Компатибилизм: свободное действие в STIT-логике

Елена Попова (НИУ ВШЭ)
(Международная лаборатория логики, лингвистики и формальной философии)

15 мая 2021 г.

План

План доклада

1. Компатибилизм и свободное действие
2. Логика ветвящегося времени
3. Stit-логика
4. Расширения stit-логики: знание и намерение
5. Логический анализ контрпримеров Франкфурта
6. Выводы

Компатибилизм

Definition

Компатибилизм – концепция совместимости свободы действия с детерминированностью мира.

Definition

Детерминизм – концепция, согласно которой в мире существует только единственное актуальное развитие событий.

Как свободное действие возможно в мире, где нельзя поступить иначе?

.

Мы проживаем единственно возможную последовательность событий, но предполагаем, что онтологически существует ряд альтернативных развитий событий.

Ветвящееся время

Теория ветвящегося времени (Prior, 1967), (Thomason, 1970) служит основой, на которой выстраиваются шкалы stit-логики.

Definition

Шкалы ветвящегося времени представляют из себя структуру

$$\mathcal{T} = (\mathrm{T}, \prec),$$

где T – непустое множество моментов, а $\prec$ – отношение временного предшествования на T, такое что оно транзитивно, иррефлексивно и асимметрично.

Для шкал ветвящегося времени работает запрет на ветвление в прошлое:

(nbb) $\forall x \forall y \forall z(((x \preceq z) \wedge (y \preceq z)) \rightarrow ((x \preceq y) \vee (y \preceq x)))$

Будем обозначать $m_1 \preceq m_2$, где $m_1, m_2 \in \mathrm{T}$, тогда и только тогда, когда $m_1 \prec m_2$ или $m_1 = m_2$.

Ветвящееся время

Definition

Историей называется максимальная линейная последовательность моментов из множества T. Будем обозначать истории как h. Если момент $\mathrm{m} \in \mathrm{h}$, то это означает, история h проходит через момент m. Модель ветвящегося времени допускает одновременное существование нескольких историй в одном моменте. Множество всех историй, проходящих через момент m, обозначается как $\mathcal{H}_{\mathrm{m}}$, и $\mathcal{H}_{\mathrm{m}} = \{\mathrm{h} \mid \mathrm{m} \in \mathrm{h}\}$. $\mathcal{H}_{\mathcal{T}}$ – множество всех историй на $\mathcal{T}$.

В структуре ветвящегося времени под возможным миром будет рассматриваться пара момент/история m/h, где $\mathrm{m} \in \mathrm{T}$ и $\mathrm{h} \in \mathcal{H}_{\mathcal{T}}$.

Ветвящееся время

Шкалы ветвящегося времени показывают этапы развития мира. За ветвлением стоит цель показать, что на определенных этапах могло быть другое положение дел.

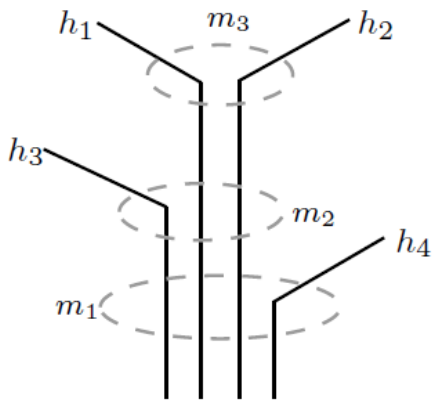


Схема ветвящегося времени для детерминизма.

Ветвящаяся темпоральая логика

Definition

Синтаксис $\mathcal{L}_{\mathrm{BT}}$ порождается следующей грамматикой:

$$\varphi, \psi ::= \mathrm{p} \mid \neg\varphi \mid (\varphi \wedge \psi) \mid \Box\varphi \mid \mathrm{F}\varphi \mid \mathrm{P}\varphi,$$

где $\mathrm{p} \in \mathrm{Var}$, Var – множество пропозициональных переменных.

$\Diamond\varphi \equiv \neg\Box\neg\varphi$, $\mathrm{G}\varphi \equiv \neg\mathrm{F}\neg\varphi$, $\mathrm{H}\varphi \equiv \neg\mathrm{P}\neg\varphi$.

Definition

Модель логики ветвящегося времени является структурой:

$$\mathcal{M} = (\mathrm{T}, \prec, \mathrm{V}),$$

где T – непустое множество моментов, $\prec$ – отношение предшествования во времени, V – функция оценки, которая отображает элементы Var на множество $\mathrm{T} \times \mathcal{H}_{\mathcal{T}}$, элементами которого являются пары m/h.

Ветвящаяся темпоральая логика

Definition

Истинность модальной формулы φ в отмеченной модели $\mathcal{M}, \mathrm{m}/\mathrm{h}$ определяется по индукции:
$\mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \mathrm{p} \Longleftrightarrow \mathrm{m}/\mathrm{h} \in \mathrm{V}(\mathrm{p}), \mathrm{p} \in \mathrm{Var}$
$\mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \neg\varphi \Longleftrightarrow \mathcal{M}, \mathrm{m}/\mathrm{h} \nvDash \varphi$
$\mathcal{M}, \mathrm{m}/\mathrm{h} \vDash (\varphi \wedge \psi) \Longleftrightarrow \mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \varphi \wedge \mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \psi$
$\mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \Box\varphi \Longleftrightarrow \forall \mathrm{h}' \in \mathrm{H}_{\mathrm{m}}(\mathcal{M}, \mathrm{m}/\mathrm{h}' \vDash \varphi)$
$\mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \mathrm{F}\varphi \Longleftrightarrow \exists \mathrm{m}' \in \mathrm{h}(\mathrm{m} \prec \mathrm{m}' \wedge \mathcal{M}, \mathrm{m}'/\mathrm{h} \vDash \varphi)$
$\mathcal{M}, \mathrm{m}/\mathrm{h} \vDash \mathrm{P}\varphi \Longleftrightarrow \exists \mathrm{m}' \in \mathrm{h}(\mathrm{m}' \prec \mathrm{m} \wedge \mathcal{M}, \mathrm{m}'/\mathrm{h} \vDash \varphi)$

Логика ветвящегося времени BTL замкнута относительно правил вывода modus ponens и Гёделя для операторов G, H, $\Box$.

Ветвящаяся темпоральая логика

Минимальный набор аксиом для BTL (Reynolds, 2002):

система K.t

(CL) Все аксиомы классической логики

($\mathrm{K_G}$) $\mathrm{G}(\varphi \to \psi) \to (\mathrm{G}\varphi \to \mathrm{G}\psi)$

(4_G) $\mathrm{G}\varphi \to \mathrm{GG}\varphi$ – транзитивность для G

(GP) $\varphi \to \mathrm{GP}\varphi$ – принцип Оккама

($\mathrm{K_H}$) $\mathrm{H}(\varphi \to \psi) \to (\mathrm{H}\varphi \to \mathrm{H}\psi)$

(4_H) $\mathrm{H}\varphi \to \mathrm{HH}\varphi$ транзитивность для H

(HF) $\varphi \to \mathrm{HF}\varphi$ – принцип Оккама

аксиомы линейности

($\mathrm{LIN_G}$) $\mathrm{F}\varphi \to \mathrm{G}(\mathrm{F}\varphi \vee \varphi \vee \mathrm{P}\varphi)$

($\mathrm{LIN_H}$) $\mathrm{P}\varphi \to \mathrm{H}(\mathrm{F}\varphi \vee \varphi \vee \mathrm{P}\varphi)$

Ветвящаяся темпоральая логика

система S5 для $\Box$

($K_\Box$) $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$

($T_\Box$) $\Box\varphi \to \varphi$ – рефлексивность для $\Box$

($B_\Box$) $\varphi \to \Box\Diamond\varphi$ – симметричность для $\Box$

($4_\Box$) $\Box\varphi \to \Box\Box\varphi$ – транзитивность для $\Box$

($\Box H$) $\Box H\varphi \equiv H\Box\varphi$

($\Box G$) $\Box G\varphi \to G\Box\varphi$

($P\Box$) $P\Box\varphi \to \Box P\varphi$

($\Box\bot$) $G\bot \to \Box G\bot$

Шкалы stit-логики

Stit-логика является формальным аппаратом для рассуждений об агентности. В ее основе лежит идея о том, что агент своими действиями гарантирует определенное положение дел в мире.

Definition

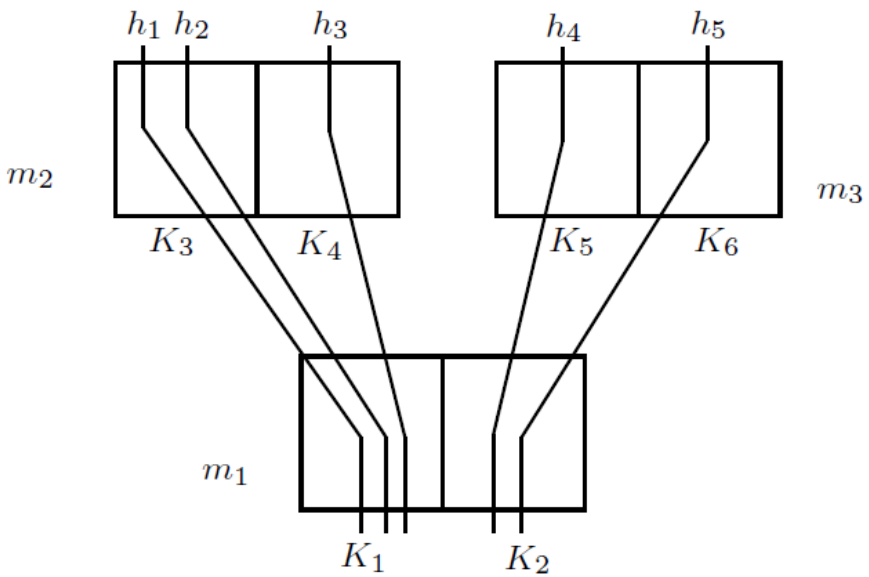
Stit-шкалы представляют из себя структуру:

$$\mathcal{F} = (\mathrm{T}, \prec, \mathrm{A}, \mathrm{Choice}),$$

где $(\mathrm{T}, \prec) = \mathcal{T}$, $\mathrm{A} \neq \varnothing$, функция выбора Choice отображает агента $\mathrm{i} \in \mathrm{A}$ и момент времени $\mathrm{m} \in \mathrm{T}$ на разбиение $\mathrm{Choice}_{\mathrm{i}}^{\mathrm{m}}$ множества историй $\mathcal{H}_{\mathrm{m}}$. Агент i способен выбирать из $\mathrm{Choice}_{\mathrm{i}}^{\mathrm{m}}$ в момент m. Ключевая идея агентности, которая стоит за этой структурой, заключается в том, что в момент m агент i выбирает один из классов эквивалентности K множества $\mathrm{Choice}_{\mathrm{i}}^{\mathrm{m}}$ и тем самым своими действиями гарантирует наступление истории из выбранного класса эквивалентности K.

Шкалы stit-логики

$K \in \text{Choice}_i^m$ концептуально может рассматриваться как действие. $\text{Choice}_i^m(h)$ отсылает к классу эквивалентности K. Элементами K являются истории, которые могут наступить при выборе агентом конкретного K.



Шкалы stit-логики

Существует несколько условий для функции Choice.

1. Независимость агентов. В мультиагентных системах в момент m выбор одним агентом конкретного K не может повлиять на ряд возможных выборов, которые доступны другим агентам.
2. Нет выбора между неразделенными историями. Выбор, доступный для агента в момент m, не предполагает какое-либо различие между неразделенными историями.

Stit-логика

Definition

Синтаксис stit-логики порождается следующей грамматикой:

$$\varphi, \psi ::= \mathrm{p} \mid \neg\varphi \mid (\varphi \wedge \psi) \mid \Box\varphi \mid \mathrm{F}\varphi \mid \mathrm{P}\varphi \mid [\mathrm{stit}]_{\mathrm{i}}\varphi,$$

где $\mathrm{p} \in \mathrm{Var}$, $\mathrm{i} \in \mathrm{A}$.

Definition

Модель stit-логики представляет из себя следующую структуру:

$$\mathcal{M} = (\mathrm{T}, \prec, \mathrm{A}, \mathrm{Choice}, \mathrm{V}),$$

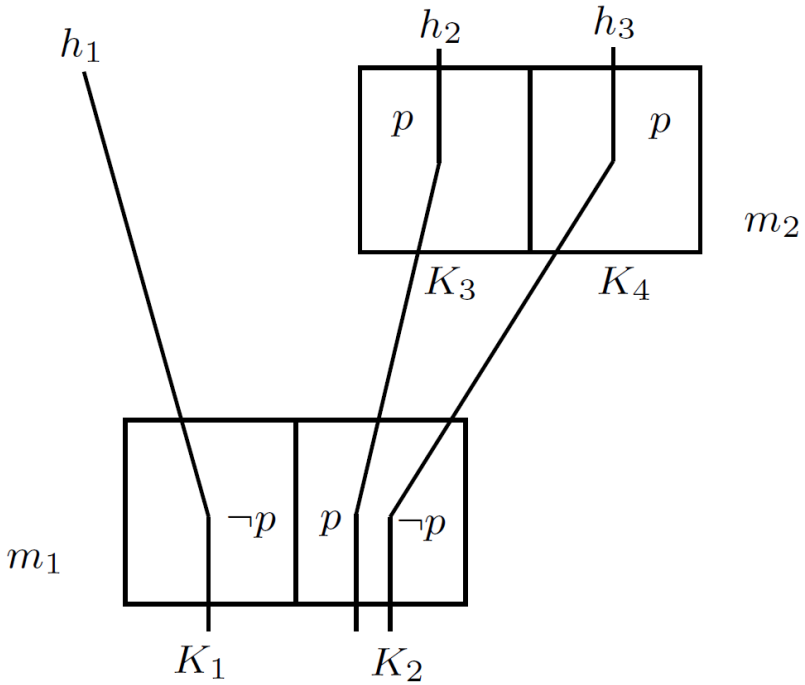
где $(\mathrm{T}, \prec, \mathrm{A}, \mathrm{Choice})$ – stit-шкала, а V – функция оценки, $\mathrm{V} : \mathrm{Var} \mapsto \mathrm{T} \times \mathcal{H}_{\mathcal{T}}$, которая отображает каждую пропозициональную переменную на множество $\{(\mathrm{m}/\mathrm{h}) : \mathrm{m} \in \mathrm{h}\}$.

Stit-логика

Definition

Семантика stit оператора:
$\mathcal{M}, \mathrm{m/h} \vDash [\mathrm{stit}]_{\mathrm{i}}\varphi \Leftrightarrow \forall \mathrm{h}'(\mathrm{h}' \in \mathrm{Choice}_{\mathrm{i}}^{\mathrm{m}}(\mathrm{h}) \rightarrow \mathcal{M}, \mathrm{m/h}' \vDash \varphi)$

Модель stit-логики



Stit-логика

Stit-логика замкнута относительно правил вывода modus ponens и Гёделя для операторов G, H, $\Box$, $[stit]_i$.

BTL + система S5 для оператора $[stit]_i$

($K_{[stit]_i}$) $[stit]_i(\varphi \to \psi) \to ([stit]_i\varphi \to [stit]_i\psi)$

($T_{[stit]_i}$) $[stit]_i\varphi \to \varphi$ – рефлексивность для $[stit]_i$

($B_{[stit]_i}$) $\varphi \to [stit]_i\langle stit\rangle_i\varphi$ – симметричность для $[stit]_i$

($4_{[stit]_i}$) $[stit]_i\varphi \to [stit]_i[stit]_i\varphi$ – транзитивность для $[stit]_i$

Аксиомы связывающие $\Box$ и $[stit]_i$

(INC) $\Box\varphi \to [stit]_i\varphi$

(IA) $(\Diamond[stit]_1\varphi_1 \wedge ... \wedge \Diamond[stit]_n\varphi_n) \to \Diamond([stit]_1\varphi_1 \wedge ... \wedge [stit]_n\varphi_n)$, где $1, ..., n \in A$

Эпистемическое расширение stit-логики

Для анализа действия не менее важно то, каким «путем» агент пришел к этому действию.

.

Совершение действия предполагает наличие у агента определенных эпистемических установок. Агент мог не знать о всех возможных последствиях своего действия или иметь некоторые пристрастные убеждения относительно них.

.

Существует целый ряд вариантов введения эпистемического расширения в stit-логику (см. Broersen, 2011), (см. Xu, 2015), (см. Horty, 2019), (см. Abarca, Broersen, 2020).

Эпистемическое расширение stit-логики

Введем эпистемическое отношение $\sim_i$ для каждого $i \in A$. $\sim_i$ является отношением эквивалентности на множестве моментов T. Доксатическое отношение правдоподобия обозначается как $\preccurlyeq_i$.

Definition

$$\mathcal{M} = (T, \prec, A, Choice, \{\sim_i\}_{i \in A}, \{\preccurlyeq_i\}_{i \in A}, V),$$

где $(T, \prec, A, Choice, V)$ – модель базовой stit-логики, $\sim_i$ – эпистемическое отношение достижимости на множестве T для $i \in A$, $\preccurlyeq_i$ – отношение правдоподобия на множестве T для $i \in A$

Отношение $\sim_i$ является отношением эквивалентности: оно рефлексивно и евклидово.
Отношение $\preccurlyeq_i$ рефлексивно и транзитивно.

Эпистемическое расширение stit-логики

Существуют свойства, связывающие отношения $\sim_i$ и $\preccurlyeq_i$ на T:

$\forall x \forall y((x \preccurlyeq_i y \vee y \preccurlyeq_i x) \to (x \sim_i y))$;

$\forall x \forall y((x \sim_i y) \to (x \preccurlyeq_i y \vee y \preccurlyeq_i x))$.

Введем несколько важных определений:

$\max_{\preccurlyeq_i}(T') := \{m \in T' \mid \forall m' \in T' : m' \preccurlyeq_i m\}$, где $T' \subseteq T$

$[m]_i := \{m' \in T \mid m \sim_i m'\}$

$[\varphi]_{\mathcal{M}} := \{m \in T \mid \mathcal{M}, m/h \vDash \varphi\}$, где $h \in \mathcal{H}_m$

Definition

В stit-логике семантика эпистемического оператора определяется следующим образом:

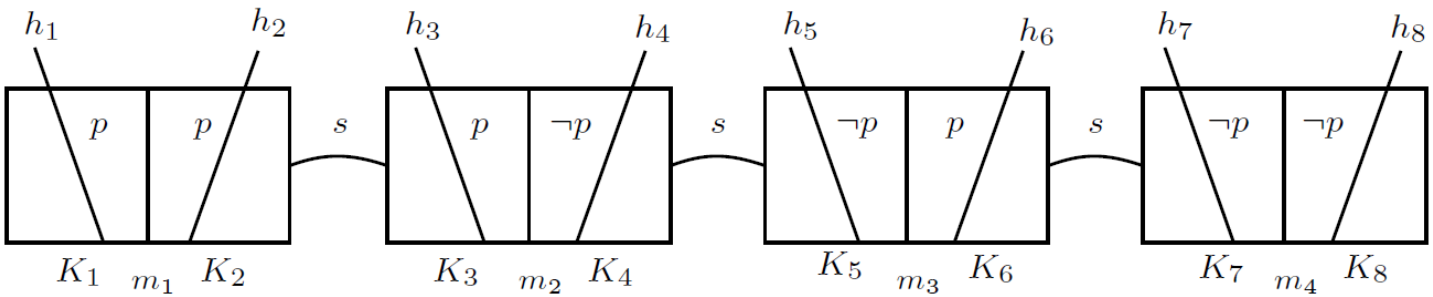
$$\mathcal{M}, m/h \vDash K_i\varphi \iff \forall m'/h' : m \sim_i m' \wedge h' \in \mathcal{H}_{m'} : \mathcal{M}, m'/h' \vDash \varphi$$

Семантика доксатического оператора определяется следующим образом:

$$\mathcal{M}, m/h \vDash B_i\varphi \iff \forall m'/h' \in \max_{\preccurlyeq_i}([m]) \wedge h' \in \mathcal{H}_{m'} : \mathcal{M}, m'/h' \vDash \varphi$$

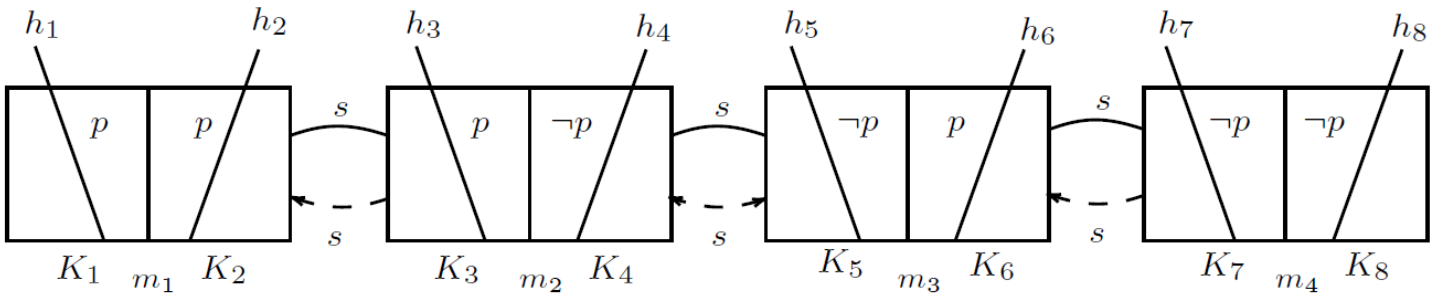
Пример "Старик и море"

Кубинский рыбак Сантьяго планирует охоту на марлина. У него есть два варианта действий: он может отплыть либо на восток от берега ($\mathrm{K}_1, \mathrm{K}_3, \mathrm{K}_5, \mathrm{K}_7$), либо на север ($\mathrm{K}_2, \mathrm{K}_4, \mathrm{K}_6, \mathrm{K}_8$). Он не знает, где именно ему удастся поймать марлина, кроме того, он допускает, что сегодняшняя охота может оказаться неудачной и он вернется домой без улова.



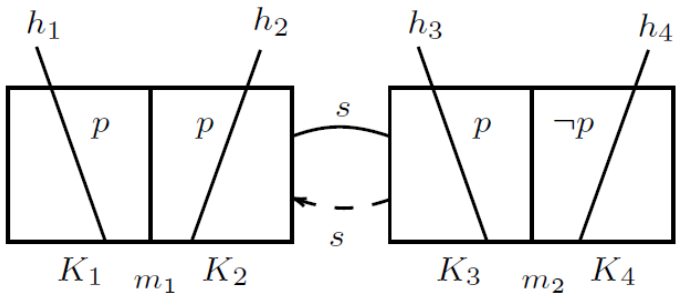
Пример "Старик и море"

Сантьяго верит, что сегодня ему обязательно повезет и он в любом случае поймает марлина вне зависимости от того, в какую сторону он решит плыть. Наименее вероятным он считает исход, в котором он ничего не поймает. А варианты, где он сможет поймать марлина только в одной стороне он расценивает как равновероятные.



Пример "Старик и море"

Предположим, что Сантьяго решил повернуть на восток от берега $(\mathrm{K}_1, \mathrm{K}_3, \mathrm{K}_5, \mathrm{K}_7)$ и поймал там марлина. Однако он все еще не знает, какой из моментов является актуальным.



Эпистемическоий индетерминизм

Несмотря на то, в детерминированном мире возможен только единственный вариант развития событий, нам свойственно предполагать, что есть альтернативные истории.

Совершая выбор и наблюдая его последствия, агент все еще не будет знать каковы альтернативные истории в его мире, и есть ли они вообще.

Индетерминизм – в том числе удобное рассуждение о будущих событиях мира, которые для нас неизвестны.

Наряду с онтологическим детерминизмом существует эпистемический индетерминизм.

Интенциональное расширение stit-логики

Теории, посвященные намерению, можно разделить на редукционистские (O'Shaughnessy, 1973), (Harman, 1998), (см. Davidson, 1980) и нередукционистские (Bratman, 1987).

.

Следует разделять этап подготовки к действию и этап совершения действия. Намерение будет относиться к этапу подготовки.
Ментальные состояния, в которые входят желания и убеждения агента, могут быть причинами интенций, но сами они не редуцируются друг к другу.
Будем понимать под намерением обязательство перед самим собой относительно совершения будущего действия.

Интенциональное расширение stit-логики

Существует несколько попыток ввести интенциональное расширение в stit-логику (Bentzen, 2010), (Broersen, 2011).

.

Добавим в модель stit-логики с эпистемическим расширением функцию намеренных последствий.

Definition

Функция намеренных последствий $\mathcal{I}$ отображает агента $i \in A$ и действие $K \in Choice_i^m$ на разбиение $\mathcal{I}_i^K \subseteq K$, где $\mathcal{I}_i^K$ – разбиение (множество) таких историй (последствий), которые агент намеревался сделать актуальными.
Элементы разбиения $\mathcal{I}$ будем обозначать Γ.

Интенциональное расширение stit-логики

Definition

Семантика оператора намерения $(\text{intit})_\text{i}$ определяется следующим образом:
$\mathcal{M}, \text{m}/\text{h} \vDash (\text{intit})_\text{i}\varphi \iff \forall \text{h}'((\text{h}' \in \mathcal{I}_\text{i}^\text{K}(\text{h}) \wedge \text{K} \in \text{Choice}_\text{i}^\text{m}) \to \mathcal{M}, \text{m}/\text{h}' \vDash \varphi) \vee \forall \text{m}''/\text{h}''((\text{m} \sim_\text{i} \text{m}'' \wedge \text{h}'' \in \mathcal{I}_\text{i}^{\text{K}'} \wedge \text{K}' \in \text{Choice}_\text{i}^{\text{m}''}) \to \mathcal{M}, \text{m}''/\text{h}'' \vDash \varphi)$

Возможна ситуация, в которой агент намеревается сделать φ, но не знает, что в актуальном мире φ невозможно. Именно поэтому важно эксплицировать эпистемический компонент намерения.

Возможные теоремы интенционального и эпистемического расширения stit-логики

Формулы, связывающие различные модальности:

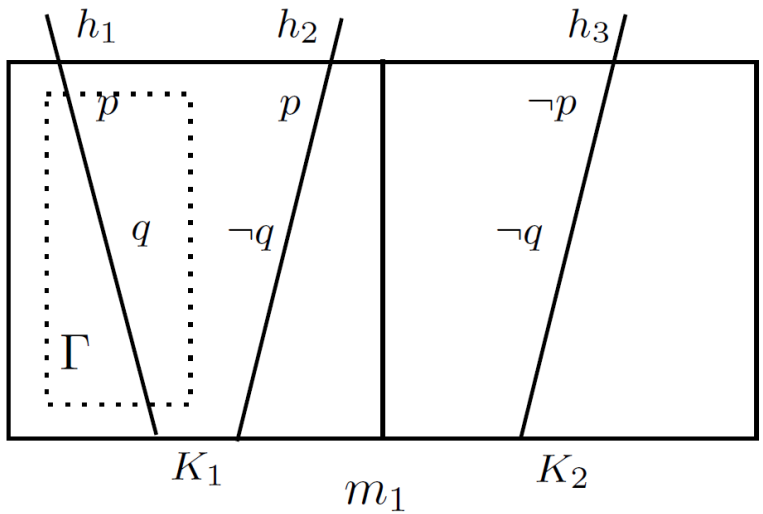
$$(\text{intit})_{\text{i}}\varphi \to \hat{\text{K}}_{\text{i}}\Diamond\varphi$$

$$(\text{intit})_{\text{i}}\varphi \to \hat{\text{K}}_{\text{i}}\langle\text{stit}\rangle_{\text{i}}\varphi$$

$$(\text{intit})_{\text{i}}\varphi \to \hat{\text{K}}_{\text{i}}\Diamond[\text{stit}]_{\text{i}}\varphi$$

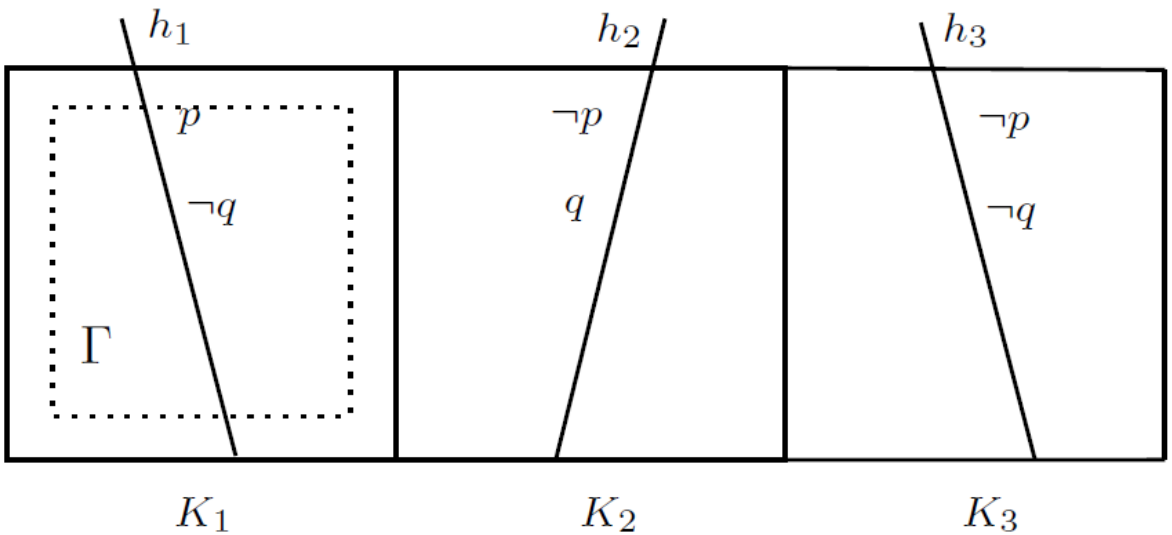
Пример "Лотерея"

В парке аттракционов предлагают сыграть в лотерею. Аня – очень азартный человек. Она хочет участвовать в лотерее и намеревается выиграть. Для Ани есть два доступных действия: участвовать (K_1) или не участвовать в лотерее (K_2). Припишем пропозициональным переменным следующие значения: p := «Аня участвует в лотерее». q := «Аня выигрывает приз».



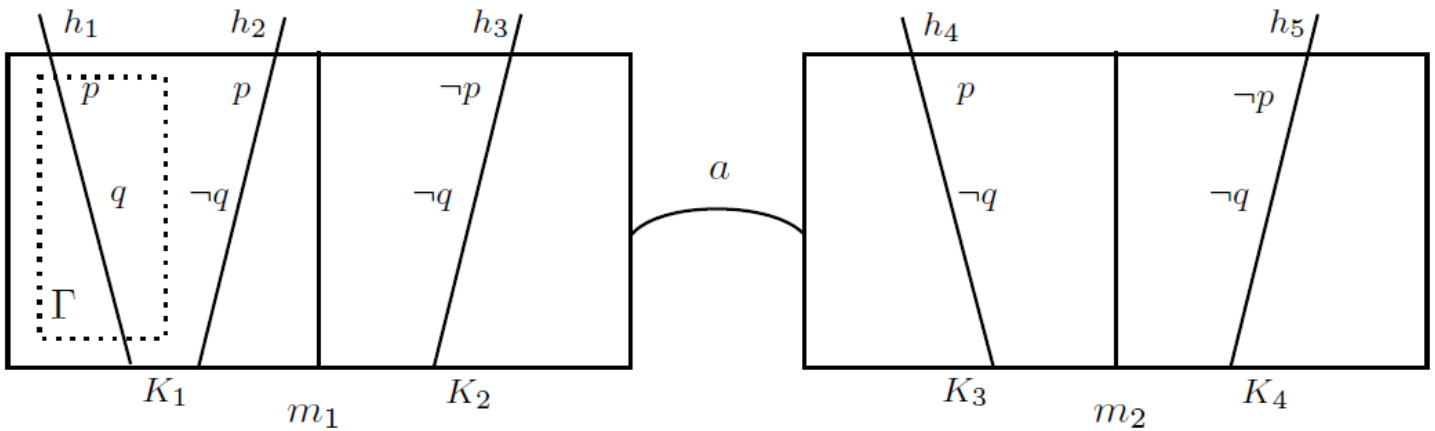
Пример "Пирожные"

Женя стоит у витрины магазина, где продают только два вида пирожных с клубникой и с черникой, денег у него хватить только на какое-то одно. Женя очень любит чернику, поэтому намерен взять пирожное с черничным вкусом.
Для Жени доступны три действия: K_1 – купить только черничное пирожное, K_2 – купить только клубничное пирожное и K_3 – ничего не покупать. p := «Женя купил черничное пирожное», q := «Женя купил клубничное пирожное».



Пример "Нечестная лотерея"

Аня собирается принять участие в другой лотерее, условия которой следующие: за участие в лотерее нужно заплатить 100 рублей. Победитель получает миллион рублей, однако в тайне от всех лотерея устроена таким образом, что в ней нет победителей. Аня не знает об этом и, покупая билет, намеревается выиграть приз. Для Ани доступны два действия: участвовать в лотерее (K_1, K_3) или не участвовать (K_2, K_4). Пропозициональная переменная p означает «Аня принимает участие в лотерее». q – «Аня выигрывает приз».



Намерение и свободное действие

Действие может иметь сложную структуру. Процесс действия следует разделить на два этапа: 1) этап подготовки и 2) этап действия. На этапе подготовки у человека могут появляться, меняться различные эмоции, намерения, знания и убеждения, на основе которых протекает второй этап.

.

Как правило первый и второй этапы связаны: человек совершает те действия, которые намеревался или желал сделать. Так, благодаря тесной связи этих двух этапов можно утверждать, что действия человека являются свободными.

Интенциональное расширение логики агентности позволяет эксплицировать один из ключевых элементов этой свободы.

Принцип альтернативных возможностей

Принцип альтернативных возможностей:
Действие агента является свободным только в том случае, если он мог поступить иначе.

.

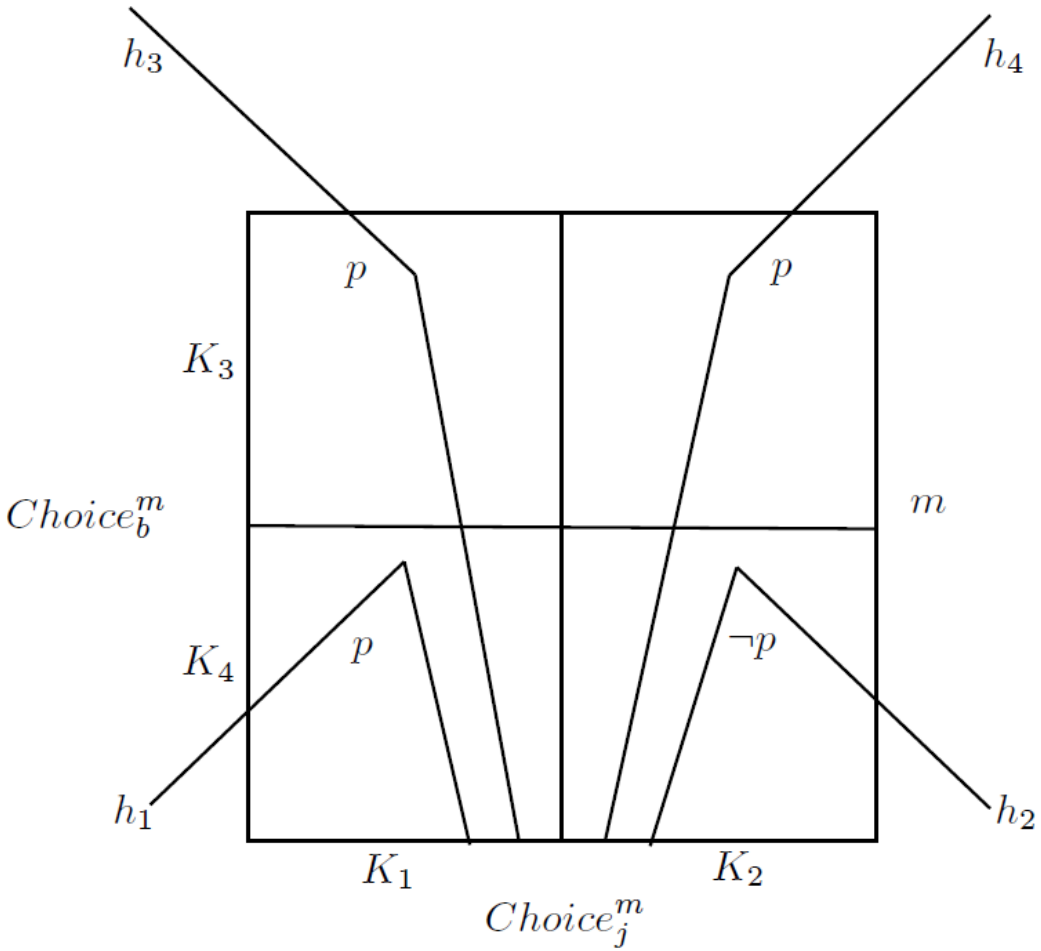
.

Контрпримеры Франкфурта (см. Frankfurt, 1969) являются, наверное, самыми известными и весомыми возражениями против принципа альтернативных возможностей.

Контрпример Франкфурта

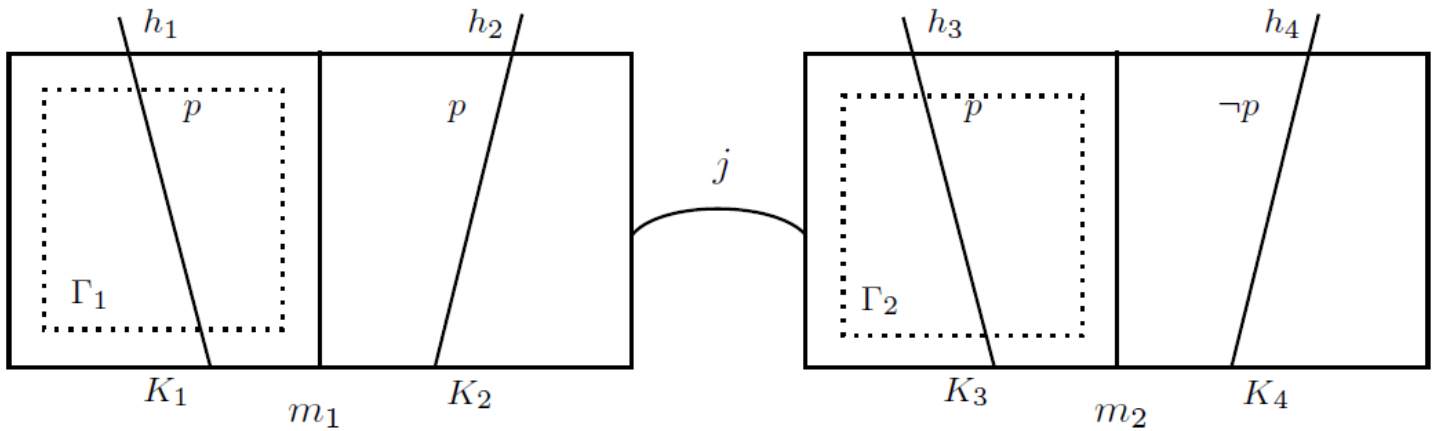
Блэк очень хочет, чтобы Джонс убил Смита. Блэк готов добиваться своей цели любой ценой, но ни в коем случае, не хочет, чтобы Джонс узнал об этом. Блэк внимательно следит за тем, как Джонс собирается поступить, и, если он решит не убивать Смита, Блэк сразу же вмешается и сделает так, что Джонс все-таки убьет его. Таким образом, если Джонс попытается или даже подумает о том, что хочет поступить иначе, Блэк не позволит ему это сделать. У Джонса нет альтернативных возможностей. Однако Джонс самостоятельно решает убить Смита и убивает его. Блэк не вмешивается и никаким образом не воздействует на Джонса. (Frankfurt, 1969)

Контрпример Франкфурта в stit-логике



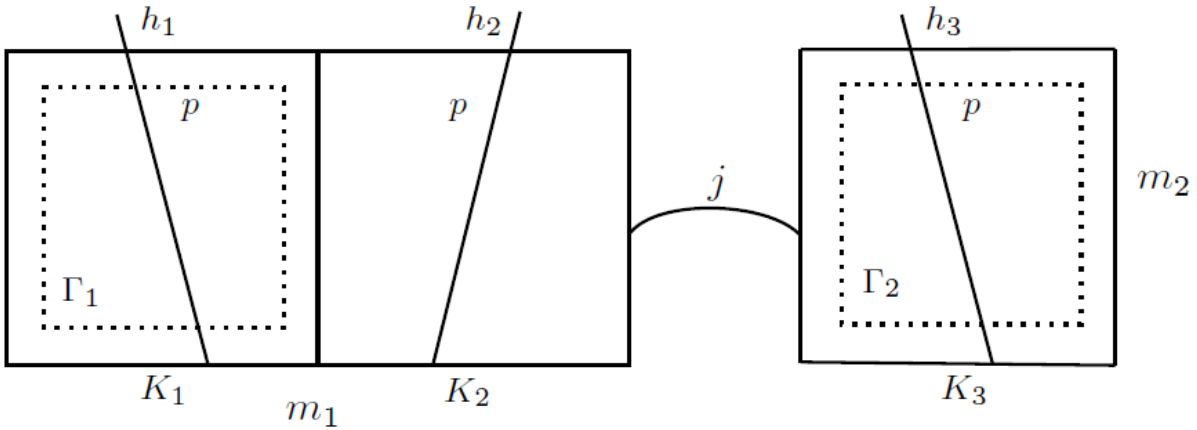
Контрпример Франкфурта в расширениях stit-логики

В контрпримере Франкфурта Джонс не различает, в каком моменте он находится.



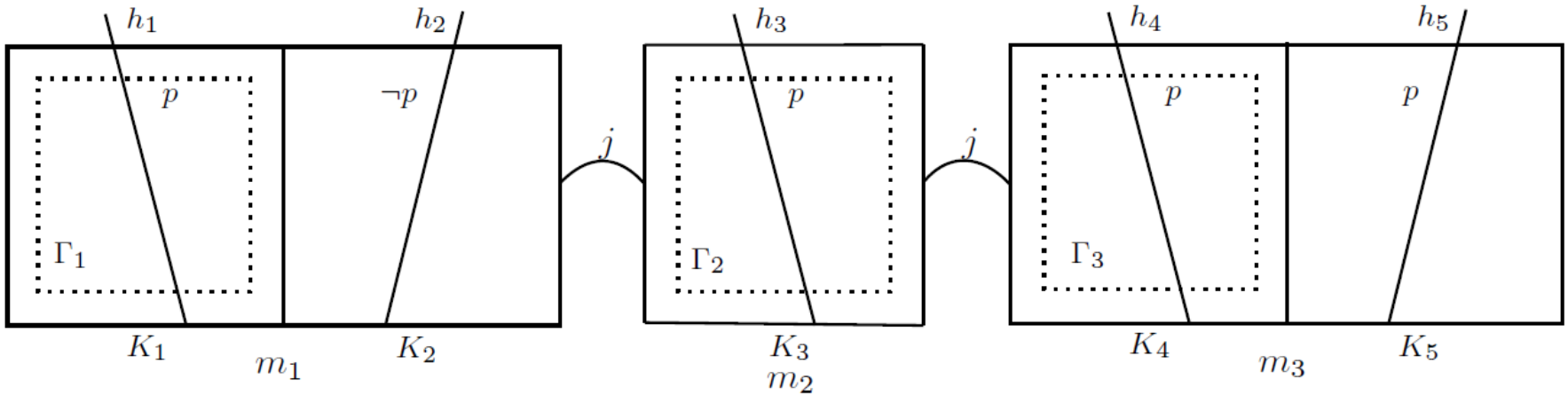
Контрпример Франкфурта в расширениях stit-логики

Даже если «онтологически» у Джонса не было возможности поступить по-другому, он не будет различать эти моменты.



Контрпример Франкфурта в расширениях stit-логики

Джонс вообще не будет различать ни один из рассмотренных моментов, если оценка намеренных последствий будет совпадать.



Выводы

1. Если агент намеренно совершает действие, неважно была ли у него онтологическая возможность поступить иначе, его действие будет свободным.
2. Рассуждения об альтернативных возможностях – следствие эпистемического индетерминизма. Под кажущейся онтологической неопределенность на самом деле стоит недостаток знания агента о мире.
3. Нам неважно даже существование альтернативных возможностей, если актуальное положение дел в мире гарантируется намеренным выбором агента.
4. Компатибилизм – красивая концепция про сочетание свободы агента с детерминированностью мира.