Головоломка с токсином: намерение как перформативный акт

Елена Попова

8 октября 2020 г.

Toxin puzzle

Kavka G. S. The Toxin Puzzle. Analysis, 1983:

> An eccentric billionaire places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon.

- В полночь сформировать намерение выпить токсин в полдень;
- До полудня деньги будут переведены на Ваш счет;
- В полдень Вы можете не пить токсин.

Решение Кавки

Выиграть пари невозможно.

-

Намерение – это предрасположенность к действию, которая основывается на причинах совершать действие – ради самого процесса или возможных последствий.

-

Нельзя разделять основания намереваться и основания совершать действие.

Рациональное решение Готье

Выиграть пари можно, сформированное в полночь намерение в полдень завершится действием.

-

Употребление токсина в полдень рационально, поскольку является частью наиболее выгодной стратегии.

-

Искреннее намерение неизбежно перетекает в действие.

Решение Меле и его экзотический агент Тэд

Искреннее намерение действительно не может существовать без вытекающего из него действия.

-

Единственным победителем пари с миллиардером может быть экзотический агент Тэд, который непреднамеренно выпивает все токсины, которые есть поблизости, если они не угрожают его жизни. Изначально зная, что неизбежно совершит действие, Тэд способен сформировать намерение.

Теоретико-игровая формализация

Кавка и Мелe

	намерение было	намерения не было
пить		b
не пить		a

Готье

	намерение было	намерения не было
пить	a	c
не пить		b

Теоретико-игровая формализация

моя гипотеза

	намерение было	намерения не было
пить		
не пить	a	b

Головоломка с токсином и парадокс Мура

Goldstein L., Cave P. A Unified Pyrrhonian Resolution of the Toxin Problem, The Surprise Examination and Newcomb's Puzzle. American Philosophical Quarterl, 2008. pp. 365 - 376:

- $\mathrm{I_i}$ – оператор намерения;
- $\mathrm{I_a}(\varphi)$ – a намеревается сделать φ;
- $\mathrm{I_a(p)} \wedge \neg \mathrm{p}$;
- $\mathrm{I_a(I_a(p)} \wedge \neg \mathrm{p})$;

Головоломка с токсином и парадокс Мура

Goldstein L., Cave P. A Unified Pyrrhonian Resolution of the Toxin Problem, The Surprise Examination and Newcomb's Puzzle. American Philosophical Quarterl, 2008. pp. 365 - 376:

- I_i – оператор намерения;
- $I_a(\varphi)$ – a намеревается сделать φ;
- $I_a(p) \wedge \neg p$;
- $I_a(I_a(p) \wedge \neg p)$;

Парадокс Мура:

- $p \wedge \neg K_a(p)$

Намерение

- Трудность в дефиниции намерения;
- Намерение – это волевой акт, который формируется, основываясь на причинах действовать определенным образом;
- Причины для формирования намерения и для совершения действия могут не совпадать;
- Появление намерения и его переход в действие предполагают темпоральный разрыв, а это значит, что связь между намерением и действем не является неразрывной.

Термины

Definition

Осуществление (fulfillment) намерения – формирование намерения в момент времени t_1 и совершение действия, на которое оно было направлено, в момент времени t_2 .

Definition

Исполнение (performance) намерения – формирование намерения в момент времени t_1 , но пока еще не совершение действия или его отмена по внутренним или внешним причинам в момент времени t_2.

Формализация

I(.) – оператор намерения;
p – действие;
I(p) – намерение совершить действие;
I(I(p)) – намерение намереваться совершить действие;
I(I(I(p))) – намерение намереваться намереваться ...

Theorem

Правило 1: $I^n(p) \rightarrow I(p)$

Theorem

Следствие 1: неверно, что ($I^n(p)$ & не-I(p))

Формализация

I(.) – оператор намерения;
p – действие;
I(p) – намерение совершить действие;
I(I(p)) – намерение намереваться совершить действие;
I(I(I(p))) – намерение намереваться намереваться ...

Theorem

Правило 1: $I^n(p) \rightarrow I(p)$

Theorem

Следствие 1: неверно, что ($I^n(p)$ & не-I(p))

Кавка: всегда верно, что $I(p) \rightarrow p$
я: нет, не всегда.

Намерение как перформативный акт

Намерение, являясь волевым актом, может рассматриваться как перформатив.

-

Намерение = обещание самому себе. Формируя намерение, я заключаю с собой сделку, что совершу р.

-

Остаётся ли агент всегда тождественным самому себе в темпоральный промежуток $[t_1; t_2\ ]$ между исполнением и осуществлением намерения?

Намерение как перформативный акт

Условия успешного обещания по Сёрлю:

1. Соблюдены условия нормального входа и выхода;
2. S при произнесении T выражает мысль, что p;
3. Выражая мысль, что p, S предицирует будущий акт говорящему S;
4. H предпочел бы совершение говорящим S акта A несовершению говорящим S акта A, и S убежден, что H предпочел бы совершение им A несовершению им A;
5. Как для S, так и для H не очевидно, что S совершит A при нормальном ходе событий;
6. S намерен совершить A

Намерение как перформативный акт

Условия успешного обещания по Сёрлю:

7 S намерен с помощью высказывания T связать себя обязательством совершить A;

8 S намерен вызвать у H посредством произнесения T убеждение в том, что условия (6) и (7) имеют место благодаря опознанию им намерения создать это убеждение, и он рассчитывает, что это опознание будет следствием знания того, что данное предложение принято употреблять для создания таких убеждений;

9 Семантические правила того диалекта, на котором говорят S и H, таковы, что T является употребленным правильно и искренне, если, и только если, условия (1)-(8) соблюдены.

Намерение как перформативный акт

В темпоральный разрыв с агентом может произойти нечто, благодаря чему он станет нетождественным в прошлом намеревавшемуся себе, что отменит основания для p.

-

Всякое изменение, которое заставило S предпочитать не-p влечет мгновенную элиминацию H, который хотел совершения p.

-

В случае с обещанием самому себе возможна ситуация, когда H перестает существовать, в связи с чем для успешного осуществления обещания как перформативного акта не хватает одного из условий, именно поэтому исчезает и обещание, за выполнение которого S берет на себя ответственность.

Радикальное обновление

Definition

Радикальное обновление ($M \Uparrow \varphi$) – смена информационного состояния агента, радикальное изменение его первоначального мнения;
где M – модель доксатической логики, φ – причина (поступившая информация), из-за которой с агентом происходит радикальное обновление.

Радикальное обновление может повлечь как незначительные, так и коренные изменения агента, которые отменят необходимые основания для совершения действия и сохранения обещания.

Возможное решение головоломки с токсином

1. Основание для исполнения (performance) намерения и осуществления (fulfillment) намерения не всегда совпадают.
2. В момент времени t_1 агент может намереваться выпить токсин и будет считать, что действительно его выпьет днем в t_3 , тем самым соблюдая условие искренности Сёрля.
3. Исполнение намерения (фиксация аппаратом) в t_2 влечет за собой радикальное обновление агента, которое отменяет основания для совершения последующего действия (употребления токсина днем в t_3).

Возможное решение головоломки с токсином

Намерение – это автономный волевой акт, который следует понимать как перформатив.

-

Для существования намерения достаточно его исполнения: намерение не обязательно должно перетекать действие.

-

Во временной промежуток между исполнением намерения и его осуществлением с агентом может произойти радикальное обновление, которое отменит основание совершать действие по «внутренним» причинам агента.

Формализация в BDI-логиках: Belief-Desire-Intention

Definition

BDI-логики помогают формализовать ментальные состояния (а иногда даже эмоции) разумного агента.

У нас есть широкий и разноплановый инструментарий для формализации, на первый взгляд, неформализуемых явлений и состояний.

Формализация в BDI-логиках: Belief-Desire-Intention

Намерение по Братману:

- это высокоуровневый план;
- оно направляет и "запускает"дальнейший процесс планирования – абстрактный план превращается во всё более точный;
- агент отказывается от намерения при условии:
  – намерение было осуществлено;
  – агент понял, что результата невозможно достичь;
  – агент отказался от некого необходимого условия для осуществления намерения.

Формализация в BDI-логиках: Belief-Desire-Intention

Подход Cohen and Levesque:

- У i есть достижимая цель φ, если i предпочитает исход, в котором φ будет истинно, и верит, что φ пока ложно.

$$\mathrm{AGoal_i}\varphi \underset{\mathrm{def}}{=} \mathrm{Pref_i F}\varphi \wedge \mathrm{Bel_i}\neg\varphi$$

- У i будет неотступная цель φ, если у i есть достижимая цель φ и он будет стремиться к ней, пока она не будет достигнута или будет считаться недостижимой.

$$\mathrm{PGoal_i}\varphi \underset{\mathrm{def}}{=} \mathrm{AGoal_i}\varphi \wedge (\mathrm{AGoal_i}\varphi)\mathrm{U}(\mathrm{Bel_i}\varphi \vee \mathrm{Bel_i G}\neg\varphi)$$

Формализация в BDI-логиках: Belief-Desire-Intention

Подход Cohen and Levesque:

- i намеревается φ, если у i есть неотступная цель φ и вера в то, что он может достичь φ своими действиями (поэтому появляется квантор существования).

$$\mathrm{Intend_i}\varphi \underset{\text{def}}{=} \mathrm{PGoal_i}\varphi \wedge \mathrm{Bel_i F}\exists\alpha\mathrm{Happ_{i:\alpha}}\varphi$$

- более слабое условие (Sadek and Bretier):

$$\mathrm{Intend_i}\varphi \underset{\text{def}}{=} \mathrm{PGoal_i}\varphi \wedge \mathrm{Pref_i F}(\exists\alpha\mathrm{Happ_{i:\alpha}F}\varphi)$$

- более сильное условие (Sadek and Bretier):

$$\mathrm{Intend_i}\varphi \underset{\text{def}}{=} \mathrm{PGoal_i}\varphi \wedge \mathrm{Pref_i}\forall\alpha(\mathrm{Bel_i Happ_{i:\alpha}F}\varphi \rightarrow \mathrm{Pref_i FHapp_{i:\alpha}}\top)$$

Формализация в BDI-логиках: Belief-Desire-Intention

Rao and Georgeff's BDI-logics (в ветвящейся темпоральной логике):

1. $\mathrm{GOAL}(\alpha) \to \mathrm{BEL}(\alpha)$
2. $\mathrm{INTEND}(\alpha) \to \mathrm{GOAL}(\alpha)$
3. $\mathrm{INTEND}(\mathrm{does(e)}) \to \mathrm{does(e)}$
4. $\mathrm{INTEND}(\varphi) \to \mathrm{BEL}(\mathrm{INTEND}(\varphi))$
5. $\mathrm{GOAL}(\varphi) \to \mathrm{BEL}(\mathrm{GOAL}(\varphi))$
6. $\mathrm{INTEND}(\varphi) \to \mathrm{GOAL}(\mathrm{INTEND}(\varphi))$
7. $\mathrm{done(e)} \to \mathrm{BEL}(\mathrm{done(e)})$
8. $\mathrm{INTEND}(\varphi) \to \mathrm{inevitable} \diamond (\neg \mathrm{INTEND}(\varphi))$

KARO логика для эмоциональных агентов

(Oatley & Jenkins):

- Счастье: в процессе достижения цели все идет по плану, как и ожидалось, т. е. подцели пока достигнуты;
- Грусть: в процессе достижения цели, что-то идет не так, как ожидалось, и подцели не достигнуты;
- Злость: возникает из-за безысходности, которая является следствием невозможности выполнить намеченный план, что тем не менее заставляет агента прилагать больше усилий;
- Страх: возникает, когда цель агента оказывается под угрозой, из-за чего он должен сначала позаботиться о ней, прежде чем переходить к другим действиям.

Литература

1 Kavka G. S. The Toxin Puzzle. Analysis, 1983. pp. 33–36.

2 Gauthier D. Assure and Threaten. Ethics, 1994. pp. 690–721.

3 Mele A. R. Intentions, Reasons, and Beliefs: Morals of the Toxin Puzzle. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 1992. pp. 171–194.

4 Broek M. You Don't Believe This Is The Title. 2018.

5 Rao A. S., M. P. Georgeff. Decision procedures for BDI logics. Journal of Logic and Computation, 1998. pp. 293-344.

6 Cohen P. R., Levesque H.J. Intention is choice with commitment. Artificial Intelligence, 1190. pp. 213-261.

Бертран Рассел:

> “A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science.”