal{I}, r, t_j) \models B_j\overline{start}$, i.e., $(\mathcal{I}, r, t_j) \models K_j(correct_j \to \overline{start})$, for all $j \in G$. Let $t' := \max\{t'_j \mid j \in G\}$. We claim that

$$(\mathcal{I}, r, t') \quad \models \quad \bigwedge_{j\in G} K_j(correct_j \to \overline{start}). \tag{6.14}$$

Indeed, for an arbitrary agent $j \in G$, consider any alternative run $\overline{r} \in R$ and time $\overline{t'} \in \mathbb{N}_0$ such that $\overline{r}_j(\overline{t'}) = r_j(t')$. Given that $t' \geq t'_j$ and the agents have perfect recall, there must exist some time $\overline{t'_j} \leq \overline{t'}$ such that $\overline{r}_j(\overline{t'_j}) = r_j(t'_j)$. Thus, $(\mathcal{I}, \overline{r}, \overline{t'_j}) \models correct_j \to \overline{start}$. Since the latter formula is stable[2], it remains true in $\overline{r}$ by the time $\overline{t'}$. We showed that $(\mathcal{I}, \overline{r}, \overline{t'}) \models correct_j \to \overline{start}$ whenever $\overline{r}_j(\overline{t'}) = r_j(t')$, meaning $(\mathcal{I}, r, t') \models K_j(correct_j \to \overline{start})$. This argument applies to every $j \in G$, hence, (6.14) is demonstrated for the group G of $2f+1$ correct agents. Correctness (C) applied at time t' ensures $(\mathcal{I}, r, t') \models \bigwedge_{i\in\mathcal{A}} \Diamond(correct_i \to \overline{fire}_i)$, and, since $t \leq t'$, we also have

$$(\mathcal{I}, r, t) \quad \models \quad \bigwedge_{i\in\mathcal{A}} \Diamond(correct_i \to \overline{fire}_i).$$

Given that r and t were chosen arbitrarily, we have proved

$$\mathcal{I} \quad \models \quad E^{\Diamond H}\overline{start} \to \bigwedge_{i\in\mathcal{A}} \Diamond(correct_i \to \overline{fire}_i). \tag{6.15}$$

Using (6.12) and monotonicity of $\Diamond$ in (6.15), we further obtain

$$\mathcal{I} \quad \models \quad E^{\Diamond H}\overline{start} \to \bigwedge_{i\in\mathcal{A}} \Diamond(correct_i \to B_i(\overline{start} \wedge E^{\Diamond H}\overline{start})),$$

[2]The proof is similar to Lemma 6.10.

i.e.,

$$\mathcal{I} \quad \models \quad E^{\Diamond H}\overline{start} \to \bigwedge_{i\in\mathcal{A}} \Diamond H_i(\overline{start} \wedge E^{\Diamond H}\overline{start}).$$

In other words, we have demonstrated

$$\mathcal{I} \quad \models \quad E^{\Diamond H}\overline{start} \to E^{\Diamond H}(\overline{start} \wedge E^{\Diamond H}\overline{start}).$$

Using the Induction Rule (5.15) with $\psi = E^{\Diamond H}\overline{start}$ and $\varphi = \overline{start}$, we conclude

$$\mathcal{I} \quad \models \quad E^{\Diamond H}\overline{start} \to C^{\Diamond H}\overline{start}. \qquad \square$$

Corollary 6.23. *Let $\mathcal{I}$ be an interpreted system and let there be at least $3f+1$ agents, where $f > 0$. If $\mathcal{I}$ is consistent with FRR, then* (6.13) *holds.*

Proof. For interpreted systems consistent with (U) and (R), property (6.12) follows from Theorem 6.15. $\square$

Lemma 6.24. *Let $\mathcal{I}$ be an interpreted system. If $\mathcal{I}$ is consistent with (C) and (U), then*

$$\mathcal{I} \models \bigvee_{\substack{G\subseteq\mathcal{A}\\ |G|=2f+1}} \bigwedge_{j\in G} B_j\overline{start} \to E^{\Diamond H}\overline{start}. \tag{6.16}$$

Proof. Using (C), (6.5), and monotonicity of $\Diamond$, we immediately obtain

$$\mathcal{I} \models \bigvee_{\substack{G\subseteq\mathcal{A}\\ |G|=2f+1}} \bigwedge_{j\in G} B_j\overline{start} \to \bigwedge_{i\in\mathcal{A}} \Diamond(correct_i \to B_i\overline{start}),$$

i.e.,

$$\mathcal{I} \models \bigvee_{\substack{G\subseteq\mathcal{A}\\ |G|=2f+1}} \bigwedge_{j\in G} B_j\overline{start} \to \bigwedge_{i\in\mathcal{A}} \Diamond H_i\overline{start}. \qquad \square$$

Corollary 6.25. *Let $\mathcal{I}$ be an interpreted system. If $\mathcal{I}$ is consistent with FRR, then* (6.16) *holds.*

Finally, in the following theorem, we establish sufficient conditions for solving FRR.

Theorem 6.26 (Sufficient conditions for solving FRR)**.** *Let $\mathcal{I}$ be an interpreted system. Assume* (6.16) *and* (6.13)*. If*

$$\begin{aligned}\mathcal{I} \quad \models \quad & \bigwedge_{i\in\mathcal{A}} \Big(\Big(\neg B_i\Big(\overline{start} \wedge E^{\Diamond H}\overline{start}\Big) \to \neg\overline{fire_i}\Big)\\ & \wedge\Big(B_i\Big(\overline{start} \wedge E^{\Diamond H}\overline{start}\Big) \to \Diamond(correct_i \to \overline{fire_i})\Big)\Big),\end{aligned} \tag{6.17}$$

then $\mathcal{I}$ is consistent with FRR.

Proof. In order to show that $\mathcal{I}$ is consistent with FRR, we need to prove that the conditions (C), (U), and (R) from Definition 6.7 are satisfied:

(C) Using (6.16) and (6.13), we obtain $\mathcal{I} \models \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G|=2f+1}} \bigwedge_{j \in G} B_j \overline{start} \to C^{\Diamond H} \overline{start}$. Since $\mathcal{I} \models C^{\Diamond H}\varphi \to \bigwedge_{i \in \mathcal{A}} \Diamond H_i(\varphi \wedge C^{\Diamond H}\varphi)$ for any formula φ according to the Fixpoint Axiom (5.14),

$$\mathcal{I} \models \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G|=2f+1}} \bigwedge_{j \in G} B_j \overline{start} \to \bigwedge_{i \in \mathcal{A}} \Diamond\Big(correct_i \to B_i(\overline{start} \wedge C^{\Diamond H} \overline{start})\Big). \tag{6.18}$$

Therefore, using monotonicity of B_i and $\Diamond$ in (6.18), we obtain that

$$\mathcal{I} \models \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G|=2f+1}} \bigwedge_{j \in G} B_j \overline{start} \to \bigwedge_{i \in \mathcal{A}} \Diamond\Big(correct_i \to B_i(\overline{start} \wedge E^{\Diamond H} \overline{start})\Big)$$

also holds since $\mathcal{I} \models C^{\Diamond H}\varphi \to E^{\Diamond H}\varphi$ for any formula φ according to (5.16). Finally, using the second conjuct of (6.17) and monotonicity of $\Diamond$, we get

$$\mathcal{I} \models \bigvee_{\substack{G \subseteq \mathcal{A} \\ |G|=2f+1}} \bigwedge_{j \in G} B_j \overline{start} \to \bigwedge_{i \in \mathcal{A}} \Diamond\Big(correct_i \to \Diamond(correct_i \to \overline{fire_i})\Big).$$

It remains to note that $\mathcal{I} \models \Diamond\Big(\varphi \to \Diamond(\varphi \to \psi)\Big) \to \Diamond(\varphi \to \psi)$ for all formulas φ and ψ.

(U) Using the first conjuct of (6.17), just like in Theorem 6.21, we obtain the desired.

(R) Using (6.13) and the above used $\mathcal{I} \models C^{\Diamond H} \overline{start} \to E^{\Diamond H} \overline{start}$, we obtain that the formulas in (6.17) and (6.10) are equivalent. Thus, the condition (R) indeed holds according to Theorem 6.21. □

Is belief in $\overline{start}$ reduntant in some cases?

If there is no reason for agents to expect START to occur, their predictions about START occurring can only rely on it already having occurred. This observation is formalized in Theorem 6.31 and the immediately following corollary.

Definition 6.27 (Potentially persistent formulas)**.** *A formula $\varphi \in \mathcal{L}_{FRR}$ is called* potentially persistent *in an interpreted system $\mathcal{I} = (R, \pi)$ if, for any run $r \in R$ and any time $t \in \mathbb{N}_0$ such that $(\mathcal{I}, r, t) \models \varphi$, there exists a run $r' \in R$ such that $r'(t) = r(t)$ — i.e., r' is an alternative continuation of the global state $r(t)$ — and such that $(\mathcal{I}, r', t) \models \Box\varphi$. In other words, a true potentially persistent formula can stay true forever.*

Lemma 6.28. *For any agent $i \in \mathcal{A}$ and formula $\varphi \in \mathcal{L}_{FRR}$ that is potentially persistent in an interpreted system $\mathcal{I}$, it holds that $\mathcal{I} \models K_i\Diamond\neg\varphi \to K_i\neg\varphi$.*

Proof. Let $\mathcal{I} = (R, \pi)$. Assume that $(\mathcal{I}, r, t) \not\models K_i\neg\varphi$ for some $r \in R$ and $t \in \mathbb{N}_0$. Then there exists another run $r' \in R$ and time $t' \in \mathbb{N}_0$ such that $r_i(t) = r'_i(t')$ and $(\mathcal{I}, r', t') \models \varphi$. By the potential persistence of φ, there exists an alternative continuation $r'' \in R$ of the prefix $r'(t')$ such that $r''(t') = r'(t')$ and $(\mathcal{I}, r'', t') \models \Box\varphi$. Thus, $(\mathcal{I}, r'', t') \not\models \Diamond\neg\varphi$. It remains to note that $r''_i(t') = r'_i(t') = r_i(t)$. Hence, $(\mathcal{I}, r, t) \not\models K_i\Diamond\neg\varphi$. □

Lemma 6.29. *$\mathcal{I} \models B_i\Diamond(correct_i \to \varphi) \leftrightarrow K_i\Diamond(correct_i \to \varphi)$ for any agent $i \in \mathcal{A}$, formula $\varphi \in \mathcal{L}_{FRR}$, and interpreted system $\mathcal{I}$, i.e., believing something eventually happens modulo one's own correctness is as strong as knowing it eventually happens modulo one's own correctness.*

Proof. The right-to-left direction is trivial as $\mathcal{I} \models K_i\varphi \to B_i\varphi$ for any agent $i \in \mathcal{A}$ and formula $\varphi \in \mathcal{L}_{FRR}$. Therefore, we prove the implication from left to right.

Firstly, $\neg correct_i \to (correct_i \to \varphi)$ is an instance of a propositional tautology. Hence,

$$\mathcal{I} \models \Box\neg correct_i \to \Box(correct_i \to \varphi).$$

Thus, using Lemma 6.9, we get

$$\mathcal{I} \models \neg correct_i \to \Box(correct_i \to \varphi).$$

Using $\mathcal{I} \models \Box\psi \to \Diamond\psi$ (which follows by seriality of temporal modalities), and knowledge necessitation, we further obtain

$$\mathcal{I} \models K_i(\neg correct_i \to \Diamond(correct_i \to \varphi)).$$

By epistemically internalized propositional reasoning, we have $\mathcal{I} \models K_i(correct_i \to \Diamond(correct_i \to \varphi)) \wedge K_i(\neg correct_i \to \Diamond(correct_i \to \varphi)) \to K_i\Diamond(correct_i \to \varphi)$. Since we have just shown the second conjunct above to be valid, we obtain the desired

$$\mathcal{I} \models K_i(correct_i \to \Diamond(correct_i \to \varphi)) \to K_i\Diamond(correct_i \to \varphi).$$ □

Corollary 6.30. *For any agent $i \in \mathcal{A}$, formula $\varphi \in \mathcal{L}_{FRR}$, and interpreted system $\mathcal{I}$, it holds that $\mathcal{I} \models B_i\Diamond H_i\varphi \leftrightarrow K_i\Diamond H_i\varphi$.*

Theorem 6.31 (Early local belief)**.** *If formula $correct_i \wedge \neg\overline{start}$ is potentially persistent in an interpreted system $\mathcal{I}$, then*

$$\mathcal{I} \models B_i\Diamond H_i\overline{start} \to B_i\overline{start}.$$

Proof. By Corollary 6.30, $\mathcal{I} \models B_i\Diamond H_i\overline{start} \rightarrow K_i\Diamond H_i\overline{start}$. Applying the factivity property of knowledge and propositional reasoning to the expanded version of $K_i\Diamond H_i\overline{start}$ yields

$$\mathcal{I} \models K_i\Diamond(correct_i \rightarrow K_i(correct_i \rightarrow \overline{start})) \rightarrow K_i\Diamond(correct_i \rightarrow \overline{start}).$$

Since $correct_i \wedge \neg\overline{start}$ is potentially persistent, and its negation is equivalent to $correct_i \rightarrow \overline{start}$, we have by Lemma 6.28 that

$$\mathcal{I} \models K_i\Diamond(correct_i \rightarrow \overline{start}) \rightarrow K_i(correct_i \rightarrow \overline{start}).$$

Combining all implications, we conclude that

$$\mathcal{I} \models B_i\Diamond H_i\overline{start} \rightarrow B_i\overline{start}. \qquad \square$$

Corollary 6.32. *If formula $correct_i \wedge \neg\overline{start}$ is potentially persistent in an interpreted system $\mathcal{I}$, then*

$$\mathcal{I} \models B_iE^{\Diamond H}\overline{start} \rightarrow B_i\overline{start}, \tag{6.19}$$

$$\mathcal{I} \models B_iC^{\Diamond H}\overline{start} \rightarrow B_i\overline{start}. \tag{6.20}$$

Proof. The proof of (6.19) follows from the definition of mutual eventual hope, Theorem 6.31 and monotonicity of B_i. The proof of (6.20) is obtained by combining (5.16) and (6.19) using the monotonicity of B_i. $\square$

Remark 6.33. *While sufficient for dropping the conjunct $\overline{start}$ from the conditions triggering and preventing* FIRE *in Theorem 6.21, the potential persistency of $correct_i \wedge \neg\overline{start}$ is not necessary. Indeed,* (6.20) *can hold even when* START *is always guaranteed to occur in every run. For instance, in an interpreted system where* START *occurs exactly once per run, no agent ever becomes byzantine faulty, and, in addition, agents never communicate, $\mathcal{I} \models \neg B_iE^{\Diamond H}\overline{start}$ automatically holds because only the agent who observed* START *can learn that it already occurred. All other agents can only be sure that* START *will occur eventually at some agent in the system. By* (5.16) *and monotonicity of B_i, $\mathcal{I} \models \neg B_iC^{\Diamond H}\overline{start}$. Thus, both implications* (6.19) *and* (6.20) *are vacuously true, allowing to drop $\overline{start}$, though admittedly in such cases agents should never fire anyways.*

6.4 Related work

The FRR problem is a problem related to the *consistent broadcasting primitive*, introduced by Srikanth and Toueg in [ST87b]. The motivation behind this communication primitive was to simulate signed communication in order to be able to convert an authenticated fault-tolerant algorithm into an equivalent non-authenticated fault-tolerant algorithm. In addition, it has been used as a pivotal building block in distributed algorithms for byzantine fault-tolerant clock synchronization [DFP+14, FS12, RS11, ST87a, WS09].

The Firing Rebels with Relay problem can be viewed as a non-synchronous version of the *Byzantine Firing Squad problem* [BL87], which is a problem of synchronizing a collection of processors (some of which might be byzantine faulty) in systems with synchronous communication only (i.e, with bounded message delays). In [BL87], the authors considered two versions of the Byzantine Firing Squad problem, namely a *permissive* and a *strict* version, and they developed protocols solving them by reductions to *Byzantine Agreement*.

In [BZM14], Ben-Zvi and Moses considered the *Ordered Response problem*, where the agents had to respond to an external START event by executing a special one-shot FIRE action in a given order $i_1, i_2, \ldots$. The authors showed that, in every correct solution of the Ordered Response problem, agent i_k has to establish nested knowledge $K_{i_k} K_{i_{k-1}} \ldots K_{i_1} \overline{occurred}(\text{START})$ whenever executing FIRE. In [BM11], the authors also identify corresponding sufficient conditions. In the conference version [BZM10] of [BZM14], the authors also considered the *Simultaneous Response problem*, where all agents had to issue FIRE at the same time. In this case, the group G of firing agents has to establish common knowledge $C_G \overline{occurred}(\text{START})$. This work was later extended to responses that are not simultaneous but tightly coordinated in time [BZM13, GM13].

Closely related to our FRR problem is *Eventual Distributed Agreement* studied in [HMW01], where the stronger notion of continual common knowledge proved its value. The latter needs to hold throughout a run, i.e., from the beginning, which makes sense in the context of [HMW01] since it is applied to conditions on the initial state only. Continual common knowledge is not readily applicable to FRR, however, as START can occur at any time in a run.

In [GM18, GM20], Goren and Moses introduced and epistemically analyzed *silent choirs* as a fundamental primitive for message-optimal protocols in synchronous crash-resilient distributed systems. In synchronous systems, where one can time-out messages, it is well-known [Lam78] that an agent can convey information also by *not* sending a message. In a system where the sender may also crash, however, not receiving a message is not informative in that sense. Still, if only up to f of the $n > f$ agents in a system may crash, a silent choir of $f + 1$ agents that aim to convey identical information suffices: at least one agent in the choir must be correct, so its silence can be relied on. In view of the reduction introduced in [MTH14], broadcasting (and hence FRR) can be seen as the byzantine analog of a silent choir.

CHAPTER 7

Summary of our accomplishments and follow-up/future work

7.1 Summary of our accomplishments

In this thesis, we illustrated how to study byzantine fault-tolerant asynchronous message-passing distributed systems using (temporal-)epistemic logic. Based on our framework for modeling such systems, we established agents' knowledge limitations in the presence of byzantine faulty agents. Since, as we showed, knowledge is not achievable by agents in most cases of interest, we explored how to best capture their actual epistemic states in those situations. On that journey, we encountered different epistemic modalities and studied them from a purely logical point of view. Given that our ultimate goal has been gaining insight into agents' decision-making process in byzantine fault-tolerant systems, we used the newly encountered epistemic modalities to analyze in full detail a canonical distributed computing problem called Firing Rebels with Relay (FRR).

Our main accomplishments can be summarized as follows:

In **Chapter 3**, we derived generic results about what asynchronous agents can(not) know in byzantine fault-tolerant message-passing distributed systems. In our central result, the Brain-in-a-Vat lemma, we showed that no matter what it observed, an asynchronous agent in a byzantine setting can never rule out the possibility of those observations being imaginary results of its malfunction. Using this result, we concluded that the Knowledge of Preconditions principle (according to which any precondition for action must be known by the acting agent) severely restricts the kinds of preconditions for actions agents can rely on in such a setting. Consequently, we investigated how the corresponding adequate preconditions for actions look like, which gave us insight into the epistemic state of an agent in systems with byzantine faults.

In **Chapter 4**, we studied the hope modality, introduced in Chapter 3, from a purely logical point of view. We proposed a separate (from knowledge) axiomatization for the individual hope modality while relying on correctness atoms. We then provided a detailed proof of strong soundness and strong completess for the proposed axiom system with respect to a newly designed class of Kripke models capturing hope in a precise way. The resulting logic turned out to violate the uniform substitution rule, however. In addition, we also provided a proof of soundness and completeness with respect to the standard S5 models for knowledge via a suitable translation function. Finally, we showed that the proposed logic of hope has the finite model property as well as that it is decidable.

In **Chapter 5**, we proposed an alternative axiomatization for the hope modality which successfully avoids the use of correctness atoms. The resulting new logic of hope turned out to be a normal multi-agent epistemic logic. We also proposed a joint logic of hope and knowledge as well as a logic extended with notions of common hope and common knowledge. The proposed systems enabled us to logically characterize byzantine fault-tolerant distributed systems. We also provided a thorough soundness and completeness proof for the joint logic of common hope and common knowledge. In addition, we showed that all of the logics presented in the chapter have the finite model property as well as that they are decidable. Finally, we described a way to introduce a particular temporal-epistemic group notion of hope called common eventual hope and proved some of its basic properties used in Chapter 6.

In **Chapter 6**, using epistemic reasoning, we analyzed a canonical distributed computing problem called Firing Rebels with Relay (FRR) within the byzantine fault-tolerant asynchronous message-passing model of distributed systems. We established the necessary epistemic state that needs to be acquired by correct agents in order to FIRE in every correct solution of the problem. The respective epistemic state turns out to involve common eventual hope, which we show to be attained already by achieving one level of mutual eventual hope in case there are at least $3f + 1$ agents in the system in total. Finally, we also identified sufficient conditions for solving FRR.

7.2 Follow-up/future work

Given that this thesis entered some scientific uncharted teritory, there are many possible directions for future research. We picked two such topics, which are particularly close to the content of this thesis and already started working on them. These are:

(1) Axiomatizing common eventual hope, and

(2) Modeling agent fault recovery using epistemic logic.

Regarding (1), coming up with a suitable sound and complete axiomatization for the common eventual hope modality seems as an especially interesting task given that it is a temporal-epistemic mixture of an operator. The challenge lies in working with the

semantics for it. As we saw in Section 5.6, in order to evaluate a formula of the form $C_G^{\Diamond H}\varphi$ at a world w of a model M, we have to check whether the world w belongs to the set

$$\bigcup\{B \subseteq W \mid B \subseteq f_{E_G^{\Diamond H}(\varphi \wedge x)}(B)\}.$$

There does not seem to be a way around this since it is not clear whether a corresponding relation $\mathcal{R}$ such that

$$M, w \models C_G^{\Diamond H}\varphi \quad \textit{iff} \quad M, v \models \varphi \quad \textit{for all} \quad v \in \mathcal{R}(w)$$

exists. So far, we have identified a candidate axiom system for the simplest form of *common eventual knowledge* (i.e., obtained by taking only the K axiom) as well as a candidate corresponding class of Kripke models. We are not aware of any work on axiomatizing common eventual knowledge directly. However, common eventual knowledge has been introduced in the literature [HM90, FHMV95] using modal mu-calculus [Koz83]. Thus, potentially, some results could be obtained from the already existing work on modal mu-calculus [Wal00, JKS08].

Questions to be addressed:

- How to go about proving soundness and completeness of the identified axiom system directly?
- Is the obtained logic of the simplest form of common eventual knowledge compact?
- Is it straightforward to adapt the obtained results for common eventual hope?
- What insights does the obtained logic provide about the process of reaching common eventual hope among agents in byzantine fault-tolerant distributed systems?

Regarding (2), building on top of the logics introduced in the thesis, we aim to model correctness change of agents using tools from dynamic epistemic logic (DEL) [vDvdHK08]. We showed that correctness of agent i can be represented using an atomic proposition $correct_i$ (as it was done in Chapter 4), or using the hope modality such that "agent i is correct" corresponds to "agent i does not hope false", i.e., $\neg H_i\bot$ (as it was done in Chapter 5).

Therefore, changing the correctness status of agent i can be modeled either

i.) as a factual change, which amounts to changing the truth value of the atom $correct_i$ (at world(s) of interest), or

ii.) as a relation update, which amounts to updating the hope relation $\mathcal{H}_i$ such that it is not empty if we wish to make the agent in question correct (at world(s) of interest), or so that it is in fact empty if we wish to make the agent in question byzantine faulty (at world(s) of interest).

We find this second option particularly interesting, as it represents a novel method of updates in DEL. So far, we have considered several different logics based on whether the relation updates are public (witnessed by all agents) or private.

Questions to be addressed:

- Are the considered logics "rich" enough to capture self-correcting agents in the intended way?
- How do the obtained results relate to *recovery* distributed systems [EAWJ02, Rus96]?

List of Figures

2.1 Axiom system $\mathscr{S}5_n$ 13
2.2 Axiom system $\mathscr{K}45_n$ 13
2.3 Details of round $t.5$ of a $\tau^B_{P_\epsilon,P}$-transitional run r. 32

4.1 Axiom system $\mathscr{H}_{\text{co}}$ 54

5.1 Axiom system $\mathscr{H}$ 78
5.2 Axiom system $\mathscr{KH}$ 83

Bibliography

[ASW88] Hagit Attiya, Marc Snir, and Manfred K. Warmuth. Computing on an anonymous ring. *J. ACM*, 35(4):845–875, 1988.

[AW04] Hagit Attiya and Jennifer Welch. *Distributed Computing*. John Wiley & Sons, 2nd edition, 2004.

[bBvEK06] Johan van Benthem, Jan van Eijck, and Barteld Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[BdRV01] Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 2001.

[BL87] James E. Burns and Nancy A. Lynch. The Byzantine Firing Squad problem. In Franco P. Preparata, editor, *Parallel and Distributed Computing*, volume 4 of *Advances in Computing Research: A research annual*, pages 147–161. JAI Press, 1987.

[BM11] Ido Ben-Zvi and Yoram Moses. On interactive knowledge with bounded communication. *J. Appl. Non Class. Logics*, 21(3-4):323–354, 2011.

[BS22] Claudia Bloeser and Titus Stahl. Hope. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

[BvDH+16] Thomas Bolander, Hans van Ditmarsch, Andreas Herzig, Emiliano Lorini, Pere Pardo, and François Schwarzentruber. Announcements to attentive agents. *Journal of Logic, Language and Information*, 25(1):1–35, 2016.

[BZM10] Ido Ben-Zvi and Yoram Moses. Beyond Lamport's happened-before: On the role of time bounds in synchronous systems. In Nancy A. Lynch and Alexander A. Shvartsman, editors, *DISC 2010*, volume 6343 of *LNCS*, pages 421–436. Springer, 2010.

[BZM13] Ido Ben-Zvi and Yoram Moses. Agent-time epistemics and coordination. In Kamal Lodaya, editor, *ICLA 2013*, volume 7750 of *LNCS*, pages 97–108. Springer, 2013.

[BZM14] Ido Ben-Zvi and Yoram Moses. Beyond Lamport's happened-before: On time bounds and the ordering of events in distributed systems. *Journal of the ACM*, 61(2:13), 2014.

[CGM14] Armando Castañeda, Yannai A. Gonczarowski, and Yoram Moses. Unbeatable consensus. In Fabian Kuhn, editor, *DISC 2014*, volume 8784 of *LNCS*, pages 91–106. Springer, 2014.

[Che80] Brian F. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1980.

[CM86] K. Mani Chandy and Jayadev Misra. How processes learn. *Distributed Comput.*, 1(1):40–52, 1986.

[CZ97] Alexander V. Chagrov and Michael Zakharyaschev. *Modal Logic*, volume 35 of *Oxford logic guides*. Oxford University Press, 1997.

[DFP+14] Danny Dolev, Matthias Függer, Markus Posch, Ulrich Schmid, Andreas Steininger, and Christoph Lenzen. Rigorously modeling self-stabilizing fault-tolerant circuits: An ultra-robust clocking scheme for systems-on-chip. *Journal of Computer and System Sciences*, 80:860–900, 2014.

[Dij74] Edsger W. Dijkstra. Self-stabilizing systems in spite of distributed control. *Communications of the ACM archive*, 17(11):643–644, 1974. http://www.auto.tuwien.ac.at/Projects/W2F/papers.html.

[DLS88] Cynthia Dwork, Nancy Lynch, and Larry Stockmeyer. Consensus in the presence of partial synchrony. *Journal of the ACM*, 35(2):288–323, 1988.

[DM90] Cynthia Dwork and Yoram Moses. Knowledge and common knowledge in a Byzantine environment: Crash failures. *Information and Computation*, 88:156–186, 1990.

[EAWJ02] E. N. (Mootaz) Elnozahy, Lorenzo Alvisi, Yi-Min Wang, and David B. Johnson. A survey of rollback-recovery protocols in message-passing systems. *ACM Computing Surveys*, 34(3):375–408, September 2002.

[FHMV95] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[Fim18] Patrik Fimml. Knowledge in distributed systems with byzantine failures. Master's thesis, Technische Universität Wien, Institut für Technische Informatik, 2018.

[FKS21] Krisztina Fruzsa, Roman Kuznets, and Ulrich Schmid. Fire! In Joseph Y. Halpern and Andrés Perea, editors, *Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge, TARK 2021, Beijing, China, June 25-27, 2021*, volume 335 of *EPTCS*, pages 139–153, 2021.

[FLP85] Michael J. Fischer, Nancy A. Lynch, and M. S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(2):374–382, 1985.

[Fru21] Krisztina Fruzsa. Hope for epistemic reasoning with faulty agents! In Alexandra Pavlova, Mina Young Pedersen, and Raffaella Bernardi, editors, *Selected Reflections in Language, Logic, and Information - ESSLLI 2019, ESSLLI 2020 and ESSLLI 2021 Student Sessions, Selected Papers*, volume 14354 of *Lecture Notes in Computer Science*, pages 93–108. Springer, 2021.

[FS12] Matthias Függer and Ulrich Schmid. Reconciling fault-tolerant distributed computing and systems-on-chip. *Distributed Computing*, 24:323–355, 2012.

[GLR22] Éric Goubault, Jérémy Ledent, and Sergio Rajsbaum. A simplicial model for $\mathbf{KB4_n}$: Epistemic logic with agents that may die. In Petra Berenbrink and Benjamin Monmege, editors, *39th International Symposium on Theoretical Aspects of Computer Science (STACS 2022)*, volume 219 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 33:1–33:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022.

[GM13] Yannai A. Gonczarowski and Yoram Moses. Timely common knowledge. In Burkhard C. Schipper, editor, *TARK XIV*, pages 79–93, 2013.

[GM18] Guy Goren and Yoram Moses. Silence. In *PODC '18*, pages 285–294. ACM, 2018.

[GM20] Guy Goren and Yoram Moses. Silence. *J. ACM*, 67(1):3:1–3:26, 2020.

[Hin62] Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, 1962.

[HM90] Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37:549–587, 1990.

[HMW01] Joseph Y. Halpern, Yoram Moses, and Orli Waarts. A characterization of eventual Byzantine agreement. *SIAM Journal on Computing*, 31:838–865, 2001.

[HS99] Maurice Herlihy and Nir Shavit. The topological structure of asynchronous computability. *J. ACM*, 46(6):858–923, 1999.

[JKS08] Gerhard Jäger, Mathis Kretz, and Thomas Studer. Canonical completeness of infinitary mu. *J. Log. Algebraic Methods Program.*, 76(2):270–292, 2008.

[Koz83] Dexter Kozen. Results on the propositional mu-calculus. *Theor. Comput. Sci.*, 27:333–354, 1983.

[KPS+19] Roman Kuznets, Laurent Prosperi, Ulrich Schmid, Krisztina Fruzsa, and Lucas Gréaux. Knowledge in Byzantine message-passing systems I: Framework and the causal cone. Technical Report TUW-260549, TU Wien, 2019.

[KPSF19] Roman Kuznets, Laurent Prosperi, Ulrich Schmid, and Krisztina Fruzsa. Epistemic reasoning with byzantine-faulty agents. In Andreas Herzig and Andrei Popescu, editors, *Frontiers of Combining Systems - 12th International Symposium, FroCoS 2019, London, UK, September 4-6, 2019, Proceedings*, volume 11715 of *Lecture Notes in Computer Science*, pages 259–276. Springer, 2019.

[Kuz08] Roman Kuznets. *Complexity Issues in Justification Logic*. PhD thesis, The Graduate Center, City University of New York, 2008.

[Lam78] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21:558–565, 1978.

[LSP82] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems*, 4:382–401, 1982.

[Lyn96] Nancy Lynch. *Distributed Algorithms*. Morgan Kaufman, 1996.

[Mic89] Ruben Michel. A categorical approach to distributed systems, expressibility and knowledge. In Piotr Rudnicki, editor, *PODS '89*, pages 129–143. ACM, 1989.

[MM91] John C. Mitchell and Eugenio Moggi. Kripke-style models for types lambda calculus. *Annals of Pure and Applied Logic*, 51(1–2):99–124, 1991.

[Mos15] Yoram Moses. Relating knowledge and coordinated action: The knowledge of preconditions principle. In R. Ramanujam, editor, *TARK 2015*, pages 231–245, 2015.

[MS93] Yoram Moses and Yoav Shoham. Belief as defeasible knowledge. *Artificial Intelligence*, 64:299–321, 1993.

[MT86] Yoram Moses and Mark R. Tuttle. Programming simultaneous actions using common knowledge: Preliminary version. In *27th Annual Symposium on Foundations of Computer Science*, pages 208–221. IEEE, 1986.

[MT88] Yoram Moses and Mark R. Tuttle. Programming simultaneous actions using common knowledge. *Algorithmica*, 3:121–169, 1988.

[MTH14] Hammurabi Mendes, Christine Tasson, and Maurice Herlihy. Distributed computability in Byzantine asynchronous systems. In *STOC 2014: 46th Annual Symposium on the Theory of Computing*, pages 704–713. ACM, 2014.

[PG96] Andrew Pessin and Sanford Goldberg, editors. *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's the "Meaning of 'Meaning'"*. Routledge, 1996.

[PT86] Kenneth J. Perry and Sam Toueg. Distributed agreement in the presence of processor and communication faults. *IEEE Transactions on Software Engineering*, SE-12(3):477–482, 1986.

[RS11] Peter Robinson and Ulrich Schmid. The Asynchronous Bounded-Cycle model. *Theoretical Computer Science*, 412:5580–5601, 2011.

[Rus96] John Rushby. Reconfiguration and transient recovery in state machine architectures. In *Proceedings of the Twenty-Sixth International Symposium on Fault-Tolerant Computing: June 25–27, 1996, Sendai, Japan*, pages 6–15. IEEE, 1996.

[SSK20] Thomas Schlögl, Ulrich Schmid, and Roman Kuznets. The persistence of false memory: Brain in a vat despite perfect clocks. In Takahiro Uchiya, Quan Bai, and Ivan Marsá-Maestre, editors, *PRIMA 2020: Principles and Practice of Multi-Agent Systems - 23rd International Conference, Nagoya, Japan, November 18-20, 2020, Proceedings*, volume 12568 of *Lecture Notes in Computer Science*, pages 403–411. Springer, 2020.

[ST87a] T. K. Srikanth and Sam Toueg. Optimal clock synchronization. *Journal of the ACM*, 34:626–645, 1987.

[ST87b] T. K. Srikanth and Sam Toueg. Simulating authenticated broadcasts to derive simple fault-tolerant algorithms. *Distributed Computing*, 2:80–94, 1987.

[Tar55] Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.

[vDFK22] Hans van Ditmarsch, Krisztina Fruzsa, and Roman Kuznets. A new hope. In David Fernández-Duque, Alessandra Palmigiano, and Sophie Pinchinat, editors, *Advances in Modal Logic, AiML 2022, Rennes, France, August 22-25, 2022*, pages 349–370. College Publications, 2022.

[vDvdHK08] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer, 2008.

[Wal00] Igor Walukiewicz. Completeness of kozen's axiomatisation of the propositional μ-calculus. *Inf. Comput.*, 157(1-2):142–182, 2000.

[WS07] Josef Widder and Ulrich Schmid. Booting clock synchronization in partially synchronous systems with hybrid process and link failures. *Distributed Computing*, 20(2):115–140, 2007.

[WS09] Josef Widder and Ulrich Schmid. The Theta-Model: achieving synchrony without clocks. *Distributed Computing*, 22:29–47, 2009.

[WSS19] Kyrill Winkler, Manfred Schwarz, and Ulrich Schmid. Consensus in directed dynamic networks with short-lived stability. *Distributed Computing*, 32(5):443–458, 2019.