Logics of Rational Agency

Notes for ESSLLI 2009 Course

Eric Pacuit*

June 18, 2009

These notes contain material for a course on "Logics of Rational Agency" taught at the European Summer School for Logic, Language and Information in Bordeaux, France on July 26 - 31, 2009 (ESSLLI 2009). This document contains an extended outline of the course including pointers to relevant literature. The main idea of this reader is to provide a bird's eye view of the literature and to highlight some of the main themes that we will discuss in this course. The course website:

ai.stanford.edu/∼epacuit/classes/esslli/log-ratagency.html

will contain the (updated) lecture notes and slides (updated each day). Enjoy the course and please remember to ask questions during the lecture and also point out any mistakes and/or omitted references in this text!

The goal of this course is to introduce the main conceptual ideas and technical tools that drive much of the research developing logics of rational agency. The course draws on a number of different sources including a recent course on "Rational Agency and Intelligent Interaction" taught at Stanford University with Yoav Shoham and Johan van Benthem. The website contains pointers to a number of relevant papers and textbooks. Below is a (very) brief outline of the course:

Day 1: Introduction, Motivation and Background

Day 2: Basic Ingredients for a Logic of Rational Agency

Day 3: Logics of Rational Agency and Social Interaction, Part I

Day 4: Logics of Rational Agency and Social Interaction, Part II

Day 5: Conclusions and General Issues

*Tilburg Institute for Logic and Philosophy of Science, University of Tilburg. Website: ai.stanford.edu/∼epacuit, Email: E.J.Pacuit@uvt.nl

Introduction and Motivation

A quick glance at the opening paragraphs in many of the classic logic textbooks reveals a common view: logical methods highlight the reasoning patterns of a single (idealized) agent engaged in some form of *mathematical* thinking[1]. However, this traditional view of the "subject matter" of logic is expanding. There is a growing literature using phrases such as "rational interaction" or "information flow" to describe its subject matter while still employing traditional logical methods. The clearest example of this can be found in the work of Johan van Benthem and others on *logical dynamics* (van Benthem, 1996; van Ditmarsch et al., 2007), Rohit Parikh and others on *social software* (Parikh, 2002; van Eijck and Vergrugge, 2009)[2]; and Samson Abramsky and others on *game semantics* (Abramsky, 2007). There are many issues driving this shift in thinking about what logic is *about* (see van Benthem, 2005, for a discussion). Most relevant for this course is the important (and sometimes controversial) role logic has played in AI (Thomason, 2009) and the analysis of distributed algorithms (Halpern et al., 2001, Section 5).

This course will introduce logics for reasoning about communities of agents engaged in some form of social interaction. Much of this work builds upon existing logical frameworks developed by philosophers and computer scientists incorporating insights and ideas from philosophy, game theory, decision theory and social choice theory. The result is a web of logical systems each addressing different aspects of rational agency and social interaction. Rather than providing an encyclopedic account of these different logical systems, we will focus on the main conceptual and technical issues that drive a logical analysis. The main objective is to see the various logical systems as a *coherent* account of rational agency and social interaction. To that end, we will focus on the following three questions:

1. How can we *compare* different logical frameworks addressing similar aspects of rational agency and social interaction (eg., how information evolves through social interaction)?

2. How should we *combine* logical systems which address *different* aspects of social interaction towards the goal of a comprehensive (formal) theory of rational agency?

[1] A (biased) sampling from my bookshelf: Shoenfield's *Mathematical Logic*: "Logic is the study of reasoning; and mathematical logic is the study of the type of reasoning done by mathematicians"; Enderton's *A Mathematical Introduction of Logic*: "Symbolic logic is a mathematical model of deductive thought"; and Chiswell and Hodges *Mathematical Logic*: "In this course we shall study some ways of proving statements."

[2] This is the topic of the ESSLLI 2009 course *Games, Actions and Social Software* taught by Rineke Verbrugge & Jan van Eijck.

3. How does a logical analysis contribute to the broader discussion of rational agency and social interaction within philosophy and the social sciences?

There are two main goals of this course. The first is to analyze various logical systems addressing different aspects of (rational) agency. Again the objective is not to provide a complete survey of *all* the relevant logical systems, but rather to develop concrete answers to the above questions. This will certainly raise a number of different methodological and technical issues (especially concerning the first two questions). Sometimes the differences between two competing logical systems are technical in nature reflecting different conventions used by different research communities. And so, with a certain amount of technical work, such frameworks are seen to be equivalent up to model transformations (cf. Halpern, 1999; Lomuscio and Ryan, 1997; Pacuit, 2007; Goranko and Jamroga, 2004). Other differences point to key conceptual issues about rational agency and social interaction (cf. van Benthem et al., 2008; van der Hoek and Wooldridge, 2003). This leads us to the second main goal of this course: to develop the main technical skills and conceptual understanding needed to navigate the fast-growing literature on logics of rational agency

Any discussion of *rationality*, *agency* and *social interaction* naturally touches on a variety of disciplines and often evokes strong opinions. This is especially true when considerations from empirically-based sciences pointing to experiments showing how "humans *really* behave" are taken into account. However, the situation is not any simpler when engaged in "pure" conceptual modeling. Here the use of formal models often generates a number of interesting technical issues. Furthermore (and often more interesting), there are questions that are not mathematical in nature but involve the meaning of the fundamental concepts employed in the formal analyses. In other words, the debates concerns the very nature of *rationality* (cf. Nozik, 1993; Harman, 1999), what constitutes an *agent* (cf. Bratman, 2007; van Benthem, 2009), underlying assumptions about *rational decision making* (cf. Skyrms, 1990; McClennen, 1990; Stalnaker, 1999; Brandenburger, 2007; Binmore, 2009) and the role formal models play in philosophy and the social sciences (cf. Arrow, 1951; Aumann, 1985; Binmore, 2007; Kreps, 1990; van Benthem, 2008).

Of course, there is no single approach that can address *all* of the complex phenomena that arise when rational and not-so rational agents interact with one another and the environment. Thus it is important to understand both the scope of a particular analysis and how different analyses from within and across the disciplines mentioned above can fit together. This will be an important theme throughout this course.

Ingredients of a logical analysis of rational agency Agents are faced with many diverse tasks when interacting with each other and the environment.

The logical systems discussed in this course describe many different aspects of social interactive situations. They not only highlight patterns of theoretical and practical reasoning, but also the dynamic processes that govern social interactions. Obviously, no formal model can address *every* issue that influences and agent's (rational) behavior in a social situation. So, there are a number of important choices that guide any logical analysis:

What are the basic building blocks? Any mathematical model of a social situation starts with a number of underlying assumptions. For example, there are often simplifying assumptions about the nature of time (continuous or discrete/branching or linear), how (primitive) *events* or *actions* are represented, how *causal* relationships are represented and what constitutes a *state of affairs*. Such assumptions typically reflect common-sense intuitions about these fundamental concepts[3].

Single agent vs. many agents. The difference between these two choices often goes beyond choosing whether or not to explicitly represent multiple agents in the formal models. There are a number of group notions that have been extensively studied including *common knowledge*, *distributed knowledge* and *coalitional ability*. Another distinction is relevant here: a logical analysis may take a first person perspective or a third person perspective. On the third person perspective (arguably the one most commonly found in the literature), the logic is intended to reason *about* a group of agents interacting with each other *from the point-of-view of the modeler*. On the first-person perspective, one agent is given a special status and the logical system represents the reasoning and/or abilities of that agent (cf. Aucher, 2008). This distinction has been extensively discussed by philosophers such as David Lewis (1996) but only recently has made its way into the logic literature (cf. Hendricks, 2009)

Which aspects of agency and social interaction are relevant? Alternatively, how should we talk about an agent's "cognitive makeup" and the normative forces that influence social behavior? The vocabulary used in many logical analyses typically reflects a common-sense, or *folk*, understanding of our states of mind[4]. The logical frameworks discussed in this course not only describe physical aspects of a social

[3]This is not to say that the extensive philosophical literature discussing each of these important issues should be ignored. Rather this reflects the somewhat crude level of abstraction where these logical frameworks reside.

[4]Again, this is not meant to suggest that important ideas from cognitive science, psychology and philosophy are or should be ignored. The point is that much of the current work discussed in these notes presumes some form of *folk psychology*. Of course, the influential work of Dennett (1987) is relevant here. See also (Ravenscroft, 2008) for a more general discussion and pointers to the relevant literature.

situation (eg., which actions are available, which events have taken place, what facts are true and what may be true in the future) but also the agents' *attitudes* towards these different descriptions. The different attitudes can be categorized according to their *direction of fit*[5]:

Informational attitudes describe an agent's view about the way the world *is*. Typical examples from this category include "knowledge", "beliefs" and "certainty".

Motivational attitudes describe what the agent may want to *change* in the world. Typical examples from this category include "preferences", "desires", and "goals"

There is a third type of attitude that has also been the center of logical analyses.

Normative attitudes do not describe what an agent wants or believes, but what the agent *should* do. Typical examples from this category include "obligations" and "permissions".

Static vs. dynamic. Much of this course will focus on how to incorporate "dynamics" into various logical frameworks, so we will will not go into details here.

Related Work Currently, there is no textbook that covers all of the issues we will highlight in this course. However, the reader is encouraged to consult the textbooks (Fagin et al., 1995; Wooldridge, 2000; van Benthem, 2009; Horty, 2001; Shoham and Leyton-Brown, 2009) and the articles (Meyer and Veltman, 2007; van der Hoek and Wooldridge, 2003) for extended discussion of some of the specific logical frameworks and issues we will discuss during the course. There are also other courses here at ESSLLI that will discuss related topics:

- Logics of Individual and Collective Intentionality taught by Anreas Herzig and Emiliano Lorini (Week 1)
- Games, Actions and Social Software taught by Rineke Verbrugge and Jan van Eijck (Week 1)
- The Workshop Logical Methods for Social Concepts (Week 1)

[5]This famous distinction, first pointed out by (Anscombe, 1963, pg. 56), has been widely discussed by philosophers.

- Dynamic Logics for Interactive Belief Revision taught by Alexandru Baltag and Sonja Smets (Week 2)
- Logic and Agent Programming Languages taught by Natasha Alechina and Brian Logan (Week 2)

This is a foundational course, so there are no prerequisites and all the material will be self-contained. Some experience with modal logic may be helpful, but this is not necessary. The appendix contains a short introduction to some of the technical details we will touch on in the course and can be used as a reference

Extended Outline

Below is an extended outline of the course. The ordering of the topics may change, so please consult the course website for the most up-to-date information. The course will be taught over 5 days during the 2nd week of ESSLLI. A detailed schedule will be maintained on the website . The course will be divided into 4 main parts with the general outline is given below (including pointers to relevant reading material):

Part 0: Introduction, Motivation and Background (*.5 meetings*) The main source will be these course notes, but also the articles and textbooks mentioned above in the related works section.

Part I: Basic Ingredients (*1.5 meetings*) We will introduce a number of logical frameworks for reasoning about

- ✓ informational attitudes (eg., knowledge and belief)
- ✓ motivational attitudes (eg., preferences)
- ✓ time, actions and ability
- ✓ group notions (eg., common knowledge and coalitional ability)
- ✓ normative attitudes (eg., obligations)

The primary objective of this part is to introduce the main logical systems that will be used to reason about different aspects of social interactive situations.

Part II: Logical Analysis of Rational Agency (*2.5 meetings*) This is the main part of the course. The primary objective in this part is to develop concrete answers to questions 1 & 2 discussed above. We will discuss the following 6 topics (a few relevant papers are listed below each topic):

1. Background: Combining Modal Logics

   Walter Carnielli and Marcelo Esteban Coniglio (Winter 2008 Edition), "Combining Logics", *The Stanford Encyclopedia of Philosophy* , Edward N. Zalta (ed.). (plato.stanford.edu/entries/logic-combining/)

2. Logics of Knowledge and Beliefs

   Yoav Shoham and Kevin Leyton-Brown (2009). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, Section 13.7. www.masfoundations.org

   Joe Halpern (1996). Should knowledge entail belief?, *Journal of Philosophical Logic*, **25:5**, pp. 483-494.

3. Reasoning about Knowledge, Actions and Abilities

   David Carr (1979). The Logic of Knowing How and Ability, *Mind* **88:351**, pp. 394 - 409.

   Renate Schmidt and Dmitry Tishkovsky (2008). On combinations of propositional dynamic logic and doxastic modal logics, *Journal of Logic, Language and Information*. **17**, pp. 109 - 129.

   Johan van Benthem and Eric Pacuit (2006). The Tree of Knowledge in Action, *Proceedings of Advances of Modal Logic*, pp. 87 - 106.

4. Comparing Logics of Information Flow

   Johan van Benthem, Jelle Gerbrandy, Tomohiro Hoshi and Eric Pacuit (2009). Merging Frameworks for Interaction, *Journal of Philosophical Logic*.

5. Entangling Knowledge, Beliefs and Preferences

   Johan van Benthem, *Logical dynamics of information and interaction*, Book manuscript (Chapter 8)

   Stephen Morris (1996). The Logic of Belief and Belief Change: A Decision Theoretic Approach, *Journal of Economic Theory*, **69**, pp. 1 - 23.

6. Planning and Intentions

   Wiebe van der Hoek, Wojciech Jamroga and Michael Wooldridge (2007). Towards a theory of intention revision, *Synthese: Knowledge, Rationality and Aciton*, **155**, pp. 265 - 290.

   Thomas Icard, Eric Pacuit and Yoav Shoham (2009). A Dynamic Logic of Belief and Intention, manuscript.

Part III: General Issues (*0.5 meetings*) Time permitting, we will conclude with a discussion of some broader issues. In particular, we will point to some initial investigations concerning question 3 discussed above.

Relevant Conferences & Online Resources

There are a number of online resources and conferences that address many of the issues discussed in this course:

- www.loriweb.org: a web portal with a number of important resources (call for papers, conference announcements, available positions, general discussions, etc.)
- LORI: *Workshop on Logic, Rationality and Interaction* is a workshop devoted to many of the themes discussed in this course.(golori.org).

  LORI-II will take place in Chongqing, China, October 8 - 11, 2009!
- TARK: *Theoretical Aspects of Rationality and Knowledge* is a bi-annual conference on the interdisciplinary issues involving reasoning about rationality and knowledge (www.tark.org)
- LOFT: *Logic and the Foundations of Game and Decision Theory* is a bi-annual conference which focuses, in part, on applications of formal epistemology in game and decision theory. (www.econ.ucdavis.edu/faculty/bonanno/loft.html)
- FEW: *Formal Epistemology Workshop* is a yearly conference aimed at general issues in formal epistemology. (fitelson.org/few/)
- KR: *Conference on the Principles of Knowledge Representation and Reasoning* is a bi-annual conference geared towards computer scientists that emphasizes both theoretical and practical applications. (www.kr.org)

References

Abramsky, S. (2007). A compositional game semantics for multi-agent logics of partial information. Volume 1 of *Texts in Logic and Games*, pp. 11–48. Amsterdam University Press.

Anscombe, G. (1963). *Intention*. Cornell University Press.

Arrow, K. (1951). Mathematical models in the social sciences. In D. Lerner and H. Lasswell (Eds.), *The Policy Sciences*. Stanford University Press.

Aucher, G. (2008). Internal models and private multi-agent belief revision. In *AAMAS '08: Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, Richland, SC, pp. 721–727. International Foundation for Autonomous Agents and Multiagent Systems.

Aumann, R. (1985). What is game theory trying to accomplish? In K. Arrow and S. Honkapohja (Eds.), *Frontiers of Economics*, pp. 28 – 76. Oxford.

Binmore, K. (2007). *Does Game Theory Work? The Bargaining Challenge*. The MIT Press.

Binmore, K. (2009). *Rational Decisions*. Princeton University Press.

Brandenburger, A. (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory 35*, 465–492.

Bratman, M. (2007). *Structures of Agency*. Oxford University Press.

Dennett, D. (1987). *The Intentional Stance*. The MIT Press.

Fagin, R., J. Halpern, Y. Moses, and M. Vardi (1995). *Reasoning about Knowledge*. Boston: The MIT Press.

Goranko, V. and W. Jamroga (2004). Comparing semantics of logics for multi-agent systems. *Synthese: Knowledge, Rationality, and Action 139*(2), 241–280.

Halpern, J. (1999). Set-theoretic completeness for epistemic and conditional logic. *Annals of Mathematics and Artificial Intelligence 26*, 1 – 27.

Halpern, J. Y., R. Harper, N. Immerman, P. G. Kolaitis, M. Y. Vardi, and V. Vianu (2001). On the unusual effectiveness of logic in computer science. *The Bulletin of Symbolic Logic 7*(2), 213 – 236.

Harman, G. (1999). *Reasoning, Meaning and Mind*, Chapter Rationality. Oxford University Press.

Hendricks, V. (2009). Axioms of distinction in social software. In J. van Benthem, A. Gupta, E. Pacuit, and R. Parikh (Eds.), *Games, Norms and Reasons*.

Horty, J. (2001). *Agency and Deontic Logic*. Oxford University Press.

Kreps, D. (1990). *Game Theory and Economic Modelling*. Clarendon Press.

Lewis, D. (1996). Elusive knowledge. *The Australian Journal of Philosophy 74*, 549–567.

Lomuscio, A. and M. Ryan (1997). On the relation between interpreted systems and kripke models. In *Proceedings of the AI97 Workshop on Theoretical and Practical Foundation of Intelligent Agents and Agent-Oriented Systems*, Volume LNCS 1441.

McClennen, E. (1990). *Rationality and Dynamic Choice : Foundational Explorations*. Cambridge University Press.

Meyer, J.-J. and F. Veltman (2007). Intelligent agents and common sense reasoining. In P. Blackburn, J. van Benthem, and F. Wolter (Eds.), *Handbook of Modal Logic*. Elsevier.

Nozik, R. (1993). *The Nature of Rationality*. Princeton University Press.

Pacuit, E. (2007). Some comments on history based structures. *Journal of Applied Logic 5*(4), 613–624.

Parikh, R. (2002, September). Social software. *Synthese 132*(3).

Ravenscroft, I. (Fall 2008). Folk psychology as a theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Shoham, Y. and K. Leyton-Brown (2009). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

Stalnaker, R. (1999). Extensive and strategic forms: Games and models for games. *Research in Economics 53*, 293 – 319.

Thomason, R. (Spring 2009). Logic and artificial intelligence. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

van Benthem, J. (1996). *Exploring Logical Dynamics*. CSLI Press.

van Benthem, J. (2005). Where is logic going and should it? *Topoi 25*, 117 – 122.

van Benthem, J. (2008). Logic and reasong: do the facts matter? *Studia Logica 88*(1), 67 – 84.

van Benthem, J. (2009). Logical dynamics of information and interaction. Book manuscript.

van Benthem, J., J. Gerbrandy, T. Hoshi, and E. Pacuit (2008). Merging frameworks of interaction. *Journal of Philosophical Logic*, Forthcoming.

van der Hoek, W. and M. Wooldridge (2003). Towards a logic of rational agency. *Logic Journal of the IGPL 11*(2), 135 – 160.

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

van Eijck, J. and R. Vergrugge (Eds.) (2009). *Discourses on Social Software*. Texts in Logic and Games.

Wooldridge, M. (2000). *Reasoning about Rational Agents*. The MIT Press.

A A Brief Introduction to Modal Logic

These short notes are intended to supplement the lectures and introduce some of the basic concepts of *Modal Logic*. The primary goal is to provide a study guide that will complement the technical material presented during the course. There are many textbooks that you can consult for more information. The following is a list of some texts (this is not a complete list, but a pointer to books that I have found particularly useful).

- *Modal Logic for Open Minds* by Johan van Benthem. A new textbook on modal logic (still in draft form) provides a modern introduction to modal logic. This book will be published sometime the end of the summer/early next fall.

- *Modal Logic* by Brian Chellas. A nice introduction to modal logic though somewhat outdated.

- The *Modal Logic* entry at the Stanford Encyclopedia of Philosophy (http://plato.stanford.edu/entries/logic-modal/). This entry was written by James Garson and provides a nice overview of the philosophical applications of modal logic.

There are also more advanced books that you should keep on your radar.

- *Handbook of Modal Logic* edited by Johan van Benthem, Patrick Blackburn and Frank Wolter. This very extensive volume represents the current stat-of-affairs in modal logic.
- *Modal Logic* by Patrick Blackburn, Maarten de Rijke and Yde Venema. An advanced, but very accessible, textbook focusing on the main technical results in the area.
- *Dynamic Epistemic Logic* by Hans van Ditmarsch, Wiebe van der Hoek and Barteld Kooi. This text presents a number of the basic technical results about dynamic versions of epistemic logics.

A.1 Syntax and Semantics of Modal Logic

What is a modal? A *modal* is anything that qualifies the truth of a sentence. There are many ways to qualify the truth of a statement in natural language. For example, each of the phrases below can be used to complete the sentence:
John _____ happy.

- is necessarily
- is possibly
- is known/believed (by Ann) to be
- is permitted to be
- is obliged to be
- is now
- will be
- can do something to ensure that he is

The basic modal language is a generic formal language with unary operators that have been used to reason about situations involving modal notions. This language is defined as follows:

Definition A.1 (The Basic Modal Language) Let $\mathsf{At} = \{p, q, r, \ldots\}$ be a set of sentence letters, or atomic propositions. We also include two special propositions '$\top$' and '$\bot$' meaning 'true' and 'false' respectively. The set of well-formed formulas of modal logic is the smallest set generated by the following grammar:

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box\varphi \mid \Diamond\varphi$$

where $p \in \mathsf{At}$. $\triangleleft$

Examples of modal formulas include: $\Box\bot$, $\Box\Diamond\top$, $p \rightarrow \Box(q \wedge r)$, and $\Box(p \rightarrow (q \vee \Diamond r) \leftrightarrow \Diamond\Box p)$.

One language, many readings. There are many possible readings for the modal operators '$\Box$' and '$\Diamond$'. Here are some samples:

- **Alethic Reading**: $\Box\varphi$ means 'φ is necessary' and $\Diamond\varphi$ means 'φ is possible'.
- **Deontic Reading**: $\Box\varphi$ means 'φ is obligatory' and $\Diamond\varphi$ means 'φ is permitted'. In this literature, typically 'O' is used instead of '$\Box$' and 'P' instead of '$\Diamond$'.
- **Epistemic Reading**: $\Box\varphi$ means 'φ is known' and $\Diamond\varphi$ means 'φ is consistent with the current information'. In this literature, typically 'K' is used instead of '$\Box$' and 'L' instead of '$\Diamond$'.
- **Temporal Reading**: $\Box\varphi$ means 'φ will always be true' and $\Diamond\varphi$ means 'φ will be true at some point in the future'.

There are many interesting arguments involving modal notions. Here I will give two examples both of which have been widely discussed by philosophers.

Example A.2 (Aristotle's Sea Battle Argument) A general is contemplating whether or not to give an order to attack. The general reasons as follows:

1. If I give the order to attack, then, necessarily, there will be a sea battle tomorrow
2. If not, then, necessarily, there will not be one.
3. Now, I give the order or I do not.
4. Hence, either it is necessary that there is a sea battle tomorrow or it is necessary that none occurs.

The conclusion is that either it is inevitable that there is a sea battle tomorrow or it is inevitable that there is no battle. So, why should the general bother giving the order? There are two possible formalizations of this argument corresponding to different readings of "if A then necessarily B":

$$\begin{array}{l} A \to \Box B \\ \neg A \to \Box \neg B \\ A \vee \neg A \\ \hline \Box B \vee \Box \neg B \end{array} \qquad\qquad \begin{array}{l} \Box(A \to B) \\ \Box(\neg A \to \neg B) \\ A \vee \neg A \\ \hline \Box B \vee \Box \neg B \end{array}$$

Are these two formalizations the same? If not, which argument is valid?

The second example, provided J. Forrester in 1984, involves the Deontic reading of modal logic.

Example A.3 (The Gentle Murder Paradox) Suppose that Jones murders Smith. Accepting the principle that 'If Jones murders Smith, then Jones ought to murder Smith gently', we can argue that, in fact, Jones *ought* to murder Smith as follows:

1. Jones murders Smith. (M)
2. If Jones murders Smith, then Jones ought to murder Smith gently. $(M \to OG)$
3. Jones ought to murder Smith gently. (OG)
4. If Jones murders Smith gently, then Jones murders Smith. $(G \to M)$
5. If Jones ought to murder Smith gently, then Jones ought to murder Smith. $(OG \to OM)$
6. Jones ought to murder Smith. (OM)

Is this argument valid? Note that reasoning from statement 4. to statement 5. follows a general modal reasoning pattern: if '$X \to Y$' has been established, then we can establish '$\Box X \to \Box Y$'.

In order to answer the questions in the examples above, we need a natural semantics for the basic modal language.

Question A.4 *Can we give a truth-table semantics for the basic modal language?* (**Hint***: there are only 4 possible truth-table for a unary operator. Suppose we want* $\Box A \to A$ *to be valid (i.e., true regardless of the truth value assigned to* A*), but allow* $A \to \Box A$ *and* $\neg\Box A$ *to be false (i.e., for each formula, there is a possible assignment of truth values to* A *which makes the formulas false)*).

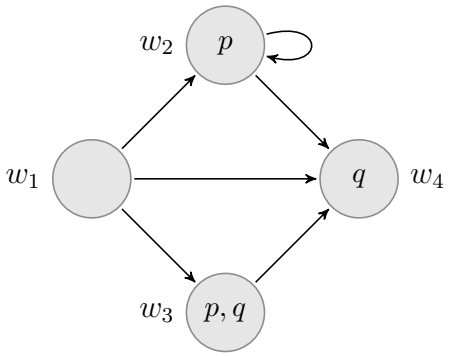
A semantics for the basic modal language was developed by Saul Kripke, Stag Kanger, Jaakko Hinitkka and others in the 1960s and 1970s. Formulas are interpreted over graph-like structures:

Definition A.5 (Relational Structure) A **Relational Structure** (also called a possible worlds model, Kripke model or a modal model) is a triple $\mathcal{M} = \langle W, R, V\rangle$ where W is a nonempty set (elements of W are called **states**), R is a relation on W (formally, $R \subseteq W \times W$) and V is a **valuation function** assigning truth values $V(p, w)$ to atomic propositions p at state w (formally $V : \mathsf{At} \times W \to \{T, F\}$ where At is the set of sentence letters). $\triangleleft$

Example A.6 (A Relational/Kripke Structure) Often relational structures are drawn instead of formally defined. For example, the following picture represents the relational structure $\mathcal{M} = \langle W, R, V\rangle$ where $W = \{w_1, w_2, w_3, w_4\}$,

$$R = \{(w_1, w_2), (w_1, w_3), (w_1, w_4), (w_2, w_2), (w_2, w_4), (w_3, w_4)\}$$

and $V(p, w_2) = V(p, w_3) = V(q, w_3) = V(q, w_4) = T$ (with all other propositional variables assigned F at the states).



Formulas of the basic modal language are interpreted at states in a relational structure.

Definition A.7 (Truth of Modal Formulas) Truth of a modal formula φ at a state w in a relational structure $\mathcal{M} = \langle W, R, V\rangle$, denoted $\mathcal{M}, w \models \varphi$ is defined inductively as follows:

1. $\mathcal{M}, w \models p$ iff $V(p, w) = T$ (where $p \in \mathsf{At}$)
2. $\mathcal{M}, w \models \top$ and $\mathcal{M}, w \not\models \bot$
3. $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$

4. $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$

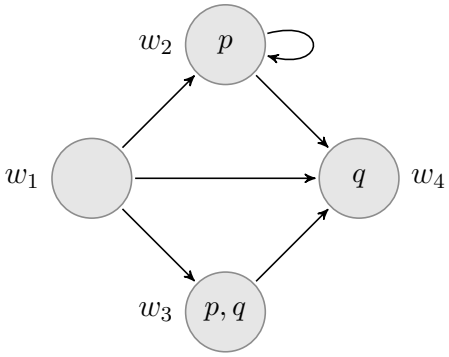
5. $\mathcal{M}, w \models \Box\varphi$ iff for all $v \in W$, if wRv then $\mathcal{M}, v \models \varphi$

6. $\mathcal{M}, w \models \Diamond\varphi$ iff there is a $v \in W$ such that wRv and $\mathcal{M}, v \models \varphi$ $\triangleleft$

Two remarks about this definition. First, note that truth for the other boolean connectives $(\rightarrow, \vee, \leftrightarrow)$ is not given in the above definition. This is not necessary since these connectives are *definable* from '$\neg$' and '$\wedge$'. [6] As an exercise, make sure you can specify the truth definition in the style of the Definition above for each of the boolean connectives not mentioned. Second, note the analogy between '$\Box$' and a universal quantifier and '$\Diamond$' and a existential quantifier.

Question A.8 *Let $\mathcal{M} = \langle W, R, V\rangle$ be a relational model. Give the recursive definition of a function $\overline{V} : WFF_{ML} \rightarrow \wp(W)$ so that $\overline{V}(\varphi) = \{w \in W \mid \mathcal{M}, w \models \varphi\}$ (recall that $\wp(W) = \{X \mid X \subseteq W\}$ is the powerset of W).*

Example A.9 *To illustrate the above definition of truth of modal formula, recall the relational structure from Example A.6:*



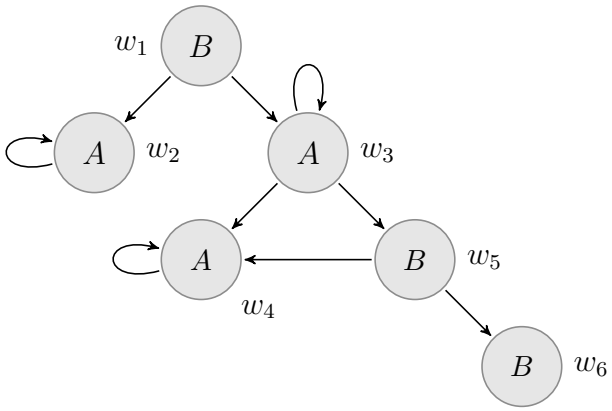
- $\mathcal{M}, w_3, \models \Box q$*: w_4 is the only worlds accessible from w_3 and q is true at w_4.*
- $\mathcal{M}, w_1 \models \Diamond q$*: there is a state accessible from w_1 (namely w_3) where q is true.*
- $\mathcal{M}, w_1 \models \Diamond\Box q$*: w_3 is accessible from w_1 and q is true in all of the worlds accessible from w_3.*
- $\mathcal{M}, w_4 \models \Box\bot$*: there are no worlds accessible from w_4, so any formula beginning with '$\Box$' will be true (this is analogous to the fact the universal sentences are true in any first-order structure where the domain is empty). Similarly,*

[6] For example, $\varphi \rightarrow \psi$ can be defined as (i.e., is logically equivalent to) $\neg(\varphi \wedge \neg\psi)$.

any formula beginning with a '$\Diamond$' will be false (again, this is analogous to the fact that existential statements are false in first-order structures with empty domains). ◁

For an extended discussion surrounding the interpreting modal formulas in relational structures, see Chapter 2 of *Modal Logic for Open Minds* by Johan van Benthem.

Question A.10 *Consider the following relational structure.*



1. $\Box A \to \Box\Box A$
2. $\Box\Box A \to \Box A$
3. $\Diamond(\Diamond A \wedge \Diamond B)$
4. $\Diamond\Box\bot$
5. $\Box(\Box A \to A) \to \Box A$

For each formula to the right, list the states where the formula is true.

A.2 Modal Validity

Definition A.11 (Modal Validity) A modal formula φ is **valid in a relational structure** $\mathcal{M} = \langle W, R, V\rangle$, denoted $\mathcal{M} \models \varphi$, provided $\mathcal{M}, w \models \varphi$ for each $w \in W$. A modal formula φ is **valid**, denoted $\models \varphi$, provided φ is valid in all relational structures. ◁

In order to show that a modal formula φ is valid, it is enough to argue informally that φ is true at an arbitrary state in an arbitrary relational structure. On the other hand, to show a modal formula φ is not valid, one must provide a counter example (i.e., a relational structure and state where φ is false).

Fact A.12 *$\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$ is valid.*

Proof. Suppose $\mathcal{M} = \langle W, R, V\rangle$ is an arbitrary relational structure and $w \in W$ an arbitrary state. We will show that $\mathcal{M}, w \models \Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$. We first show that if $\mathcal{M}, w \models \Box\varphi$ then $\mathcal{M}, w \models \neg\Diamond\neg\varphi$. If $\mathcal{M}, w \models \Box\varphi$ then for all $v \in W$, if wRv then $\mathcal{M}, v \models \varphi$. Suppose (to get a contradiction) that $\mathcal{M}, w \models \Diamond\neg\varphi$. Then

there is some v' such that wRv' and $\mathcal{M}, v' \models \neg\varphi$. Therefore, since wRv' we have $\mathcal{M}, v' \models \varphi$ and $\mathcal{M}, v' \models \neg\varphi$ which means $\mathcal{M}, v' \not\models \varphi$. But this is a contradiction, so $\mathcal{M}, w \not\models \Diamond\neg\varphi$. Hence, $\mathcal{M}, w \models \neg\Diamond\neg\varphi$.

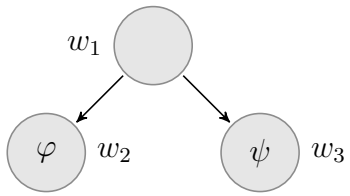
We now show that if $\mathcal{M}, w \models \neg\Diamond\neg\varphi$ then $\mathcal{M}, w \models \Box\varphi$. Suppose that $\mathcal{M}, w \models \neg\Diamond\neg\varphi$. Then there is no state v such that wRv and $\mathcal{M}, v \models \neg\varphi$. Let v be any element of W such that wRv. Then $\mathcal{M}, w \models \varphi$ (since otherwise there would be an accessible state satisfying $\neg\varphi$). Therefore, $\mathcal{M}, w \models \Box\varphi$. QED

Fact A.13 $\Box\varphi \wedge \Box\psi \rightarrow \Box(\varphi \wedge \psi)$ *is valid.*

Proof. Suppose $\mathcal{M} = \langle W, R, V\rangle$ is an arbitrary relational structure and $w \in W$ an arbitrary state. We will show $\mathcal{M}, w \models \Box\varphi \wedge \Box\psi \rightarrow \Box(\varphi \wedge \psi)$. Suppose that $\mathcal{M}, w \models \Box\varphi \wedge \Box\psi$. Then $\mathcal{M}, w \models \Box\varphi$ and $\mathcal{M}, w \models \Box\psi$. Suppose that $v \in W$ and wRv. Then $\mathcal{M}, v \models \varphi$ and $\mathcal{M}, v \models \psi$. Hence, $\mathcal{M}, v \models \varphi \wedge \psi$. Since v is an arbitrary state accessible from w, we have $\mathcal{M}, w \models \Box(\varphi \wedge \psi)$. QED

Fact A.14 $(\Diamond\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\varphi \wedge \psi)$ *is not valid.*

Proof. We must find a relational structure that has a state where $(\Diamond\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\varphi \wedge \psi)$ is false. Note that without loss of generality we can assume that φ and ψ are atomic propositions. Consider the following relational structure:



Call this relational structure $\mathcal{M}$. We have $\mathcal{M}, w_1 \models \Diamond\varphi \wedge \Diamond\psi$ (why?), but $\mathcal{M}, w_1 \not\models \Diamond(\varphi \wedge \psi)$ (why?). Hence, $\mathcal{M}, w_1 \not\models (\Diamond\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\varphi \wedge \psi)$. QED

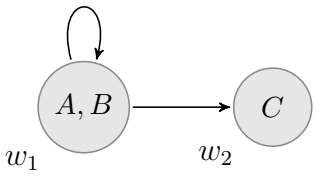
Question A.15 *Determine which of the following formulas valid (prove your answers):*

1. $\Box\varphi \rightarrow \Diamond\varphi$
2. $\Box(\varphi \vee \neg\varphi)$
3. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$
4. $\Box\varphi \rightarrow \varphi$

5. $\varphi \to \Box\Diamond\varphi$

6. $\Diamond(\varphi \vee \psi) \to \Diamond\varphi \vee \Diamond\psi$

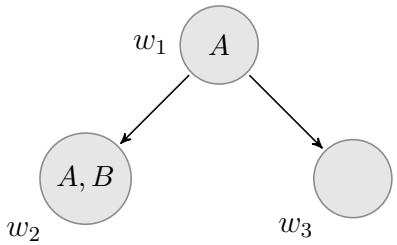
We can now see why the two formalizations of Aristotle's Sea Battle Argument (cf. Exercise A.2) are not "equivalent". They would be the "same" if $\Box(A \to B)$ is (modally) equivalent to $A \to \Box B$. That is if $\Box(A \to B) \leftrightarrow (A \to \Box B)$ is valid. The following relational structure shows that this is not the case:



Here $\Box(A \to B)$ is true at w_1 but $A \to \Box B$ is not true at w_1 (why?). Furthermore, the second formalization of Aristotle's Sea Battle Argument is not valid:

$$\frac{\begin{array}{l}\Box(A \to B)\\ \Box(\neg A \to \neg B)\\ A \vee \neg A\end{array}}{\Box B \vee \Box\neg B}$$

To show this, we must find a relational structure that has a state where all of the premises are true but the conclusion ($\Box B \vee \Box\neg B$) is false. The following relational structure does the trick (w_1 satisfies all of the premises but not the conclusion):

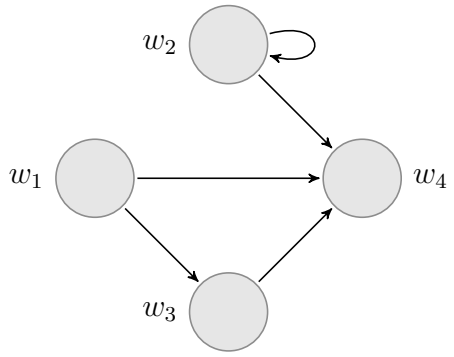


A.3 Definability

Question A.8 shows that we can assign to every modal formula φ a set of states in a relational structure $\mathcal{M} = \langle W, R, V\rangle$ (i.e., the set $\overline{V}(\varphi)$ of states where φ is true in $\mathcal{M}$). We sometime write $(\varphi)^{\mathcal{M}}$ for this set. What about the converse: given and arbitrary set, when does a formula uniquely pick out that set?

Definition A.16 (Definable Subsets) Let $\mathcal{M} = \langle W, R, V\rangle$ be a relational structure. A set $X \subseteq W$ is **definable in** $\mathcal{M}$ provided $X = (\varphi)^{\mathcal{M}} = \{w \in W \mid \mathcal{M}, w \models \varphi\}$ for some modal formula φ. ◁

Example A.17 (Defining states with modal formulas) All four of the states in the relational structure below are uniquely defined by a modal formula:



- $\{w_4\}$ is defined by $\Box\bot$
  (w_4 *is the only "dead-end" state*)
- $\{w_3\}$ is defined by $\Diamond\Box\bot \wedge \Box\Box\bot$
  (w_3 *can only see a "dead-end" state*)
- $\{w_2\}$ is defined by $\Diamond\Diamond\Diamond\top$
  (w_2 *is the only state where 3 steps can be taken*)
- $\{w_1\}$ is defined by $\Diamond(\Diamond\Box\bot \wedge \Box\Box\bot)$
  (w_1 *is the only state that can see* w_3)

Given the above observations, it is not hard to see that *all* subsets of $W = \{w_1, w_2, w_3, w_4\}$ are definable (why?). However, note that even in finite relational structures, not all subsets may be definable. A problem can arise if states cannot be distinguished by modal formulas. For example, if the reflexive arrow is dropped in the relational structure above, then w_2 and w_3 *cannot* be distinguished by a modal formula (there are ways to formally prove this, but see if you can informally argue why w_2 and w_3 cannot be distinguished).

The next two definitions make precise what it means for two states to be *indistinguishable* by a modal formula.

Definition A.18 (Modal Equivalence) Let $\mathcal{M}_1 = \langle W_1, R_1, V_1\rangle$ and $\mathcal{M}_2 = \langle W_2, R_2, V_2\rangle$ be two relational structures. We say $\mathcal{M}_1, w_2$ and $\mathcal{M}_2, w_2$ are **modally equivalent** provided

$$\text{for all modal formulas } \varphi,\ \mathcal{M}_1, w_1 \models \varphi \text{ iff } \mathcal{M}_2, w_2 \models \varphi$$

We write $\mathcal{M}_1, w_1 \leftrightsquigarrow \mathcal{M}_2, w_2$ if $\mathcal{M}_1, w_1$ and $\mathcal{M}_2, w_2$ are modally equivalent. (Note that it is assumed $w_1 \in W_1$ and $w_2 \in W_2$) ◁

Definition A.19 (Bisimulation) Let $\mathcal{M}_1 = \langle W_1, R_1, V_1\rangle$ and $\mathcal{M}_2 = \langle W_2, R_2, V_2\rangle$ be two relational structures. A nonempty relation $Z \subseteq W_1 \times W_2$ is called a **bisimulation** provided for all $w_1 \in W_1$ and $w_2 \in W_2$, if $w_1 Z w_2$ then

1. (atomic harmony) For all $p \in \mathsf{At}$, $V_1(w_1, p) = V_2(w_2, p)$.
2. (zig) If $w_1 R_1 v_1$ then there is a $v_2 \in W_2$ such that $w_2 R_2 v_2$ and $v_1 Z v_2$.

3. (zag) If $w_2 R_2 v_2$ then there is a $v_1 \in W_1$ such that $w_1 R_1 v_1$ and $v_1 Z v_2$.

We write $\mathcal{M}_1, w_1 \leftrightarrows \mathcal{M}_2, w_2$ if there is a bisimulation relating w_1 with w_2. ◁

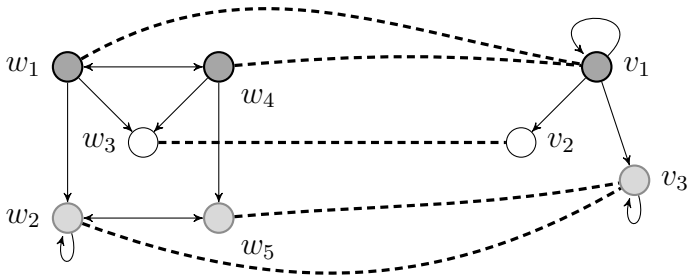
Definition A.18 and A.19 provide two concrete ways to answer the question: *when are two states the same*? The following questions are straightforward consequences of the relevant definitions.

Question A.20 *1. Prove $\leftrightsquigarrow$ and $\leftrightarrows$ are equivalence relations.*

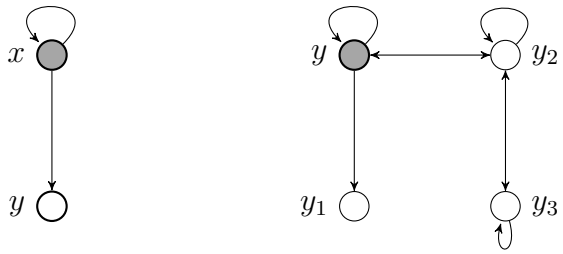
2. Prove that if X is a definable subset of $\mathcal{M} = \langle W, R, V \rangle$, then X is closed under the $\leftrightsquigarrow$ relation (if $w \in X$ and $\mathcal{M}, w \leftrightsquigarrow \mathcal{M}, v$ then $v \in X$).

3. Prove that there is a largest bisimulation: given $\{Z_i \mid i \in I\}$ a set of bisimulations relating the relational structures $\mathcal{M}_1 = \langle W_1, R_1, V_1 \rangle$ and $\mathcal{M}_2 = \langle W_2, R_2, V_2 \rangle$ (i.e., for each $i \in I$, $Z_i \subseteq W_1 \times W_2$ satisfies Definition A.19), show the relation $Z = \bigcup_{i \in I} Z_i$ is a bisimulation.

Example A.21 (Bisimulation) The dashed lines is a bisimulation between the following two relational structures (for simplicity, we do assume that all atomic propositions are false):



On the other hand, there is no bisimulation relating the state x and y in the following two relational structures:



Using Lemma A.22 below, we can *prove* that there is no bisimulation relating x and y. We first note that $\Box(\Diamond\Box\bot \vee \Box\bot)$ is true at state x but not true at state y. Then by Lemma A.22, x and y cannot be bisimilar.

Lemma A.22 (Modal Invariance Lemma) *Suppose $\mathcal{M}_1 = \langle W_1, R_1, V_1\rangle$ and $\mathcal{M}_2 = \langle W_2, R_2, V_2\rangle$ are relational structures. For all $w \in W_1$ and $v \in W_2$, if $\mathcal{M}_1, w \leftrightarrows \mathcal{M}_2, v$ then $\mathcal{M}_1, w \leftrightsquigarrow \mathcal{M}_2, v$.*

Proof. The proof can be found on pages 27 and 28 in *Modal Logic for Open Minds*. QED

Lemma A.23 *Suppose $\mathcal{M}_1 = \langle W_1, R_1, V_1\rangle$ and $\mathcal{M}_2 = \langle W_2, R_2, V_2\rangle$ are* finite *relational structures. If $\mathcal{M}_1, w_1 \leftrightsquigarrow \mathcal{M}_2, w_2$ then $\mathcal{M}_1, w_1 \leftrightarrows \mathcal{M}_2, w_2$.*

Proof. The proof can be found on page 29 in *Modal Logic for Open Minds*. QED

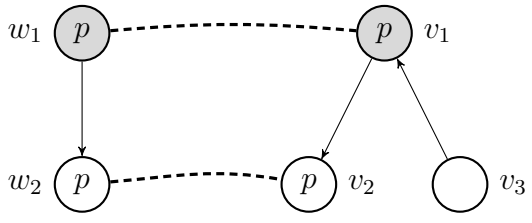
The modal invariance Lemma (Lemma A.22) can be used to prove what can and cannot be expressed in the basic modal language.

Fact A.24 *Let $\mathcal{M} = \langle W, R, V\rangle$ be a relational structure. The* universal operator *is a unary operator $\mathsf{A}\varphi$ defined as follows:*

$$\mathcal{M}, w \models \mathsf{A}\varphi \text{ iff for all } v \in W,\ \mathcal{M}, v \models \varphi$$

The universal operator A is not definable in the basic modal language.

Proof. Suppose that the universal operator is definable in the basic modal language. Then there is a basic modal formula $\alpha(\cdot)$ such[7] that for any formula φ and any relational structure $\mathcal{M}$ with state w, we have $\mathcal{M}, w \models \mathsf{A}\varphi$ iff $\mathcal{M}, w \models \alpha(\varphi)$. Consider the relational structure $\mathcal{M} = \langle W, R, V\rangle$ with $W = \{w_1, w_2\}$, $R = \{(w_1, w_2)\}$ and $V(w_1, p) = V(w_2, p) = T$. Note that $\mathcal{M}, w_1 \models \mathsf{A}p$. Since the universal operator is assumed to be defined by $\alpha(\cdot)$, we must have $\mathcal{M}, w_1 \models \alpha(p)$. Consider the relational structure $\mathcal{M}' = \langle W', R', V'\rangle$ with $W' = \{v_1, v_2, v_3\}$, $R' = \{(v_1, v_2), (v_3, v_1)\}$ and $V'(v_1, p) = V'(v_2, p) = T$. Note that $Z = \{(w_1, v_2), (w_2, v_2)\}$ is a bismulation relating w_1 and v_1 (i.e., $\mathcal{M}, w_1 \leftrightarrows \mathcal{M}', v_1$). These relational structures and bisimulation is pictured below:



[7]The notation $\alpha(\cdot)$ means that α is a basic modal formula with "free slots" such that $\alpha(\varphi)$ is a well formed modal formula with φ plugged into the free slots.

By Lemma A.22, $\mathcal{M}, w_1 \leftrightsquigarrow \mathcal{M}', v_1$. Therefore, since $\alpha(p)$ is a formula of the basic modal language and $\mathcal{M}, w_1 \models \alpha(p)$, we have $\mathcal{M}', v_1 \models \alpha(p)$. Since $\alpha(p)$ defines the universal operator, $\mathcal{M}', v_1 \models \mathsf{A}p$, which is a contradiction. Hence, A is not definable in the basic modal language. QED

Fact A.25 *Let $\mathcal{M} = \langle W, R, V\rangle$ be a relational structure. Define the "exists two" operator $\Diamond_2\varphi$ as follows:*

$\mathcal{M}, w \models \Diamond_2\varphi$ iff there is $v_1, v_2 \in W$ such that $v_1 \neq v_2$, $\mathcal{M}, v_1 \models \varphi$ and $\mathcal{M}, v_2 \models \varphi$

The exist two $\Diamond_2$ operator is not definable in the basic modal language.

Proof. Suppose that the $\Diamond_2$ is definable in the basic modal language. Then there is a basic modal formula $\alpha(\cdot)$ such that for any formula φ and any relational structure $\mathcal{M}$ with state w, we have $\mathcal{M}, w \models \Diamond_2\varphi$ iff $\mathcal{M}, w \models \alpha(\varphi)$. Consider the relational structure $\mathcal{M} = \langle W, R, V\rangle$ with $W = \{w_1, w_2, w_3\}$, $R = \{(w_1, w_2), (w_1, w_3)\}$ and $V(w_2, p) = V(w_3, p) = T$. Note that $\mathcal{M}, w_1 \models \Diamond_2 p$. Since $\Diamond_2$ is assumed to be defined by $\alpha(\cdot)$, we must have $\mathcal{M}, w_1 \models \alpha(p)$. Consider the relational structure $\mathcal{M}' = \langle W', R', V'\rangle$ with $W' = \{v_1, v_2\}$, $R' = \{(v_1, v_2)\}$ and $V'(v_2, p) = T$. Note that $Z = \{(w_1, v_2)\}$ is a bismulation relating w_1 and v_1 (i.e., $\mathcal{M}, w_1 \leftrightarrows \mathcal{M}', v_1$). By Lemma A.22, $\mathcal{M}, w_1 \leftrightsquigarrow \mathcal{M}', v_1$. Therefore, since $\alpha(p)$ is a formula of the basic modal language and $\mathcal{M}, w_1 \models \alpha(p)$, we have $\mathcal{M}', v_1 \models \alpha(p)$. Since $\alpha(\cdot)$ defines $\Diamond_2$, $\mathcal{M}', v_1 \models \Diamond_2 p$, which is a contradiction. Hence, $\Diamond_2$ is not definable in the basic modal language. QED

A.3.1 Defining Classes of Structures

The basic modal language can also be used to define *classes* of structures.

Definition A.26 (Frame) A pair $\langle W, R\rangle$ with W a nonempty set of states and $R \subseteq W \times W$ is called a **frame**. Given a frame $\mathcal{F} = \langle W, R\rangle$, we say the model $\mathcal{M}$ **is based on the frame** $\mathcal{F} = \langle W, R\rangle$ if $\mathcal{M} = \langle W, R, V\rangle$ for some valuation function V. ◁

Definition A.27 (Frame Validity) Given a frame $\mathcal{F} = \langle W, R\rangle$, a modal formula φ is **valid on** $\mathcal{F}$, denoted $\mathcal{F} \models \varphi$, provided $\mathcal{M} \models \varphi$ for all models $\mathcal{M}$ based on $\mathcal{F}$. ◁

Suppose that P is a property of relations (eg., reflexivity or transitivity). We say a frame $\mathcal{F} = \langle W, R\rangle$ has property P provided R has property P. For example,

- $\mathcal{F} = \langle W, R\rangle$ is called a **reflexive frame** provided R is reflexive, i.e., for all $w \in W$, wRw.

- $\mathcal{F} = \langle W, R\rangle$ is called a **transitive frame** provided R is transitive, i.e., for all $w, x, v \in W$, if wRx and xRv then wRv.

Definition A.28 (Defining a Class of Frames) A modal formula φ **defines the class of frames with property** P provided for all frames $\mathcal{F}$, $\mathcal{F} \models \varphi$ iff $\mathcal{F}$ has property P. $\triangleleft$

Remark A.29 (N) ote that if $\mathcal{F} \models \varphi$ where φ is some modal formula, then $\mathcal{F} \models \varphi^*$ where φ^* is any **substitution instance** of φ. That is, φ^* is obtained by replacing sentence letters in φ with modal formulas. In particular, this means, for example, that in order to show that $\mathcal{F} \not\models \Box\varphi \rightarrow$ it is enough to show that $\mathcal{F} \not\models \Box p \rightarrow p$ where p is a sentence letter. (This will be used in the proofs below).

Fact A.30 $\Box\varphi \rightarrow \varphi$ *defines the class of reflexive frames.*

Proof. We must show for any frame $\mathcal{F}$, $\mathcal{F} \models \Box\varphi \rightarrow \varphi$ iff $\mathcal{F}$ is reflexive.

($\Leftarrow$) Suppose that $\mathcal{F} = \langle W, R\rangle$ is reflexive and let $\mathcal{M} = \langle W, R, V\rangle$ be any model based on $\mathcal{F}$. Given $w \in W$, we must show $\mathcal{M}, w \models \Box\varphi \rightarrow \varphi$. Suppose that $\mathcal{M}, w \models \Box\varphi$. Then for all $v \in W$, if wRv then $\mathcal{M}, v \models \varphi$. Since R is reflexive, we have wRw. Hence, $\mathcal{M}, w \models \varphi$. Therefore, $\mathcal{M}, w \models \Box\varphi \rightarrow \varphi$, as desired.

($\Rightarrow$) We argue by contraposition. Suppose that $\mathcal{F}$ is not reflexive. We must show $\mathcal{F} \not\models \Box\varphi \rightarrow \varphi$. By the above Remark, it is enough to show $\mathcal{F} \not\models \Box p \rightarrow p$ for some sentence letter p. Since $\mathcal{F}$ is not reflexive, there is a state $w \in W$ such that it is not the case that wRw. Consider the model $\mathcal{M} = \langle W, R, V\rangle$ based on $\mathcal{F}$ with $V(v, p) = T$ for all $v \in W$ such that $v \neq w$. Then $\mathcal{M}, w \models \Box p$ since, by assumption, for all $v \in W$ if wRv, then $v \neq w$ and so $V(v, p) = T$. Also, notice that by the definition of V, $\mathcal{M}, w \not\models p$. Therefore, $\mathcal{M}, w \models \Box p \wedge \neg p$, and so, $\mathcal{F} \not\models \Box p \rightarrow p$. QED

Fact A.31 $\Box\varphi \rightarrow \Box\Box\varphi$ *defines the class of transitive frames.*

Proof. We must show for any frame $\mathcal{F}$, $\mathcal{F} \models \Box\varphi \rightarrow \Box\Box\varphi$ iff $\mathcal{F}$ is transitive.

($\Leftarrow$) Suppose that $\mathcal{F} = \langle W, R\rangle$ is transitive and let $\mathcal{M} = \langle W, R, V\rangle$ be any model based on $\mathcal{F}$. Given $w \in W$, we must show $\mathcal{M}, w \models \Box\varphi \rightarrow \Box\Box\varphi$. Suppose that $\mathcal{M}, w \models \Box\varphi$. We must show $\mathcal{M}, w \models \Box\Box\varphi$. Suppose that $v \in W$ and wRv. We must show $\mathcal{M}, v \models \Box\varphi$. To that end, let $x \in W$ be any state with vRx. Since R is transitive and wRv and vRx, we have wRx. Since $\mathcal{M}, w \models \Box\varphi$, we have $\mathcal{M}, x \models \varphi$. Therefore, since x is an arbitrary state accessible from v, $\mathcal{M}, v \models \Box\varphi$.

Hence, $\mathcal{M}, w \models \Box\Box\varphi$, and so, $\mathcal{M}, w \models \Box\varphi \rightarrow \Box\Box\varphi$, as desired.

($\Rightarrow$) We argue by contraposition. Suppose that $\mathcal{F}$ is not transitive. We must show $\mathcal{F} \not\models \Box\varphi \rightarrow \Box\Box\varphi$. By the above Remark, it is enough to show $\mathcal{F} \not\models \Box p \rightarrow \Box\Box p$ for some sentence letter p. Since $\mathcal{F}$ is not transitive, there are states $w, v, x \in W$ with wRv and vRx but it is not the case that wRx. Consider the model $\mathcal{M} = \langle W, R, V\rangle$ based on $\mathcal{F}$ with $V(y, p) = T$ for all $y \in W$ such that $y \neq x$. Since $\mathcal{M}, x \not\models p$ and wRv and vRx, we have $\mathcal{M}, w \not\models \Box\Box p$. Furthermore, $\mathcal{M}, w \models \Box p$ since the only state where p is false is x and it is assumed that it is not the case that wRx. Therefore, $\mathcal{M}, w \models \Box p \wedge \neg\Box\Box p$, and so, $\mathcal{F} \not\models \Box p \rightarrow \Box\Box p$, as desired. QED

Question A.32 *Determine which class of frames are defined by the following modal formulas (prove your answer).*

1. $\Box\varphi \rightarrow \Diamond\varphi$
2. $\Diamond\varphi \rightarrow \Box\varphi$
3. $\varphi \rightarrow \Box\Diamond\varphi$
4. $\Box(\Box\varphi \rightarrow \varphi)$
5. $\Diamond\Box\varphi \rightarrow \Box\Diamond\varphi$

A.4 The Minimal Modal Logic

For a complete discussion of this material, consult Chapter 5 of *Modal Logic for Open Minds* by Johan van Benthem.

Definition A.33 (Substitution) A **substitution** is a function from sentence letters to well formed modal formulas (i.e., $\sigma : \mathsf{At} \rightarrow WFF_{ML}$). We extend a substitution σ to all formulas φ by recursion as follows (we write φ^σ for $\sigma(\varphi)$):

1. $\sigma(\bot) = \bot$
2. $\sigma(\neg\varphi) = \neg\sigma(\varphi)$
3. $\sigma(\varphi \wedge \psi) = \sigma(\varphi) \wedge \sigma(\psi)$
4. $\sigma(\Box\varphi) = \Box\sigma(\varphi)$
5. $\sigma(\Diamond\varphi) = \Diamond\sigma(\varphi)$ $\triangleleft$

For example, if $\sigma(p) = \Box\Diamond(p \wedge q)$ and $\sigma(q) = p \wedge \Box q$ then

$$(\Box(p \wedge q) \to \Box p)^\sigma = \Box((\Box\Diamond(p \wedge q)) \wedge (p \wedge \Box q)) \to \Box(\Box\Diamond(p \wedge q))$$

Definition A.34 (Tautology) A modal formula φ is called a **(propositional) tautology** if $\varphi = (\alpha)^\sigma$ where σ is a substition, α is a formula of propositional logic and α is a tautology. $\triangleleft$

For example, $\Box p \to (\Diamond(p \wedge q) \to \Box p)$ is a tautology because $a \to (b \to a)$ is a tautology in the language of propositional logic and

$$(a \to (b \to a))^\sigma = \Box p \to (\Diamond(p \wedge q) \to \Box p)$$

where $\sigma(a) = \Box p$ and $\sigma(b) = \Diamond(p \wedge q)$.

Definition A.35 (Modal Deduction) A **modal deduction** is a finite sequence of formulas $\langle \alpha_1, \ldots, \alpha_n \rangle$ where for each $i \leq n$ either

1. α_i is a tautology
2. α_i is a substitution instance of $\Box(p \to q) \to (\Box p \to \Box q)$
3. α_i is of the form $\Box\alpha_j$ for some $j < i$
4. α_i follows by modus ponens from earlier formulas (i.e., there is $j, k < i$ such that α_k is of the form $\alpha_j \to \alpha_i$).

We write $\vdash_{\mathbf{K}} \varphi$ if there is a deduction containing φ. $\triangleleft$

The formula in item 2. above is called the **K axiom** and the application of item 3. is called the rule of **necessitation**.

Fact A.36 *If* $\vdash_{\mathbf{K}} \varphi \to \psi$ *then* $\vdash_{\mathbf{K}} \Box\varphi \to \Box\psi$

Proof.

1.	$\varphi \to \psi$	assumption
2.	$\Box(\varphi \to \psi)$	Necessitation 1
3.	$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$	Substitution instance of K
4.	$\Box\varphi \to \Box\psi$	MP 2,3

QED

Fact A.37 $\vdash_{\mathbf{K}} \Box(\varphi \wedge \psi) \rightarrow (\Box\varphi \wedge \Box\psi)$

Proof.

1.	$\varphi \wedge \psi \rightarrow \varphi$	tautology
2.	$\Box((\varphi \wedge \psi) \rightarrow \varphi)$	Necessitation 1
3.	$\Box((\varphi \wedge \psi) \rightarrow \varphi) \rightarrow (\Box(\varphi \wedge \psi) \rightarrow \Box\varphi)$	Substitution instance of K
4.	$\Box(\varphi \wedge \psi) \rightarrow \Box\varphi$	MP 2,3
5.	$\varphi \wedge \psi \rightarrow \psi$	tautology
6.	$\Box((\varphi \wedge \psi) \rightarrow \psi)$	Necessitation 5
7.	$\Box((\varphi \wedge \psi) \rightarrow \varphi) \rightarrow (\Box(\varphi \wedge \psi) \rightarrow \Box\psi)$	Substitution instance of K
8.	$\Box(\varphi \wedge \psi) \rightarrow \Box\psi$	MP 5,6
9.	$(a \rightarrow b) \rightarrow ((a \rightarrow c) \rightarrow (a \rightarrow (b \wedge c)))$	tautology ($a := \Box(\varphi \wedge \psi)$, $b := \Box\varphi, c := \Box\psi$)
10.	$(a \rightarrow c) \rightarrow (a \rightarrow (b \wedge c))$	MP 4,9
11.	$\Box(\varphi \wedge \psi) \rightarrow \Box\varphi \wedge \Box\psi$	MP 8,10

QED

Definition A.38 (Modal Deduction with Assumptions) Let Σ be a set of modal formulas. A **modal deduction of φ from** Σ, denoted $\Sigma \vdash_{\mathbf{K}} \varphi$ is a finite sequence of formulas $\langle \alpha_1, \ldots, \alpha_n \rangle$ where for each $i \leq n$ either

1. α_i is a tautology
2. $\alpha_i \in \Sigma$
3. α_i is a substitution instance of $\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q)$
4. α_i is of the form $\Box\alpha_j$ for some $j < i$ and $\vdash_{\mathbf{K}} \alpha_j$
5. α_i follows by modus ponens from earlier formulas (i.e., there is $j, k < i$ such that α_k is of the form $\alpha_j \rightarrow \alpha_i$). ◁

Remark A.39 (Applying Necessitation) Note that the side condition in item 4. in the above definition is crucial. Without it, one application of Necessitation shows that $\{p\} \vdash_{\mathbf{K}} \Box p$. Using the general fact (cf. Exercise #4, Section 1.2 of Enderton) that $\Sigma; \alpha \vdash_{\mathbf{K}} \beta$ implie $\Sigma \vdash_{\mathbf{K}} \alpha \rightarrow \beta$, we can conclude that $\vdash_{\mathbf{K}} p \rightarrow \Box p$. But, clearly $p \rightarrow \Box p$ cannot be a theorem (why?).

Definition A.40 (Logical Consequence) Suppose that Σ is a set of modal formulas. We say φ is a **logical consequence** of Σ, denoted $\Sigma \models \varphi$ provided for all frames $\mathcal{F}$, if $\mathcal{F} \models \alpha$ for each $\alpha \in \Sigma$, then $\mathcal{F} \models \varphi$. ◁

Theorem A.41 (Soundness) *If* $\Sigma \vdash_{\mathbf{K}} \varphi$ *then* $\Sigma \models \varphi$.

Proof. The proof is by induction on the length of derivations. QED

Theorem A.42 (Completeness) *If* $\Sigma \models \varphi$ *then* $\Sigma \vdash_{\mathbf{K}} \varphi$.

Proof. See any textbook on modal logic for a proof. QED

Remark A.43 (Alternative Statement of Soundness and Completeness) Suppose that Σ is a set of modal formulas. Define the minimal modal logic as the smallest set $\Lambda_{\mathbf{K}}(\Sigma)$ of modal formulas extending Σ that (1) contains all tautologies, (2) contains the formula $\Box(p \to q) \to (\Box p \to \Box q)$, (3) is closed under substitutions, (4) is closed under the Necessitation rule (i.e., if $\varphi \in \Lambda_{\mathbf{K}}$ is derivable *without premises* – $\vdash_{\mathbf{K}} \varphi$ – then $\Box\varphi \in \Lambda_{\mathbf{K}}$) and (4) is closed under Modus Ponens. Suppose $\mathfrak{F}(\Sigma) = \{\varphi \mid \Sigma \models \varphi\}$. Then, soundness and completeness states that $\Lambda_{\mathbf{K}}(\Sigma) = \mathfrak{F}(\Sigma)$.