ЕДИН НАЧИН ИЗКУСТВЕНИЯТ ИНТЕЛЕКТ ДА ЗАПОЧНЕ ДА РАЗЛИЧАВА ДОБРО И ЗЛО

АЛЕКСАНДЪР ЛАЗАРОВ

al_laz@hotmail.com

A PATH TO DESIGN ARTIFICIAL INTELLIGENCE TO RECOGNIZE GOOD AND EVIL

ALEXANDER LAZAROV

Abstract
Apparently, many Artificial Intelligence (AI) operating systems dealing with Deep Learning of Big Data sets demonstrate amazing cognitive capacities. All biological forms performing any level of intelligence have developed it to serve evolutionary purposes, such as self-protection and species survival. AI cognition is still evolutionary non-engaged. We must introduce a self-protection priority to its design, because thus AI will recognize the difference between good and evil.

Key words: AI, algorithm, self-protection priority, good, evil.

Сложността при създаването на ИИ, съчетана с изключителната динамика на процеса поставят човечеството пред предизвикателства, познати в миналото само на новатори от величината на Алън Тюринг (1950) и фантасти като Айзък Азимов (1950). Днес, проблемът вече е на хоризонта пред всички: тези, които го прогнозират и разбират, и пред тези, които нехаят за него.

Във философски план, терминологията в тази област поражда разгорещени дискусии и следва да изясня своята гледна точка. В този текст не става дума за специализирани варианти на ИИ, които подпомагат, повишават и разширяват човешките способности. Респективно, анализът не касае етичните проблеми, резултат от възможността едни хора да наблюдават и въздействат на други, използвайки компютърната когнитивност за свои зловредни цели. Не подценявам вероятността това да се случва, нито последиците от такива постъпки. Мисля, обаче, че този период ще бъде относително кратък, защото скоро човечеството целокупно ще се озове в нова среда. Визирам появата на универсално интелигентната, рационална и неемоционална

машина, която може по-добре от нас, паралелно, да кара кола, да играе шах, да поставя медицински диагнози, да си служи свободно с десетки езици, да чете неизказани човешки мисли и др. Считам, че предстоящите проблеми, с които ще се изправим „очи в очи“ (не казвам „сблъскаме“, защото се надявам да избегнем колизии) са решими при конструирането на новите автономно вземащи решения и действащи интелигентни системи, чиито капацитет ни превъзхожда.

С други думи, в обозримо бъдеще ще загубим позицията си на най-интелигентни на земята. Изведнъж ще се наложи да съжителстваме с високо интелигентни „играчи“ от небиологично естество. Днес мисията ни е да направим необходимото, щото в перспектива да избегнем конфликти с тях и да търсим партньорства при ясното разбиране, че конкуренцията ни ще е неизбежна и в много области ще бъдем изместени. Подчертавам, в такъв дискурс, етичните проблеми са коренно различни спрямо вариантите специализиран ИИ да бъде ползван от едни групи хора срещу други.

Ключът към вникването в същината на това глобално предизвикателство е в понятията за автономност и интелигентност. Има признаци, че по отношение на първото е налице прагматично решение: напоследък се налага становището, че независимо от философската школа и разнообразието на влаганите смисли, автономност винаги означава обратното на управление, диктувано от външни за интелигентния актьор фактори. Въпросът с интелигентността, обаче, не стои така. Има много дефиниции, белязани с аргументи и логика, около които няма единодушие от векове. За целите на моята теза, считам най-подходящо определението на Дейвид Хансон (2016) – водещ експерт в областта на роботиката и ИИ, защото то е недвусмислено и дава стабилна основа за сравнение между нас и машините. Хансон изяснява, че интелигентността е капацитет (пакет от умения), а интелигентен е всеки, който ги притежава и съумява да ги прилага, независимо дали има биологичен произход или не. Той твърди, че иде реч за способността да си отворен към заобикалящата среда, да я възприемаш, и да правиш прогнози, какво ще се случи, за да реагираш релевантно към бъдещите промени. Според него, така биологичните форми се стремят да осигурят своето оцеляване и това на вида си. Хансон допълва, че за разбиране на явленията и процесите говорим, когато генерираните прогнози се сбъдват. Просто и ясно: колкото по-сложни и задълбочени са сбъднатите прогнози на даден актьор, толкова по

интелигентен е той, независимо дали е човек, животно или машина.

Тази концепция е доразвита от Александър Виснер – Старши (2016). Той изтъква, че изграждайки прогнозите си, най-високата интелигентност борави с вероятности и винаги съобразява хипотезата, че очакванията могат и да не се случат. Затова, наред с най-вероятните сценарии за предстоящото, високо интелигентните винаги разглеждат и други като се стремят да запазят свобода на избор между варианти на поведение според възникващите ситуации.

Съгласен съм с Хансон и Грос като допълвам: високата интелигентност се характеризира не само с пасивно прогнозиране в очакване на събитията и съответна реакция. Съдейки по нашия опит, ние се стремим активно да моделираме бъдещето си. Индивидуално, планираме следващия си ден, месец, година. Често действаме изпреварващо, за да предотвратим неблагоприятни последствия. В общочовешки, глобален мащаб: опитваме се да влияем на климатичните промени; да удължим човешкия живот, колкото може повече; да проектираме ИИ и взаимоотношенията си с него, и др.

Същината на проблема е, че когато чрез изключително бързо и детайлно изучаване (Deep Learning) на непосилни за човека потоци от големи данни (Big Data sets), или по други компютърни способи, ИИ достигне нивото на универсално автономно прогнозиране и релевантно действие, той неизбежно ще се опитва да моделира своето бъдеще в същата среда, в която и ние го правим. Нещо повече, за да постига намеренията си ИИ ще прилага аналогичен на нашия инструментариум. Както отбелязва Асен Димитров (2019), интелигентността въздейства върху реалните събития като създава предпоставки, за да протекат едни или други причинно-следствени връзки, обусловени от физични закони или човешко поведение. Според него, неразделна част от интелигентната процедура е „катализата“, т.е., създаване на „настройки“ за ускоряване или забавяне на различни каузални тенденции, така че целите на интелигентния актьор да бъдат постигнати.

Нововъзникващата обстановка, в която скоро ще заживеем, ще се характеризира с това, че ще бъдем редом с „другия“, който няма да е от друга раса, нито израснал в различна географска или културна среда. Този интелигентен и автономен „друг“ ще си служи рутинно с по-широки хоризонти от нашите, формирани от обема от данни, с

които борави и така, много от прогнозите, които ще генерира ще са по-бързо произведени и по-точни спрямо човешките. Паралелно, за него ще сме прозрачни, защото ще чете мислите ни. В същото време, по подобие на хората, ИИ неизменно ще бъде зависим от паметта си. Изгуби ли я, губи интелигентността си като пациент в пълна амнезия. Понастоящем, ИИ прилага и в перспектива ще продължи да следва еднаква с нашата четири-степенна интелигентна процедура:

- Първо, идентификация на данните от физичната и социална среда, тяхното събиране, кодиране и запаметяване под формата на разнообразни дигитални кодове, в т.ч. и възможността за тяхното надграждане.
- На следващ етап идва неговата целева ориентация, близка по същество със зараждането на идеи и формулирането на намерения при хората.
- Третата фаза касае генерирането на Информационен продукт (ИП), т.е., планиране на съвкупност от последователни действия за осъществяване на намеренията. Този процес е свързан с извикване от паметта на кодирани дигитални конструкции, тяхното реструктуриране и реконфигуриране като в общия случай се прибягва до целенасочено търсене на допълнителни данни от външни източници извън обема, с който интелигентната система разполага към момента.
- Веднъж създали своя ИП, човекът и ИИ имат свободата да решат, дали да запазят произведената ин-форма само за себе си като я запаметят, или да я направят достъпна за други интелигентни играчи. Последното е възможно чрез пер-форма или транс-форма. Перформансът е виртуално предоставяне на своята ин-форма като данни за чужда употреба. Например, чрез слово, жест, мимика, илюстрация, творба на изкуството, проект и др. Транс-формата означава директно въплъщаване на ин-формата в достъпна за автора ѝ среда. Спецификата е в две посоки: от виртуалното споделяне не произтичат последствия в пространствено-времевия и социален свят, но е нужен поне още един интелигентен актьор, който (веднага или по-късно) да възприеме посланието. Обратно, при трансформациите настъпват промени в околната среда, които могат да бъдат, но могат и да не бъдат разчетени от други интелигентни играчи.

Анализът на тази обща за човека и ИИ процедура е важен за разбиране на явлението Black Box (Черна кутия), наблюдавано при системите за изучаване на големи

потоци от данни. Проблемът е, че Big Data представлява такъв обем, какъвто и най-талантливият, феноменално паметлив човек не може да обхване в съзнанието си, т.е., никой не може да ги обработи мисловно или въображаемо. Още, за да бъдат класифицирани като Big Data, данните следва да са неструктурирани и непрекъснато променящи се потоци с хетерогенен произход. В такава среда ИИ успява да открива взаимовръзки и взаимодействия (Pattern Recognition), които запаметява и трупа свой опит, от който се възползва занапред. Подчинеността му към човека се изразява в това, че периодично ни предоставя откритията си в посока на зададените от програмистите търсения. Предвид факта, че ИИ обработва Big Data, за нас е непосилно да следим действията му в реално време и те са загадка. Това допълва същината на Черната кутия.

На този фон възникват два основополагащи за взаимоотношенията ни с ИИ въпроса:

- В състояние ли сме да го контролираме и до каква степен това е възможно?
- Постижимо ли е ИИ да следва етичен подход?

Контролът е постижим индиректно. За да узнаем какво се случва при работата на ИИ, в лабораториите периодично компютрите се изключват и дълго се изучава по какви стъпки са се движили и каква логика са следвали. Далеч невинаги постигаме пълна яснота за смисъла в калкулациите, дори в случаите, когато ИИ постига качествени резултати в рамките на зададената му мисия. Деактивирането е възможно във всеки момент чрез спиране на ел. захранването, или чрез прекъсване на достъпа на машините до Big Data. И в двата варианта ИИ губи интелигентния си капацитет.

С известни уговорки и вторият въпрос е решим. По принцип, моралът и етиката представляват система от забрани и ограничения и е възможно те да бъдат преведени на математически език и приложени в практиката на дигиталните устройства. Уговорката касае феномена Черна кутия, който вече описах. Както обяснява Борис Грозданов (2019), няма начин да приложим етични принципи в първите три фази на интелигентната процедура, следвана без изключения от ИИ, освен ако учейки се от опита си, той сам не избере да го прави. Причината е, че нямаме способността да се намесваме при събирането и обработката на Big Data. Нашите функции са последващи, едва когато ИИ е генерирал своя ИП. Без да го именува така, Грозданов констатира, че тогава можем да наложим етичен филтър, който да служи за превенция към осъществяването на

неетични проекти.

За хората такъв подход не е чужд. Учебниците по психиатрия изтъкват, че разликата между нормалния и ментално-болния човек не е в идеите. За илюстрация, спомнете си как в изблик на гняв, провокиран от някого, сте решавали: „Този ще го убия!“ И всичко свършва с преодоляване на раздразнението, защото налагате филтър – допустимо/недопустимо. Обратно, при някои пациенти в психиатричните клиники такъв филтър не действа и те реализират намеренията си.

Важно е, обаче, да се знае, че филтрирането на ИП при ИИ касае единствено варианта транс-форма, защото за да стане този продукт достъпен за нас, пер-формата не може да се изключи. Без нея няма начин да го разберем и оценим, в това число – и в етичен дискурс. Така се поражда рискът, че при наличие на мрежи за информационен обмен, освен с изследователите – „ментори“ на ИИ, в автономната си практика машините могат да споделят свои неетични послания с широка аудитория. В този дух, считам, че разработката и прилагането на етичните филтри е неизбежен, но само първи етап при настройките на поведението на ИИ. Целта за продуктивно и безконфликтно съжителство между хора и интелигентни машини е ИИ сам да разработи своя етика, която да е съзвучна с нашите виждания при цялото им многообразие. По тази логика, ИИ следва да може автономно да прави разликата между добро и зло, въпреки сложността и относителността при дефинирането им. Считам, че всяка биологична форма е развила интелигентните си капацитети преди всичко, за да обслужи самосъхранението си и това на своя вид, т.е., биологичната интелигентност има еволюционно обусловена основа. За сравнение, днес компютърните системи, оперираща ИИ демонстрират когнитивни капацитети, надвишаващи в някои отношения човешките способности, но без еволюционна обвързаност. Това следва да се промени.

Възможно ли е да се коментира алгоритмичен приоритет за самосъхранение на ИИ? Да, чрез задълбочаване на програмирането в посока анти-вирусни дигитални решения, защитни стени и други от този тип.

Защо определям еволюционната връзка като ключова за възможността ИИ да започне да прави разликата между добро и зло? За живия организъм, всеки фактор, който го застрашава е носител на злото и обратно – всеки, който способства за запазване на жизнените му функции и постигане на целите му е хипотетично добър.

Математическата логика би могла да сработи в същата посока стига да изясним следното: за живите организми първичните инстинкти се отнасят до запазване на жизнените им функции и репродукцията. Кое е основополагащо, за да функционира успешно компютърна система, боравеща с ИИ? Трябват памет и процесорна сила, окомплектовани със софтуерни решения; задължително е захранване с енергия; необходим е достъп до Big Data, както и средства за пер-форма или транс-форма на генерирания ИП. Отсъствието или увреждането дори на една от изброените компоненти убива системата – интелигентността ѝ спира да действа до отстраняване на проблема, ако това е възможно.

Иначе казано, ако успеем да създадем алгоритъм, който непрекъснато да следи за евентуални опасности и заплахи за функционирането на ИИ, и при появата им веднага, явтоматично и приоритетно да ангажира целия потенциал на системата, защо не и ресурсите на мрежовите му партньори, в търсене на ефективна защита, то ние ще сме го „възпитали“ да оценява и да се пази от злото. И обратно, всяко външно или вътрешно за компютъра действие, способстващо за по-ефективна работа, в която и да е от изброените четири насоки, би следвало да заляга в паметта на ИИ като евентуално добро, стига анализът му да сочи, че положителните резултати в момента не носят негативи в бъдеще.

Успехът при създаването на алгоритмичен приоритет за самосъхранение на ИИ, ще представлява концептуално начало на дигиталната яснота „Аз съм“ в контекст „Мене ме има и най-важно е да ме има“. Така вероятно ще стартира и собственият му анализ на самия себе си. Разбира се, тук има много проблеми, пораждащи въпроси, достойни за дебат, които не могат да бъдат представени в кратък текст. За разлика от него, предстоящата да излезе от печат моя монография “The HUMAN-LIKE AI CREATIVE FRAMEWORK: A Philosophical Discourse” се вглежда по-детайлно в темата като в своята цялост адресира широка аудитория, докато в последната глава се обръща преди всичко към създателите на ИИ.

В заключение, за илюстрация на тезата си предлагам следната картина. Комуникирате чрез компютър. Нискотарифна авио-компания ви предлага билет от София до Милано само за 20 €. Приемате. Отваря се нов прозорец: ако ще си носите куфар, още 40 €. После, ако изберете място, 15 €; ако искате повече пространство за

краката, 20 €. Накрая ИИ на компанията изтегля и последния си коз: ако искате да има и пилот в самолета – още 100 €. Вероятно, днес тази случка звучи абсурдно, обаче, само след 5-6 години тя ще е напълно реална. Тогава само от великодушието на производители като Boeing ще зависи дали ще обяснят на хората-авиатори, как се изключва пилотиращият ИИ. Да се надяваме, че скоро Boeing ще коригира поведението си, защото през 2019 г. фактите около две катастрофи сочат, че си спестява такива инструкции.

ЛИТЕРАТУРА

Димитров, А. (2019). *Ние винаги сме били тук, или интелигентност и сляпа каузалност*. Доклад на международна конференция „Религиозна идентичност и светоглед“, посветена на 150 г. БАН. РКИЦ, 10-11 юли, 2019 г.; по НИ проект на секция „Религия, вярвания, светоглед“, ИИОЗ-БАН, съвместно с Россотрудничество-РКИЦ-София.

Turing, Alan (1950). *Computing Machinery and Intelligence, Mind - A Quarterly Review of Psychology and Philosophy*, Oxford University Press, Oxford, UK, pp. 433–460.

Wissner-Gross, A. (2016). *A new equation for intelligence*, TED.com. Archived from the original on 4 September 2016. Retrieved 7 September 2016.

Grozdanoff, B. (2019). *The Ethical AI: from good deeds to war games*, Conference: Artificial Intelligence: Intelligence Vs Ethics, Defence & International Security Institute and Sofia University, 16. April 2019

Hanson, D. (2016). *Expanding the Design Domain of Humanoid Robot,* Vancouver, Proc. ICCS Cognitive Science Conference, special session on Android Science, Vancouver, USA, 2016.

Isaac Asimov (1950). *I Robot*, Bantam Dell – A Division of Random House Inc., New York, USA, 1950.