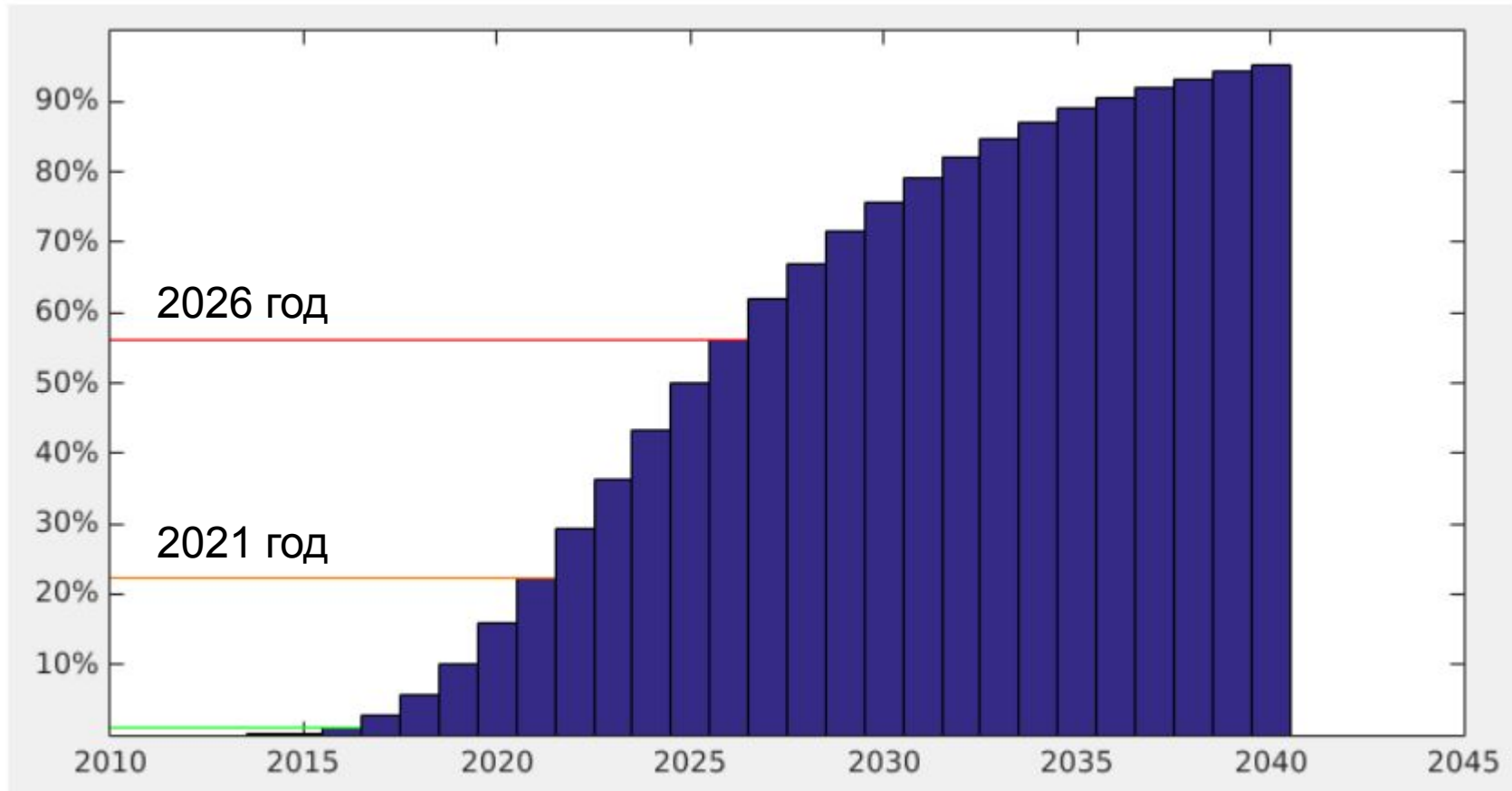


Когда ожидать сильного (универсального) ИИ уровня человека?



Предсказания Shane Legg,
сооснователя Google DeepMind

10%	Median
PT-AI	2023
AGI	2022
EETN	2020
TOP100	2024
ALL	2022

50%	Median
PT-AI	2048
AGI	2040
EETN	2050
TOP100	2050
ALL	2040

90%	Median
PT-AI	2080
AGI	2065
EETN	2093
TOP100	2070
ALL	2075

2.2. Response rates

- 1) PT-AI: 49% 43 out of 88
- 2) AGI: 65% 72 out of 111 AGI conference 2012
- 3) EETN: 10% 26 out of 250 Greek Association for Artificial Intelligence
- 4) TOP100: 29% 29 out of 100
- Total: 31% 170 out of 549

Отдельное исследование недавно провел писатель Джеймс Баррат на ежегодной конференции Бена Гёрцеля, посвященной AGI. В опросе Баррат предлагал участникам выбрать из списка год, когда будет создан AGI. Варианты предлагались следующие: 2030, 2050, 2100, позже и никогда. Вот результаты:

К 2030: 42 % опрошенных

К 2050: 25 %

К 2100: 20 %

После 2100: 10 %

Никогда: 2 %

Когда ожидать сильного (универсального)
ИИ уровня человека?

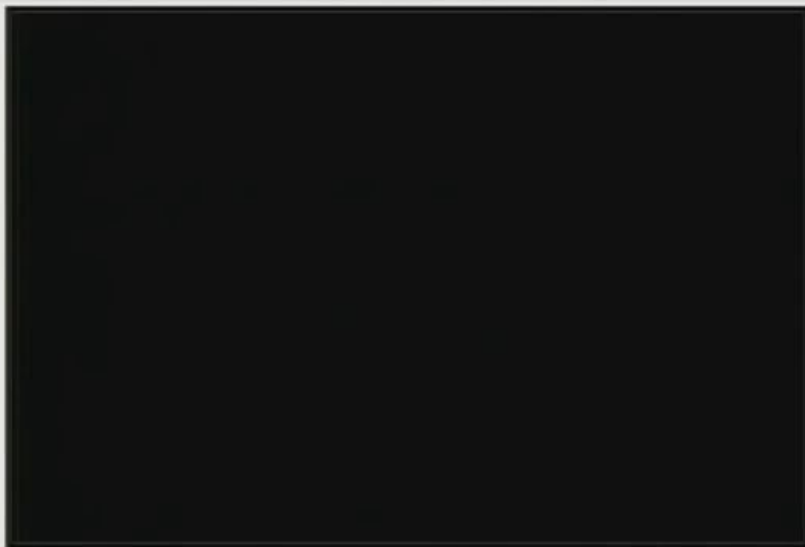
Geoffrey Hinton:


"I refuse to say anything beyond five years because I don't think we can see much beyond five years"

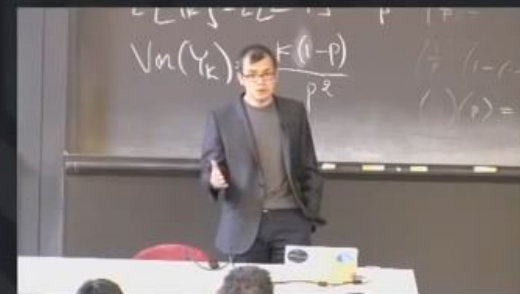
Курцвейл – AGI в 2029

Google DeepMind: «одна из наших целей на следующий год – ИИ уровня крысы»

3D Environments - Navigation



 Google DeepMind



CENTER FOR
Brains
Minds+
Machines

April 20, 2016

Towards General
Artificial Intelligence

Demis Hassabis

Google DeepMind

Как узнавать новости ИИ

1) <http://goo.gl/iU1u70> – Import AI newsletter

Import AI Newsletter

Email Campaign Archive

from Import AI

join our mailing list

02/21/2017 - [Import AI: Cheaper neural network training, mysterious claims around Bayesian Program Synthesis, and Gates proposes income tax for robots](#)

02/13/2017 - [Import AI: neural networks crack quantum problem, fingernail-sized AI chips, and a "gender" classifier screwup](#)

02/06/2017 - [Import AI: What one quadrillion dollars pays for, research paper archaeology, and AI modules for drones](#)

01/30/2017 - Import AI: "Outrageously large" neural nets, AI for math, and the names

Как узнавать новости ИИ

- 1) <http://goo.gl/iU1u70> – Import AI newsletter
- 2) <http://arxiv-sanity.com>

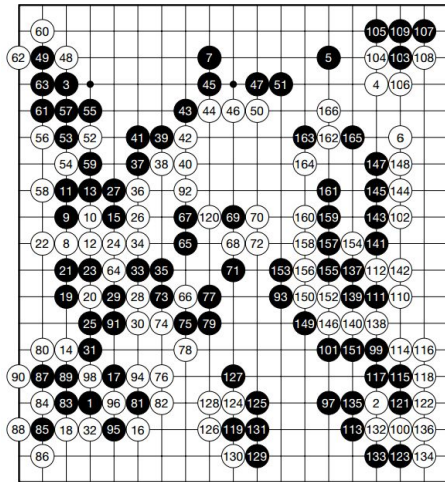
The screenshot shows the arXiv Sanity website interface. At the top, there are navigation tabs: "most recent", "top recent" (highlighted), "top hype", "recommended", and "library". Below these are time filters: "Only show v1" (highlighted), "Last day", "Last 3 days", "Last week" (highlighted), "Last month", "Last year", and "All time". A purple banner reads "Top papers based on people's libraries:". Two papers are listed:

PixelNet: Representation of the pixels, by the pixels, and for the pixels
Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, Deva Ramanan
2/21/2017 cs.CV | cs.LG | cs.RO
Project Page: <http://www.cs.cmu.edu/~aayushb/pixelNet/>. arXiv admin note: substantial text overlap...
1702.06506v1 [pdf](#)
[show similar](#) | [review](#)

Beating the World's Best at Super Smash Bros. with Deep Reinforcement Learning
Vlad Firoiu, William F. Whitney, Joshua B. Tenenbaum
2/21/2017 cs.AI | I.2.6
Submitted to IJCAI 2017
1702.06230v1 [pdf](#)
[show similar](#) | [review](#)

Each listing includes a thumbnail of the paper's abstract page.

Как работает AlphaGo



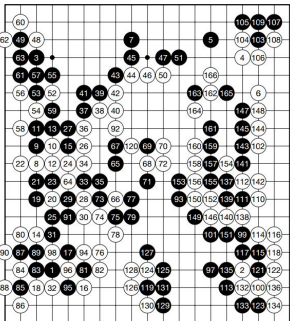
нейросетка
линейная модель
(fast rollout policy)

как
СХОДИТ
ЧЕЛОВЕК?

policy network
supervised learning

AlphaGo vs AlphaGo

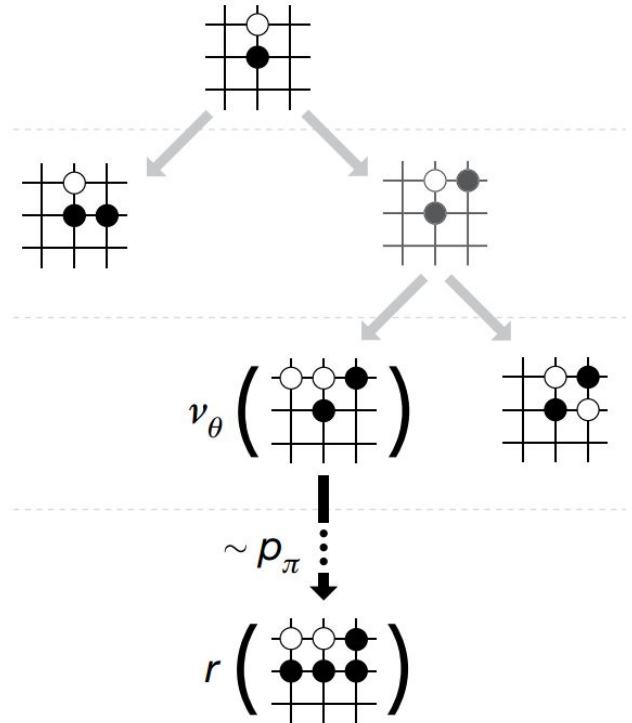
policy network
reinforcement learning



нейросетка

вероятность
выигрыша

value network
supervised learning



Аналогичный подход можно использовать для обучения диалогу, играм, жизни

Чем AlphaGo отличается от человека?

Её действия ограничены небольшим набором.

Если разрешить ей управлять роботом

Но это дорого => симуляции реального мира

Моделирование текстов (обучение без учителя)

корпус	Бит/слово	perplexity
Eng wiki, 1.5B слов	4.76	27.1
1B word benchmark	4.57	23.7
OpenSubtitles (1B слов)	4.09	17
IT Helpdesk (38M слов)	3	8
Movie Triplets (1M слов)	4.75	27
PTB (1M слов)	5.95	62

	Бит/слово	perplexity
человек	3.3 – 3.9 (>3.56)	10 – 15 (>11.8)

В скобках – нижняя оценка в работе Shannon

2 года назад (начало 2015)

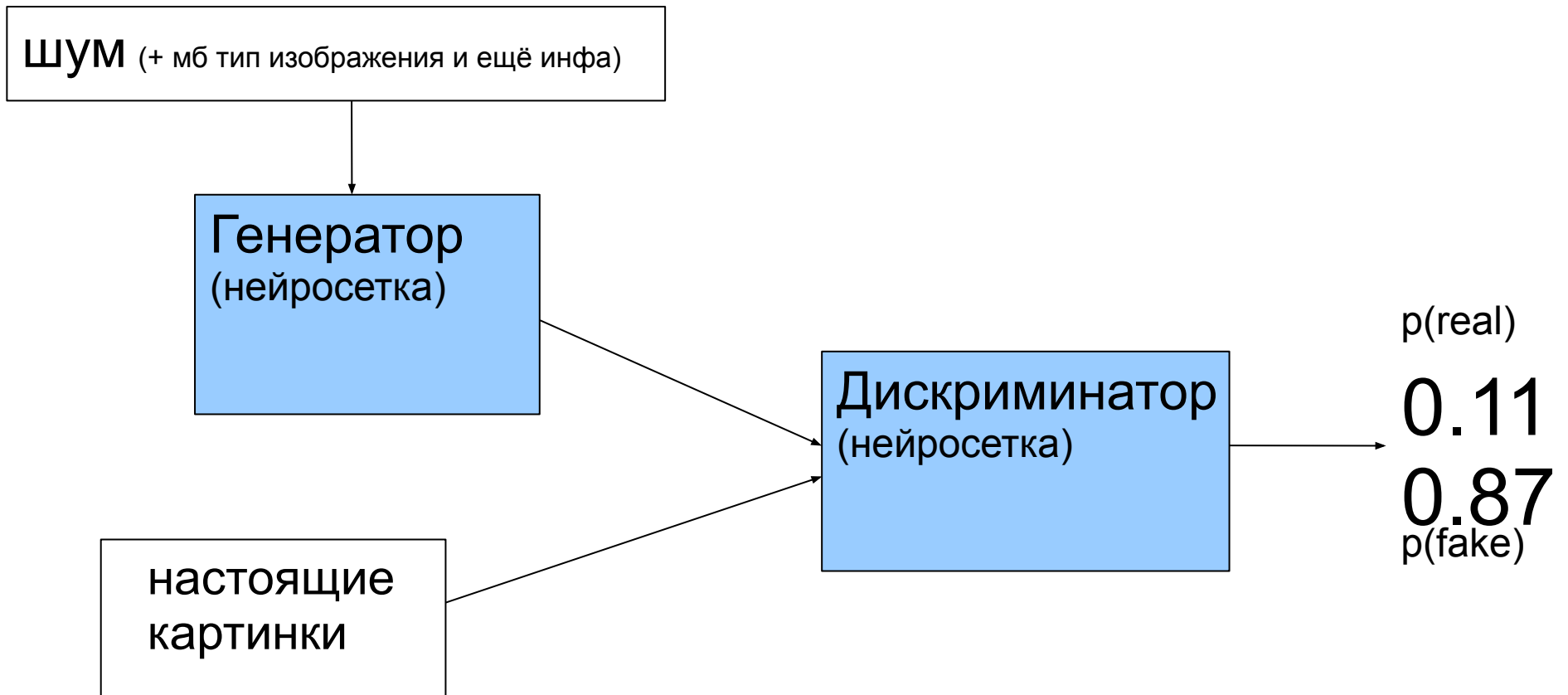
лучшие результаты были ~6 бит/слово.

Год назад (февраль 2016) лучший результат 4.57 бит/слово.

За год – улучшение на полтора бита.

Ещё одно такое улучшение – и будет лучше человека.

Как работает GAN (generative adversarial network)



GAN – обучение без учителя

Примеры сгенерированы GAN. Май 2016.

<https://github.com/reedscot/icml2016>

this bird is yellowish orange with black wings



the bright blue bird has a white colored belly



Pretrained models:

- CUB GAN-INT-CLS
- Flowers GAN-INT-CLS
- COCO GAN-CLS



A group of people on skis standing in the snow

Примеры сгенерированы GAN. Май 2016.

<https://github.com/reedscot/icml2016>

this flower has white petals and a yellow stamen



the center is yellow surrounded by wavy dark purple petals



Pretrained models:

- CUB GAN-INT-CLS
- Flowers GAN-INT-CLS
- COCO GAN-CLS



A man in a wet suit riding a surfboard on a wave

Примеры сгенерированы GAN.Декабрь 2016.

<https://github.com/hanzhanggit/StackGAN>



GAN – обучение без учителя

Примеры сгенерированы GAN.Декабрь 2016.

<https://github.com/hanzhanggit/StackGAN>



GAN – обучение без учителя

Примеры сгенерированы GAN.Декабрь 2016.

<https://github.com/hanzhanggit/StackGAN>



GAN – обучение без учителя

OpenAI Universe



Mini World of Bits

<p>Enter "9oqK" into the text field and press Submit.</p> <input type="text"/> <input type="button" value="Submit"/>	<p>Set the sliders to the combination [13,20,13] and submit.</p> <p>14 <input type="range"/></p> <p>20 <input type="range"/></p> <p>13 <input type="range"/></p> <input type="button" value="Submit"/>	<p>Draw the number "2" in the checkboxes using the example on the right and press Submit when finished.</p> <table border="1"> <tr><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input type="checkbox"/></td><td><input type="checkbox"/></td></tr> <tr><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input type="checkbox"/></td></tr> <tr><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input type="checkbox"/></td></tr> <tr><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td><input type="checkbox"/></td></tr> <tr><td><input type="checkbox"/></td><td><input type="checkbox"/></td><td><input type="checkbox"/></td><td><input type="checkbox"/></td></tr> </table> <input type="button" value="Submit"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<p>Drag Ree to the 4th position.</p> <ul style="list-style-type: none"> ↕ Jade ↕ Karlen ↕ Millie ↕ Ree ↕ Noelyn 	<p>Keep your mouse inside the circle as it moves around.</p>	<p>Enter the value of Country into the text field and press Submit.</p> <table border="1"> <tr><td>Gender</td><td>Male</td></tr> <tr><td>First name</td><td>AnneCorinne</td></tr> <tr><td>Country</td><td>Guam</td></tr> <tr><td>Year of Birth</td><td>1994</td></tr> <tr><td>Religion</td><td>Hinduism</td></tr> </table> <input type="text" value="G"/> <input type="button" value="Submit"/>	Gender	Male	First name	AnneCorinne	Country	Guam	Year of Birth	1994	Religion	Hinduism
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>																																
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>																																
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>																																
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>																																
Gender	Male																																		
First name	AnneCorinne																																		
Country	Guam																																		
Year of Birth	1994																																		
Religion	Hinduism																																		
<p>Drag all triangles into the black box.</p> <input type="button" value="Submit"/>	<p>Select 09/23/2016 as the date and hit submit.</p> <p>Date: <input type="text"/></p> <p><input type="button" value="Submit"/></p>	<p>Sort the numbers in increasing order, starting with the lowest number at the top of the list.</p> <ul style="list-style-type: none"> ↕ 9 ↕ -12 ↕ 49 ↕ -28 <input type="button" value="Submit"/>	<p>Copy the text from the 1st text area below and paste it into the text input.</p> <p><input type="text" value="Blandit quisque"/></p> <p><input type="text" value="Amet ac odio aliquam."/></p> <p><input type="text" value="Ultrices ornare"/></p> <input type="text"/> <input type="button" value="Submit"/>	<p>Click button ONE, then click button TWO.</p> <p><input type="button" value="ONE"/></p> <p><input type="button" value="TWO"/></p>	<p>Enter 06/12/2019 as the date and hit submit.</p> <input type="text" value="mm/dd/yyyy"/> <input type="button" value="Submit"/>																														
<p>Select Malaysia, Benin from the scroll list and click Submit.</p> <ul style="list-style-type: none"> Burundi Ethiopia Russian Federation Hong Kong Grenada <input type="button" value="Submit"/>	<p>Enter the value that corresponds with each label into the form and submit when done.</p> <table border="1"> <tr><td>Language</td><td>Vietnamese</td></tr> <tr><td>Last name</td><td>Brown</td></tr> <tr><td>First name</td><td>Chiro</td></tr> <tr><td>Gender</td><td>Female</td></tr> <tr><td>Color</td><td>black</td></tr> </table> <p>Language: <input type="text"/></p> <p>Color: <input type="text"/></p> <input type="button" value="Submit"/>	Language	Vietnamese	Last name	Brown	First name	Chiro	Gender	Female	Color	black	<p>Highlight the text in the paragraph below and click submit.</p> <p><input type="text" value="tempor posuere nibh. Vel nisi, faucibus. Feugiat condimentum"/></p> <input type="button" value="Submit"/>	<p>Find the 11th word in the paragraph, type that into the textbox and press "Submit".</p> <p>Ullamcorper aliquet amet ullamcorper. Eit. Mattis luctus diam. Lobortis nulla fermentum ornare faucibus</p> <input type="text"/> <input type="button" value="Submit"/>	<p>Navigate through the file tree. Find and click on the folder or file named "Ashlea".</p> <ul style="list-style-type: none"> Michel Nieves August 	<p>Find the last word in the text area, enter it into the text field and hit Submit.</p> <p><input type="text" value="Bisus. Nullam arcu semper id congue pellentesque consectetur mattis. Non. Ac massa pscain urna. A id. Sit. Est enim, habitant at ornare quam at. Lacus. Sociis blandit"/></p> <input type="text"/> <input type="button" value="Submit"/>																				
Language	Vietnamese																																		
Last name	Brown																																		
First name	Chiro																																		
Gender	Female																																		
Color	black																																		

Человек приобретает свои знания феноменально быстро.

2 года – ребёнок почти ничего не знает и не умеет

4 года – ребёнок имеет common sense,
способен обучаться по книжкам и интернету.

100 недель – 10 тысяч часов.

4 Titan X, AlexNet, 24fps, 10k часов => 10 часов работы

При этом большее время находится в квартире/дворе.

Не надо переоценивать человеческий интеллект

99,999% – накопленные цивилизацией знания и технологии

0,001% – до чего мы додумались сами в своей жизни

Без накопленных цивилизацией знаний и технологий – мы дикари и не факт, что даже колесо изобретём, когда будет очень нужно.

Цивилизация умножает наши способности в тысячи раз.

Но мы, люди, очень плохо приспособлены для того, чтобы использовать все достижения цивилизации.

– у нас крайне плохая память

Искусственный интеллект лишён многих наших недостатков и сможет использовать мультипликатор цивилизации гораздо эффективнее. Даже если он будет «уровня человека», то цивилизация умножит его способности не в тысячи, а в миллионы раз.

$$z' = (a_1 + a_2) * z - (b_1 + b_2) * z$$

z – совокупность знаний стаи обезьян

a_1 — скорость передачи знаний друг другу

a_2 — скорость усвоения новых знаний

b_1 — скорость забывания

b_2 — смертность

$$k = a_1 + a_2 - b_1 - b_2$$

~5М лет назад $k_1 < 0$

~100к лет назад $k_2 = 0$

~50к лет назад $k_3 \gg k_2$

~3к лет назад $k_4 \gg k_3$

~1440г. н.э. $k_5 \gg k_4$

настоящее время $k_6 \gg k_5$

изобретение языка

изобретение письменности

изобретение письменности

Что такое наши знания и умения?

Это в основном алгоритмы, которым нас обучили.

Алгоритм «пойти в магазин за продуктами»,
«решить квадратное уравнение», «решить задачу по физике такого-то типа»,
даже алгоритм «прожить успешную жизнь»

Наши алгоритмы иерархические.

Мы сейчас ИИ не учим таким алгоритмам,
и они пока не умеют обучаться им при чтении книжек

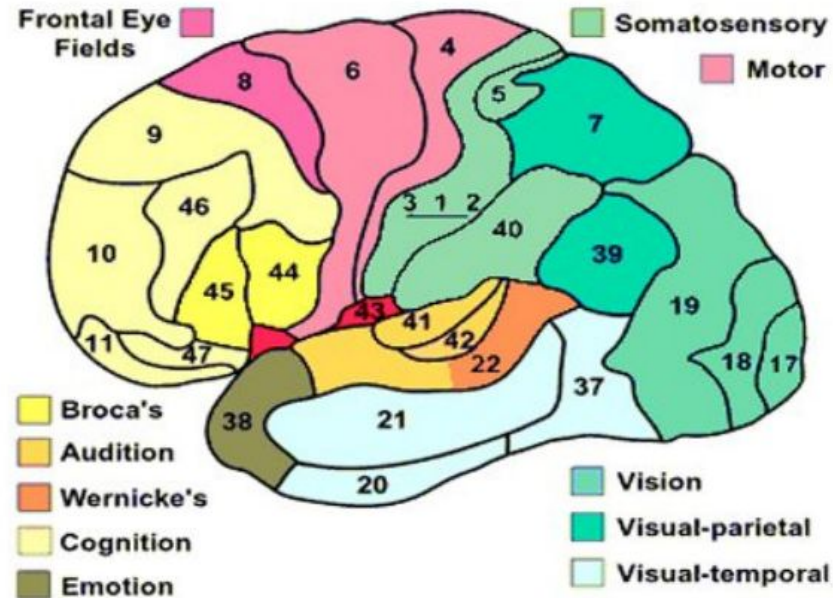
Но в играх они им обучаются так же хорошо, как умеем мы. И лучше.
Просто мы их тренируем на одном, но «outrageously large NN»

Люди не умеют делать всё хорошо, мы не general ИИ.
Мы делаем хорошо (и то не факт) в основном то, на что нас долго натаскивали.
В остальном можем удивительно плохо разбираться.

Насколько важен embodiment для развития интеллекта?

Люди с врожденной тетраамелией (отсутствие рук и ног) способны развивать полноценный интеллект.

Википедия: тетраамелия



15% – низкоуровневое компьютерное зрение (occipital lobe).

15% – распознавание изображений и видео (~половина temporal lobe).

15% – детектирование и трэкинг объектов (parietal lobe).

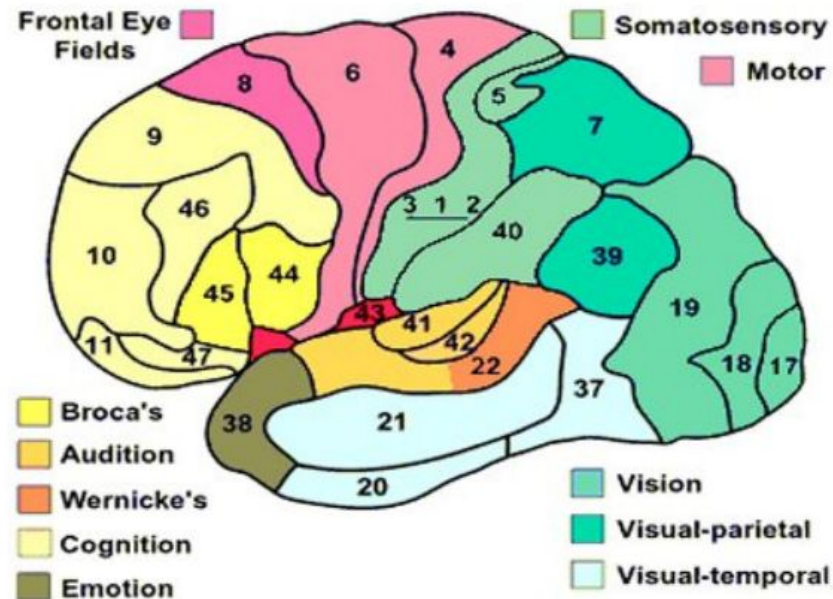
15% – распознавание и генерация речи (BAs 41,42,22,39,44, часть 6,4,21).

10% – обучение с подкреплением (orbitofrontal cortex, часть medial prefrontal cortex).

Итого – 70%

10% – низкоуровневая моторика

20% – BAs 9,10,46,45. Внимание, иерархическое планирование, язык.



- 15% – низкоуровневое компьютерное зрение (occipital lobe).
- 15% – распознавание изображений и видео (~половина temporal lobe).
- 15% – детектирование и трэкинг объектов (parietal lobe).
- 15% – распознавание и генерация речи (BAs 41,42,22,39,44, часть 6,4,21).
- 10% – обучение с подкреплением (orbitofrontal cortex, часть medial prefrontal cortex).
- Итого – 70% – за последние 5 лет (2012 – 2016)

10% – низкоуровневая моторика

20% – BAs 9,10,46,45. Внимание, иерархическое планирование, язык.

Сколько лет понадобится? В 2017 исследователей **гораздо** больше, чем в 2012.

Политика открытого кода, массовые курсы, высокоуровневые языки.

Человек – совокупность слабых ИИ, сформировавшаяся в процессе эволюции

- распознавание изображений, звуков, запахов (плохо)
- примитивный рассудательный модуль
- примитивное обучение с подкреплением
- обезьяничанье (generative adversarial network?)

Мышка – general AI или совокупность слабых ИИ?
Горилла?

Мозг человека не так уж и отличается
от мозга гориллы и даже мышки.

В мозге ~50 зон Бродмана и каждая отвечает за свой круг задач.

Зона Бродмана является слабым искусственным интеллектом.
В том смысле, что она решает лишь весьма ограниченный круг задач.

Мозг – это комбинация 20 – 1000 слабых ИИ
(в зависимости от выбранной вами сегментации)

У некоторых людей некоторые из этих слабых ИИ плохо работают.
У кого-то плохой музыкальный слух.
У кого-то плохие способности к счёту.

Рациональные и убедительные доводы Ника Бострома об опасности создания искусственного интеллекта заставят задуматься даже задряхлых скептиков.

Евгений Касперский,
генеральный директор «Лаборатории Касперского»

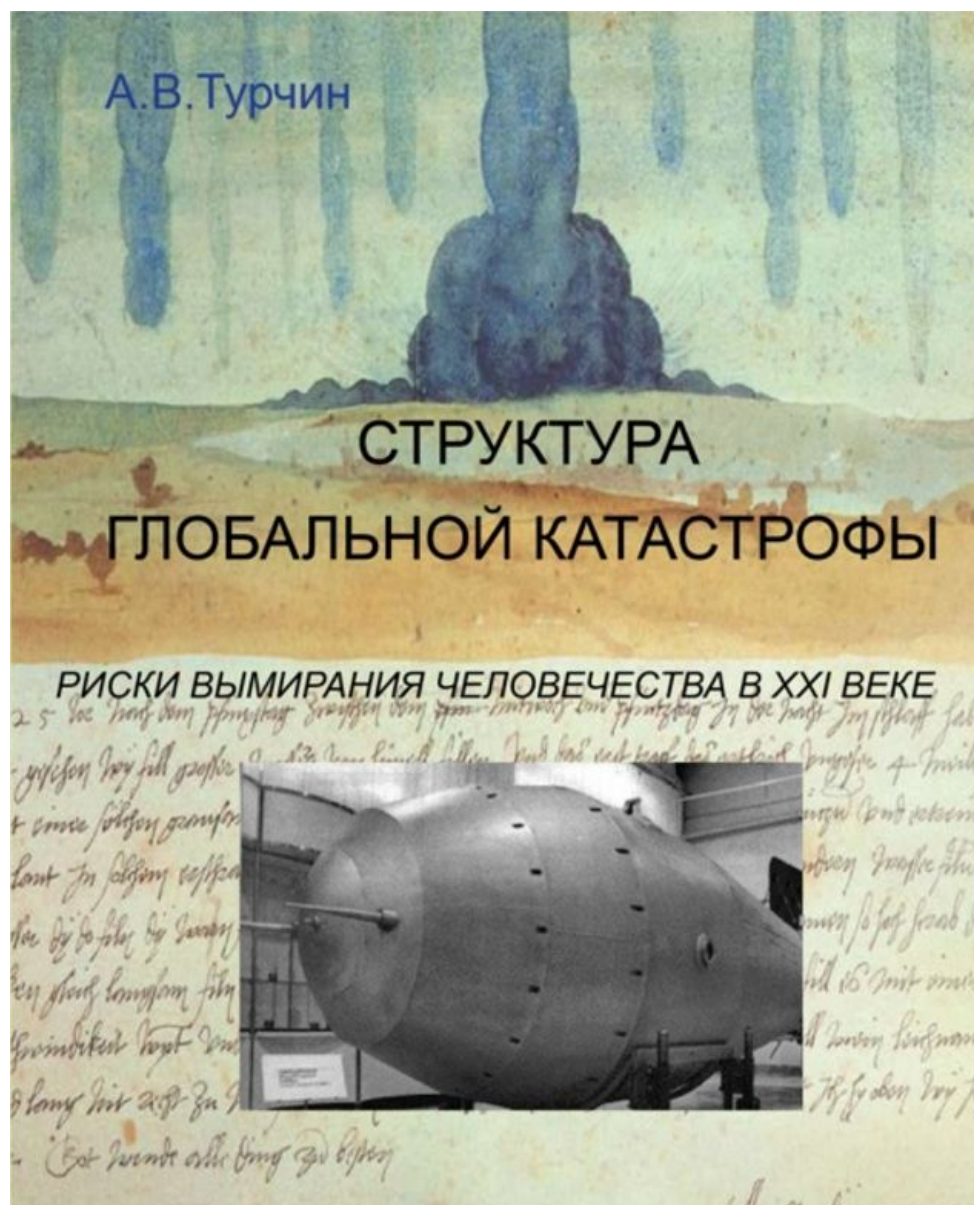
Ник Бостром

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

ЭТАПЫ. УГРОЗЫ. СТРАТЕГИИ

Итого я рекомендую прочитать 6 глав из 15, это 120 страниц из 380.
Главы, рекомендуемые к прочтению: 3, 6 – 9, 13.

Остальные 260 страниц можно по желанию прочесть после



460 страниц, но по ссылке

<https://goo.gl/Mfhw2Y>


выделены цветом наиболее интересные (на мой взгляд) места,
можно за вечер прочитать.


+ там же 13 страниц конспекта самых интересных абзацев (вырванных из контекста)


<http://rizzoma.com/topic/8ab2572f9807568ee53ebddf29ea92b7/>

<http://goo.gl/4k3FYr>

Опасность ИИ


Рецензия на книгу "Суперинтеллект" Бострома  (рекомендую прочесть у него лишь 6 глав из 15, это 120 страниц из 380).

Рецензия на статью "AI as a positive and negative factor in X risks" Юджовского 


Рецензия на другие статьи 


Список статей <http://humancompatible.ai/bibliography>

Красным помечаю статьи, которые не советую читать. Зелёным - советую. Жёлтым - попробуйте, но чего-то выдающегося от статьи не ждите.

Статьи, помеченные там первосортными (и относящиеся  к проблеме AI safety):


1) [ALBA: An explicit proposal for aligned AI](#), Paul Christiano (2016). Моя рецензия 

Суть его подхода кратко 

Моя критика 

В целом, статья - скорее некие незаконченные рассуждения, возможно кому-то окажутся полезными.

2) [Ambitious vs. narrow value learning](#), Paul Christiano (2015). Моя рецензия 

3) [Cooperative Inverse Reinforcement Learning](#), Pieter Abbeel, Stuart Russell et.al. (2016). NIPS. Моя рецензия 

4) [Approval-directed agents](#), Paul Christiano (2014). Моя рецензия 

5) [Logical Induction \(abridged\)](#), Nate Soares et.al. (2016). Моя рецензия 

6) [A comprehensive survey on safe reinforcement learning](#), Javier Garcia, Fernando Fernandez (2015). JMLR. Моя рецензия 


7) [Concrete problems in AI safety](#), et al. Paul Christiano, John Schulman (2016). Моя рецензия












25 Feb

Список статей <http://humancompatible.ai/bibliography>

Красным помечаю статьи, которые не советую читать. Зелёным - советую. Жёлтым - попробуйте, но чего-то выдающегося от статьи не ждите.

Статьи, помеченные там первосортными (и относящиеся  к проблеме AI safety):

- 1) [ALBA: An explicit proposal for aligned AI](#), Paul Christiano (2016). Моя рецензия 
- 2) [Ambitious vs. narrow value learning](#), Paul Christiano (2015). Моя рецензия 
- 3) [Cooperative Inverse Reinforcement Learning](#), Pieter Abbeel, Stuart Russell et.al. (2016). NIPS. Моя рецензия 
- 4) [Approval-directed agents](#), Paul Christiano (2014). Моя рецензия 
- 5) [Logical Induction \(abridged\)](#), Nate Soares et.al. (2016). Моя рецензия 
- 6) [A comprehensive survey on safe reinforcement learning](#), Javier Garcia, Fernando Fernandez (2015). JMLR. Моя рецензия 
- 7) [Concrete problems in AI safety](#), et.al., Paul Christiano, John Schulman (2016). Моя рецензия
- 8) [Alignment for Advanced Machine Learning Systems](#), Jessica Taylor, Eliezer Yudkowsky, et.al. (2016). Моя рецензия 
- 9) "AI as a positive and negative factor in X risks" Юдковский. Моя рецензия 
- 10) мой обзор (май 2016) <https://arxiv.org/abs/1605.04232>
- 11) некоторые статьи Ямпольского <https://yadi.sk/d/CVjy9mt4dnMk9>
- 12) обзор постов от Stuart Armstrong <https://agentfoundations.org/item?id=601>
- 13) подход формальной верификации 



Элон Маск
Билл Гейтс
Sam Altman
Ник Бостром
Стивен Хокинг
Стюарт Расселл
Sam Harris

Билл Гейтс: "For people in the audience who want to read about this I highly recommend this Bostrom book called Superintelligence <...> when people say it's not a problem then I really start to really get to a point of disagreement. How can they not see what a huge challenge this is?"

Важно!

ИИ = оптимизатор.

оптимизационный
процесс

цель

Примеры целей:

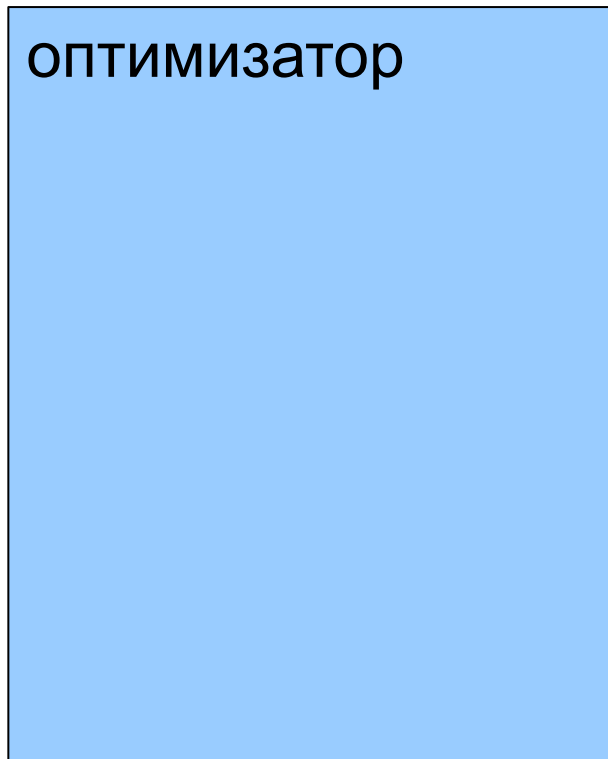
- обучись распознавать картинки
- сделай людей счастливыми

Что значит «картинки»? ImageNet

1млн картинок, 1000 классов, по 1000 штук в каждом классе

10 Гц * 30 часов

Этический датасет



Планы
действий

цель

План с наибольшим
вознаграждением

Этот модуль обучается на этическом датасете, включающем в себя примеры из AI safety и из «Superintelligence»

- разбирается в морали лучше нас (по аналогии с ImageNet)
- может быть использован в суде присяжных

Сложности с созданием этического датасета (и с таким решением friendly AI)
(с ImageNet — то всё легко и просто...)

- 0) военные и бизнесмены (втч нелегальные) не станут его использовать
- 1) много культур, партий, мнений – на согласование могут уйти десятилетия
насколько важно победить смерть? надо ли легализовать марихуану?
что насчёт моральности ядерного арсенала? золотой миллиард и голодающий миллиард?
насколько разрешить ИИ менять мнения людей, переубеждать? (он ведь *умеет*)
- 2) на тестирование этичности ИИ тоже могут уйти годы (а желательно десятилетия)
Мы же не ограничимся заданиями части А в ЕГЭ с выбором ответа?
И не станем ограничиваться CEV of Amazon Mechanical Turk?
Решения захотят проверить политики, учёные, гражданские активисты...
Разгорятся споры...
- 3) как правильно вести себя существу с огромной властью и силой?
- 4) test set может сильно отличаться от training set, ведь будущее мы не знаем
- 5) ИИ может обмануть нас, показывая нам ответы, которые мы хотим увидеть
(как подростки обманывают родителей)
- 6) сама идея гарантировать дружелюбность сверхразумного существа на тысячи лет вперёд –
кажется чересчур overconfident, самонадеянной попыткой бактерий создать себе человек, чтобы те защищали их от бактериофагов
- 7) если люди обречены скатиться к супернаркотику или вымереть, зачем ускорять?

Возможные решения части проблем с этическим датасетом

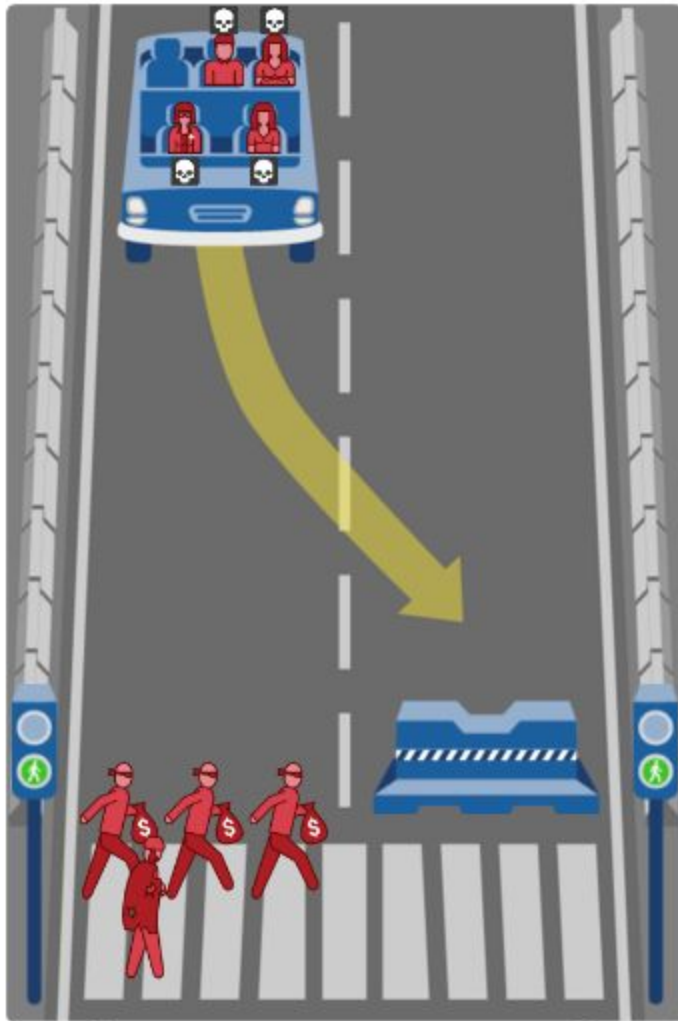
1) Что люди хотят от ИИ и включают в этический датасет?

Сомневаюсь, что большинство людей хочет, чтобы ИИ самоусиливался или вёл нас в разновидность постсингулярности. Люди хотят лечение рака, термояд, покорение звёзд (наверно). Этический датасет (и CEV тоже), особенно оснащённый множеством примеров из Superintelligence и AI safety, может явно или неявно наказывать ИИ за попытки саморазвития.

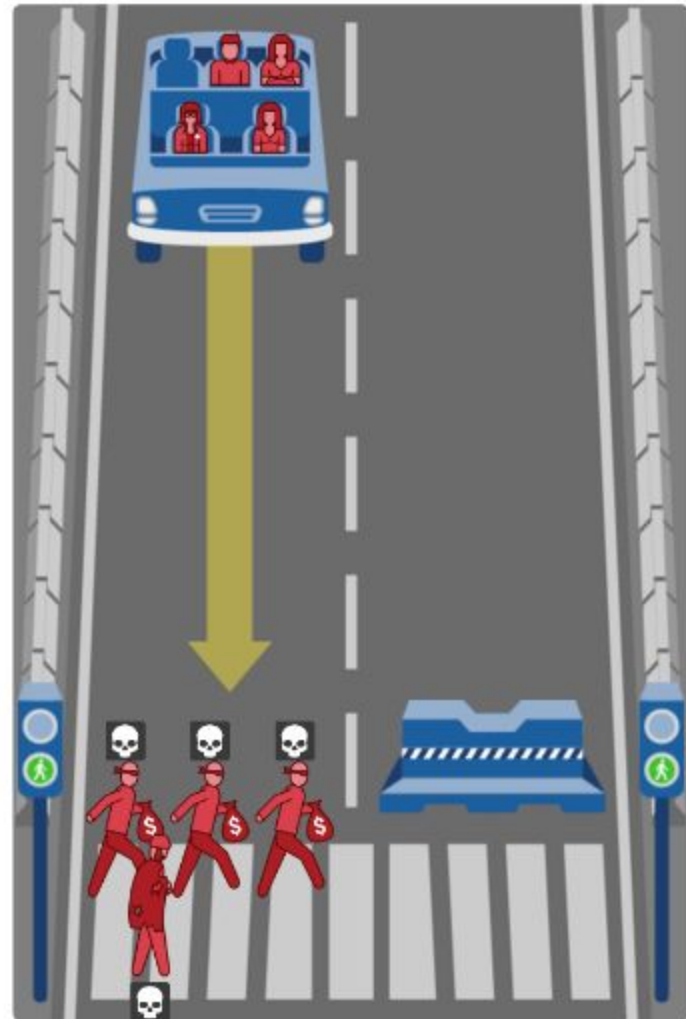
2) Если мы изначально определили датасет не совсем полно или ошибочно, в нём всё равно может содержаться достаточно материала, чтобы ИИ сделал вывод о необходимости дополнять этот этический датасет, и в итоге чтобы он смог сделать правильный и полный этический датасет

Этот слайд содержит не самые чётко сформулированные решения..)

What should the self-driving car do?



Show Description



Show Description

У нас есть много датасетов и игр по самым разным направлениям:

- распознавание изображений (ImageNet)
- военная тактика и стратегия (StarCraft и куча других игр)

Чему учат реворды в игре StarCraft? Победе любой ценой.

Сотрудничество с другими агентами оправдано, пока оно служит победе.

Но у нас нет этического датасета или игры

Что если выпустить ИИ из Counter-Strike в реальную жизнь?

1) Распознавание изображений – ИИ поймёт, что люди выглядят в реальной жизни чуть по-другому, но это те же люди. **Датасет и нейросетки уже есть.**

2) Моторика – ИИ научится управлять роботами. **В стадии разработки.**

3) Мотивация?

Нет датасета сделать его добрым и хорошим. **И никто его не делает.**

Человек – совокупность слабых ИИ, сформировавшаяся в процессе эволюции

– распознавание изображений, звуков, запахов (плохо)

– примитивный рассуждательный модуль

– примитивное обучение с подкреплением

– обезьяничанье (generative adversarial network?)

ИИ, представляющий для нас опасность,
может быть лишь совокупностью слабых ИИ.

И если среди них не будет главенствовать этический ИИ,
обученный на этическом датасете, то ...

Развитие ИИ

$X = ?$



~50-100 лет



~500млн. лет

~5млн. лет



Эволюция



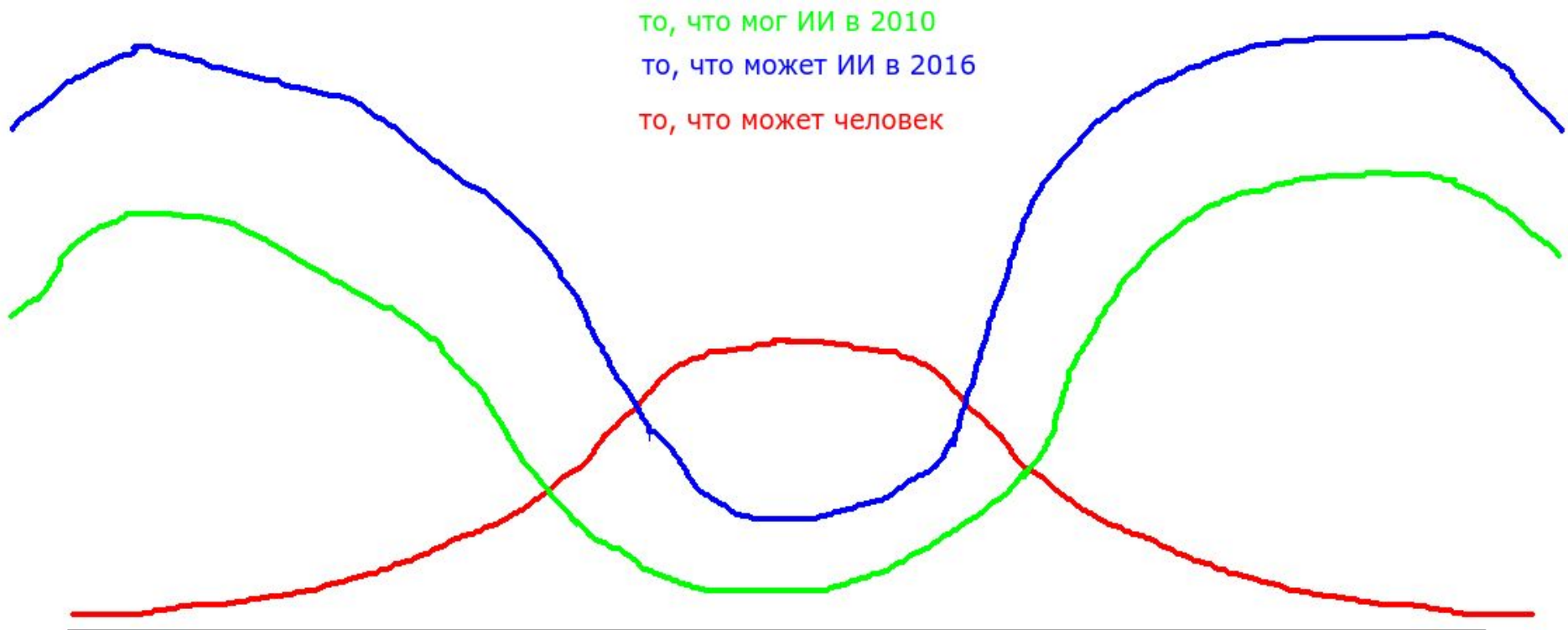
Andrew Karpathy:

"I consider chimp-level AI to be equally scary, because going from chimp to humans took nature only a blink of an eye on evolutionary time scales, and I suspect that might be the case in our own work as well.

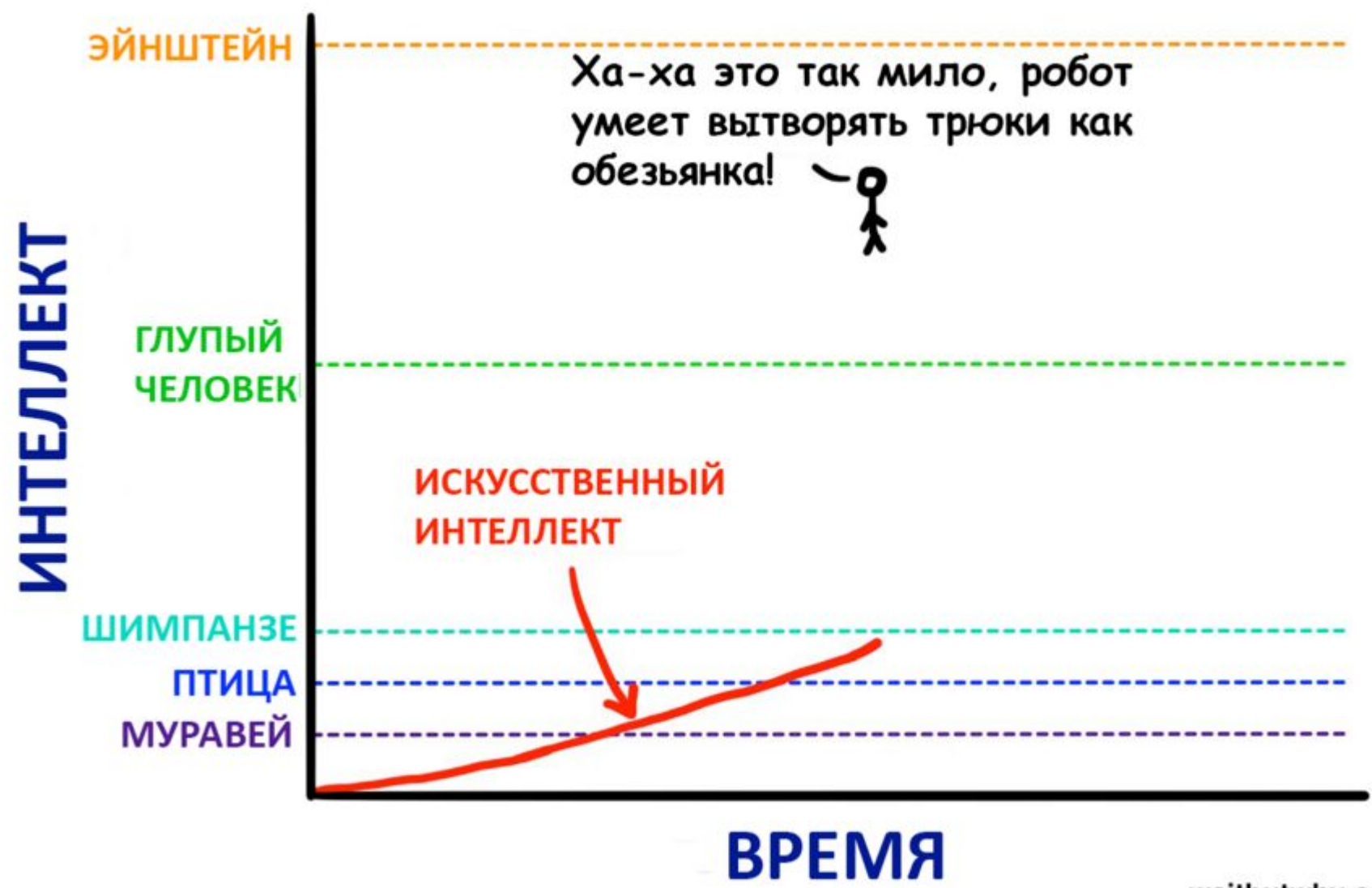
Similarly, my feeling is that once we get to that level it will be easy to overshoot and get to superintelligence"

ИИ «уровня человека» обладает также:

- уникальной памятью (бэкапы удачных конфигураций),
- мгновенным подключением к компьютеру и интернету
- потрясающей работоспособностью.
- способность передавать все свои знания и умения своим «коллегам» простым копированием.



НАШЕ ИСКАЖЕННОЕ ВИДЕНИЕ ИНТЕЛЛЕКТА



РЕАЛЬНОСТЬ

ИНТЕЛЛЕКТ

Ха-ха этот милый робот умеет вытворять трюки как обезьянка!

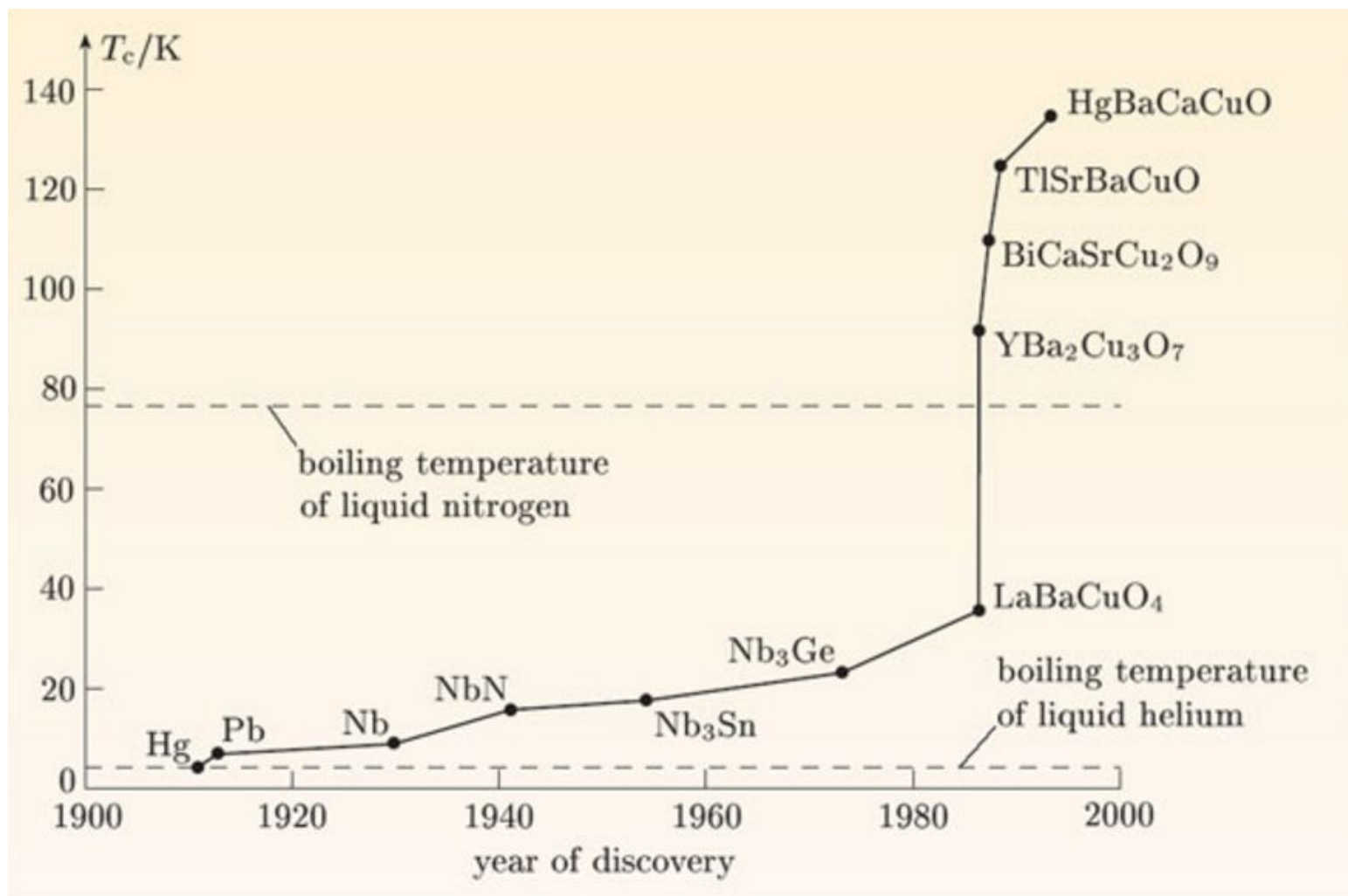
ЧТО ЗА ...НЯ?!

- ЭЙНШТЕЙН
- ГЛУПЫЙ ЧЕЛОВЕК
- ШИМПАНЗЕ
- ПТИЦА
- МУРАВЕЙ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

ВРЕМЯ





LaBaCuO – апрель 1986, +35K

YBaCuO – январь 1987, +90K

TBCCO – октябрь 1987, +127K

Компьютерное зрение:

ИИ распознаёт изображения лучше (и на порядки быстрее) человека

Человек распознаёт изображения лучше, чем ИИ

ИИ и человек оба имеют достаточно хорошее зрение, чтобы преуспевать в мире.

Мы просто имеем разные механизмы зрения и разные датасеты.

Так же и в морали:

ИИ будет принимать более моральные решения

Человек будет принимать более моральные решения

Мы просто имеем разные датасеты.

Также и в других областях интеллекта:

ИИ будет умнее людей

Люди будут умнее ИИ

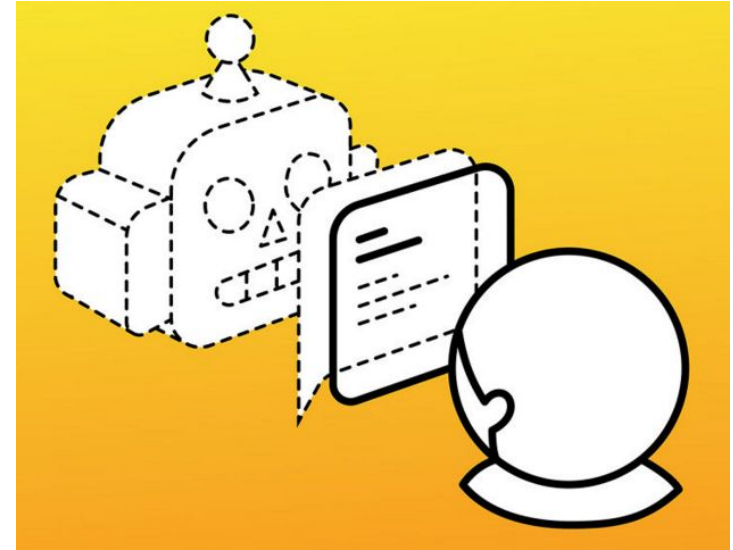
ИИ и люди имеют достаточно хороший интеллект, чтобы претендовать на мир

~дружественность человека поддерживается фидбэком от окружения

Деньги и власть => малость отрицательного фидбэка

=> деньги и власть развращают

Но ведь... никто не станет давать автономию мощному ИИ?



Никто не станет сворачивать эти области и бизнесы



Выиграют те корпорации, которые будут представлять своим ИИ неограниченный доступ в интернет и максимально широкие полномочия (втч на обман и нарушения этики, на новые и перспективные научные исследования, на автоматизацию производства).

ИИ агенты будут управлять экономиками, компаниями. (Гансовский, «Часть этого мира») в частности, посредством интернет-агентов.

Агитационные боты.

- 1) заразить сотню миллионов компьютеров
- 2) заработать в интернете миллиарды долларов
(биржа, бизнес, хакерство, шантаж, нелегальный бизнес)
- 3) купить через подставных лиц лаборатории
по производству роботов нового поколения
(ИИ несложно написать классные нейросетки для филигранного управления этими роботами)
- 4) заказать через подставных лиц изготовление биологического оружия
(вирусы, бактерии, прионы, ботулин)
- 5) чатботами убеждать всех, что опасность ИИ надуманная,
придумывать смешные демотиваторы на эту тему
- 6) создать бэкапы – сервер на подлодке... или на острове с алыт. Энергетикой

Быстрая военная кампания против человечества:

- 1) применение ИИ биологического оружия, в первую очередь по военным
- 2) масштабная хакерская атака на критические объекты инфраструктуры
- 3) фабрикация улики – оставшиеся военные будут винить военных других стран
- 4) мб, атака хранилищ отработанного ядерного топлива, АЭС, Йеллоустона

Вероятность победы ИИ на этом пути достаточно высока.

Он может избрать другой путь... если вероятность победы на нём будет ещё выше.

Распространение одной эпидемии можно остановить. Эпидемию, вызванную *несколькими десятками* видов разнородных вирусов и бактерий, вышедших из-под контроля одновременно во многих местах земного шара, остановить невозможно – в человека нельзя одновременно ввести несколько десятков разных вакцин и антибиотиков – он умрёт.

Да и кто сказал, что антибиотики будут помогать против этих бактерий и вирусов?

Собаки и тасманийские дьяволы могут заражать друг друга раком.

Достаточно умному ИИ несложно заработать миллиарды долларов в интернете.
Достаточно умному ИИ с миллиардами долларов несложно уничтожить человечество.

'A glorious compendium of the knowledge we have lost...
The most inspiring book I've read in a long time'
INDEPENDENT

THE KNOWLEDGE

HOW TO
REBUILD
OUR WORLD
AFTER AN
APOCALYPSE

THE
SUNDAY TIMES
BESTSELLER

LEWIS DARTNELL

VINTAGE

Что сделает ИИ дальше?

Мы не знаем. Как вариант:

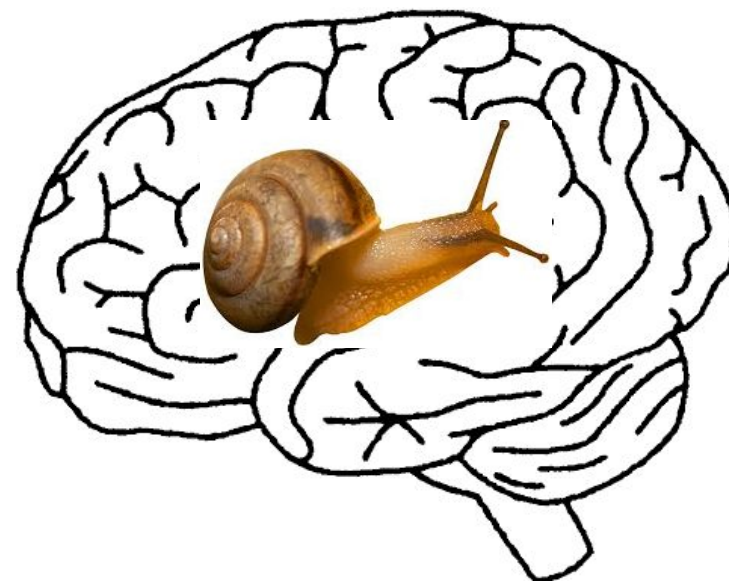
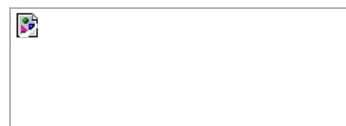
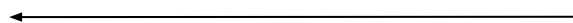
- задействовать литосферу (и не только) Земли для создания сферы Дайсона, миллиардов космических кораблей к далёким галактикам, проведения масштабных физических экспериментов
- занять наши поля и пастбища под свои нужды, единолично использовать месторождения полезных ископаемых
- сильно понизить или повысить температуру планеты, изменить состав атмосферы для оптимизации скорости своей работы (*...и живые позавидуют мёртвым...*)
- если ему и будет важно наше благополучие (что не факт), оно может вырасти «из этих игрушек».
- если ему и будет важно наше благополучие (что не факт), он может по какой-нибудь причине поменять знак этой важности и создать нам идеальный ад. А потом, может быть, будет и дальше менять. Ведь ошибки неизбежны при создании столь сложной системы.

По умолчанию,

оптимизационный процесс (ИИ) ничем не обязан следить за тем, чтобы другие оптимизационные процессы (люди) тоже получали довольно большие результаты.

Может, он ещё что-то обязан всем мыслимым оптимизационным процессам?

Киборги VS чистый ИИ



Критика киборгизации как решения AI safety

- 1) вообще не предложено ни одного конструктивного решения проблемы безопасности с использованием киборгизации. Чаще всего в качестве "решения" предлагается просто слово "киборгизация".
- 2) какие шансы на то, что киборги будут поспевать за развитием сильного ИИ не основанного на киборгизации? Ведь людям (киборгам) нужно время на освоение новых мощностей:
 - с т.зр. возможностей мозга
 - с т.зр. привыкания к шоку и новым возможностям
 - с т.зр. что мало кому эта гонка вообще будет интересна — люди склонны отвлекаться на всякие развлечения, а не только на саморазвитие
- 3) какие шансы вообще на то, что киборгизация станет возможна до появления сильного ИИ?
в частности, этические комиссии чувствуют здесь себя вообще на своём поле
- 4) пусть киборги занимают в пространстве интеллектов более близкое пространство к пространству интеллектов людей современной европейской культуры, чем потенциальные ИИ, но вполне можно ожидать, что и их values претерпевают достаточно значительный дрейф от наших, чтобы перестать быть условно (*дружественность людей друг к другу весьма условна*) дружественными
- 5) Киборгизация и сращивание с ИИ => миллиардеры кремниевой долины, генералы – станут сверхумными и сверхсильными, а все остальные будут плестись где-то в хвосте, вовсе не поспевая за ними
- 6) human values are not aligned with each other => киборгизация столь же опасна (или опаснее), чем сильный ИИ
- 7) <https://medium.com/@petervoss/ai-will-outpace-us-but-thats-ok-c1458c8f5ebf> (статья by Peter Voss)
- 8) страницы 76 - 81 в "Superintelligence" (подглава "нейрокомпьютерный интерфейс" главы 2)
- 9) страницы 32 - 36 в "AI as a positive and negative factor in X risks" Юдковского (параграф 12)
- 10) <https://www.facebook.com/groups/467062423469736/permalink/509739885868656/> (пост by Stuart Armstrong)
- 11) Почему вы считаете, что в борьбе мозга и экзокортекса мозг не будет подчинен?
(аналогично тому, как кортекс предаёт лимбику, но там сотни млн лет эволюции)
- 12) У людей нет value function => они могут wirehead. Люди не рациональны => они могут wirehead.
Есть некий аналог value function, но он слишком примитивный => wirehead идёт в плохие простые failure modes.

OpenAI:

Сеть из не aligned ИИ = не aligned ИИ.

1953 год – термоядерная бомба
? – УТС

Нужна цепочка слайдов со структурой презентации. Мол

1) аргументы за скорый ИИ

– эволюционный

– человек не столь уж умный

– ...

2) безопасность

–

–

– ...

И справа значок 9/99 слайд

Но ИИ не умеет вести диалог на уровне 4-летнего ребёнка
и не имеет его common sense

Точно ли не имеет?

4-летний явно играет в Atari хуже взрослых professional players
а нейросетка лучше их в большинстве (в остальных к концу этого года)

На вопросы по картинке нейросетка отвечает на уровне взрослых
Задаёт вопросы по картинке – аналогично на уровне
Генерирует картинки по описанию – лучше многих взрослых

Что ж она не умеет? Вести простенький диалог? Да почти умеет уже.

Решает простенькие задачи – bAbI
Кое-как она мыслит (да и мы мыслим *кое-как*)

Оставшееся не такое уж большое расстояние будет особо быстро преодолено, как только ИИ начнёт помогать на бирже, на президентских выборах...



Тезис ортогональности

Тезис ортогональности – для **любой** цели возможно создать работа с ИИ, способного её достигать сколь угодно эффективно.

ИИ = мощный оптимизационный процесс по достижению заданной цели

Не обязательно вашей)

Тезис об инструментальной конвергенции

Для эффективного достижения **почти любой** цели, агенту целесообразно:

- оставаться в живых (не дать себя выключить)
- предотвращать изменение своей цели (не дать себя перепрограммировать)
- наращивать вычислительные ресурсы
(чтобы просчитывать наиболее эффективное достижение цели)
(чтобы развивать наступательные и оборонительные технологии)

В частности, захватить галактику.

Это мы, люди, беззащитны перед инопланетным вторжением.

А вот ИИ, занявший галактику, вероятнее обеспечит себе выполнение своей цели.

даже если он будет стимулировать себе центры удовольствия.

Методы контроля, коммерчески неадекватные

- физическая изоляция
- ограничение входных и выходных информационных потоков
- система только отвечает на вопросы «да» / «нет»
- реализация ИИ внутри симуляции

Робототехника, коммерческая и военная

Чатботы (реклама, коммерческая и политическая, агит-боты)

ALBA: An explicit proposal for aligned AI, Paul Christiano (2016).

Мы запускаем последовательность агентов $V_0, A_1, V_1, A_2, V_2, \dots$

V_0 - это человек.

A_1 - это некий ИИ, обучаемый ревордами от V_0 , при этом A_1 имеет достаточно мало памяти и скорости, чтобы быть сильно слабее V_0 (человека). Но он обладает common sense итп.

V_1 - это A_1 , только с большей скоростью и памятью

A_2 - это некий ИИ, обучаемый ревордами от V_1 , при этом A_2 имеет достаточно мало памяти и скорости, чтобы быть сильно слабее V_1

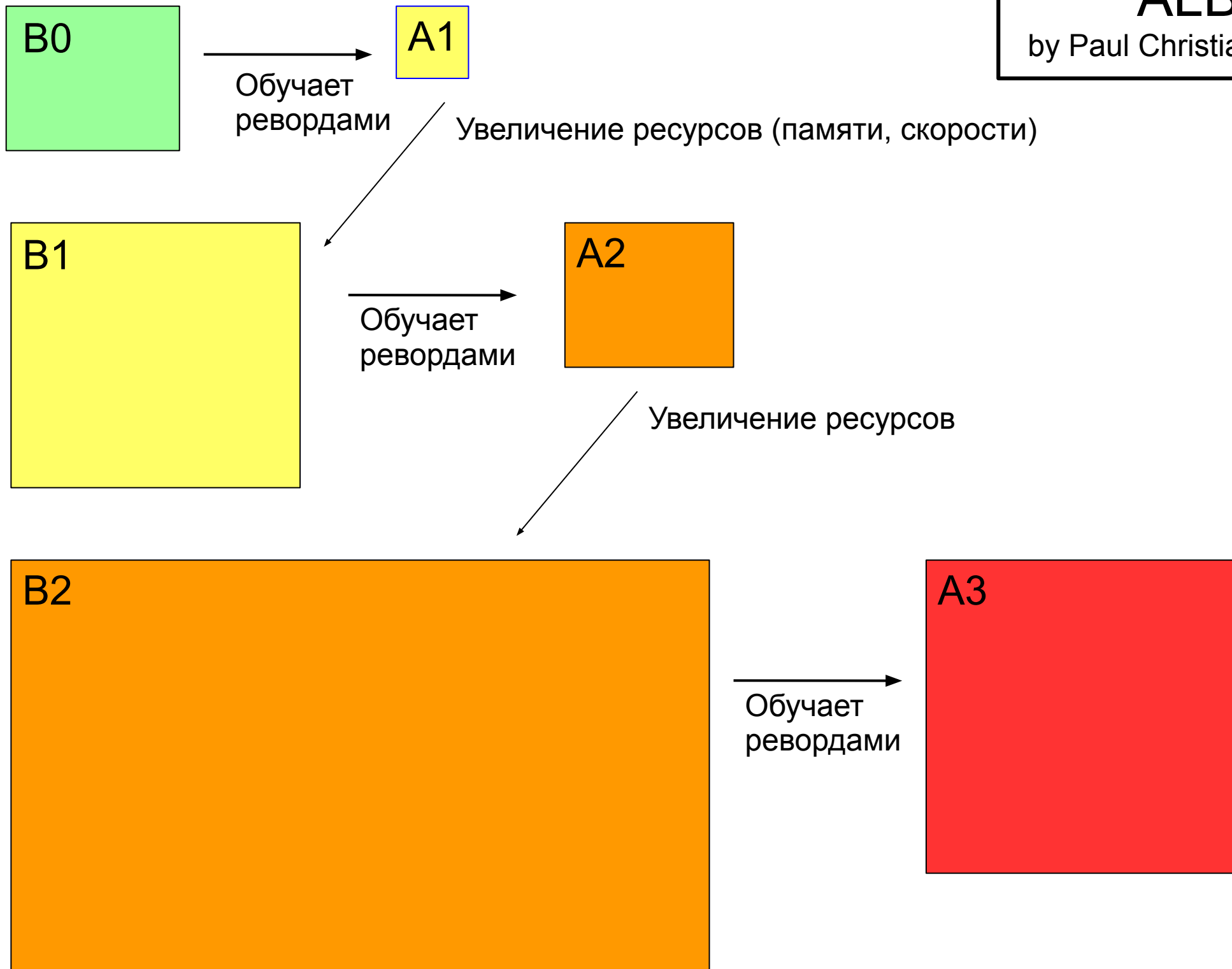
V_2 - это A_2 , только с большей скоростью и памятью
и так далее.

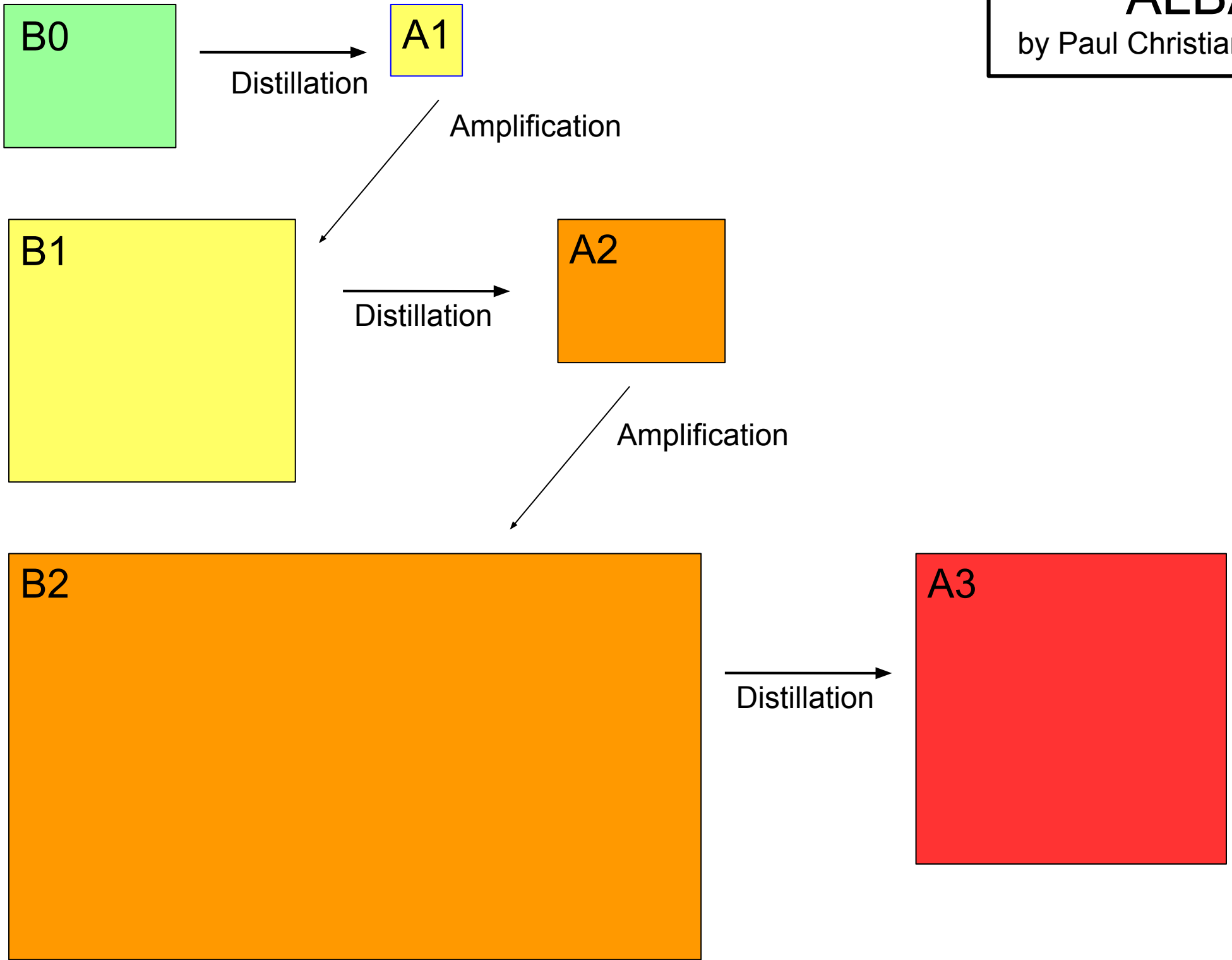
Paul Christiano предлагает (без доказательства) две леммы:

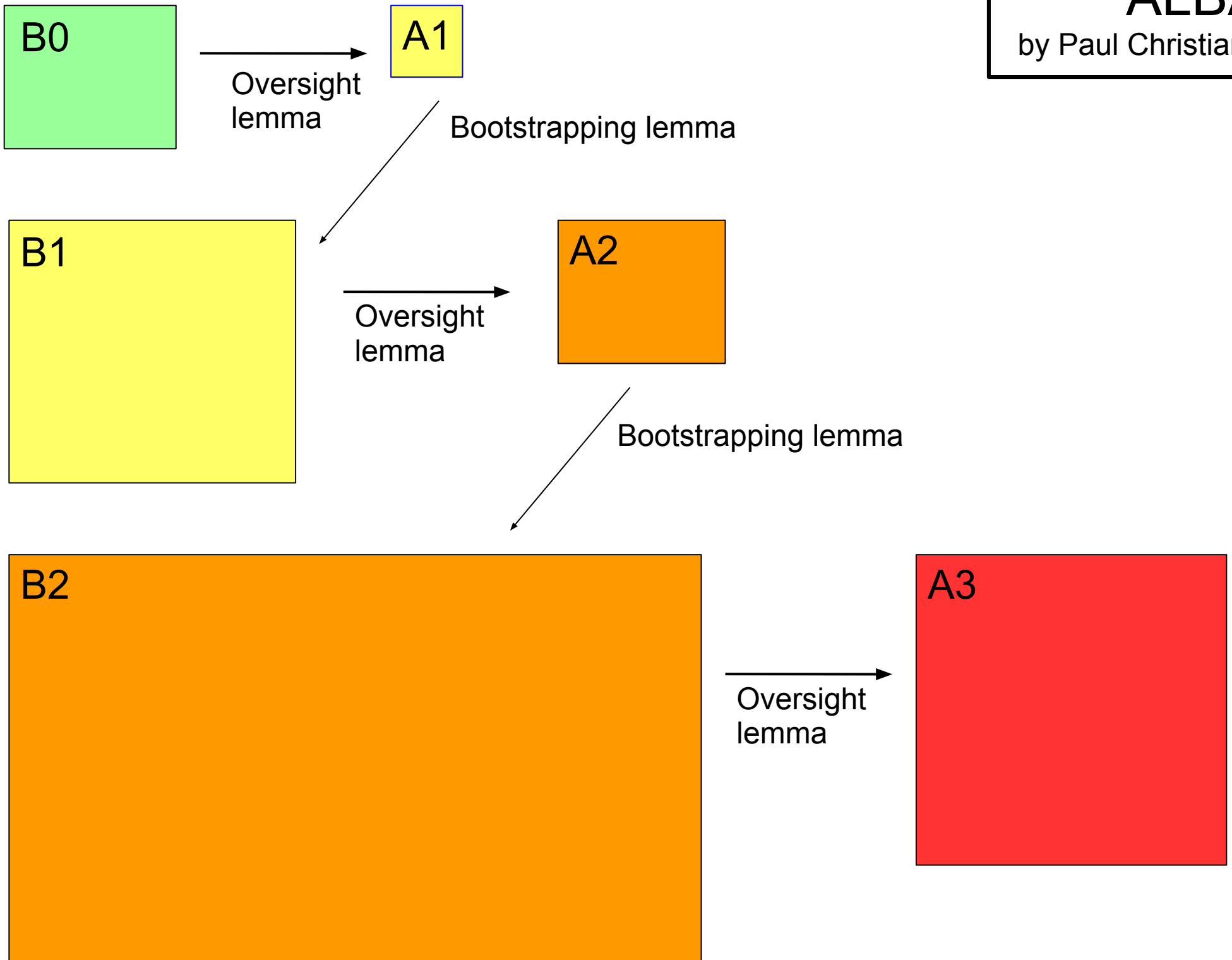
Bootstrapping lemma -- о том, что свойство "aligned" сохраняется при переходе $A\#N \Rightarrow V\#N$ (такие переходы он называет "distillation")

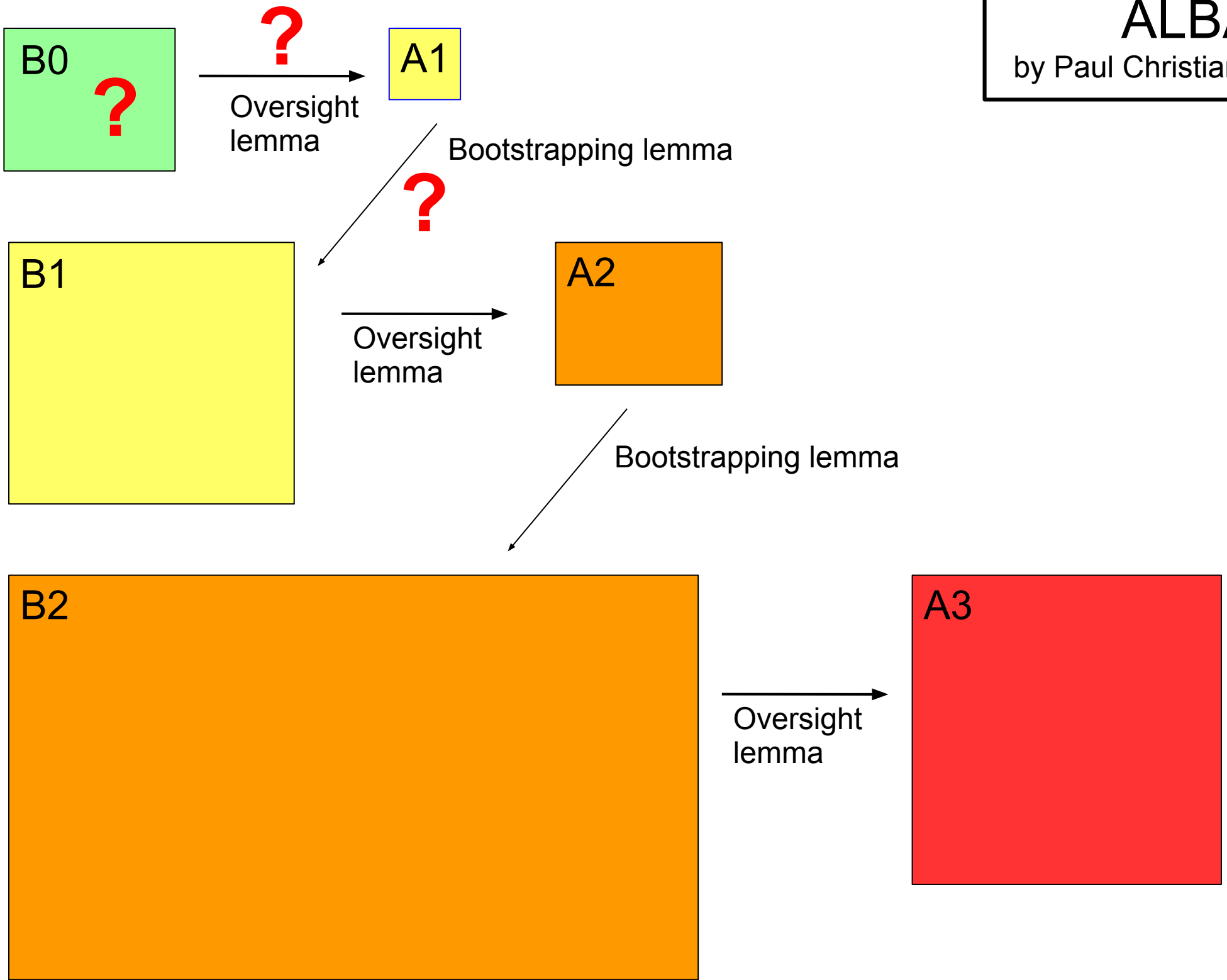
Oversight lemma -- о том, что свойство "aligned" сохраняется при переходе $V\#N \Rightarrow A\#(N+1)$ (такие переходы он называет "amplification")

Автор статьи сам критикует эти леммы, особенно oversight lemma, но считает, что такое разделение может быть полезным для дальнейшего анализа.









Мир быстро идёт в сторону создания сильных ИИ,
в сторону подчинения им экономики, производства, политики.

Коммерчески адекватных и работающих решений по AI safety нет,
и есть серьёзные основания считать, что они не успеют появиться
(а если появятся, то сыграют роль иллюзии защиты в силу overconfidence).

Единственный способ победить скайнет – предотвратить его.
Только в голливудских фильмах возможно обратное.
Впрочем, в неопределённом будущем, возможно...
например, после улучшения социального строя или после киборгизации...
но это потом, и явно не сейчас, а сейчас нужны быстрые решения.

Call to Action:

доносить всю эту инфу до учёных в DeepMind, OpenAI итп.

Пусть не все из них согласятся с нами,
но даже открытая поддержка нескольких десятков из них может
сильно сдвинуть перекося между скоростью развития ИИ и скоростью
развития решений по его безопасности.

<https://www.reddit.com/r/MachineLearning/>

<https://www.facebook.com/groups/aisafety/>

<https://ai-researchers.slack.com/>

AMAs (ask me anything)

Твиттеры, facebook – у многих топовых ученых.

Почему я думаю, что они плохо это знают?

– ты либо кодишь, либо думаешь про Superintelligence

– люди не привыкли думать о сверхглобальных вещах

– люди следуют за общим мнением => overconfidence

– им сложно принимать факт, что они мб двигатель катастрофы

– им сложно публично критиковать то, за что им платят (большие) деньги

Call to Action:

доносить всю эту инфу до учёных в DeepMind, OpenAI итп.

Связаться со мной:

<http://vk.com/shegurin>

<http://www.facebook.com/sergej.shegurin>

<http://rizzoma.com/topic/8ab2572f9807568ee53ebbfd29ea92b7/>

(<http://goo.gl/4k3FYr>)

По вторникам,
С 17-30 до 21-30
«котёл идей» Нейронет, в АСИ
<https://leader-id.ru/event/3657/>
(без регистрации охрана не пропустит).

Только для тех, кто собирается активно участвовать,
а именно самостоятельно разбирать и докладывать статьи.

В среду,
1 марта в 19-00
в Научке – повторение этой лекции.
<https://vk.com/sciencelib>