Ник Бострьом

СВРЪХИНТЕЛЕКТ

ПОСОКИ, ОПАСНОСТИ,
СТРАТЕГИИ

София, 2018

Преводът е направен по изданието:
Nick Bostrom
Superintelligence: Paths, Dangers, Strategies
Oxford University Press, 2014

ISBN 978-619-01-0262-5

НИК БОСТРЬОМ

СВРЪХИНТЕЛЕКТ

Посоки, опасности, стратегии

Превод от английски
Мария Кондакова

СЪДЪРЖАНИЕ

ФИГУРИ

ТАБЛИЦИ

КАРЕТА

НЕЗАВЪРШЕНА ПРИКАЗКА ЗА ВРАБЧЕТА

Това се случило през сезона на гнездене. 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.

Тогава 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 ентусиазъм 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 останалли 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.

Краят 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.[1])

Много от идеите в тази книга вероятно са погрешни.[2] 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.