Метарассуждения: логико-когнитивный подход

Д. В. Зайцев[1]

В работе предлагается использовать синтетический подход к формальному представлению рассуждений в искусственном интеллекте. Основу это подхода составляют положения теории двойного процесса, позволяющие различить элементарные, быстрые, интуитивные базовые рассуждения и контролируемые, осознаваемые, медленные рассуждения, основанные на размышлениях. Другим источником являются развиваемые в логике и формальной аргументации идеи представления метарассуждений как трансформации исходных рассуждений.

Ключевые слова: человеко-совместимый искусственный интеллект, метарассуждения, трансформация рассуждений.

С годами парадигма исследований искусственного интеллекта (далее ИИ) претерпевает существенные изменения. Одна из последних концепций представлена в книге С. Расселла [1], опубликованной в 2019 г. Автор считает, что проект ИИ зашел в тупик, а причину этого видит в неверной трактовке интеллекта. Согласно стандартному подходу, основы которого были заложены Н. Винером, машины разумны в той степени, в которой в какой от их действий можно ожидать достижения их целей, предварительно заложенных в них людьми. Стандартному пониманию интеллекта противопоставляется идея "человеко-совместимого" (human compatible) ии: машины полезны в той степени, в какой от их действий можно ожидать достижения наших целей. Такой подход предполагает способность ии понимать действия людей, в той степени, в которой они понятны нам самим, то есть со значительной мерой неопределенности.

Естественно эволюция понимания ИИ сопровождается изменениями в трактовке рассуждений как одной из главных функций ИИ. Логические рассуждения, формализуемые в соответствии с жесткими критериями классической логики, уступили место формальным моделям естественных рассуждений, существенными чертами которых являются неопределенность, правдоподобность, модифицируемость (немонотонность) и т.п. Человеко-совместимому ИИ соответствует идея метарассуждений (см. например, [2], [3], [4]). Метарассуждения понимаются как рассуждения о рассуждениях. Это в свою очередь предполагает рациональный выбор вычислительный действий из определенного предзаданного

[1] *Зайцев Дмитрий Владимирович* — д.ф.н., профессор каф. логики философского ф-та МГУ, e-mail: zaitsev@philos.msu.ru

Zaitsev Dmitry Vladimirovich — D.Sc., Professor, Lomonosov Moscow State University, Faculty of Philosophy, Deapartment of Logic.

набора и оценку ожидаемой полезности этих действий, то есть выход на метауровень, что позволяет осуществлять так называемый «контроль за обдумыванием» (control of deliberation).

Следует отметить, что метарассуждения не являются темой, обсуждаемой исключительно в рамках логики или компьютерной науки. В когнитивных науках существует давняя традиция исследования метарассуждений в контексте метапознания (metacognition, [5], [6]). Важно, что в рамках этой традиции метарассуждения анализируются в терминах теории двойного процесса. Не вдаваясь в подробности этой интерпретации, зафиксируем важную для дальнейшего рассмотрения идею – интеллектуальные процедуры, которые мы традиционно называем рассуждениями, распадаются на два принципиально различных типа когнитивных процессов: автоматические, быстрые, интуитивные, встроенные механизмы рассуждений (система1) и управляемые, медленные, основанные на рефлексии и размышлении механизмы (система 2), подробнее см. [7].

Программа метарассуждений включает в себя множество различных направлений исследований. В данном случае акцент делается на рассмотрении перспектив трактовки метарассуждений как процесса перехода от одних (исходных, атомарных) рассуждений к другим.

Одним из снований развиваемого подхода служат наши совместные исследования с сотрудниками кафедры нейрофизиологии факультета психологии МГУ [8], направленные на выявление когнитивных критериев различения непосредственных автоматических переходов от посылок к заключениям, называемых в традиционной логике умозаключениями, и опосредованных рассуждений, понимаемых как процедуры пошагового обоснования некоторого высказывания. Успешная реализация этого проекта позволит отказаться от конвенционального разведения элементарных умозаключений и состоящих из них рассуждений, и выделить некоторое множество атомарных непосредственных переходов от посылок к заключениям как когнитивный базис человеческих рассуждений.

Другой источник обнаруживается в логике и формальной аргументации. Идея трансформации рассуждений в виде специальных правил, описывающих допустимые способы перестройки вывода отнюдь не нова. Достаточно вспомнить такие принципы как сведение к абсурду или рассуждение от противного:

$$\frac{\Gamma, A \vdash \bot}{\Gamma \vdash \neg A},$$

или структурные правила генценовского типа, например, ослабление или сечение:

$$\frac{\Gamma \vdash \Delta}{\Gamma, A \vdash \Delta}, \qquad \frac{\Gamma \vdash \Delta, A \quad A, \Theta \vdash \Sigma}{\Gamma, \Theta \vdash \Delta, \Sigma}.$$

Следующий важный шаг в представлении метарассуждений как трансоформации рассуждений делают идеологи формальной аргументации. Особенно хорошо это заметно на примере так называемого «абстрактного аргуменативного каркаса» а-ля Данг (abstract argumentation framework), названного так благодаря основополагающей работе [9]. Главные идеи абстрактной аргументации состоят в том, чтобы абстрагироваться от внутренней структуры рассуждений и рассматривать аргументативные рассуждения как атомарные объекты, на которых задано единственное отношение атаки, соответствующее принятому в неформальной теории аргументации понятию критики. В результате аргументативная структура – это пара, включающая множество рассуждений (arguments) и двухместное отношение атаки. Утверждение о том, что рассуждение A атакует рассуждение B обычно представляется как формула вида $A \longrightarrow B$. Примечательно, что абстрактное отношение атаки оказывается фундаментальным и позволяет выразить другие отношения между рассуждениями, например отношение поддержки или восстановления (одним рассуждением другого):

$$A \longrightarrow B \longrightarrow C$$

В этом примере рассуждение B атакует рассуждение C. Но само рассуждение B находится под атакой рассуждения A, которого ничто не атакует. Это позволяет интерпретировать отношение между A и C как поддержку или восстановление: A как бы нейтрализует негативное влияние B.

Таким образом в рамках формальной аргументации разработан эффективный механизм представления и оценки аргументации, предполагающий фиксацию исходных рассуждений как неделимых объектов.

Подводя итог, перспектива соединения описанных выше идей для проекта ИИ видится в следующем.
1. Следуя примеру разработчиков формальной аргументации, я предлагаю рассматривать метарассуждения как способы допустимых трансформаций исходных атомарных умозаключений.
2. Набор атомарных умозаключений выявляется на основе когнитивно-логических исследований и представляет собой множество интуитивно приемлемых автоматически осуществляемых переходов от посылок к заключениям.

Исследование выполнено в рамках научного направления «Философия когнитивных наук и искусственного интеллекта» научно-образовательной школы Московского государственного университета

имени М. В. Ломоносова «Мозг, когнитивные системы и искусственный интеллект».

Metareasonings: logical-cognitive approach
Zaitsev D.V.

The paper promotes a synthetic approach to the formal representation of reasoning in artificial intelligence. The basis of this approach is the provisions of the dual process theory, which make it possible to distinguish between elementary, fast, intuitive basic reasoning and controlled, conscious, slow reasoning based on reflection. Another source of my approachi are developed in logic and formal argumentation ideas of the representation of meta-reasoning as a transformation of the original reasoning.

Keywords: human compatible artificual intellegence, metareasoning, transformation of reasoning.

References

[1] Russell S., *Human compatible: Artificial intelligence and the problem of control*, Penguin, New York City, 2019, 352 pp.

[2] Brazdil P. B., Konolige K. (ed.), *Machine learning, meta-reasoning and logics*, Springer-Verlag, New York City, 1990, 348 pp.

[3] Russell S., Wefald E., "Principles of metareasoning", *Artificial intelligence*, **49**:1-3 (1991), 361–395

[4] Cox M. T., Raja A. (ed.), *Metareasoning: Thinking about thinking*, MIT Press, Cambridge, 2011, 340 pp.

[5] Flavell J. H., "Metacognitive aspects of problem solving", *The nature of intelligence*, ed. L. B. Resnick, Lawrence Erlbaum, Hillsdale, NJ, 1976, 231–235.

[6] Ackerman R., Thompson V. A., "Meta-reasoning: Monitoring and control of thinking and reasoning", *Trends in Cognitive Sciences*, **21**:8 (2017), 607–617.

[7] Thompson, V.A., "Dual-process theories: A metacognitive perspective", *In Two Minds: Dual Processes and Beyond*, ed. Evans, J. Frankish, K., Oxford University Press, Oxford, 2009, 171–195.

[8] Kovalev A., Kiselnikov A., Sukhotina K., Zaitsev D., Zaitseva N., "Oculomotor indicators can differentiate various types of inference processes", *Psychophysiology*, **58**:51 (2021), S53–S53.

[9] Dung P.M., "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games", *Artificial Intelligence*, **77**:2 (1995), 321–357.