Коммюнике онтологического саммита 2019: ОБЪЯСНЕНИЕ[1]

(Фрагмент проекта ´ Ontology Summit 2019 Communiqu□e: Explanations□)

Июнь 2019

1 Введение

В последние годы наблюдается значительное увеличение использования методов машинного обучения (ML) для решения проблем в области искусственного интеллекта (AI). Эти успехи обусловлены наличием огромного количества уже накопленных данных, на основе обработки которых методы ML строят сложные статистические модели. К сожалению, разработчикам этих моделей сложно объяснить, как на их основе приходят к тем или иным выводам, поскольку каждое решение, в принципе, является результатом работы программы, которая включает весь набор данных, используемых для разработки модели. Поэтому встал вопрос о том, что такое объяснение, а также какие критерии могут использоваться для оценки точности объяснения.

Саммит онтологов 2019 года стремился исследовать, идентифицировать и сформулировать, как онтология может принести пользу проблеме автоматизации объяснений сложных систем в целом. Саммит решил эту задачу, сначала изучив понятие объяснения в серии сессий осенью 2018 года. Наиболее важными областями были определены: (1) *здравый смысл и знание* и (2) *повествование*. Затем они были более подробно изучены на последующих сессиях в 2019 году. Кроме того, было решено изучить некоторые конкретные области, чтобы определить виды проблем, с которыми сталкиваются практики в отношении объяснений. Выбранными предметными областями были: финансы и медицина. Наконец, так как объяснения AI были первоначальной мотивацией, изучена проблема Объяснимого AI.

Саммит онтологов 2019 года посвящён роли онтологий для объяснения аргументации работы программной системы. В частности, основное внимание было уделено критическим пробелам в объяснении и роли онтологий для устранения этих пробелов. На сессиях рассмотрены существующие технологии и реальные потребности, обусловленные рисками и требованиями соответствия правовым и другим стандартам.

2 Предпосылки

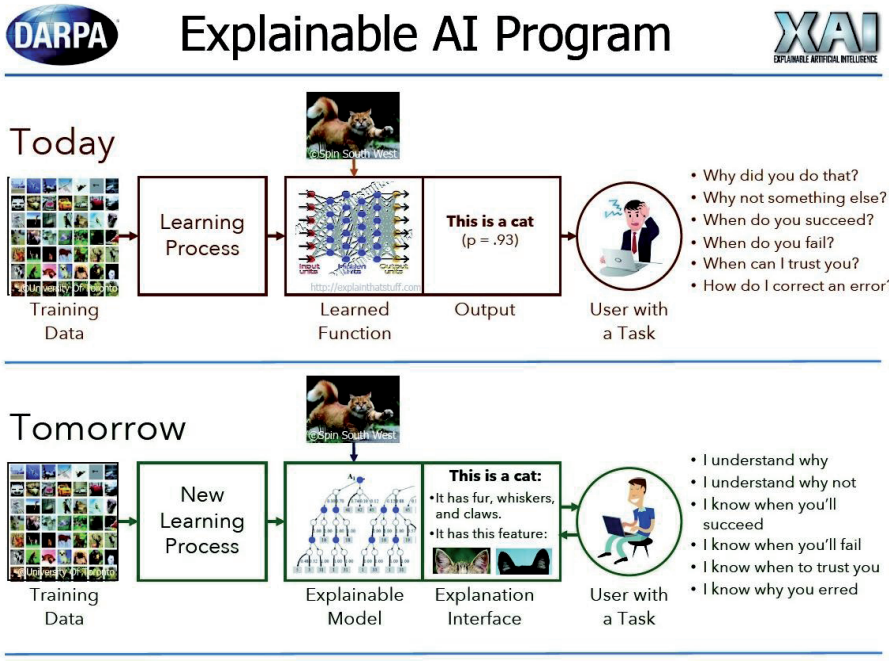
Объяснение - это ответ на вопрос «Почему?». Соответственно, объяснения обычно происходят в контексте процесса, который может быть диалогом между человеком и системой или процессом связи между агентами двух систем. Объяснения также происходят в социальных взаимодействиях, когда излагают точку зрения или интерпретируют поведение.

Первые известные попытки понять причину объяснений были у греческих и индийских философов. Например, чтобы понять и объяснить почему произошла Пелопоннесская война, Фукидид определил объяснения как процесс, в котором факты (неоспоримые данные) наблюдаются, оцениваются на основе общих знаний о человеческой природе. В трудах Платона объяснение базируется на знаниях универсальных форм, которые являются абстракциями сущностей мира. Факты с этой точки зрения являются происшествиями или ситуациями и могут быть лишь описательной частью объяснения, а не причиной. Взгляд Аристотеля даёт уже знакомое представление об объяснении как части логического, дедуктивного процесса для достижения выводов. После Декарта, Лейбница и Ньютона современная детерминистская причинность стала центральной для объяснений. Знать, что является причиной события, значит использовать естественные законы в качестве основного средства, чтобы понять и объяснить, почему это произошло.

Тема онтологического саммита, вдохновлённая проектом DARPA Explainable AI (XAI) [5, 6], рассматривалась как общая проблема объяснения. На саммите рассматривались не только системы AI, которые могут объяснить их действия, но также и интеллектуальные инженерные системы, которые могут взаимодействовать друг с другом. С ростом количества программного обеспечения, предназначенного для промышленной автоматизации и управления процессами, эта возможность становится важной. Объяснения включают в себя объяснение причин, описание сильных и слабых сторон и проецирование поведения систем в будущем.

Онтологии могут играть важную роль в объяснениях, поскольку интеллектуальные инженерные системы должны представлять концептуальную основу, которая поддерживает объяснения. Версия естественного языка может использоваться для описания состояний и действий в терминах, которые люди легко понимают, а также в концептуальных структурах, в которых происходит диалог.

[1]Редакция журнала решила не дожидаться финальной версии очередного Коммюнике Онтологического саммита за 2019 год, который обсуждается с осени 2018 года, но ещё не принят, а уже сейчас познакомить своих читателей с его концепцией - https://s3.amazonaws.com/ontologforum/OntologySummit2019/Symposium/communique2019.pdf.



Преимущество использования онтологий в объяснениях заключается в возможности улучшения взаимодействия между системами. Опасность текущих усилий по объяснению состоит в том, что они разрознены. Это приведёт к разнообразию несовместимых методов объяснения, которые индивидуально могут удовлетворить требованию предоставления объяснений, но при интеграции в крупномасштабные системы – будут мало полезными.

Онтологии - это спроектированные артефакты знаний, существующие в вычислительной среде, которая позволяет рассуждать, и поэтому должна включать способность объяснять то, что они «знают» и как «обосновать эти знания». Они должны быть в состоянии выразить обоснование выбранного использования соответствующих частей онтологии или набора онтологий; объяснить сильные и слабые стороны выбранной онтологии; и объяснить данные в этой онтологии.

3 Объяснимый AI [6]

4 Здравый смысл и знание [8]

5 Роль повествования [1, 4, 7, 9-12]

6 Финансовые объяснения [3]

7 Медицинские объяснения [13,14]

8 Результаты [12]

9 Вызовы и возможности [2]

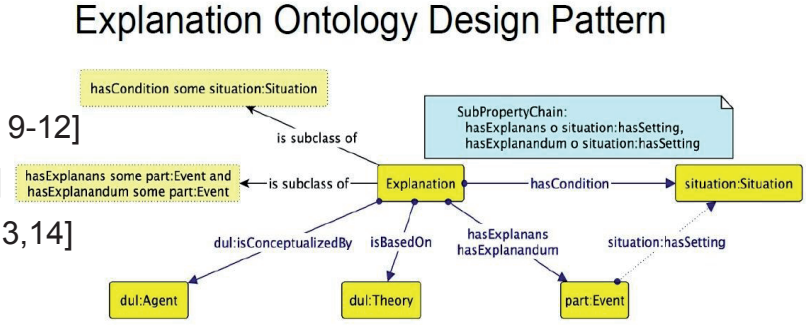


Figure 2: The Explanation Ontology Design Pattern.

Tiddi, et al. (2015)

Список источников

[1] K. Baclawski. Proof as explanation and narrative, January 2019. Retrieved December 9, 2017 from http://bit.ly/2RqQJQ5.

[2] K. Baclawski, M. Bennett, G. Berg-Cross, C. Casanave, D. Fritzsche, J. Ring, T. Schneider, R. Sharma, J. Singer, J. Sowa, R.D. Sriram, A. Westerinen, and D. Whitten. Ontology Summit 2018 Communiqu´e: Contexts in Context. Journal of Applied Ontology, July 2018. DOI: 10.3233/AO-180200.

[3] K. Ford, A. Canas, and J. Coffey. Participatory explanation. In FLAIRS 93: Sixth Florida Artificial Intelligence Research Symposium, pages 111–115, Ft. Lauderadale, FL, April 18-21 1993. Retrieved on June 5, 2019 from http://bit.ly/318aUbX.

[4] H.P. Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, Speech Acts, pages 41–58. Academic Press, New York, 1975.

[5] D. Gunning. DARPA Explainable Artificial Intelligence: Program Update, 2017. Retrieved on June 4, 2019 from http://bit.ly/2ITKwfD.

[6] D. Gunning. DARPA Explainable Artificial Intelligence, 2018. Retrieved on December 3, 2018 from http://bit.ly/2s9d4pH.

[7] F.C. Keil. Explanation and understanding. Annu. Rev. Psychol., 57:227–254, 2006.

[8] J. McCarthy. Programs with common sense. In M. Minsky, editor, Semantic Information Processing, pages 403–418. MIT Press, 1968. Originally published in 1959.

[9] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 2018.

[10] T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. arXiv preprint, 2017.arXiv:1712.00547.

[11] S. Rodriguez, J. Schaffer, J. O'Donovan, and T. H¨ollerer. Knowledge complacency and decision support systems. In IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2019.

[12] I. Tiddi, M. d'Aquin, and E. Motta. An ontology design pattern to define explanations. In Proceedings of the 8th International Conference on Knowledge Capture. ACM, October 2015. Article no. 3.

[13] E. Topol. In Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Hatchett Book Group, New York, 2019.

[14] E. Topol. High-performance medicine: roles in a high-level knowledge representation language. In *Artificial intelligence review*. doi: 10.1007/s10462-017-9571-5.