Глава 1

Введение

Данный текст представляет собой конспект вводного курса “Формальная философия” прочитанный В.В. Долгоруковым в ВШЭ. Лекции, к сожалению, уже не доступны. Сам конспект иногда вольно излагает содержание курса.

1.1 Что такое формальная философия?

Формальная философия (=математическая философия) — направление внутри философии, предполагающее использование точных формальных методов в анализе философской проблематики.

Под точными формальными методами имеется ввиду, прежде всего, математическая логика, теория вероятностей, теория игр и другие математические дисциплины.

Философское рассуждение часто предполагает словесную форму. Приведём в пример классический парадокс лжеца, формулирующейся так: «Критянин сказал, что все критяне лжецы».

Анализ подобного высказывания в философии осуществляется тоже в словесной форме, опираясь на семантику высказывания. Однако, мы могли бы записать этот парадокс в виде формулы логики первого порядка. Пусть K множество всех критян, $P(x)$ означает «x говорит правду». Тогда наш парадокс является следующей формулой

$$\exists k \in K \,:\, P(k) \wedge \forall x \in K \, \neg P(x).$$

Очевидно, что эта формула влечет противоречие, которое мы будем обозначать как $\bot$.

Анализ этого парадокса лежит вне наших целей, поэтому мы его оставим. И всё же, чем будет отличаться формальное выражение этого парадокса от обычного?

Во-первых, формальные предложения можно «вычислять». Формальные выражения можно переписывать по учрежденным правилам и приходить к нетривиальным выводам. Пример такого вывода и является главным героем этих конспектов.

Подобные рассуждения мы будем называть «синтаксическими». Дело в том, что мы формируем синтаксис, набор формальных правил и уходим от интерпретации. Этот процесс крайне похож на математическое рассуждение (им он и является). Подобно тому, как Евклид сформировал свои «Начала», строя геометрию на основе аксиом, так сделаем и мы по отношению к философским понятиям. Начнем же мы с рассуждений о знании.

1.2 Парадокс Фитча

Парадокс Фитча — это парадокс в формальной эпистемологии (науки о знании), появившийся в 1963 году в статье [1]. Суть его располагается в структуре предложения «знать x». Мы не знаем что такое знание, но можем предположить свойства знания, поддающиеся формализации.

Для Фитча эти свойства довольно просты, например, «знать x и y» = «знать x и знать y». Также, если верно «знать x», то x истинно. И самое важное, если «x истинно», то «возможно знать x».

Если принять такие свойства, то понятие «знать» будет проблематично интерпретировать, что и называется парадоксом Фитча, но мы к этому придем позже, сейчас же обсудим интуитивность этих свойств. С первым свойством довольно трудно спорить, в самом деле, знание можно выносить из под конъюнкции (сможете ли Вы привести пример когда нельзя?). Второе свойство опирается на классическое представление о знании, в самом деле, «знать» более сильное утверждение, чем «верить» (подробности можно посмотреть здесь). Третье свойство самое интересное, если что-то истинно, то это «возможно знать», что можно было бы назвать «гносеологическим оптимизмом». Понятие «возможно» мы пока не раскрываем; для этого нам нужно продвинутся дальше, что будет сделано далее в курсе, когда мы подойдем к основам модальной логики, также можно посмотреть здесь.

Важно заметить, что мы здесь не оговариваем конкретное значение «знать» и не проводим никакого семантического анализа. Для нас прежде всего важны свойства, а не понятия. Этот подход несколько напоминает математический тип рассуждения. Редкий математик будет спрашивать «что такое число?», поскольку результативным вопросом будет «какое число?», i.e. каковы его свойства.

В самом деле, вопрос «что такое число 2?» не прокладывает дорогу к результативному анализу, другое дело, спросить «какое это число?». Немедленным ответом будет: 2 — это натуральное число, элемент полугруппы ($\mathbb{N}$,
, $+$) и т.д.

1.2.1 Формулировка

Итак, сформулированные свойства понятия «знать» мы не доказали и не будем этого делать. Мы берём их за основу нашей модели «знать» в качестве аксиом. Последующий анализ этих аксиом приведёт нас к некоторому глубокому выводу, касательно природы подобных систем.

Аксиомы:

1. $\varphi \to \Diamond K\varphi$ (Если φ истинен, то возможно «знать» φ)
2. $K\varphi \to \varphi$ («знать φ» значит φ истинно)
3. $(K(\varphi \wedge \psi) \to (K\varphi \wedge K\psi)$ («знать φ и ψ» значит «знать φ и знать ψ»)

Пользуясь этими аксиомами мы будем анализировать очень естественное утверждение «p истинен и я не знаю p». В самом деле, мы не знаем существуют ли решения (за исключением тривиального, конечно) в целых числах для уравнения $J_3(n) = \mathrm{rad}(n)^m$, где $J_k(n)$ — Жорданов тотиент, $\mathrm{rad}(n)$ — наибольший бесквадратный делитель числа n.

Что же, пусть p такой неизвестный нам факт, тогда возможно узнать, что есть такой факт, которого мы не знаем по аксиоме 1. Дальнейший вывод будет состоять в упрощении выражения.

Вывод:

1. $p \wedge \neg Kp$ — гипотеза
2. $(p \wedge \neg Kp) \to \Diamond K(p \wedge \neg Kp)$ — аксиома 1.
3. $\Diamond K(p \wedge \neg Kp)$ — modus ponens 1, 2
4. $K(p \wedge \neg Kp) \to (Kp \wedge K\neg Kp)$ — аксиома 3.
5. $\Diamond K(p \wedge \neg Kp) \to \Diamond(Kp \wedge K\neg Kp)$ — из 4 по монотонности $\Diamond$
6. $\Diamond(Kp \wedge K\neg Kp)$ — modus ponens 3, 5
7. $K\neg Kp \to \neg Kp$ — аксиома 2
8. $Kp \wedge K\neg Kp \to K\neg Kp$

9. $(Kp \wedge K\neg Kp) \to (Kp \wedge \neg Kp)$
10. $\Diamond(Kp \wedge K\neg Kp) \to \Diamond(Kp \wedge \neg Kp)$
11. $\Diamond(Kp \wedge \neg Kp)$
12. $\Diamond \bot$
13. $\bot$
14. $\neg(p \wedge \neg Kp)$ из 1-13
15. $p \to Kp$ из 14

В шаге 13 мы пришли к противоречию, следовательно, по закону «об исключённом третьем», мы отвергаем нашу посылку. Верно её отрицание. Но тогда получается, что для фактов верна формула $p \to Kp$, то есть если p — факт, то мы его «знаем».

Иными словами, наше формальное рассуждение привело нас к довольно интересному результату. В любой модели понятия «знать», где мы встречаем аксиомы 1–3, мы приходим к парадоксу интерпретации. Наше «знание» любого факта актуально.

С помощью подобного анализа мы можем сформулировать более тонкие и сложные модели, включающие мультиагентность, уверенность в знании и т.п.

1.2.2 Как воспринимать синтаксические аргументы?

Есть две крайние позиции по интерпретации синтаксических аргументов:

- Излишняя впечатлительность: своего рода фатализм, вера в абсолютную силу формальных рассуждений.
- Необоснованный скептицизм: представление о формальном методе, как о приложимом к чему угодно, звучащее так «всё что угодно можно доказать формально, подобрав нужные правила вывода».

Обе позиции ложны. В самом деле, даже если мы доказали некоторое сильное утверждение, например, существует только один вариант развития будущего (философский детерминизм), то мы его доказали в определённой формальной системе. Следовательно, мы можем сделать выводы об конкретной формальной системе, но не о мире в целом.

Для построения формальной системы мы используем посылки/аксиомы, методы вывода и сам вывод — это и есть доказательство в строгом смысле. Однако, сами посылки уже не являются частью доказательства в строгом смысле, поскольку сформированы на основе интуиции, что верно и для методов вывода (мы увидим подобное в заметке про формальную этику).

Кроме того, даже если предположить, что проблем с самой формальной системой нет, у нас возникает проблема интерпретации. Понятие «знать» тоже проблематично интерпретировать (подробнее можно почитать в [2].

И, в тоже время, мы не можем создать формальную систему, доказывающую что угодно (в некотором смысле можем, введя $\bot$ как аксиому). Существует очень ограниченный набор подобных доказательств и их изучение может приводить к весьма необычным результатам.

1.2.3 Идеальные агенты

В психологии существует понятие интроспекции (от лат. introspecto — смотреть внутрь), обозначающее метод, заключающийся в наблюдении собственных психологических процессов. В эпистемической логике мы можем сформулировать принцип позитивной и негативной интроспекции:

$$(P) \quad K\varphi \to KK\varphi$$
$$(N) \quad \neg K\varphi \to K\neg K\varphi$$

Принцип позитивной интроспекции (P) означает, что знание факта ведёт знание о факте знания.

Для негативной интроспекции (N) означает, что не знание факта ведёт к знанию факта не знания. Интуитивно ясно, что негативная интроспекция более проблематична, поскольку читатель этих строк, быть может, не знает теорему о свободе воли из квантовой механики, тогда мало вероятно, что читатель осведомлён о своём не знании.

Однако, негативная интроспекция может быть существенна для «идеальных агентов», e.g. роботов, управляющих атомной электростанцией. Нам желательно, чтобы программа или робот не имели «бессознательного», иными словами, они осведомлены обо всех факторах, которые требуют узнавания...

Внимательный читатель мог заметить, что замета аксиомы 2 на (N) сохраняет парадокс. В самом деле, аксиома 2 используется в шаге 7, но не отличается от (N). Словом, парадокс Фитча можно получить разными способами, для разных агентов. В далеком будущем мы попробуем сформулировать парадокс Фитча в логике объявлений, но уже не с такой ультимативной силой.

1.3 Аргумент Диодора Крона

Диодор Крон (ок. 350 – ок. 284 до н.э.) — представитель Мегарской школы, известен своим аргументом в пользу детерминизма. Аргумент Диодора Крона — это известный философский аргумент в пользу детерминизма. До наших дней от него ничего не сохранилось, кроме посылок и вывода, однако существует реконструкция А. Прайора (отеца современной темпоральной и модальной логики). Он сконструировал формально как этот аргумент мог бы выглядеть, если заполнить все пробелы в рассуждениях. Давайте пройдёмся по этой реконструкции.

Собственно, сам аргумент: Три следующих утверждения несовместимы.

1. Всякое истинное суждение относительно прошлого необходимо.
2. Невозможное не следует из возможного.
3. Возможно нечто, что не существует и не будет существовать.

Вместе с Диодором этот аргумент обсуждали Клеанф и Хрисип. Каждый полагал какое-то из этих утверждений ложным. Клеанф отрицал первое утверждение, Хрисип — второе и Диодор отрицал третье.

Первое означает, что вариант прошлого только один, у прошлого нет развилок. Второе до некоторой степени ясно, но третье обозначает некоторое событие, возможное, но неактуализируемое (его не было и никогда не наступит). Можно привести примеры такого события с разной юмористической окраской вроде наступления мирового коммунизма, приход демократии в России, создание Евангелиона и т. п.

1.3.1 Темпоральная логика

Для изложения аргумента нам будут нужны некоторые обозначения, используемые в темпоральной логике (от лат. tempora «времена»). Как и в прошлый раз, мы не будем уделять много времени формализации этого языка — это мы оставляем на потом. Вместо этого наша задача состоит лишь в построении формализма, в котором присутствует выше означенный парадокс.

- $P\varphi$ — когда-то было, что φ.
- $F\varphi$ — когда-то будет, что φ.
- $H\varphi$ — всегда было, что φ.
- $G\varphi$ — всегда будет, что φ.

Значение этих символов точно неизвестно, предположительно они произошли от словосочетаний **past**, **future**, **has been** и **going to**.

1.3.2 Формализация

Диодор использовал следующие три допущения:

- Всякое истинное суждение относительного прошлого необходимо:

$$A1. \quad P\varphi \to \Box P\varphi$$

- Невозможное не следует из возможного:

$$A2. \quad \Box(\varphi \to \psi) \to (\neg\Diamond\psi \to \neg\Diamond\varphi)$$

Формулировка второй аксиомы довольна необычна если сравнить ее со словесной формой, поэтому давайте докажем, что они действительно эквивалентны.

Доказательство. «Невозможное не следует из возможного» $= (\neg\Diamond\psi \wedge \Diamond\varphi) \to \neg\Box(\varphi \to \psi)$, поскольку $\neg a \vee b \equiv a \to b$ мы получаем формулу
$(\neg\Diamond\psi \wedge \Diamond\varphi) \to \neg\Box(\varphi \to \psi)$
Пользуясь законом $a \to b \implies \neg b \to \neg a$ (проверьте!) мы получаем требуемое

$$\neg(\neg\Diamond\psi \to \neg\Diamond\varphi) \to \neg\Box(\varphi \to \psi)$$

□

- Возможно нечто, что не существует и не будет существовать

$$A3. \quad \Diamond p_0 \wedge \neg p_0 \wedge G\neg p_0$$

Ещё нам нужны дополнительные допущения, которые «likely to have been taken for granted both by Diodorus and by his main opponent» (Prior A. Diodoran Modalities. p. 210): - (Принцип Оккама относительно будущего) Если φ и всегда будет φ, то был момент в прошлом, когда стало φ: - $A4. \quad (\varphi \wedge G\varphi) \to PG\varphi$ - $A5. \quad \Box(\varphi \to HF\varphi)$
Итак, сам аргумент.

1. $\Diamond p_0$ — из $A3$
2. $\Box(p_0 \to HFp_0)$ — из $A5$
3. $\Diamond HFp_0$ — 1, 2, $A2$
4. $\neg p_0 \wedge G\neg p_0$ — $A3$
5. $PG\neg p_0$ — 4, $A4$
6. $\Box PG\neg p_0$ — 5, $A1$
7. $\neg\Diamond HFp_0$ — 6

Таким образом, если заключить этот аргумент верным, нам предстоит либо отбросить одно из трех посылок, либо создать другую модель, в которой формализация не приведёт к парадоксу. Более подробное про эти аспекты темпоральной логики можно посмотреть в [13].

Глава 2

Модальная метафизика

Под модальностью (лат. modus — способ, вид) понимается способ, вид бытия или некоторого события, которые, традиционно, подразделяются на категории необходимых и возможных. Мы будем их обозначать следующим образом

- «$\Box\varphi$» — «метафизически необходимо, что φ»,
- «$\Diamond\varphi$» — «(возможно,) что φ».

Примеры: «$\Box(2+2=4)$», «$\neg\Box$ Москва — столица РФ».

Некоторые факты метафизически необходимы. Иными словами, во всех возможных исходах, обстоятельствах (в терминологии Крипке «во всех возможных мирах») какие-то утверждения будут истинны. Математические теоремы относят к такого рода фактам. Все зависимости от обстоятельств нашей жизни $3^2+4^2=5^2$ будет истинно, поэтому мы говорим, что оно верно с *необходимостью*. С другой стороны если что-то возможно, мы предполагаем, что существуют иные обстоятельства (иной возможный мир), в которых верно иное. Например, если вы подбросите монетку, у вас выпадет решка. Это возможно, но не необходимо.

Наши модальности связаны следующими законами двойственности:

- $\Box\varphi \equiv \neg\Diamond\neg\varphi$
- $\Diamond\varphi \equiv \neg\Box\neg\varphi$

Интуиция, стоящая за ними, проста. Необходимость означает невозможность иного исхода, i.e. во всех обстоятельствах, во всех вариантах будущего необходимое утверждение будет истинным. Возможность означает не необходимость иного исхода, i.e. противоположный исход не обязателен. Дополнительный пример, $\Box(\vec{F}=m\vec{a})$ является, строго говоря, ложным утверждением. Ведь существуют среды и скорости, для которых второй закон Ньютона не выполняется. В «Диалогах о естественной религии» Юма поднимается следующий вопрос «являются ли физические законы необходимыми»? В самом деле, для любого предмета, лежащего на земле (покоящегося) мы знаем формулу потенциальной энергии $P=mgh$, верно ли, что $\Box(P=mgh)$? В отличии от математических законов, физические законы не содержат внутри себя какого-то доказательства, следовательно, у нас нет никаких веских доводов считать, что эти законы являются необходимыми. Иными словами $\Diamond\neg(P=mgh)$. На самом деле этот вопрос можно обсуждать более широко с точки зрения проблемы Юма (физические законы не являются необходимыми), но это выходит за рамки наших целей.

Кроме двойственности нам нужны ещё два самоочевидных утверждения:

- $\Box T$
- $\Box(\varphi\wedge\psi) \equiv (\Box\varphi\wedge\Box\psi)$

2.1 Необходимость тождества С. Крипке

Сол Крипке (1940–2022) считается одним из самых влиятельных философов.

имена или дескрипции

- $a = b$ или $\Box(a = b)$
- $P(a)$ или $\Box P(a)$

Обладание свойством P и необходимое обладание свойством P — это разные вещи. Однако, аргумент Крипке показывает, что свойство «быть равным чему-то» имлицирует «необходимо быть равным чему-то». В случае с равенством мы можем сказать, что равенство необходимо, иными словами,

$$a = b \to \Box(a = b).$$

В общем же случае закон $\varphi \to \Box\varphi$ неверен.

Равенство определяется принципом Г. Лейбница, который формулируется как «если объекты тождественны, то они не различимы» или формально:

1. $\forall x \forall y(x = y \to (P(x) \leftrightarrow P(y)))$
2. $\forall x \forall y(x = y \to (P(x) \to P(y)))$

Вторая формулировка слабее первой, ее мы и будем использовать.

Упражнение 1. Докажите, что $A \to (B \to C) \implies B \to (A \to C)$.

Мы будем пользоваться принципом Лейбница и правилами Гёделя для доказательства необходимости тождества. Само доказательство довольно простое.

Доказательство. 1. $a = b \to (\Box(a = b) \to \Box(a = b))$ — принцип Лейбница

2. $\Box(a = a) \to ((a = b) \to \Box(a = b))$ — по перестановке антецедентов
3. $\Box(a = a)$
4. $(a = b) \to \Box(a = b)$

□

Разные философы осмысляли этот аргумент по разному:

- У. Куайн: Квантификация в модальных объектах порочна. Такого рода доказательство является аргументом против самой возможности квантифицированной модальной логики.
- С. Крипке: имена — особенные конструкции, имена являются «жесткими десигнаторами», то есть, во всех возможных мирах отсылают к одному и тому же объекту непосредственным образом.

2.1.1 Два метафизических различия

Давайте посмотрим внимательно на два эпистемически-метафизических различия:

- априорные и апостериорные суждения:
  - Вода кипит при 100° Цельсия — это апостериорное суждение, истинность которого проверяется на опыте.
  - $a \wedge \neg a$ — это априорное суждение, его истинность не связанна с опытом.
- контингентные и необходимые суждения:
  - p является необходимым суждением, если $\Box p \vee \Box\neg p$ истинно, то есть необходимо либо p, либо не p.

– p является контингентным суждением, если $\Diamond p \wedge \Diamond \neg p$ истинно, то есть возможно и p, и не p.

Традиционно в философии априорные суждения отождествлялись с необходимыми, а апостериорные с контингентными. Однако, Крипке приводит примеры и иных комбинаций:

- Контингетные и априорные:
  - Рассмотрим утверждение: «Линия перпендикулярна плоскости». Это знание априорно, так как его можно понять из геометрических принципов без обращения к опыту, но оно контингентно, так как могло бы быть иначе в другой системе геометрии (например, в неевклидовой геометрии).
  - Рассмотрим утверждение: «Длина эталона метра составляет один метер». Это утверждение является контингентным, поскольку оно зависит от исторического и культурного контекста, в котором было принято это определение И оно является априорным, потому что верно по определению: эталон метра определяет что такое один метр.
  - «Гагарин полетел в космос как раз в День Космонавтики». Разберите в качестве упражнения.
- Необходимые и апостериорные:
  - Рассмотрим утверждение: «Вода есть H_2O». По мнению Сола Крипке, это утверждение апостериорное, так как мы узнали его через научное исследование, но оно также является необходимым, потому что в любом возможном мире вода (если она есть) будет иметь молекулярную структуру H_2O.
  - Рассмотрим утверждение: «Геспер (вечерняя звезда, Венера) есть Фосфор (утренняя звезда, Венера)». Древние греки не знали, что наблюдаемая ими утренняя звезда и вечерняя звезда — это один и тот же объект, планета Венера. Иными словами, референт Геспера и референт Фосфора есть планета Венера, один и тот же объект, но $\Box(a = a)$, поэтому это утверждение необходимое, однако, оно и апостериорное, так как мы узнали об этом вследствие опыта.

2.2 Невозможность нечётких объектов Г. Эванса.

Об этом аргументе говорить чуть сложнее, поскольку сначала нужно определить понятие «нечёткого объекта».

Обозначения у нас будут следующие

- $\nabla(a = b)$ — «истинностное значение $a = b$ не детерминировано» (sentence «$a = b$» is of indeterminate truth value),
- $\lambda x[P(x)]$ — «свойство P».

Правила формального вывода:

- $\lambda x[P(x)](a) = P(a)$ — β-редукция
- $\neg\nabla(a = a)$ — аксиома

Мы вновь будем использовать принцип Лейбница для доказательства.

Доказательство. 1. $\nabla(a = b)$ допущение

2. $\lambda x[\nabla(a = x)](b)$ из 1
3. $\neg\nabla(a = a)$ аксиома
4. $\neg\lambda x[\nabla(a = x)](a)$

5. $a \neq b$ из 2, 3 по принципу Лейбница

6. противоречие 1, 5

□

Как указано у Эванса «contradicting the assumption, with which we began that the identity statement 'a = b' is of indeterminate truth value», что, конечно, не является прямым противоречием в логическом смысле, поскольку $a \neq b$ не является отрицанием допущения. Мы можем получить прямое противоречие, если воспользуемся аргументом Крипке о необходимости тождества. Подробнее можно посмотреть в [5].

Глава 3

Эпистемическая логика

Настала пора формализовать всё то, что мы обсуждали в первой части конспектов. Мы начали с анализа классических синтаксических аргументов, но не вводили какую-то семантику строгим образом. Мы брали некоторые содержательные аксиомы и записывали их формально, предполагая, что эта запись является носителем той же мысли, что и словесная форма. Таким образом у нас получалось исчисление, описывающее правила манипуляции символами. Теперь будет более строгий инструмент вместе с семантикой.

Будем вводить этот аппарат в эпистемической логике (от грек. ἐπιστήμη — знание), начиная со следующих обозначений.

- «$K_i\varphi$» — «агент i знает, что φ»
- «$\hat{K}_i\varphi$» — «агент i допускает, что φ»
- $\hat{K}_i\varphi := \neg K_i \neg\varphi$ (буквально «я не знаю обратного»)

У вершин будет хорошая интерпретация, например $K_i^?\varphi := K_i\varphi \vee K_i\neg\varphi$ — агент информирован о φ. $\neg K_i^?\varphi = \hat{K}_i\varphi \wedge \hat{K}_i\neg\varphi$ — агент не информирован.

Теперь поговорим о значении стрелок:

- черные — следствия ($a \to a \vee b$ или $a \wedge b \to a$)
- зелёные — субконтрарность (не совместимы по ложности, но по истинности)
- синие — контрарность (не совместимы по истине, i.e. не могут быть одновременно истинными, но могут быть одновременно ложными)
- красные — контрадикторные (не совместимые)

3.0.1 Правильные формулы

Мы будем использовать форму BNF. Правильно построенная формула языка эпистемической логики определяется следующей грамматикой:

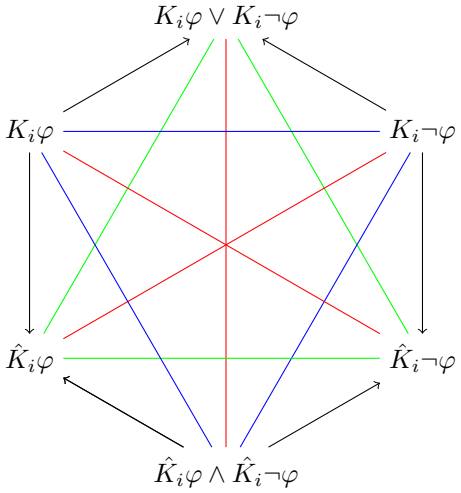
$$\varphi, \psi ::= p|\neg\varphi|(\varphi \wedge \psi)|(\varphi \to \psi)|K_i\varphi|\hat{K}_i\varphi.$$

3.0.2 Модель Крипке

Моделью Крипке называется $\mathcal{M} = (W, (\sim_i)_{i\in\mathcal{A}}, V)$, где $\mathcal{A}$ — непустое конечное множество агентов, W — непустое множество возможных миров, $\sim_i$ — отношение эквивалентности на W для агента i, $V : Var \mapsto \mathcal{P}(W)$ — функция оценки, i.e. $V(p) \subseteq W$ (мы указываем миры, где переменная истинна).

Тогда у нашей модели будет следующая семантика

Рис. 3.1: Эпистемический шестиугольник



1. $\mathcal{M}, x \models p \Leftrightarrow x \in V(p)$
2. $\mathcal{M}, x \models \neg\varphi \Leftrightarrow \mathcal{M}, x \not\models \varphi$
3. $\mathcal{M}, x \models \varphi \wedge \psi \Leftrightarrow \mathcal{M}, x \models \varphi \wedge \mathcal{M}, x \models \psi$
4. $\mathcal{M}, x \models K_i\varphi \Leftrightarrow \forall y(x_i \sim_i y \implies \mathcal{M}, y \models \varphi)$

Мы говорим про семейство эпистемических логик. можно немного изменить начальные положения. вопрос какие свойства знания нам нужны остаётся открытым. мы можем изменить отношение достижимости. мы полагаем, что в нашей модели есть отношение эквивалентности: рефлексивность + симметричность + транзитивность = рефлексивность + евклидовость.

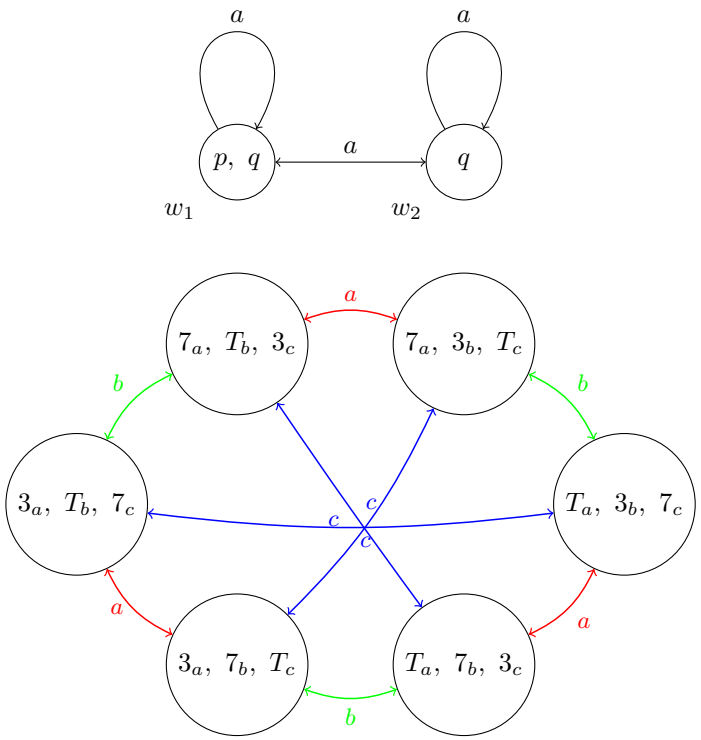
Евклидовость $x \sim y \wedge x \sim z \to y \sim z$

Соответствие между свойствами отношения достижимости (классом шкал, шкала — это модель крипке без оценки) и эпистемическими формулами:

- Фактичность: $K_i\varphi \to \varphi$ — общезначима там, где есть рефлексивность
- (Нет названия) — $\varphi \to K_i\hat{K}_i\varphi$ — симметричность
- Позитивная интроспекция: $K_i\varphi \to K_iK_i\varphi$ — общезначима там, где есть транзитивность
- Негативная интроспекция: $\neg K_i\varphi \to K_i\neg K_i\varphi$ — евклидовость

Мы можем изучать логику систематически анализируя свойства отношение достижимости. Например, транзитивность порождает парадокс кучи и проблему веса монетки. Пусть у меня есть монетка и я не знаю сколько она весит. Очевидно, что я не могу (без точных весов) различить миры, где эта монетка весит 6 грамм или 7 грамм. На самом деле я не могу различить мир, где монетка весит n грамм, от мира где монетка весит $n + 1$ грамм и если верна транзитивность, то мы приходим к абсурдному заключению: я не могу различить монетку весом в 6 грамм от монетки весом в 6 килограмм. По этому примеру можно убедиться, что транзитивность подходит не для всех логических моделей.

Есть большая разница в эпистемических операторах (объективное знание) и доксатических (субъективное знание). Понятие «допускаю» можно понимать в эпистемическом смысле, как знание, которое не может быть ложным априори. С другой стороны, Джордано Бруно верил в то, что звёзды — это живые существа, как и люди, и животные. В данном случае он не допускал верное утверждение «Звезды не живые существа» в доксатическом смысле, что мы бы выразили



оператором B (от анг. believe). Конечно, различение знания факта φ как $K\varphi$ и $B\varphi$ не отменяет вопроса об их связи. Возникающая здесь проблема ещё не решена в эпистемологии.

Сейчас мы занимаемся первым приближением знания, которое хорошо работает в описании идеального агента (быть осведомлённым о всех известных нам фактах и всех неизвестных). Для описания реальных агентов нам понадобятся более сложные модели.

Давайте рассмотрим простой пример.

У нас есть два мира, со следующими оценками: $V(p) = w_1$, $V(q) = W$. Давайте посмотрим что истино в этой модели. Получается

$$M, w_1 \models K_a q \wedge \neg K_a p,$$

то есть агент знает q (а значит q верно во всех возможностях), но не знает p, иными словами, он допускает $\hat{K}_a \neg p$.

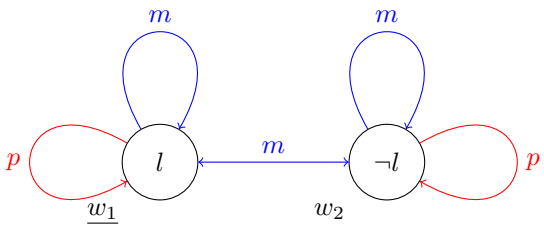
Теперь более содержательный пример. Итак, у нас три агента, каждый из которых выбирает карту из трех вариантов: туза, семёрки, тройки. Поскольку агенты выбирают карты случайно, они не знают какие карты у кого. Однако, если агент a получил семёрку, он не различает миры, где у него семёрка. Иными словами, он может допускать мир $(7_a,\ T_b,\ 3_c)$ или мир $(7_a,\ 3_b,\ T_c)$, почему мы и связываем их красным ребром, что обозначает достижимость этих двух миров.

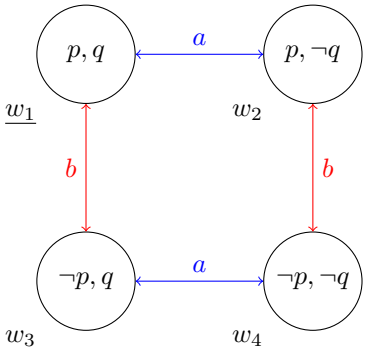
Важно, что конкретный агент не знает «истинного положения вещей», но знает факт, позволяющий ему рассматривать меньшее семейство возможных миров (=возможностей).

Отметим, что петли пропущены для удобства. Мы всегда предполагаем, что агент не различает мир, в котором он находится.

Теперь простой мультиагентный пример. Пьеро любит Мальвину (l), но она не знает об этом: $l \wedge K_p l \wedge \neg K_m l \wedge \neg K_m \neg l = l \wedge K_p l \wedge K_m^? l$.

Теперь рассмотрим двух агентов: Бенедикта и Беатриче. Бенедикт любит Беатриче, но она не знает об этом, а Беатриче любит Бенедикта, но он не знает об этом. Формулы, верные в нашей модели $\mathcal{M}, w_1 \models K_a p \wedge \neg K_b p \wedge \neg K_a q \wedge K_b q$.





3.0.3 Формы коллективного знания

Важное сокращение. $K_i^n\varphi = \underbrace{K_i \dots K_i}_{n \text{ раз}}\varphi$.

Теперь мы будем работать с двумя дополнительными операторами. Первый у нас оператор распространённого знания E (everybody knows), определяемого как $E_G\varphi := \bigwedge_{i\in G} K_i\varphi$.

Второй оператор будет оператором общего знания C (common knowledge), определённый как $C_G\varphi := \bigwedge_{i=0}^{\infty} E_G^i\varphi$. Конечно, тут у нас возникает проблема с интерпретацией оператора $C_G\varphi$, поскольку субъективно можно сомневаться в том, что агент «знает, что знает, что знает, что...». Ещё не очень понятно как вычислять бесконечную конъюкцию. Но в языке нет бесконечных конъюкций, поэтому мы введем семантическое определение этого оператора

Для разрешения этой проблемы, мы введём семантические определения этих операторов (что они значат для нашей модели).

$\mathcal{M}, x \models E_G\varphi \Leftrightarrow \forall y(x(\bigcup_{i\in G} \sim_i)y \implies M, y \models \varphi)$

$\mathcal{M}, x \models C_G\varphi \Leftrightarrow \forall y(x(\bigcup_{i\in G} \sim_i)^+y \implies M, y \models \varphi)$, где R^+ есть транзитивное замыкание отношения R.

Упражнение. Почему мы используем в этом определении объединение, когда с конъюнкцией, обычно, дружит пересечение?

Пример про Деда Мороза. Андрюша уже знает, что Деда Мороза не существует и он знает, что папа Борис знает, что Деда Мороза не существует. Папа знает, что Андрюша знает правду о Деде Морозе. Но Андрюша пока не знает, знает ли папа, что он уже знает правду.

$$E_{ab}^2 p \wedge \neg E_{ab}^3 p \wedge \neg C_{ab} p \wedge \neg K_a K_b K_a p$$

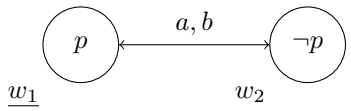
Другой пример. Скороговорка.

Король Пото играл в лото. Королева Пото знала про то, что граф Пото играл в лото, а король Пото не знал про то, что королева Пото знала про то, что король Пото играл в лото.

Проблема Византийских генералов.

Пример сплетен. Аня рассказала Боре страшный секрет (p). Боря рассказал его Свете, но попросил Свету не рассказывать об этом Ане. Света ничего не сказала Ане.

$$E_{abc} p \wedge C_{ab} p \wedge \neg C_{abc} p$$



У общего и распространенного знания имеются следующие свойства. Пусть G является группой агентов, тогда

1. $C_G(\varphi \to \psi) \to (C_G\varphi \to C_G\psi)$ (Нормальность)
2. $C_G\varphi \to (\varphi \wedge E_G C_G\varphi)$ (Неподвижная точка)
3. $C_G(\varphi \to E_G\varphi) \to (\varphi \to C_G\varphi)$ (Индукция)

   Новая аксиоматизация в [10] (K + 4 + $C_G\varphi \to E_G\varphi + C_G E^?_G\varphi \to C^?_G\varphi$), где $E^?_G\varphi := E_G\varphi \vee \neg E_G\varphi$.

Динамическая эпистемическая логика Эпистемические протоколы
Русская карточная задача Проблема победы всех криптографов

3.1 Логика публичных объявлений (PAL)

Логика публичных объявлений (Public announcement logic) состоит в следующем. Пусть у нас есть предложение φ, тогда $!\varphi$ есть публичное сообщение из надежного источника. Мы добавим оператор динамической модальности $[!\varphi]\psi$ — всегда после публичного объявления φ верно, что ψ. Тогда $\varphi, \psi ::= p|\neg\varphi|(\varphi \wedge \psi)|K_i\varphi|[!\varphi]\psi$ и наша эпистемическая модель обновится следующим образом. Пусть $\mathcal{M} = (W, (\sim_i)_{i\in Ag}, V)$ — модель эпистемической логики, тогда $\mathcal{M}^{!\varphi} = (W^{!\varphi}, (\sim_i^{!\varphi})_{i\in Ag}, V^{!\varphi})$ — обновленная модель относительно $!\varphi$, где $W^{!\varphi} = [\varphi]_{\mathcal{M}} = w \in W|\mathcal{M}, w \models \varphi$, $\sim_i^{!\varphi} = \sim_i \cap([\varphi]_{\mathcal{M}} \times [\varphi]_{\mathcal{M}})$ и $V^{!\varphi}(p) = V(p) \cap [\varphi]_{\mathcal{M}}$.

Семантика оператора публичного объявления следующая $\mathcal{M}, w \models [!\varphi]\psi$ если и только если $\mathcal{M}, w \models \varphi \implies \mathcal{M}^{!\varphi}, w \models \psi$.

Упражнение 2. Пусть $\langle!\varphi\rangle\psi := \neg[!\varphi]\neg\psi$. Выпишите семантику для этого оператора.

Рассмотрим пример раскрытия конверта.

Лежит неоткрытый конверт. Пусть агенты вместе открыли конверт и прочитали, что p истинно. Следующие формулы будут истины в этом примере.

- $\mathcal{M}, w_1 \models p$
- $\mathcal{M}, w_1 \models \neg K_a p \wedge \neg K_b p$
- $\mathcal{M}, w_1 \models \neg K_a \neg p \wedge \neg K_b \neg p$
- $\mathcal{M} \models \bigwedge\limits_{i\in a,b} \neg K_i^? p$
- $\mathcal{M} \models C_{ab}\left(\bigwedge\limits_{i\in a,b} \neg K_i^? p\right)$
- $\mathcal{M}^{!\varphi}, w_1 \models C_{ab} p$
- $\mathcal{M}, w_1 \models [!p] C_{ab} p$

Упражнение 3. Всего в колоде три карты 3, 7 и туз. Анна, Борис и Семен вытаскивают по одной карте и у них завязывается следующая беседа, — У меня нет '7' — говорит Анна. — Тогда я знаю, что у тебя 'туз' — отвечает Семен.

У кого какие карты? Воспользуйтесь предыдущей моделью для карт чтобы выписать истинные утверждения используя логику публичного объявления (Подсказка. Анна своим объявлением перенесла нас в мир $\mathcal{M}^{!\neg 7a}$, в какой мир мы попадаем после объявления Семена?)

У PAL есть следующие законы

1. $[!\varphi](\psi \wedge \chi) \leftrightarrow ([!\varphi]\psi \wedge [!\varphi]\chi)$
2. $[!\varphi]\neg\psi \leftrightarrow (\varphi \rightarrow \neg[!\varphi]\psi)$
3. $[!\varphi]K_i\psi \leftrightarrow (\varphi \rightarrow K_i[!\varphi]\psi)$
4. $[!\varphi][!\psi]\chi \leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\chi$
5. $(\varphi \wedge [!\varphi]\psi) \leftrightarrow \langle!\varphi\rangle\psi$

По этим законам можно заметить, что динамическую логику можно сводить к статической.

Всегда ли верно, что $[!\varphi]\varphi$? Тут есть следующий интересный парадокс, принадлежащий Муру: "Идет дождь, но я в это не верю" (или "не знаю"). Формула $p \wedge \neg K_i p$ не является противоречием. Но она самоопровержимая в динамическом смысле (в самом деле, если идет дождь, то как можно в это не верить), значит формула не может быть содержанием успешного публичного объявления. Пусть $\varphi = (p \wedge \neg K_i p)$, тогда

$$[!\varphi]\neg\varphi.$$

Такого рода парадоксы называют иллокутивным самоубийством (по З. Вендлеру), например, "я сплю", "я не намекаю, что..."

Таким образом нужно различать предложения и высказывания.

Упражнение 4 (Чумазые дети)**.** Аня, Борис и Семен вернулись с прогулки. Папа им говорит: хотя бы у одного из вас чумазый лоб. Сейчас я Вам буду задавать вопросы, тот, кто догадается чумазый он или нет — должен всем сказать, что он догадался (но не говорить какой он)

Папа: Кто-то из вас знает, чумазый он или нет?

Дети: Нет!

Папа: А теперь?

Аня: Я знаю?

Борис: И я знаю!

Семен: Ну тогда все понятно: я — ...

Чумазый Семен или нет? А Борис и Аня? Как они догадались? (Подсказка, рассмотрите пространство Хэмминга для 3-битных кодов)

Упражнение 5 (День рождения Шерил)**.** Альберт и Бернард только что познакомились с Шерил, и захотели узнать, когда у нее день рождения. Шерил перечислила список из 10 возможных дат: 15 мая, 16 мая, 19 мая, 17 июня, 18 июня, 14 июля, 16 июля, 14 августа, 15 августа и 17 августа. Потом Шерил сказала Бернарду только ее день рождения, а Альберту — месяц.

— Я не знаю, когда у Шерил день рождения, но я точно знаю, что Бернард тоже не знает — сказал Альберт.

— Сначала я не знал, когда у Шерил день рождения, но теперь знаю, — возразил Бернард.

Когда же у Шерил день рождения?

3.1.1 APAL и парадокс Фитча

Теперь введем язык APAL (arbitrary PAL) и поговорим о парадоксе Фитча (подробнее в [11] и [**?**]). Напомним, что $\langle!\varphi\rangle\psi := \neg[!\varphi]\neg\psi$. Сигнатура языка APAL выглядит следующим образом

$$\varphi, \psi ::= p|\neg\varphi|(\varphi \wedge \psi)|K_i\varphi|[!\varphi]\psi|[!]\psi,$$

где $[!]\psi$ означает, что для любой формулы, после ее объявления истинно ψ; определим $\langle!\rangle\varphi\psi := \neg[!\varphi]\neg\psi$. Семантика следующая

$$\mathcal{M}, x \models \langle!\rangle\varphi \text{ если и только если } \exists\psi \in \text{PAL} : \mathcal{M}, x \models \langle!\psi\rangle\varphi.$$

Тогда у такой системы будут следующие законы

1. $[!](\varphi \wedge \psi) \leftrightarrow ([!]\varphi \wedge [!]\psi)$
2. $[!]\varphi \rightarrow \varphi$
3. $[!]\varphi \rightarrow [!][!]\varphi$
4. $[!]p \leftrightarrow p$
5. $\langle!\rangle[!]\varphi \rightarrow [!]\langle!\rangle\varphi$ (Черча-Росса)
6. $[!]\langle!\rangle\varphi \rightarrow \langle!\rangle[!]\varphi$ (МакКинзи)
7. $K_i[!]\varphi \rightarrow [!]K_i\varphi$
8. $[!]\varphi \rightarrow [!\psi]\varphi$, где ψ не содержит $[!]$

Теперь мы можем воссоздать парадокс Фитча снова. Мы описывали познаваемость как $p \rightarrow \Diamond Kp$. Теперь мы можем воспользоваться динамическим оператором: $p \rightarrow \langle!\rangle K_i p$. Иными сломами, если p истинно, то найдется объявление, после которого агент будет знать p. Заметим, что это будет работать только с атомарными формулами p, поскольку в ином случае мы можем поставить формулу Мура, о который мы говорили ранее. Тогда $(p \wedge \neg K_i p) \rightarrow \langle!\rangle K_i(p \wedge \neg K_i p)$. Наше обобщение не проходит.

Упражнение 6. Докажите, что $(p \wedge \neg K_i p) \rightarrow \langle!\rangle K_i(p \wedge \neg K_i p)$ является противоречивой формулой. Подсказка. Посмотрите как ведет себя формула $(p \wedge \neg K_i p) \rightarrow \neg\langle!\rangle K_i(p \wedge \neg K_i p)$.

В APAL существуют два вида познаваемости: th-познаваемость $\models \varphi \rightarrow \langle!\rangle K_i\varphi$ и wh-познаваемость $\models \langle!\rangle K_i\varphi \vee \langle!\rangle K_i\neg\varphi$ (wh-познаваемость — это "что познаваемость"). Wh-познаваемость не верна для любой формулы, но и не является парадоксальной в смысле Мура (проверьте!). Словами wh-познаваемость означает, что существует такое объявление, после которого агент i знает что p или не p.

Давайте обсудим связь между K_i, $[!\varphi]$ и $[!]$. У нас имеются следующие виды высказываний.

- Позитивные:
$$\varphi, \psi ::= p|\neg p|(\varphi \vee \psi)|(\varphi \wedge \psi)|K_i\varphi|[!\neg\varphi]\psi|[!]\varphi$$
- Сохранные (preserved):
$$\models \varphi \rightarrow [!]\varphi$$
- Успешные (successful):
$$\models [!\varphi]\varphi$$
- Познаваемые:
$$\models \varphi \rightarrow \langle!\rangle K_i\varphi$$

Тогда все позитивные формулы — сохранны, все сохранные формулы — успешные, все успешные формулы — познаваемы, не все успешные формулы позитивны (например, $\neg K_i p$). Также, не все познаваемые формулы успешны (например, $K_a(p \wedge K_b\neg p)$).

Теперь у нас появился инструментарий, который позволяет выразить разные типы познаваемости и прагматическую оценку познаваемости (сохраняется ли познаваемость при объявлении и успешный ли он.)

Упражнение 7. Воспользуйтесь аксиомами редукции, чтобы показать, что $p \rightarrow \langle!\rangle K_i p$ является законом для атомарной формулы p.

Глава 4

Формальная этика

В этой главе мы будем говорить о формальной этике и формальной философии действия.

Деонтическая логика

Начнем с введения языка деонтической логики, мы будем работать с модальной логикой и тремя дополнительными операторами

- $O\varphi$ — обязательно, что φ.
- $P\varphi$ — разрешено, что φ.
- $F\varphi$ — запрещено, что φ.

Конечно они выражаются друг через друга. Часто, когда мы работали с модальной логикой, мы принимали аксиому (T), утверждающую что $\Box\varphi \to \varphi$. Часто мы читаем необходимость в этике как обязательность, однако тогда мы столкнемся с проблемой. Утверждение $O\varphi \to \varphi$ является проблематичным. Поэтому мы принимаем принцип Юма: $O\varphi \not\to \varphi$. В реальном мире, далеком от совершенства, часто нельзя из рассуждений о сущем перейти к рассуждениям о должном Аксиома (D) $O\varphi \to P\varphi$

$O\varphi \to \neg P\neg\varphi$

Аксиома нормальности (K) $O(\varphi \to \psi) \to (O\varphi \to O\psi)$ (с ней связаны проблемы логического всевиденья)

Затруднения общего характера

- Дилемма Йоргенсена:
  - императивы не являются ни истинными, ни ложными
  - логика изучает следование, то есть, сохранение истинности В этике мы имеем дело с императивами. Во-первых, существует скепсис можно ли вообще построить деонтическую модальную логику, поскольку логика норм отличается от обычной логики. Здесь мы имеем дело с императивами: “принеси чай”, “переходи дорогу на зеленый свет”. Такие высказывания ни истины, ни ложны сами по себе, это не ассертив, не репрезентатив. Соответственно возникает проблема следования, ведь мы говорим о конструкции несохраняющей истинность в привычном смысле.

    Для решения этой проблемы мы можем модифицировать понятие следования.
- Принцип Канта
  - $O\varphi \to \Diamond\varphi$

  Если что-то нормативно обязательно, то это должно быть возможно. Допустим Вася взял в долг 1000 рублей, но вчера он проиграл все свои деньги в казино. Получается Вася должен, но не может.

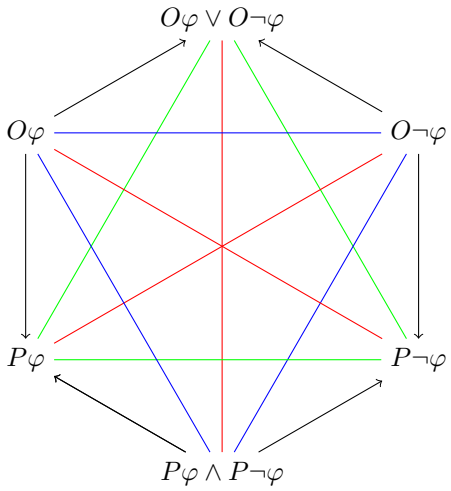


Рис. 4.1: Деонтический шестиугольник

Следующее затруднение это множество Деонтических парадоксов. Парадоксы здесь понимаются в риторическом смысле, а не как антиномии. Начнем с парадокса А. Росса.

$$Op \to O(p \vee q)$$

Очень часто парадоксами в деонтических рассуждениях являются просто формулы без какого-то трудного вывода, который мы делали ранее. Вспомним аксиому (K) $O(\varphi \to \psi) \to (O\varphi \to O\psi)$ и правило Геделя (G)

$$\frac{\varphi}{O\varphi}$$

Доказательство. 1. $p \to (p \vee q)$

2. $O(p \to (p \vee q))$
3. $O(p \to (p \vee q)) \to Op \to O(p \vee q)$
4. $Op \to O(p \vee q)$

□

По каким содержательным причинам мы можем этот вывод атаковать? Росс приводит такой пример, пусть я обязан отправить письмо. Тогда я должен отправить письмо или сжечь его. Неприятный вывод.

Логика свидетельств, есть не просто модальности, но и свидетельство на основании чего мы знаем что-то, то. есть вместо $\Box\varphi$ мы пишем $s : \varphi$, то есть у меня есть основание s для знания или веры факта φ. Парадокс Росса является, скорее, лингвистически-деонтическим

Теперь рассмотрим парадокс доброго самаритянина. Обязательно, что Смит помогает Джону, которого ограбили

$$O(p \wedge q) \to Op$$

Эта формула опасна следующим. Обязательно, что Джона ограбили. Не самый приятный вывод.

$$OK_i p \to Op$$

Ты обязательно должен знать о пожаре, следовательно, обязательно есть пожар.

Парадокс А. Прайора. У меня есть некоторая формула из логики высказываний. Пусть у нас есть противоречие, тогда из него следует все, что угодно. В деонтической логиче это вызывает проблему

$$Op \to (O\neg p \to q).$$

Это деонтический аналог противоречия в логике высказываний. Проблема, которая здесь возникает следующая. Если $(Op \wedge \neg Op) \to q$. То есть, если вы прочитали какой-нибудь маленький нормативный акт про какой-нибудь водопровод и нашли там противоречие, то обязательно грабить, убивать и есть детей. Нам такого логического закона бы не хотелось. Либо мы убираем подобный закон, либо мы работаем с подмножеством непротиворечивых высказываний. Это важное направление в деонтической логике, особенно в прикладных проектах.

Парадокс Р. Чизхолма. Самый сложный парадокс про безусловные и условные обязательства. В практике нормативной у нас часто возникают условные обязательства, i.e. при условии, что Вы гражданин США и имейте годовой доход в $\$11,000$, Ваша ставка налога 10%. Большинство обязательств условные, безусловные обязательства довольно редкие.

Как формализовать условные обязательства? Пусть $p :=$ «Джон идет помогать соседям», $q :=$ «Джон говорит соседям, что идет к ним» (Предупреждает о своем визите).

1. Op (Обязательно, что Джон помогает соседям)
2. $O(p \to q)$ (Обязательно, что если Джон помогает соседям, он их предупреждает)
3. $\neg p \to O\neg q$ (Если Джон не помогает, то обязательно не предупреждает)
4. $\neg p$ (Джон не помогает)

Наша интуиция состоит в том, что эти высказывания не противоречивы. Более того, они должны быть независимы. Однако в действительности мы получим либо противоречие, либо зависимость.

$$Op, O(p \to q), \neg p \to O\neg q, \neg p \vdash_{KD} \bot$$

Доказательство. По modus ponens и аксиоме K из 1 и 2 имеем

$$\dfrac{Op \qquad \dfrac{O(p \to q)}{Op \to Oq}}{Oq}$$

Тогда, применяя modus ponens к 3 и 4 получаем

$$\dfrac{\neg p \qquad \neg p \to O\neg q}{O\neg q}$$

Таким образом мы получаем $Oq \wedge O\neg q$, что означает противоречие в логике KD.

□

Мы можем формализовать высказывания другим способом, но они нарушат независимость (одно высказывание будет следовать из другого). Получается при интерпретации условного обязательства как $O(\varphi \to \psi)$ или как $\varphi \to O\psi$ возникают неминуемые проблемы. Чтобы их разрешить необходимо ввести более тонкий оператор $O(\varphi|\psi)$ (Обязательно φ при условии ψ, подобно условной вероятности). Это разрешит парадокс Чизхолма, но породит новые парадоксы, связанные с условными обязательствами.

Парадокс Д. Форестера («парадокс нежного убийцы»). Часто бывает так, что норма градуирована. Вообще, обязательно p, но если не p, то q. Есть идеальный вариант, но «все бывает».

- Смиту запрещено убивать Джона.
- Если Смит убьет Джона, то он должен сделать это нежно.

- Смит убьет Джона.

Упражнение 8. Формализуйте этот парадокс.

Связь деонтический и алетических операторов: Редукция А. Андерсона [7].

Мы можем попробовать свести деонтические модальности к модальностям необходимости и возможности в политическом смысле. Для этого нам нужно ввести особую константу s, называемой «санкция» (i.e. произошло что-то плохое). Тогда $O\varphi \equiv \Box(\neg\varphi \to s)$. Следовательно $P\varphi \equiv \Diamond(\varphi \wedge \neg s)$ и $F\varphi \equiv \Box(\varphi \to s)$.

Упражнение 9. Сформулируйте деонтические операторы с использованием положительной константы g (идеальный вариант). То есть не обязательно понимать деонтические модальности в наказательном смысле.

В деонтической логике уже на уровне языка у нас возникает множество проблем.

4.1 Философия действия

Философию действия можно разделить, условно, на две части: логику действия и метафизика действия.

Логика действий отвечает на вопросы про логическую структуру действий. Например, если я совершаю действие p, могу ли я совершить $\neg p$. Или если a и b совершают действие p, является ли это действие p конъюнкцией действий a и b? Верно ли, что отказ от действия p ведет к совершению действия $\neg p$? Какие бывают виды действий (по их дедуктивным свойствам)?

С другой стороны, метафизика действия занимается вопросом «что такое действие?». Из каких компонентов состоит действие? Является ли намерение агента причиной для действия? Какие бывают виды действий (по их структуре)?

Базовый формализм, который мы будем использовать, как самый распространенный, это stit-логика. Акроним stit означает *sees to it that* (обеспечивает, что…следит за тем, чтобы…, контролирует соблюдение). Мы введем $[\text{stit}_a\varphi]$ — «агент a обеспечивает, что φ» (альтернативная нотация: $[a\,\text{stit} : \varphi]$).

Но это не значит, что агент совершает действие. Я могу обеспечивать работоспособность моего компьютера, но не быть агентом, который его чинит. stit-модальность — обобщение действия, объединяющее алетические, темпоральные и деонтические модальности.

Эта логика дает изящное решение принципа Канта, более того, у нас возникнет структура группы на этом операторе, можно будет рассуждать о формальной философии ответственности (подробнее в [8]).

Пусть H_m — множество историй, проходящих через момент m. Множество историй, в которых в момент m истинно φ обозначается как $|\varphi|^{\mathcal{M}}_m := h' \in H_m | \mathcal{M}, m/h' \models \varphi$. Мы задаем разбиение H_m как Choice^m_a, где a — агент, m — момент времени. Напомним определение разбиения множества

Определение 4.1.1. Разбиением множества $A = a_1, a_2, \ldots$ называется множество $B = A_1, A_2, \ldots$ такое, что

1. $A_1, A_2, \cdots \subseteq A$,
2. $A_i \cap A_j = \varnothing$, если $i \neq j$ и
3. $A_1 \cup A_2 \cup \cdots = A$.

Пусть$\mathcal{M}$ — множество моментов, на котором задано отношение $<$,

Вместо возможного мира и фактов в нем у нас присутствует история и моменты в ней, хотя, эта разница может быть чисто внешняя и зависит от параметров модели. Введем действие по Челлосу (Chellas) как оператор $[\text{cstit}]_a\varphi$, который устроен таким образом. Множество $\text{Choice}^m_a(h) \subseteq |\varphi|^{\mathcal{M}}_m|$, то есть $\forall h'(h' \in \text{Choice}^m_a(h) \to \mathcal{M}, m/h' \models \varphi)$. В любой истории, которая попадает в наше разбиение, истинно φ.

У действия по Челлосу следующие свойства

- $[\text{cstit}]_a\varphi \to \varphi$ (аналог аксиомы T)
- $[\text{cstit}]_a\varphi \to [\text{cstit}]_a[\text{cstit}]_a\varphi$ (аналог аксиомы 4)
- $\varphi \to [\text{cstit}]_a\neg[\text{cstit}]_a\neg\varphi$ (аналог аксиомы B)
- $\Diamond[\text{cstit}]_{a_1}\varphi_1 \wedge \Diamond[\text{cstit}]_{a_2}\varphi_2 \wedge \cdots \wedge \Diamond[\text{cstit}]_{a_n}\varphi_n \to \Diamond[\text{cstit}]_{a_1,\ldots,a_n}(\varphi_1 \wedge \cdots \wedge \varphi_n)$

Добавим оператор дилебиративности [dstit] (deliberative stit). Для множества моментов $\mathcal{M}$ верно, что $\mathcal{M}, m/h \models [\text{dstit}]_a\varphi$ если, и только если, $\text{Choice}_a^m(h) \subseteq |\varphi|_m^{\mathcal{M}}$ и $H_m \neq |\varphi|_m^{\mathcal{M}}$. Недостаток оператора $[cstit]$ в том, что он не выделяет необходимые факты. Тривиально верно, что $[cstit]_a(E = mc^2)$, вне зависимости от решений агента, что не соотносится с нашей интуицией понятия «обеспечивать». Поэтому имеет смысл убрать из историй безальтернативные факты. Например, $\varphi :=$ «читатель этих строк живет в XXI веке», тогда $[cstit]_a\varphi$ будет, весьма вероятно, тривиально верен. Но для человека, родившегося в начале XXI века этот факт, увы, безальтернативен.

Выпишем stit-модель с деонтическими операторами. Пусть $\mathcal{M} = (\mathcal{A}, T, <, V, \text{Choice}, \text{Value})$, где $\mathcal{A}$ множество агентов, T моменты времени, $<$ отношение на моментах времени, V функция оценки, Choice функция разбиения и $\text{Value} : H_m \mapsto \mathbb{R}$ функция оценки истории. Тогда $\mathcal{M}, m/h \models O\varphi$ если, и только если, $\exists h' \in H_m$:

1. $\mathcal{M}, m/h' \models \varphi$
2. $\forall h''((h'' \in H_m \wedge \text{Value}_m(h') \leq \text{Value}_m(h'')) \to \mathcal{M}, m/h'' \models \varphi$

Тогда у нас возникает тонкое различие между $O\varphi$ и $O[\text{cstit}]_a\varphi$
Свойства

- $O[[cstit]]_a\varphi \to \Diamond[[cstit]]_a\varphi$ (принцип Канта)
- $O[\text{cstit}]_a\top$
- $O[\text{cstit}]_a(\varphi \wedge \psi) \equiv O[\text{cstit}]_a\varphi \wedge O[\text{cstit}]_a\psi$
- $\neg(O[\text{cstit}]_a\varphi \wedge O[\text{cstit}]_a\neg\varphi)$
- $\neg(O[\text{cstit}]_a\varphi \wedge O[\text{cstit}]_b\neg\varphi)$
- $O[\text{cstit}]_a\varphi \to O\varphi$

Коллективные действия $\mathcal{M}, m/h \models [\text{cstit}]_{ab}\varphi$ если и только если $\text{Choice}_a^m(h) \cap \text{Choice}_a^m(h) \subseteq |\varphi|_m^{\mathcal{M}}$.

Глава 5

Онтологический аргумент

5.1 Онтологический аргумент Курта Геделя

Мы попали на территорию аналитической теологии и модальной метафизики. Основные вопросы в этой сфере: формализация онтологического аргумента, логические модели всеведения, логические модели всемогущества. Мы сосредоточимся на логических аспектах доказательства.

Первый онтологический аргумент был сформулирован Ансельмом Кентерберийским (1033–1109), затем разрабатывался Рене Декартом (1596–1650) и Готфридом Лейбницем (1646–1716). После критики Иммануила Канта (1724–1804) считалось, что аргумент не обладает былой убедительной силой. Но Курт Гедель (1906–1978) возродил его, создав чисто логическое доказательство. С тех пор онтологический аргумент стал одним из самый важных и актуальных аргументов аналитической теологии и получил различные версии, например: А. Плантинга, ван Инваген и пр.

У Геделя очень сложная конструкция доказательства, использующая и модальную логику, и логику предикатов высшего порядка. Поэтому до сих пор нет доказательства, что аргумент Геделя не противоречив. Поскольку аргумент модальный, очень важно специфицировать в какой модальной логике мы работаем. Повар vs. Хороший повар.

Гедель использует позитивность (соответсвенно оператор Pos), как основу всего аргумента. Эту позитивность можно понимать как что-то “хорошее”, либо как “присутствие”. Теперь приступим к рассмотрению определений и аксиом, которые мы будем использовать.

5.1.1 Определения

1. $G(x) := \forall X(\mathrm{Pos}(X) \to X(x))$ (Быть Богом — обладать всеми позитивными свойствами)
2. $X \sqsubseteq Y := \forall x(X(x) \to Y(x))$ (Свойство X влечет свойство Y)
3. $\mathrm{ess}(X, x) := X(x) \wedge \forall Y(Y(x) \to \Box(X \sqsubseteq Y))$ (X является существенным свойством объекта x)
4. $E(x) := \forall X(\mathrm{ess}(X, x) \to \Box\exists x X(x))$ (Объект x необходимо существует, i.e. все существенные свойства x с необходимостью экземплифицируются)

Обратите внимание на конъюнкцию в определении *mathrmess*. Это и есть знаменитая поправка Скотта. Если бы ее не было, то мы могли бы подставить пустое свойство, которое стало бы существенным свойством объекта.

5.1.2 Аксиомы

$\mathrm{Pos}(X)$ = свойство X является позитивным, «Positive means positive in the moral aesthetic sense (independently of the accidental structure of the world). Only then are] the axioms true. It may also

mean pure "attribution' as opposed to 'privation' (or containing privation)». Цит. по. (Fitting 2002, p. 146 (Аксиологически)

1. $\mathrm{Pos}(G)$
2. $\mathrm{Pos}(E)$
3. $\neg\mathrm{Pos}(S) \equiv \mathrm{Pos}(\neg S)$
4. $\mathrm{Pos}(S) \to \Box\mathrm{Pos}(S)$
5. $[\mathrm{Pos}(S) \wedge \Box(S \sqsubseteq Q)] \to \mathrm{Pos}(Q)$

Далее выпишем всю цепочку утверждений, которые возможно доказать из этих аксиом, которые и станут доказательством существования Бога.

- $\mathrm{Pos}(S) \to \Diamond\exists x S(x)$ (Если свойство позитивно, то, возможно, существует объект, обладающий этим свойством)
- $\Diamond\exists x G(x)$ (Следовательно, Бог, возможно, существует. Что интересно, в ранних версиях онтологического аргумента это была аксиома)
- $G(x) \to \forall S(\mathrm{Pos}(S) \equiv G(x))$
- $G(x) \to \mathrm{ess}(G, x)$
- $G(x) \to \Box\exists y G(y)$
- $\Box\exists x G(x)$

Упражнение 10. Фильтром над A называют такое семейство подмножеств $F \subset \mathcal{P}(A)$, что

1. $\varnothing \notin F$,
2. $X \in F, X \subseteq Y \implies Y \in F$,
3. $X, Y \in F \implies X \cap Y \in F$.

Если для любого $X \in \mathcal{P}(A)$ верно одно из двух, либо $X \in F$, либо $X^c \in F$. Докажите, что свойство позитивности образует ультрафильтр над множеством позитивных свойств.

Ультрафильтры можно еще заметить в знаменитой теореме К. Эрроу о диктаторе, подробнее в [14].

5.1.3 Доказательство

Общелогический инструментарий. Мы будем использовать следующие правила вывода:

$$(\mathrm{MP}) \quad \frac{\varphi \qquad \varphi \to \psi}{\psi}$$

$$(\Diamond \to) \quad \frac{\varphi \to \psi}{\Diamond\varphi \to \Diamond\psi}$$

1. исключение квантора всеобщности
2. генерализация
3. внесение $\forall$ в скобку

Для доказательства нам понадобятся следующие леммы:

1. $\forall X(\mathrm{Pos(X)} \to \Diamond\exists \mathrm{x X(x)})$

2. $G(x) \to \Box\exists x G(x)$

Сначала рассмотрим как доказательство собирается из лемм, а потом отдельно докажем сами леммы.

$$\dfrac{\dfrac{\dfrac{}{\mathrm{Pos}(G)}\,\mathrm{A4} \qquad \dfrac{\forall X(\mathrm{Pox}(X) \to \Diamond\exists x X(x))}{\mathrm{Pos}(G) \to \Diamond\exists x G(x)}}{\Diamond\exists x G(x)}\,\mathrm{MP} \qquad \dfrac{\dfrac{\dfrac{\dfrac{G(x) \to \Box\exists x G(x)}{\forall y(G(y) \to \Box\exists x G(x))}}{\exists y G(y) \to \Box\exists x G(x)}}{\exists x G(x) \to \Box\exists x G(x)}}{\Diamond\exists x G(x) \to \Diamond\Box\exists x G(x)}\,\Diamond\to}{\Diamond\Box\exists x G(x)}\,\mathrm{MP} \qquad \dfrac{}{\Diamond\Box\exists x G(x) \to \Box\exists x G(x)}\,\mathrm{S5}$$
$$\Box\exists x G(x)\,\mathrm{MP}$$

Литература

[1] Fitch F. A. *Logical Analysis of Some Value Concepts.* The Journal of Symbolic Logic. Vol. 28, No. 2. p. 135–142, 1963.

[2] Lewis, David. *Elusive knowledge* Australasian Journal of Philosophy 74 (4):549–567, 1996.

[3] Шрамко Я. *Некоторые проблемы аналитической эпистемологии.* http://www.ruthenia.ru/logos/number/52/01.pdf

[4] Brogaard, Berit and Joe Salerno. *Fitch's Paradox of Knowability.* The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2019/entries/fitch-paradox/.

[5] Evans G. *Can there be a vague object?* Analysis. 1978. Vol. 38, No. 4, p. 208

[6] Prior A. N. *Diodoran Modalities* The Philosophical Quarterly. 1955. Vol. 5, Issue 20. p. 205–213.

[7] Anderson, A. *A reduction of deontic logic to alethic modal logic.* Mind. Vol. 22. p. 100–103.

[8] Duijf H. *The Logic of Responsibility.* Voids. Dordrecht: Springer, 2022.

[9] Frankfurt, Harry G. *Alternate Possibilities and Moral Responsibility.* Journal of Philosophy 66 (23):829-839, 1969.

[10] Andreas, Herzig and Elise Perrotin. *On the axiomatisation of common knowledge.* 13th Conference on Advances in Modal Logic, AiML 2020, Helsinki, Finland, August 24-28, 2020, 309–328.

[11] van Benthem, J. *What One May Come to Know.* Analysis, vol. 64, no. 2, 2004, 95–105.

[12] Balbiani, P., Baltag, A., Ditmarsch, H. Van, Herzig, A., Hoshi, T., and de Lima, T. *'Knowable' as 'Known After an Announcement'.* The Review of Symbolic Logic, 1(3), 305–334.

[13] Goranko V. *Temporal Logics.* Cambridge University Press; 2023.

[14] Odifreddi P. *Ultrafilters, Dictators, and God.* Finite Versus Infinite. Discrete Mathematics and Theoretical Computer Science. Springer, London, 2000. url: https://doi.org/10.1007/978-1-4471-0751-4_16