ВикипедиЯ

Юдковский, Элиезер

Материал из Википедии — свободной энциклопедии

Элиéзер Шлóмо Юдкóвский (англ. *Eliezer S. Yudkowsky*, 11 сентября 1979) — американский специалист по искусственному интеллекту, исследующий проблемы технологической сингулярности и выступающий за создание дружественного ИИ[2][3]. Ключевая фигура сообщества рационалистов.

Элиезер Юдковский

англ. *Eliezer Yudkowsky*

Элиезер Юдковский на Стэнфордском саммите сингулярности в 2006 году.

Имя при рождении	англ. *Eliezer Shlomo Yudkowsky*
Дата рождения	11 сентября 1979[1] (43 года)
Место рождения	Чикаго
Страна	США
Научная сфера	Искусственный интеллект
Место работы	Machine Intelligence Research Institute
Известен как	автор книги Гарри Поттер и методы рационального мышления
Сайт	yudkowsky.net (англ.)

Медиафайлы на Викискладе

Содержание

Биография

Научные интересы

Сочинения

Примечания

Ссылки

Биография

Элиезер Юдковский родился 11 сентября 1979 года в семье ортодоксальных евреев[4].

Научные интересы

Юдковский — сооснователь и научный сотрудник Института Сингулярности по созданию Искусственного Интеллекта Singularity Institute for Artificial Intelligence (SIAI)[5]. Он — автор книги «Создание дружественного ИИ»[6], статей «Уровни организации универсального интеллекта»[7], «Когерентная экстраполированная воля»[8] и «Вневременная теория принятия решений»[9][10]. Его последними научными публикациями являются две статьи в сборнике «Риски глобальной катастрофы» (2008) под редакцией Ника Бострома, а именно «Искусственный интеллект как позитивный и негативный фактор глобального риска» и «Когнитивные искажения в оценке глобальных рисков»[11][12][13]. Юдковский не обучался в вузах и является автодидактом без формального образования в области ИИ[14].

Юдковский исследует те конструкции ИИ, которые способны к самопониманию, самомодификации и рекурсивному самоулучшению (Seed AI), а также такие архитектуры ИИ, которые будут обладать стабильной и позитивной структурой мотивации (Дружественный искусственный интеллект). Помимо исследовательской работы, Юдковский известен своими объяснениями сложных моделей на неакадемическом языке, доступном широкому кругу читателей, например, см. его статью «Интуитивное объяснение теоремы Байеса»[15][16].

Юдковский был вместе с Робином Хансоном одним из главных авторов блога *Overcoming Bias (http://www.overcomingbias.com/about)* (преодоление предубеждений). В начале 2009 года он участвовал в организации блога *LessWrong*, нацеленного на «развитие рациональности человека и преодоление когнитивных искажений». После этого Overcoming Bias стал личным блогом Хансона. Материал, представленный на этих блогах, был организован в виде цепочек постов, которые смогли привлечь тысячи читателей — см. например, цепочку «теория развлечений»[17].

Юдковский — автор нескольких научно-фантастических рассказов, в которых он иллюстрирует некоторые темы, связанные с когнитивной наукой и рациональностью. В неакадемических кругах больше известен как автор фанфика «Гарри Поттер и методы рационального мышления» под эгидой Less Wrong[18].

Сочинения

- *Our Molecular Future: How Nanotechnology, Robotics, Genetics and Artificial Intelligence Will Transform Our World* by Douglas Mulhall, 2002, p. 321.
- *The Spike: How Our Lives Are Being Transformed By Rapidly Advancing Technologies* by Damien Broderick, 2001, pp. 236, 265—272, 289, 321, 324, 326, 337—339, 345, 353, 370.

Статьи на русском

- Систематические ошибки в рассуждениях, потенциально влияющие на оценку глобальных рисков (http://www.proza.ru/texts/2007/03/08-62.html)
- Искусственный интеллект как позитивный и негативный фактор глобального риска (http://www.proza.ru/texts/2007/03/22-285.html)
- Вглядываясь в Сингулярность (http://www.proza.ru/texts/2007/07/08-42.html)
- Таблица критических ошибок Дружественного ИИ (http://www.proza.ru/texts/2007/07/09-228.html)
- Три школы сингулярности (http://www.proza.ru/2009/02/05/432)
- Уровни организации универсального интеллекта (http://www.proza.ru/2009/10/08/1136)

Примечания

1. http://www.nndb.com/lists/517/000063328/ (http://www.nndb.com/lists/517/000063328/)
2. *Russell, Stuart.* Artificial Intelligence: A Modern Approach / Stuart Russell, Peter Norvig. — Prentice Hall, 2009. — ISBN 978-0-13-604259-4.
3. *Leighton, Jonathan.* The Battle for Compassion: Ethics in an Apathetic Universe. — Algora, 2011. — ISBN 978-0-87586-870-7.
4. Avoiding Your Belief's Real Weak Points (https://www.lesswrong.com/posts/dHQkDNMhj692ayx78/avoiding-your-belief-s-real-weak-points). *LessWrong*. Дата обращения: 31 мая 2021.
5. *Ray Kurzweil.* The Singularity Is Near (неопр.). — New York, US: Viking Penguin, 2005. — С. 599. — ISBN 0-670-03384-7.
6. Creating Friendly AI (https://intelligence.org/files/CFAI.pdf), 2001

7. Levels of Organization in General Intelligence (https://intelligence.org/files/LOGI.pdf), 2002
8. Coherent Extrapolated Volition (https://intelligence.org/files/CEV.pdf), 2004
9. Timeless Decision Theory (https://intelligence.org/files/TDT.pdf), 2010
10. Eliezer Yudkowsky Profile (https://web.archive.org/web/20101204173732/http://www.acceleratingfuture.com/people/Eliezer-Yudkowsky/). Accelerating Future. Дата обращения: 15 ноября 2010. Архивировано из оригинала (http://www.acceleratingfuture.com/people/Eliezer-Yudkowsky/) 4 декабря 2010 года.
11. Artificial Intelligence as a Positive and Negative Factor in Global Risk (https://web.archive.org/web/20130302173022/http://intelligence.org/files/AIPosNegFactor.pdf). Singularity Institute for Artificial Intelligence. Дата обращения: 28 июля 2009. Архивировано из оригинала (https://intelligence.org/files/AIPosNegFactor.pdf) 2 марта 2013 года.
12. Cognitive Biases Potentially Affecting Judgement of Global Risks (https://web.archive.org/web/20150507171632/http://intelligence.org/files/CognitiveBiases.pdf). Singularity Institute for Artificial Intelligence. Дата обращения: 29 октября 2018. Архивировано из оригинала (https://intelligence.org/files/CognitiveBiases.pdf) 7 мая 2015 года.
13. Global Catastrophic Risks (https://archive.org/details/globalcatastroph00bost) (англ.) / Bostrom, Nick. — Oxford, UK: Oxford University Press, 2008. — P. 91 (https://archive.org/details/globalcatastroph00bost/page/n6)—119, 308—345. — ISBN 978-0-19-857050-9.
14. GDay World #238: Eliezer Yudkowsky (http://gdayworld.thepodcastnetwork.com/2007/05/17/gday-world-238-eliezer-yudkowsky/). The Podcast Network. Дата обращения: 26 июля 2009. Архивировано (https://web.archive.org/web/20070717003906/http://gdayworld.thepodcastnetwork.com/2007/05/17/gday-world-238-eliezer-yudkowsky/) 17 июля 2007 года.
15. «An Intuitive Explanation of Bayes' Theorem» (http://yudkowsky.net/rational/bayes)
16. перевод (http://translatedby.com/you/an-intuitive-explanation-of-bayes-theorem/into-ru/trans/)
17. Sequences — Lesswrongwiki (http://wiki.lesswrong.com/wiki/Sequences#The_Fun_Theory_Sequence)
18. Yudkowsky — Fiction (http://yudkowsky.net/other/fiction)

Ссылки

- Personal web site (http://yudkowsky.net/)
- Biography page at KurzweilAI.net (https://web.archive.org/web/20100626002340/http://www.kurzweilai.net/bios/frame.html?main=%2Fbios%2Fbio0053.html)
- Biography page at the Singularity Institute (https://web.archive.org/web/20120606221844/http://singinst.org/aboutus/team)
- Downloadable papers and bibliography (https://web.archive.org/web/20070927150545/http://unjobs.org/authors/eliezer-yudkowsky)
- Overcoming Bias (http://www.overcomingbias.com/) — A blog to which Yudkowsky contributed regularly until 2009.
- Less Wrong (http://lesswrong.com/) — «A community blog devoted to refining the art of human rationality» founded by Yudkowsky.
- Переводы статей по рациональному мышлению на русский (http://lesswrong.ru/)
- Predicting The Future :: Eliezer Yudkowsky, NYTA Keynote Address — Feb 2003 (http://www.imminst.org/forum/index.php?s=&act=ST&f=67&t=1097&st=0)
- Eliezer Yudkowsky on The Agenda with Steve Paikin discussion panel, «Robotics Revolution and the Future of Evolution» (https://web.archive.org/web/20100217194306/http://www.q2cfestival.com/play.php?lecture_id=8014) at the Quantum to Cosmos Festival, with Hod Lipson, Michael Belfiore, Cory Doctorow

- Less Wrong Q&A with Eliezer Yudkowsky: Video Answers (http://lesswrong.com/lw/1lq/less_wrong_qa_with_eliezer_yudkowsky_video_answers/)
- Глава о Юдковском в книге «21st Century Technology and Its Radical Implications for Mind, Society and Reality» (https://books.google.ru/books?id=_YkfoiKC4PcC&pg=PA410&dq=Eliezer+Yudkowsky&hl=ru&sa=X&ei=DP9VU6KcEuak4gTu5IHgCQ&redir_esc=y#v=onepage&q=Eliezer%20Yudkowsky&f=false)
- *Ben Goertzel.* Superintelligence: Fears, Promises and Potentials (https://jetpress.org/v25.2/goertzel.htm) // Journal of Evolution and Technology. — 2015. — Vol. 24, no. 2. — P. 55—87.
- Фанфик *Harry Potter and the Methods of Rationality (https://www.fanfiction.net/s/5782108/1/Harry_Potter_and_the_Methods_of_Rationality)* (перевод: Гарри Поттер и Методы Рационального Мышления (http://www.fanfics.me/index.php?section=3&id=40982/))
