

Ник Бостром

---

**СВРЪХИНТЕЛЕКТ**

ПОСОКИ, ОПАСНОСТИ,  
СТРАТЕГИИ

София, 2018

ПРЕВОДЪТ Е НАПРАВЕН ПО ИЗДАНИЕТО:  
Nick Bostrom  
*Superintelligence: Paths, Dangers, Strategies*  
Oxford University Press, 2014

Copyright © Nick Bostrom, 2014

SUPERINTELLIGENCE was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. Iztok-Zapad Publishers is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

Всички права на български език запазени. Нито една част от тази книга не може да бъде възпроизвеждана или предавана под каквато и да е форма и по какъвто и да било начин без изричното съгласие на „Изток-Запад“.

© Мария Кондакова, превод, 2018  
© Издателство „Изток-Запад“, 2018

ISBN 978-619-01-0262-5



БИБЛИОТЕКА  
— КРАСИВ УМ —

НИК БОСТРЪОМ

# СВРЪХИНТЕЛЕКТ

ПОСОКИ, ОПАСНОСТИ, СТРАТЕГИИ

Превод от английски  
*Мария Кондакова*





# СЪДЪРЖАНИЕ

---

---

<b>Незавършена приказка за врабчета .....</b>	<b>11</b>
<b>Предговор .....</b>	<b>13</b>
<b>Глава 1. Развитие в миналото и настоящи възможности .....</b>	<b>15</b>
Форми на растеж и исторически преглед .....	15
Големите надежди .....	18
Сезони на надежда и отчаяние .....	20
Съвременно състояние .....	28
Възгледи за бъдещето на машинния интелект .....	37
<b>Глава 2. Пътища към свръхинтелекта .....</b>	<b>41</b>
Изкуствен интелект .....	42
Цялостна емуляция на човешки мозък .....	51
Когнитивни възможности на биологичния мозък .....	60
Мозъчно-компютърни интерфейси .....	71
Мрежи и организации .....	75
Резюме .....	77
<b>Глава 3. Форми на свръхинтелект .....</b>	<b>80</b>
Скоростен свръхинтелект .....	81
Колективен свръхинтелект .....	82
Качествен свръхинтелект .....	85
Пряк и непряк обseg .....	87
Източници на преимущества за цифровия интелект .....	88
<b>Глава 4. Кинетика на взривното развитие на интелект .....</b>	<b>92</b>
Момент и скорост на подема .....	92
Съпротивляемост .....	97
Пътища без участието на машинен интелект .....	97
Пътища на емуляция и ИИ .....	99
Сила на оптимизация и взривно развитие .....	107

<b>Глава 5. Решаващо стратегическо преимущество .....</b>	<b>113</b>
Ще получи ли главният претендент решаващо стратегическо преимущество? ...	114
Колко голям ще бъде успешният проект? .....	119
Наблюдение .....	120
Международно сътрудничество .....	123
От решаващо стратегическо преимущество до сингългън .....	124
<b>Глава 6. Когнитивни свръхспособности .....</b>	<b>128</b>
Функционални възможности и свръхспособности .....	129
Сценарий за превземане на властта от ИИ .....	133
Власт над природата и агентите .....	138
<b>Глава 7. Волята на свръхинтелекта .....</b>	<b>146</b>
Връзката между интелект и мотивация .....	146
Инструментална конвергенция .....	150
Самосъхранение .....	151
Интегритет на комплекса от цели .....	151
Когнитивно усъвършенстване .....	153
Технологично усъвършенстване .....	155
Придобиване на ресурси .....	156
<b>Глава 8. Гибелта ли е логичният изход? .....</b>	<b>159</b>
Екзистенциалната катастрофа – логичен резултат от взривното развитие на интелект? .....	159
Вероломният обрат .....	161
Случаи на злокачествена дисфункция .....	165
Превратна реализация .....	165
Инфраструктурен излишък .....	168
Престъпление срещу разума .....	172
<b>Глава 9. Проблемът за контрола .....</b>	<b>174</b>
Два агентски проблема .....	174
Методи за контрол на възможностите .....	177
Изолиращи методи .....	177
Стимулни методи .....	179
Задържане на развитието .....	185
Предпазители .....	186
Методи за мотивационен подбор .....	189
Пряка спецификация .....	189
Опитомяване .....	191
Непряка нормативност .....	192
Усъвършенстване .....	193
Резюме .....	194

<b>Глава 10. Оракули, джинове, суверени, инструменти .....</b>	<b>197</b>
Оракули .....	197
Джинове и суверени .....	201
ИИ инструменти .....	204
Сравнение .....	211
<b>Глава 11. Многополярни сценарии .....</b>	<b>214</b>
За конете и хората .....	215
Заплати и безработица .....	215
Капитал и обществено благосъстояние .....	217
Малтусианският принцип в историческа перспектива .....	219
Ръст на населението и инвестиции .....	221
Живот в алгоритмична икономика .....	223
Доброволно робство, делнична смърт .....	224
Ще носи ли радост максимално ефективната работа? .....	227
Несъзнателни работници с крайно тясна специализация .....	230
Еволюцията не води непременно нагоре .....	232
Формиране на сингълтън след прехода? .....	236
Втори преход .....	236
Суперорганизми и ефектът на мащаба .....	238
Обединение на договорен принцип .....	240
<b>Глава 12. Придобиване на ценности .....</b>	<b>246</b>
Задачата за внедряване на ценности .....	246
Еволюционен отбор .....	249
Обучение с утвърждаване .....	250
Асоциативно усвояване на ценности .....	251
Мотивационно скеле .....	253
Ценностно обучение .....	255
Модулиране на емуляция .....	267
Институционална конструкция .....	268
Резюме .....	274
<b>Глава 13. Избор на критерии за избор .....</b>	<b>277</b>
Необходимостта от непряка нормативност .....	277
Кохерентна екстраполирана волеизява .....	280
Някои обяснения .....	281
Доводи в полза на КЕВ .....	283
Други забележки .....	286
Модел на морална основа .....	288
Прави това, което имам предвид .....	291
Списък с компоненти .....	293
Съдържание на целта .....	294

Теория на решенията .....	295
Епистемология .....	296
Ратификация .....	298
Достатъчно близко .....	300
<b>Глава 14. Стратегически изглед .....</b>	<b>302</b>
Научна и технологична стратегия .....	303
Диференцирано технологично развитие .....	303
Предпочитан ред на възникване .....	305
Скорост на промяната и усъвършенстването на когнитивните способности .....	308
Технологични двойки .....	313
Манипулативно предвиждане .....	315
Пътища и способстващи фактори .....	317
Последици от прогреса в областта на хардуера .....	317
Трябва ли да се насърчават изследванията в областта на цялостната емуляция на мозък? .....	319
От субективна гледна точка бързината е за предпочитане .....	324
Сътрудничество .....	325
Състезателната динамика и нейните рискове .....	325
За ползите от сътрудничеството .....	329
Съвместна работа .....	334
<b>Глава 15. Критичният момент .....</b>	<b>336</b>
Философия с краен срок .....	336
Какво да се прави? .....	337
Търсене на стратегическа яснота .....	339
Изграждане на достатъчен капацитет .....	339
Специални мерки .....	340
Най-доброто в човешката природа да излезе напред .....	341
<b>Послеслов .....</b>	<b>343</b>
<b>Благодарности .....</b>	<b>347</b>
<b>Бележки .....</b>	<b>349</b>
<b>Библиография .....</b>	<b>407</b>
<b>Частичен речник .....</b>	<b>428</b>
<b>Показалец .....</b>	<b>434</b>



## ФИГУРИ

---

---

1. Крива на световния БВП за продължителен исторически период .....	17
2. Прогнози за дългосрочните последици от появата на машинен интелект на човешко ниво .....	40
3. Производителност на суперкомпютрите .....	48
4. Реконструиране на триизмерната невроанатомия по изображения от електронен микроскоп .....	53
5. Сценарий за осъществяване на цялостна емуляция на човешки мозък .....	67
6. Усредени човешки лица като метафора на изчистените от грешки геноми .....	67
7. Форма на подема .....	93
8. Една не толкова антропоморфна скала .....	102
9. Един от простите модели за взривно развитие на интелект .....	111
10. Фази в сценария за превземане на властта от ИИ .....	134
11. Схематични илюстрации на някои възможни траектории за хипотетичен здравомислещ сингълтън .....	143
12. Резултати от антропоморфизирането на мотивацията на извънземните .....	148
13. Кое да бъде първо – изкуственият интелект или емуляцията на мозък? .....	321
14. Нива на риска в технологичната надпревара за ИИ .....	327

## ТАБЛИЦИ

---

---

1. Изкуственият интелект в игрите .....	29
2. Кога машинният интелект ще достигне човешко ниво? .....	38
3. В какъв срок след създаването на МИЧН ще се появи свръхинтелект? .....	40
4. Необходими възможности за цялостна емуляция на човешки мозък .....	54
5. Максимално увеличение на КИ при селекция измежду група ембриони .....	62
6. Възможни въздействия на генната селекция при различни сценарии .....	65
7. Примери за стратегически важна технологична надпревара .....	117
8. Свърхспособности: някои стратегически важни задачи и свързаните с тях комплекси от умения .....	132
9. Различни видове предпазители .....	187
10. Методи за контрол .....	195
11. Характеристики на различните части от системи .....	212
12. Обобщение на методите за влагане на ценности .....	275
13. Списък с компоненти .....	293

---

**КАРЕТА**

---

1. Оптимален байсов агент .....	26
2. Борсовият срив от 2010 г. ....	35
3. Какво ще ни коства да възпроизведем еволюцията? .....	45
4. Подробно за кинетиката на взривното развитие на интелект .....	108
5. Технологична надпревара: няколко исторически примера .....	115
6. Сценарият с поръчка на ДНК по пощата .....	137
7. Колко голям е космическият фонд? .....	140
8. Антропичен плен .....	183
9. Странни решения в резултат от сляпо търсене .....	209
10. Математически подход към ценностното обучение .....	257
11. Изкуствен интелект с приятелски намерения .....	262
12. Две скорошни (недоизбистрени) идеи .....	264
13. Рискована надпревара без задръжки .....	325

# НЕЗАВЪРШЕНА ПРИКАЗКА ЗА ВРАБЧЕТА

---

**Т**ова се случило през сезона на гнезденето. След няколко дни усилена работа врабчетата седели на вечерното слънце, отдъхвали си и цвърчали.

– Ние сме толкова дребни и слаби. Представете си колко по-лек щеше да е животът, ако си имахме една сова, която да строи гнездата ни!

– Да – казало друго врабче. – Освен това би могла да се грижи за възрастните и за малките ни.

– Би могла да ни дава съвети и да се оглежда за кварталната котка – добавило трето.

Тогаво старшият врабец Пастус взел думата:

– Нека изпратим наши разузнавачи във всички посоки, за да потърсят някое изоставено малко или яйце от сова. Врана или невестулка вероятно също ще свършат работа. Това може да се окаже най-хубавото нещо в живота ни, откакто в задния двор отвориха Павилиона с неизчерпаемото зърно.

Ятото било обхванато от ентузиазъм и всички врабчета зацвърчали колкото им глас държи.

Само Скронфинкъл, едноок врабец със свадлив нрав, не бил убеден, че начинанието е разумно.

– Това със сигурност ще е краят ни – рекъл той. – Дали да не помислим как могат да се опитомяват и дресират сови, преди да доведем сред нас подобно същество?

А Пастус отвърнал:

– Опитомяването на сова изглежда нещо изключително трудно. Няма да е лесно дори да намерим яйце от сова. Така че нека започнем с това. След като успеем да отгледаме сова, можем да помислим да се заемем и с тази нелека задача.

– Този план е погрешен! – изцвърчал Скронфинкъл, но възраженията му останали напразни, понеже ятото вече било отлетяло, за да изпълнява нарежданията на Пастус.

Останали само две-три врабчета. Заедно те се заели да измислят как могат да бъдат опитомени или дресирани совите. Скоро разбрали, че Пастус е имал право: това било извънредно трудна задача, особено след като не разполагали

с истинска сова, с която да се упражняват. Въпреки това полагали всички възможни усилия, като постоянно се страхували, че ятото може да се върне с някое яйце от сова, преди да са намерили решение на проблема за контрола.

Крайт на приказката е неизвестен, но авторът посвещава тази книга на Скронфинкъл и неговите последователи.

# ПРЕДГОВОР

---

---

**В** мозъка ви се намира нещото, с което четете. Това нещо, човешкият мозък, притежава някои способности, от които мозъците на другите животни са лишени. На тези отличителни способности ние дължим господстващото си положение на планетата. Други животни имат по-здрави мускули и по-остри нокти, но ние притежаваме по-умни мозъци. Благодарение на нашето скромно преимущество в областта на общия интелект сме развили език, технология и сложна социална организация. Това преимущество с времето се увеличава, тъй като всяко поколение доразвива постиженията на предходните.

Ако някой ден изградим машинни мозъци, които превъзхождат човешките по обща интелигентност, тогава този нов свръхинтелект може да стане много мощен. И както съдбата на горилите днес зависи повече от нас, хората, отколкото от самите горили, така и съдбата на нашия вид ще зависи от действията на машинния свръхинтелект.

Наистина разполагаме с едно преимущество: неговото създаване зависи от нас. По принцип можем да изградим такъв тип свръхинтелект, който да защитава човешките ценности. Със сигурност имаме безспорни основания да го направим. На практика проблемът за контрола – проблемът как да контролираме това, което свръхинтелектът ще извършва – изглежда твърде труден. Освен това по всичко личи, че ще имаме само един шанс. Стигне ли се до съществуването на враждебно настроен свръхинтелект, той няма да допусне да го заменим или да въздействаме върху предпочитанията му. Съдбата ни ще бъде решена.

В тази книга се опитвам да анализирам предизвикателството, пред което ни поставя перспективата за възникване на свръхинтелект, и най-добрия възможен отговор от наша страна. Твърде възможно е това да се окаже най-важното и най-страховитото предизвикателство, което е стояло пред човечеството. И независимо дали ще се справим, или няма да успеем, това вероятно ще е последното предизвикателство, пред което някога ще се изправим.

Тази книга няма за цел да доказва, че сме на прага на голям пробив в областта на изкуствения интелект или че може с точност да се предскаже кога ще бъде направена тази крачка. Вероятно това ще се случи през настоящия век, но не знаем със сигурност. В първите две глави се разглеждат възможните пътища и се коментира накратко развитието на нещата във времето. По-голямата част от

книгата обаче е за събитията, които ще последват. Анализира се кинетиката на експлозивната поява на изкуствен разум, формите и способностите на свръхинтелекта и стратегическите възможности за избор пред свръхинтелигентен агент, придобил решаващо преимущество. След това изместваме фокуса към проблема за контрола и въпроса какво можем да направим, за да създадем началните условия за постигане на благоприятен резултат, гарантиращ оцеляването ни. В края на книгата разглеждаме по-обхватно пълната картина, очертана от нашите проучвания. Предлагат се известни насоки какво трябва да направим, за да увеличим шансовете си да избегнем екзистенциална катастрофа по-късно.

Писането на тази книга не беше лесно. Надявам се, че разчистеният от нея път ще позволи на други изследователи да достигнат до нови предели по-бързо и по-удобно, със свежи сили и готови да се включат в работата за разширяване на нашето разбиране. (А ако прокараният път е малко неравен и криволичещ, надявам се критиците, когато оценяват резултата, да не подценяват първоначалната трудност на терена!)

Книгата не беше лесна за писане; опитах се да я направя лесна за четене, но според мен без особен успех. Докато пишех, си представях читателя ѝ като едно мое по-ранно „аз“ и се опитах да създам книга, която навремето щях да прочета с удоволствие. Може да се окаже, че тази аудитория е твърде ограничена. Въпреки това смятам, че съдържанието ѝ ще е достъпно за много хора, ако вложат мисъл в четенето и устоят на изкушението да дават автоматично погрешно тълкуване на всяка нова идея, като я приравняват към най-близкото до нея клише от своя културен запас. Читателите, които нямат технически познания, не бива да се обезсърчават от математиката и специализираната терминология, които се срещат на места, защото винаги могат да схванат основната мисъл от допълнителните обяснения. (И обратно – читателите, които искат повечко сухи факти, ще намерят много такива в бележките в края на книгата.<sup>1</sup>)

Много от идеите в тази книга вероятно са погрешни.<sup>2</sup> Възможно е да има и съображения от критично значение, които съм пропуснал да взема предвид и които обезсилват някои (или всички) мои заключения. Постарал съм се да маркирам нюансите и степените на несигурност навсякъде в текста, задръствайки го с неприятна мътилка от „вероятно“, „може би“, „изглежда“, „по всяка вероятност“ и „почти сигурно“. Всяко от тези определения е поставено на мястото му внимателно и обмислено. Тези конкретни прояви на епистемологично смирение обаче не са достатъчни; те трябва да бъдат системно подсилени от допускането за несигурност и погрешност. Това не е лъжлива скромност: защото макар и да смятам, че моята книга вероятно е твърде погрешна и подвеждаща, мисля, че алтернативните възгледи, представени в литературата, са значително по-лоши – включително и онзи по подразбиране, или „нулевата хипотеза“, според която можем засега спокойно или основателно да игнорираме перспективата за свръхинтелект.