Българска академия на науките
Институт по математика и информатика
Секция „Алгебра и логика“

Автореферат

на
дисертационен труд

Изкуствен интелект – дефиниция, реализация и последствия

Какво е това, как да го направим и какво ще правим,
след като го направим?

от

Димитър Димитров Добрев

по професионално направление „Математика“ (4.5)
научна специалност „Математическа логика“

Научен ръководител: доц. д-р Любомир Иванов

София, 2024

Авторефератът съдържа 37 страници, от които 3 страници уводни раздели, 30 страници основен текст и 4 страници заключителен материал, включително библиография с 32 заглавия.

Съдържание

Увод ... 4
Кратко описание на дисертацията ... 7
1 Какво е изкуствен интелект? ... 11
1.1 Неформална дефиниция ... 11
1.1.1 ИИ – Какво е това? ... 11
1.2 Формална дефиниция ... 12
1.2.1 Дефиниция на ИИ и програма, която удовлетворява тази дефиниция ... 12
1.2.2 Въведение ... 12
1.2.3 Литературен обзор ... 13
1.2.3.1 Генерален интелект ... 13
1.2.3.2 Интуитивна дефиниция ... 14
1.2.4 Постановка на задачата ... 15
1.2.5 Оценка ... 17
1.2.6 Очакваната оценка ... 17
1.2.7 Най-добрата стратегия ... 18
1.2.8 Дефиниция на ИИ ... 21
1.2.9 Програма, удовлетворяваща дефиницията ... 21
1.2.10 Дали това е ИИ? ... 22
1.2.11 Заключение ... 22
2 Как да го направим? ... 23
2.1 Език за описание на светове ... 23
2.2 Въведение ... 23
2.3 Играта шах ... 25
2.3.1 Шах с един играч ... 26
2.3.2 Цел ... 26
2.4 Заключение ... 27
3 Какво ще правим, след като го направим? ... 28
3.1 ИИ не трябва да е *Open Source Project* ... 28
3.1.1 Въведение ... 28
3.1.2 Какво може да се случи? ... 29
3.1.3 Заключение ... 30
3.2 Как ще изглежда животът ни след появата на ИИ? ... 31
Заключение ... 33
Публикации ... 34
Декларация ... 35
Литература ... 36

Увод

Целта на тази дисертация е да разсее две заблуди, свързани с изкуствения интелект, и да зададе трите най-важни въпроса, свързани с тази тема. Двете заблуди са, че ИИ е функция без памет и че ИИ е псевдонаука. Трите въпроса, които ще разгледаме, са „Какво е ИИ?“, „Как да го създадем?“ и „Какви ще са последствията от създаването му?“.

Първата заблуда е, че ИИ е функция без памет. Тази заблуда се нарича *Full Observability*. Идеята е, че ИИ вижда всичко, което му е нужно, и няма смисъл да помни какво се е случило до момента. В литературата се споменава и за обратното, т.е. за *Partial Observability*, но по тази тема почти не се работи. Повечето автори приемат, че допускането на скрито състояние (памет) силно ще усложни задачата и ще я направи нерешима. Всъщност интересният случай е *Partial Observability* и отказът да се разглежда този случай е сериозна грешка.

Не сме точни, когато казваме, че повечето автори разглеждат ИИ като функция без памет. По-точно би било твърдението, че те приемат обучения ИИ като функция без памет. Може да се приеме, че ИИ е програмата, която търси обучения ИИ (т.е. която се обучава). Резултатът от обучението е обученият ИИ и въпросът, който ни интересува, е какво представлява обученият ИИ. Повечето автори разглеждат множество от двойки $<x, y>$ (x е въпросът, а y е отговорът). На базата на това множество те търсят функция, която на всеки въпрос x връща съответния отговор y. Тази задача е известна като „интерполация на функция“. Задачата се нарича „невронни мрежи“, когато интерполиращата функция се търси в множеството на невронните мрежи.

Не искаме да подценим разработката на невронните мрежи. Този подход дава удивителни резултати, които го правят основния метод в областта на ИИ. Все пак невронните мрежи са ограничени от липсата на памет, което е съществен недостатък.

Тайната на ИИ се крие в скритото състояние на света. Ако се откажем от невидимата информация, ще пропуснем най-интересното. В основата на ИИ е „разбирането“ на света, а да разберем света, означава да го опишем чрез някакъв език за описание на светове. Затова в тази дисертация се говори толкова много за езици за описание на света. Описанието на скритото състояние не е еднозначно, защото има много възможни обяснения на света. Различните описания може да разглеждаме като различни хипотези и можем да правим експерименти, за да проверим коя е вярната хипотеза. Търсенето на различни описания (хипотези) можем да наречем „въображение“. Тоест ИИ без скрито състояние е ИИ без въображение.

Втората заблуда е, че ИИ е псевдонаука и че сериозните учени не трябва да се занимават с това. Става дума за истинския (генералния) изкуствен интелект. Повечето хора са убедени, че генералният изкуствен интелект е нещо невъзможно (нещо от научната фантастика). Когато стане дума за ИИ, повечето хора приемат, че се говори за слабия ИИ, който не е ИИ, а е имитация на ИИ. Самата идея за слаб ИИ е създадена от хора, които не вярват в истинския ИИ и са дълбоко убедени, че това е нещо напълно невъзможно.

Това е причината, поради която обществото гледа на работещите в областта на силния ИИ като на луди. Нас ни слагат в един кюп заедно с търсачите на НЛО и изобретателите на перпетуум-мобиле. Аз самият приемам търсачите на перпетуум-мобиле за луди и не трябва да се сърдя, че и на мен се лепва този етикет.

Все пак в науката никоя идея не трябва да се отхвърля априори. Много неща са изглеждали невъзможни, но се е оказало, че просто още не им е дошло времето. Основната причина хората да не успяват да реализират една идея е това, че не вярват, че това е

възможно. В момента, в който повярват, те успяват много бързо да решат проблема. Обикновено виждат, че някой друг вече е реализирал идеята и те веднага повтарят неговия успех. Навремето хората са смятали, че машините никога няма да могат да играят шах. Сега програмата за игра на шах е нещо елементарно, което може да се обясни и напише за няколко часа (аз съм го правил в курса си по програмиране).

Заблудата, че сериозните учени не трябва да се занимават с ИИ, е сериозен проблем, защото това води до изоставане в тази област. Преди време подобна заблуда е имало спрямо кибернетиката, което е довело до сериозно изоставане в областта на компютрите (особено на страните от Източния блок).

Дори и сред специалистите по математическа логика има известен скептицизъм относно ИИ, а логиката е областта на математиката, която е най-близка и най-подходяща за създаването на ИИ. Може да се каже, че математиката е в основата на всички точни науки, а логиката е математиката на математиката. Всяка математическа дисциплина е изградена върху фундамента на логиката.

Логиката се занимава със същността на доказателството. Какво е доказателство? Това е начинът, по който хората мислят. Но защо само хората, машините също могат да мислят и да извършват доказателства. Затова логиката е най-близката област до ИИ и логиците са най-добре подготвените специалисти за извършване на сериозни изследвания в тази област.

За всяка задача си има специалисти, които са най-подготвени и най-подходящи за решаването ѝ. Ако искате картина, ще потърсите художник, ако искате песен, ще потърсите музикант. В програмирането най-добри се оказват математиците, защото тяхното формално и абстрактно мислене е много подходящо за писането на програми. За решаването на задачата ИИ най-подготвени са логиците, защото те са запознати с абстракцията на различните светове, на логическите формули и кога един свят е модел на една формула.

Първият въпрос е свързан с дефиницията на ИИ. Почти никой не задава този въпрос! Има огромен брой книги и статии, посветени на ИИ, но в тези книги и статии много рядко се задава въпросът „Какво е ИИ?“. Все пак, нека да отбележим две фундаментални публикации, където този въпрос е централен. Това са Wang (1995) и Hutter (2000). Ще отбележим и една обзорна статия написана в нашия институт (Angelova et al., 2021).

Интересна е дефиницията, дадена от Council of Europe (2022). Тя гласи:

AI is actually a young discipline of about sixty years, which brings together sciences, theories and techniques (including mathematical logic, statistics, probabilities, computational neurobiology and computer science) and whose goal is to achieve the imitation by a machine of the cognitive abilities of a human being.[1]

Интересното в тази дефиниция е, че Съветът на Европа признава, че математическата логика е в основата на ИИ, макар че много от логиците не са съгласни с това.

Най-известната дефиниция на ИИ е дадена от Тюринг и това е „Тестът на Тюринг“ (Turing, 1950). Идеята е, че зад завеса са скрити човек и компютър, който имитира човек.

[1] ИИ всъщност е млада дисциплина от около шестдесет години, която обединява науки, теории и техники (включително математическа логика, статистика, теория на вероятностите, невробиология и информатика), чиято цел е да постигне имитация на когнитивните способности на човека чрез машина.

Ако компютърът имитира човека достатъчно добре, така че да не можем да го отличим от човека, тогава това е ИИ. Дефиницията на Тюринг е много добра, но тя има един съществен недостатък. Това е дефиниция на обучен интелект, а ние искаме да дефинираме необучен интелект. Например новороденото бебе е интелект, но не е обучен интелект.

За да бъде отстранен този недостатък на дефиницията на Тюринг, беше създадена една неформална дефиниция на ИИ (Dobrev, 2000). Тя гласи следното: „**ИИ ще наричаме такава програма, която в произволен свят би се справила не по-зле от човек**“.

Тази дефиниция беше публикувана в популярен вид в научно-популярно списание и тази публикация е включена като част от дисертацията (параграф 1.1) По късно същата дефиниция беше публикувана в научно списание (Dobrev, 2005a). Тази дефиниция получи голямо разпространение, защото в момента в *Google* на въпроса *Definition of Artificial Intelligence* първият резултат, който се появява, е тази статия.

Проблем на тази дефиниция е, че тя е неформална. За да се формализира дефиницията, беше въведен коефициент на интелигентност (*IQ*). На всяка програма се съпоставя едно число, което е нейното *IQ*, и за ИИ се приемат тези програми, чието *IQ* е достатъчно високо. Този подход е в основата на статията Dobrev (2019b). Тази статия не е част от дисертацията, защото има някои недостатъци. Например там се предполага, че дължината на живота и сложността на света са ограничени (т.е. че са фиксирани). Друг проблем е, че се оказва, че тази идея вече е реализирана в публикациите Hernández-Orallo et al., 1998 и Hutter, 2000.

По-късно беше приет друг подход (Dobrev, 2022a). Този друг подход е добавен като част от тази дисертация. Предимство на новия подход е, че дължината на живота и сложността на света не са ограничени. Друго предимство на новия подход е, че се разглеждат различни езици за описание на света, докато при първия подход се разглеждаше само описание чрез изчислими функции.

Вторият въпрос. За да се създаде ИИ, трябва да се направи програма, която може да „разбере“ света и на базата на това, което е разбрала, да планира бъдещото си поведение. Да „разберем“ света означава да го опишем чрез някакъв език за описание на светове. Затова създаването на подходящ език за описание на света е съществена част от създаването на ИИ.

Втората част на дисертацията е посветена на описанието на един такъв език, който претендира да е подходящият език за описание на светове (Dobrev, 2022b и 2023). Важна част от този език са алгоритмите. Пример за алгоритъм е „Чакам, докато дойде“. Статията Dobrev (2023) включва дефиниция на алгоритъм, която обобщава стандартната дефиниция на изчислима функция. Според новата дефиниция алгоритъмът е последователност от действия, която може да бъде извършена в произволен свят. В частния случай, когато светът се състои от безкрайна лента и глава, която чете от лентата и пише върху нея, тогава алгоритъмът е машина на Тюринг.

Третият и най-важен въпрос е свързан с последствията от създаването на ИИ. Този въпрос много малко хора си го задават. Хората са като любопитни деца, които обичат да си играят с кибрит, без да мислят, че може да предизвикат пожар. Макар и малко, все пак има научни статии, които обсъждат бъдещето, което ще дойде след създаването на ИИ. Пример за такава статия е Alfonseca et al. (2021). Друг пример е Ivanova et al. (2020), където се обсъжда влиянието на ИИ върху пазара на труда и как трябва да се промени образованието в тази връзка.

Основният въпрос е: „Трябва ли да създаваме ИИ и можем ли да не го създаваме?“. Отговорът е, че не можем да не го създаваме, защото ако ние не го създадем, ще го създаде някой друг. По-добре е да помислим и да се подготвим за последствията и все пак да внимаваме да не изтървем духа от бутилката.

Отговорът на третия въпрос ще ни е нужен чак когато вече сме направили ИИ. Това не значи, че трябва да чакаме появата на ИИ и чак тогава да си зададем въпроса „И сега какво ще правим?“. Ако чакаме до момента, в който този въпрос стане наложителен, ще допуснем огромна грешка. Когато ИИ се появи, вече ще е много късно да се питаме: „А сега накъде?“. Този въпрос трябва да сме си задали много преди реалната поява на тази машина (т.е. на тази програма, защото ИИ не е машина, а е програма, както ще видите по-долу).

Приноси на дисертацията:

1. Предлага се неформална дефиниция на ИИ и тази неформална дефиниция днес е широко разпространена (в момента това е първият резултат, който *Google* връща при въпроса *Definition of Artificial Intelligence*).

2. Дава се строга математическа дефиниция на ИИ. Предложената строга дефиниция не е изцяло наша заслуга, защото тя е едно подобрение на дефиницията на ИИ, която първоначално е създадена от Hernández-Orallo през 1998 (Hernández-Orallo et al., 1998) и която съществено е подобрена от Marcus Hutter през 2000 (Hutter, 2000).

3. Въвежда се език за описание на светове, при който описанието може да се търси автоматично без помощта на човек. Има и други езици за описание на светове, които са създадени преди нашия език, но предимство на предложения от нас език е, че при него описанието на света може да се търси автоматично, докато при другите подходи се предполага, че има човек, който е разбрал света и който ще го опише на съответния език.

Кратко описание на дисертацията

Класът на изчислимите функции е един от основните обекти на изследване в областта на математическата логика. Макар изчислимите функции да са много важни за логиците, на тях винаги им е било тясно в това множество и те винаги са се опитвали да излязат от него.

Например при Тюринговите и при номерационните степени на неразрешимост се добавя оракул и по този начин се излиза от множеството на изчислимите функции. В тази дисертация също ще използваме оракули, но тези оракули ще са по-различни. От една страна, тук оракулите ще са по-малко, защото нас ни интересуват само такива, чрез които може да се предвижда няколко стъпки напред; от друга страна, ще използваме нестандартни оракули, които не се разглеждат при номерационните степени.

Задачата, която решаваме в тази дисертация, е практическа. Имаме произволна функция f и искаме да я опишем. Функцията f ще наречем „свят“, а езикът, чрез който ще я опишем, ще наречем „език за описание на светове“. Тези езици ще са изброими, което означава, че чрез тях не можем да опишем всички функции. Множеството от функциите, описани чрез един език, ще наречем „интерпретация на езика“. Тоест задачата е да намерим описание, такова че съответната функция от интерпретацията на езика да е близо до търсената от нас функция f.

За функцията f знаем много малко. Ние познаваме само една нейна крайна част (жизнения опит). Затова не е невъзможно да намерим функция, близка до f в

интерпретацията на езика. Функцията, която ще намерим, дори ще съвпада с f върху жизнения опит.

Описанието на света за нас не е самоцел. Ние търсим такова описание, което ще ни даде възможност да погледнем в бъдещето и да видим какво ще се случи след няколко стъпки. На базата на тази прогноза ще можем да изберем най-добрия ход. Затова оракулите, които ще използваме, ще са нестандартни.

Например оракулът „враг“. Това е агент, който играе срещу нас и винаги избира възможно най-лошия за нас ход. Този оракул е много нестандартен, защото зависи от функцията f, в която сме го вложили. В различни светове врагът се държи различно. Действията, които пречат в един свят, може да не пречат и дори да помагат в друг свят.

Хубавото на оракула „враг“ е, че той ни позволява да предвиждаме бъдещето. Можем да използваме алгоритъма *Min-Max* и да погледнем няколко стъпки напред. Тук *Min* отговаря на нашето очакване оракулът „враг“ да избере възможно най-лошия за нас ход, а *Max* отговаря на предположението, че ние ще играем най-доброто за нас.

Друг пример е оракулът „случайност“. Този оракул можем да си го мислим като зар, който връща няколко възможности. Всяка от тези възможности си има вероятност. С този оракул пак можем да предвиждаме бъдещето, но вместо алгоритъма *Min-Max*, ще използваме *Max-Average*. Тук не можем да предвидим точно как ще постъпи оракулът, но ще предвидим няколко възможности и ще изчислим средното.

Трети пример е оракулът „шум“. Това е агент, който е константа, но константа с шум, без да е ясна вероятността на този шум. В този случай ще предвиждаме бъдещето, като разглеждаме едно-единствено възможно действие на агента „шум“, но ще допускаме, че прогнозата може да е изкривена от този шум.

„Какво е език за описание на светове?“ е въпрос, чийто отговор няма да намерите в тази дисертация. Тук ще намерите само примери на такива езици. В първата глава на дисертацията се дават две дефиниции на ИИ. Първата дефиниция е неформална и тя казва, че ИИ е програма, която ще се справи не по-зле от човек. Втората дефиниция е по-интересна. Тази дефиниция е формална и в нея се използва език за описание на светове.

Формалната дефиниция на ИИ първо определя най-добрата стратегия, като това определение е на базата на един език за описание на светове. След това дефиницията определя ИИ като изчислима стратегия, която е достатъчно близо до най-добрата.

В дисертацията се изказва предположението, че формалната дефиниция на ИИ не зависи от езика за описание на светове, който сме използвали. Това предположение не е доказано и вероятно въобще не може да бъде доказано.

На базата на избрания език за описание на светове се прави програма, която удовлетворява дефиницията на ИИ. Тази програма е толкова неефективна, че тя на практика се нуждае от безкрайно бърз компютър, макар че теоретично това е програма, завършваща след краен брой стъпки.

За създаването на програмата, удовлетворяваща дефиницията на ИИ, е нужно езикът за описание на светове да е такъв, че чрез него да можем да предвидим бъдещето. Тук е моментът, където се вижда, че не всяко разширение на множеството на изчислимите функции е подходящо и не всяко ще ни свърши работа.

В дисертацията се изказва предположението, че макар дефиницията на ИИ да не зависи от езика за описание на светове, ефективността на програмата, която удовлетворява дефиницията и която сме направили на базата на определен език, силно зависи от езика за описание на светове, който сме избрали.

Във втората част на дисертацията се разглежда един конкретен свят и един конкретен език за описание на светове, чрез който е описан този свят.

Конкретният свят е светът на играта шах, но в този свят има една интересна особеност. Интересното в случая е, че една стъпка в този свят не е местенето на една фигура. За да се премести една фигура, трябва да се направят много стъпки. Агентът „вижда“ само едно от квадратчетата на таблото. За да премести фигура, той трябва да премести погледа си върху тази фигура, да я вдигне, да премести погледа си върху новата позиция и да спусне там вдигнатата фигура. По този начин местенето на една фигура е изпълнение на алгоритъм, състоящ се от много стъпки. Това е основното предимство на новия език за описание на светове. В този нов език бъдещето не се предвижда, като се гледа няколко стъпки напред, а се гледа няколко големи стъпки напред. Голямата стъпка е изпълнението на алгоритъм.

Един от основните приноси на дисертацията е дефиницията на понятието „алгоритъм“. В литературата обикновено понятието „алгоритъм“ не се отделя от понятието „изчислима функция“. Обикновено на всяка програма се съпоставя изчислимата функция, която се изчислява от програмата. В тази дисертация алгоритъм е последователност от действия, които се извършват в някакъв свят. Резултатът от изпълнението на тези действия зависи от света, в който ще ги изпълним. Например машината на Тюринг обикновено се разглежда като глава и безкрайна лента, които са неделима система. В тази дисертация алгоритъм е главата на машината на Тюринг, а безкрайната лента е светът, в който този алгоритъм се изпълнява. Ако променим света, ще се промени и резултатът от изпълнението на алгоритъма. Например, ако заменим безкрайната лента с крайна, резултатът от работата на алгоритъма ще бъде крайна функция. Алгоритъмът за сортиране в различни светове може да сортира различни обекти. Например би могъл да подрежда аптекарски шишета.

Важното е при предвиждането на бъдещето малките стъпки да бъдат заменени с големи (т.е. с изпълнението на алгоритми). По този начин може да се погледне по-далеч в бъдещето, защото, ако мислим в малки стъпки, ще се получи комбинаторна експлозия и няма да можем да видим по-далеч от носа си.

В третата част на дисертацията се разглеждат някои философски въпроси, свързани с последствията от създаването на ИИ. Тази част не е математическа и вътре няма нито една теорема и дори нито една дефиниция. Все пак това е важна част от дисертацията, защото математиката не се състои само от построяването на формални конструкции и извършването на формални доказателства. Математикът е длъжен да зададе и въпроса: „Какви ще са последствията от тези формални разсъждения?“.

Един от въпросите, които задаваме в края на дисертацията, е „Трябва ли технологията ИИ да бъде общодостъпна?“. Отговорът, който даваме, е, че това е опасна технология и че сериозните статии в областта на ИИ трябва да бъдат засекретени. Този отговор поставя в неудобно положение бъдещия рецензент на дисертацията. Ако той приеме тази наша теза и ако смята, че тази дисертация е сериозен принос в областта на ИИ, то той трябва да се опита да я скрие от обществеността, като даде отрицателна рецензия. От друга страна, ако рецензентът приеме, че дисертацията не е сериозен принос в областта на ИИ, то той пак трябва да даде отрицателна рецензия.

В дисертацията няма нито едно съществено математическо твърдение, което да е доказано. В първата част има пет съществени твърдения, но те не са доказани и дори нашето предположение е, че някои от тях въобще не могат да бъдат доказани. В дисертацията има две дефиниции на ИИ, като първата е неформална. Втората дефиниция е формална с изключение на това, че искаме стратегията да е достатъчно близо до най-добрата. Понятието „достатъчно близо“ не е формално, но ние разглеждаме множество от алгоритми, които са безкрайно близо, което вече е формално. (Безкрайно близо означава,

че за всяко ε съществува стойност на параметъра h такава, че при тази стойност на параметъра h съответният алгоритъм е на разстояние, по-малко от ε от най-добрата стратегия.)

Причината, поради която дисертацията съдържа много неформални твърдения, е, че областта на ИИ е една сравнително нова област, в която основният въпрос не е как да се решат известни формални задачи, а как да се формализират термините и задачите, които се появяват в тази област. Изследванията в областта ИИ започват от Алън Тюринг, което означава, че тази област не е чак толкова нова, но все още ИИ сериозно се нуждае от формализация и уточняване на основните понятия.

1 Какво е изкуствен интелект?

1.1 Неформална дефиниция

В тази дисертация ще си зададем въпросите: „Трябва ли ни да знаем какво е ИИ?“ и „Какво е интелект?“. След това ще дадем една дефиниция за изкуствен интелект. Накрая от тази дефиниция ще получим алгоритъм, който след краен брой стъпки ще открие ИИ.

1.1.1 ИИ – Какво е това?

Трябва ли ни да знаем какво е ИИ? На този въпрос ще отговорим просто: *Да, ако искаме да го открием, то определено ще е по-лесно да го намерим, ако знаем какво търсим.* В противен случай ще се окажем в положението на алхимиците, които са търсели философския камък, но почти не са имали представа какво е това.

Най-известната дефиниция на ИИ е така нареченият Тюрингов тест. Тюринг е английски математик, известен както с машините на Тюринг, така и с разбиването на немските кодове по време на Втората световна война.

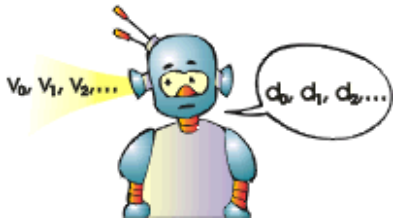
Тюринговият тест е доста прост. Поставяме нещо зад една завеса и то разговаря с нас. Ако не можем да го различим от човек, то това е ИИ. Тази дефиниция е по-стара от петдесет години и затова ще се опитаме да направим нова, по-съвременна.

Дефиницията на Тюринг предполага, че *интелект* е човек с натрупаните през годините знания. А *интелект* ли е бебето, което току-що се е появило на бял свят? Нашият отговор ще е „да“. Тоест нашата дефиниция за интелект ще е това, което не знае нищо, но може да се научи. Тук се различаваме от повечето хора, които като чуят *интелект*, си представят професор от университета.

Преди да дадем формална дефиниция на ИИ, ще направим уговорката, че приемаме тезиса на Чърч, който гласи, че всяко изчислително устройство може да се моделира с програма. Т.е. ние ще търсим ИИ в множеството на програмите. Ще предположим, че ИИ е стъпково устройство, което живее в някакъв свят, на всяка стъпка получава информация (от света) и въздейства (на света) чрез информацията, която извежда. Ще допуснем още, че информацията, която получава и предава на всяка стъпка, е крайно много. Например получава n бита и предава m бита.

След тези уговорки можем да изкажем неформално нашата дефиниция. **ИИ ще наричаме такава програма, която в произволен свят би се справила не по-зле от човек.**

Следващата задача ще е да формализираме тази дефиниция, за да можем да я използваме и с нея да търсим ИИ. Първо, какво за нас ще е свят? Това ще са две функции **World(s, d)** и **View(s)**. Първата ще взима като аргументи състоянието на света и въздействието, което оказва на света нашето устройство. Като резултат тази функция ще връща новото състояние на света (което той ще има на следващата стъпка). Втората функция ще ни казва какво вижда нашето устройство. Аргумент на тази функция ще е състоянието на света, а получената стойност ще е информацията, която ще получава устройството (на дадена стъпка). Трябва да добавим и един елемент $\mathbf{s_0}$, това ще е състоянието на света, когато нашето устройство се е родило. По време на живота му светът ще премине през състоянията $\mathbf{s_0}$, $\mathbf{s_1}$, $\mathbf{s_2}$, ... Устройството ще въздейства на света с информацията, която извежда на всяка стъпка $\mathbf{d_0}$, $\mathbf{d_1}$, $\mathbf{d_2}$, ... Също така ИИ ще получава информация от света $\mathbf{v_0}$, $\mathbf{v_1}$, $\mathbf{v_2}$, ... Ясно е, че $\mathbf{s_{i+1}} = \mathbf{World(s_i, d_i)}$ и че $\mathbf{v_i} = \mathbf{View(s_i)}$.



Дотук имаме всичко. Имаме свят и устройство, което живее в него. Липсва само едно, смисъла на живота. Какво е животът без болка и радост, ще кажат философите. Затова ще въведем смисъл на живота. Това ще е оценка, която ни казва дали една редица $\mathbf{v_0}$, $\mathbf{v_1}$, $\mathbf{v_2}$, ... е по-добра от друга.

За повечето хора животът е протекъл по-добре, ако са видели повече швейцарски курорти и по-малко каменовъглени мини. Горе-долу и ние така ще дефинираме смисъла на живота. Ще отделим от $\mathbf{v}_i$ два бита, които ще наречем „победа“ и „загуба“, и смисълът ще е да се получат повече победи и по-малко загуби.

1.2 Формална дефиниция

1.2.1 Дефиниция на ИИ и програма, която удовлетворява тази дефиниция

Ще разгледаме всички възможни стратегии на агента и ще покажем, че една от тях е най-добрата. Тази стратегия не е изчислима, но има близки до нея изчислими стратегии. Ще дефинираме ИИ като изчислима стратегия, която е достатъчно близо до най-добрата. За да определим най-добрата стратегия на агента, ни е нужен език за описание на света. Чрез този език ще направим и програма, удовлетворяваща дефиницията на ИИ. Програмата първо ще разбере света, като го опише чрез избрания език, после на базата на това описание ще предскаже бъдещето и ще избере възможно най-добрия ход. Тази програма е изключително неефективна и на практика неизползваема, но тя може да бъде подобрена чрез усъвършенстване на езика за описание на света и чрез подобряване на алгоритъма за предсказване на бъдещето. По този начин може да се получи ефективна програма, удовлетворяваща дефиницията на ИИ.

1.2.2 Въведение

Преди време разговарях с колега, който ми каза следното: „Може би един ден ние ще създадем ИИ, но това ще бъде една изключително неефективна програма и за нея ще ни е нужен безкрайно бърз компютър“. Моят отговор беше: „Ти ми дай тази неефективна програма, която е ИИ, а аз ще я подобря и ще направя от нея истински ИИ, който може да работи на реален компютър“.

Днес, в тази дисертация, аз давам това, което навремето поисках от моя колега. Давам неефективна програма, удовлетворяваща дефиницията на ИИ, и дори давам идеи и насоки как тази неефективна програма може да се подобри и да стане реална програма, работеща в реално време. Надявам се някой от читателите на тази дисертация да направи това и да създаде търсения от нас ИИ.

Доколко неефективна е тук описаната програма? На теория има само два вида програми – завършващи и зациклящи, но на практика някои от завършващите програми са толкова неефективни, че можем да мислим за тях като за зациклящи. Тази програма е такава и поради това е неизползваема (освен ако компютърът е безкрайно бърз или светът е съвсем елементарен).

Каква е дефиницията на ИИ? Ще дефинираме ИИ като стратегия. Ако агентът следва тази стратегия, то той ще се справи достатъчно добре. Това важи за произволен

свят, в който няма фатални грешки. Ако в света фатална грешка е възможна, тогава агентът може и да не се справи добре в този свят, но средното му представяне върху всички възможни светове пак ще е достатъчно добро.

Кои ще са възможните светове? Стратегиите на света са континуум много. Ако нямаме никакво очакване за това какъв е светът, то няма как да имаме очакване за това какъв е очакваният успех на агента. Ще предположим, че светът може да се опише и че неговото описание е възможно най-просто (това предположение е известно под името „бръснач на Окам“). Тоест избираме един език за описание на светове и се ограничаваме до световете, описани чрез този език. Световете, чието описание е по-просто (по-кратко), ще бъдат предпочитани (ще имат по-голяма тежест).

В тази дисертация ще разгледаме няколко езика за описание на света. Първият език ще описва детерминирани светове. В този език светът ще се описва чрез изчислима функция, която от състоянието на света и действието на агента изчислява новото състояние на света и следващото наблюдение. В този случай, ако знаем началното състояние на света и действията на агента, получаваме живота на агента в този свят.

Вторият език ще описва недетерминирани светове. Отново имаме изчислима функция. Този път изчислимата функция има още един аргумент и това е случайността. Тук, за да получим живота на агента в този свят, ни трябва да знаем още каква е била случайността.

Ще дефинираме ИИ чрез тези два езика и ще направим предположението, че тези две дефиниции съвпадат. Ще направим дори предположението, че дефиницията на ИИ не зависи от това кой език за описание на света сме избрали и че всички езици дават една и съща дефиниция на ИИ.

На базата на тези два езика ще направим две програми, които удовлетворяват дефиницията на ИИ. Тези две програми изчисляват приблизително една и съща стратегия, но тяхната ефективност е съвсем различна. Тоест езикът за описание на света не влияе върху дефиницията на ИИ, но силно влияе върху ефективността на получения чрез този език ИИ.

Приноси.

Тази дисертация подобрява дефиницията на ИИ, която първоначално е създадена от Hernández-Orallo през 1998 (Hernández-Orallo et al., 1998) и съществено е подобрена от Marcus Hutter през 2000 (Hutter, 2000). Подобренията са две.

1. Дефиниция на ИИ, която не зависи от дължината на живота. В (Orallo 1998 and Hutter 2000) е дадена дефиниция на ИИ, но там се предполага, че дължината на живота е ограничена с някаква константа, която се явява параметър на дефиницията.

2. Дефиниция на ИИ, която не зависи от езика за описание на света. В Hutter (2000) езикът е фиксиран. Тоест в Hutter (2000) се предполага, че има само един възможен начин, по който може да се опише светът.

1.2.3 Литературен обзор

1.2.3.1 Генерален интелект

Първо ще отбележим, че в тази дисертация, когато говорим за „изкуствен интелект“, имаме предвид „генерален изкуствен интелект“. В литературата се разглеждат два вида ИИ, които са наречени тесен и генерален (понякога ги наричат слаб и силен). Според мен по-правилно е тези два вида ИИ да бъдат наречени фалшив и истински ИИ.

Нека вземем като пример диамантите. И при интелекта, и при диамантите имаме деление на естествени и изкуствени. Изкуствените диаманти се делят на два вида – истински (състоящи се от въглерод) и фалшиви (направени от стъкло). Днес, когато говорим за изкуствен диамант, имаме предвид такъв, който е направен от въглерод. Нека си представим, че живеем в XIX век и все още никой не може да направи изкуствен диамант от въглерод. През XIX век, когато хората са говорели за изкуствен диамант, са имали предвид диамант от стъкло. Това е нещо, което прилича на диамант, но не е диамант. Днес наричаме диамантите от стъкло фалшиви диаманти.

Истинският изкуствен диамант по нищо не отстъпва на естествения. Като твърдост и прозрачност тези два диаманта са еднакви. Все пак имаме разлика в цената, защото изкуственият е много по-евтин от естествения, но той може да го превъзхожда по големина и по чистота.

Същото е с изкуствения интелект. Ако вземем истинския изкуствен интелект, той по нищо не отстъпва на естествения, но може да го превъзхожда в бързина, памет и „интелигентност“. Естествено, цената на изкуствения интелект ще е много по-ниска от тази на естествения. Днес, през XXI век, естественият интелект дори е безценен, защото не се продава.

Що се отнася до слабия или тесния изкуствен интелект, той прилича на интелект, но не е. Когато вече имаме истински изкуствен интелект, тези програми ще се наричат „фалшив изкуствен интелект“ или „програми, имитиращи интелигентност“.

Днес в повечето статии, когато се говори за ИИ, се има предвид тесен или фалшив ИИ. В тази дисертация, когато говорим за ИИ, имаме предвид генерален или истински ИИ.

1.2.3.2 Интуитивна дефиниция

Нека да направим литературен обзор и да разгледаме статиите, които са посветени на дефиницията на ИИ. Тази дефиниция е много важна и това е основният въпрос в ИИ. Въпреки това, тези статии са много малко, защото повечето изследователи въобще не си задават въпроса „Какво е ИИ?“ – малцина са тези, които си го задават. Причината е, че нашите колеги не вярват в ИИ. Ако не вярваш в призраци, не си задаваш въпроса „Какво е призрак?“. Наскоро присъствах на лекцията на един от водещите специалисти в областта на ИИ (Solar-Lezama, 2023). Той каза: „Колкото и умен да е ИИ, винаги ще се намери някой човек, който да е по-умен от него“. Явно този наш колега не вярва в ИИ и не може да си представи, че един ден ИИ ще е по-умен от всеки човек.

Статиите, посветени на дефиницията на ИИ, са малко, но все пак има такива. Много добър обзор на тези статии е направен в Wang 2019 и в статиите на Hernández-Orallo (2012, 2014a, 2014b, 2014c, 2017). Тук ще направим един по-кратък обзор, като ще се опитаме да кажем неща, които не са казани в споменатите обзорни статии.

Първата интуитивна (неформална) дефиниция на ИИ е направена от Алан Тюринг (Turing, 1950) и е известна под името „тестът на Тюринг“. Тази дефиниция е перфектна в своята простота. Все пак в нея има един съществен проблем. Тестът на Тюринг дефинира обучен интелект (тоест интелект плюс образование). Бихме искали да имаме дефиниция на необучен интелект (тоест интелектът без образованието). Първата дефиниция на чист необучен интелект е дадена от Pei Wang през 1995 г. (Wang, 1995). Нейното съдържание е следното:

Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources.[2]

По-късно, през 2000 г., дефиницията на Pei Wang е подобрена. Това подобрение е публикувано в Dobrev (2000). Днес това е първият резултат, който *Google* дава за дефиниция на ИИ. При заявката *Definition of Artificial Intelligence* първият резултат е статията Dobrev (2005a), която е подобрена версия на Dobrev (2000). Ето как изглежда подобреният вариант на дефиницията на Pei Wang:

AI will be such a program which in an arbitrary world will cope not worse than a human.[3]

В какво се състои подобрението? Първо, Pei Wang дефинира интелект, а подобрения вариант дефинира изкуствен интелект. Това подобрение не е съществено, защото истинският въпрос е „Какво е интелект?“. Това, че ИИ е програма, следва директно от тезиса на Чърч (Church, 1941), който казва, че всяка информационна система може да се емулира с компютърна програма.

Същественото подобрение на дефиницията на Pei Wang е, че там се иска интелектът да може да се справи в един конкретен свят (в своята среда), докато в подобрения вариант се иска ИИ да може да се справи добре във всеки свят. Защо това подобрение е съществено? В крайна сметка, за нас е важно ИИ да се справи добре в своята среда, защото това е важната среда, която ни интересува. ИИ обаче не трябва да зависи от средата, защото ние искаме да можем да го поставяме в различни среди (светове) и искаме ИИ да може да се справи добре навсякъде. Да кажем, че за нас е важен реалният свят, но този свят не е единствен. Има голямо значение къде и кога си се родил. Ако се променят някои от тези параметри, светът ни може да изглежда по съвсем различен начин. Очевидно Pei Wang е съзнавал, че светът не е единствен, и затова е добавил към дефиницията *while operating with insufficient knowledge and resources*. Тоест Pei Wang иска ИИ да може да се справи и когато е трудно, като идеята е, че щом се справя, когато е трудно, ще се справи и когато е лесно. Разбира се, когато си необразован и беден е трудно. Много по-лесно би било, ако разполагаш със знания и ресурси.

Още едно подобрение на дефиницията на Pei Wang е, че там не се казва колко добре трябва да се справи ИИ. Там се предполага, че или се справя, или не се справя, но знаем, че някои се справят по-добре, а други – по-лошо. Тоест важно е колко добре се справя ИИ и това е неговото ниво на интелигентност. В подобрения вариант на дефиницията се казва, че ще се справи не по-зле от човек. Сравнението с човек прави дефиницията неформална, но то е важно, защото трябва да отбележим кое ниво на интелигентност е достатъчно, за да приемем, че дадена програма покрива изискванията за ИИ.

1.2.4 Постановка на задачата

Нека агентът има n възможни действия и m възможни наблюдения. Нека Σ и Ω са множествата на действията и на наблюденията. Две от наблюденията ще са специални. Това ще са *good* и *bad*, които ще дават наградите 1 и -1. Останалите наблюдения от Ω ще дават награда 0.

Ще добавим още едно специално наблюдение и това ще е наблюдението *finish*. Агентът никога няма да вижда това наблюдение ($finish \notin \Omega$), но то ще ни е нужно при

[2] Интелигентността е способността на една информационна система да се адаптира към своята обкръжаваща среда, дори когато оперира с недостатъчно знания и ресурси.

[3] ИИ ще бъде такава програма, която в произволен свят би се справила не по-зле от човек.

дефиницията на модел на света. Моделът ще предсказва *finish*, когато се е счупил и повече нищо не може да предскаже. Няма да гледаме на наблюдението *finish* като край на живота, а като скок в неизвестното. Наблюдението *finish* ще дава награда -1, защото искаме нашият ИИ да избягва такива скокове в неизвестното.

Дефиниция 1: Дървото на всички възможности е едно безкрайно дърво. Върховете, които са на четна дълбочина и не са листа, ще наричаме „върхове за действие“, а тези, които са на нечетна дълбочина, ще наречем „върхове за наблюдение“. От всеки връх за действие излизат n стрелки, на които съответстват n-те възможни действия на агента. От всеки връх за наблюдение излизат $m+1$ стрелки, на които съответстват m-те възможни наблюдения на агента и наблюдението *finish*. Стрелката, която съответства на *finish*, води до листо. Останалите стрелки водят до върхове, които не са листа.

Дефиниция 2: „Свят“ ще наричаме тройката $<S, s_0, f>$, където:

1. S – крайно или изброимо множество от вътрешните състояния на света.

2. $s_0 \in S$ – началното състояние на света.

3. $f: S \times \Sigma \rightarrow \Omega \times S$ – функция, която на състояние и действие връща наблюдение и ново състояние на света.

Функцията f не може да върне наблюдението *finish* (то се предсказва само когато f не е дефинирана и когато няма следващо състояние на света). Дали функцията f е изчислима, детерминирана или тотална? Отговорът на всеки един от тези три въпроса може да е „Да“, а може да е „Не“.

Дефиниция 3: Детерминирана стратегия на агента е функция, която на всеки връх за действие ни съпоставя по едно действие.

Дефиниция 4: Недетерминирана стратегия на агента е функция, която на всеки връх за действие ни съпоставя едно или повече възможни действия.

Когато в един връх (момент) стратегията ни дава всички възможни действия, ще казваме, че в този момент стратегията не знае какво да прави. Няма да правим разлика между агент и стратегия на агента. Обединение на две стратегии ще наричаме стратегията, която се получава, когато избираме една от двете стратегии и я изпълняваме, без да я сменяме. Ако разрешим избраната стратегия да се сменя, би се получило нещо друго.

Дефиниция 5: „Живот“ ще наричаме път в дървото на всички възможности, започващ от корена.

Всеки живот може да се представи чрез редицата от действия и наблюдения:

$$a_1, o_1, \ldots, a_t, o_t, \ldots$$

Няма да правим разлика между краен живот и връх от дървото на всички възможности, защото има взаимно еднозначно съответствие между двете.

Дефиниция 6: Дължината на живота ще бъде t (броят на наблюденията). Тоест дължината на живота ще е дължината на пътя върху две.

Дефиниция 7: Завършен живот ще е такъв, който не може да се продължи. Тоест това ще е безкраен живот или живот, завършващ с наблюдението *finish*.

Когато пуснем един агент в един свят, резултатът ще е завършен живот. Ако агентът е недетерминиран, тогава резултатът не е единствен. Същото, ако светът е недетерминиран.

1.2.5 Оценка

Искаме да определим най-добрата стратегия на агента. За целта ни трябва оценка на живота. Тази оценка ще ни даде една линейна наредба, която за всеки два живота ще ни каже кой е по-добрият.

Първо, ще кажем какъв е успехът на живота L. Когато животът е краен, ще преброим колко пъти сме наблюдавали *good* – това число ще означим с $L_{good}(L)$. Аналогично за *bad* и *finish*. По този начин успехът за краен живот ще бъде:

$$Success(L) = \frac{L_{good}(L) - L_{bad}(L) - L_{finish}(L)}{|L|}$$

Ще означим с L_i началото на живота L, което е с дължина i. За безкраен живот L ще дефинираме $Success(L)$ като границата на $Success(L_i)$, когато i клони към безкрайност. Ако тази редица не е сходяща, ще вземем средното между точната долна и точната горна граница.

$$Success(L) = \frac{1}{2}.\left(\liminf_{i\to\infty}\left(Success(L_i)\right) + \limsup_{i\to\infty}\left(Success(L_i)\right)\right)$$

По този начин на всеки живот съпоставихме едно число, което е в интервала *[-1, 1]* и което е успехът на този живот. Дали успехът на живота да не бъде оценката, която търсим? Това не е добра идея, защото, ако в един свят няма фатални грешки, тогава на най-добрата стратегия ще ѝ е все едно какво играе. Ще има един максимален успех и този максимален успех винаги ще може да бъде постигнат, без значение колко грешки са направени в началото. Бихме искали, ако има две възможности, които в безкрайност дават еднакъв успех, най-добрата стратегия да предпочете тази възможност, която по-скоро ще даде положителен резултат. Затова дефинираме оценка на завършен живот по следния начин:

Дефиниция 8: Оценката на безкрайния живот L ще бъде редицата, започваща с успеха на живота и продължена с наградите, получени на i-тата стъпка:

$$Success(L), reward(o_1), reward(o_2), reward(o_3), \ldots$$

Дефиниция 9: Оценката на крайния завършен живот L ще бъде същата редица, но за $i>t$ членовете $reward(o_i)$ ще бъдат заместени със $Success(L)$:

$$Success(L), reward(o_1), \ldots, reward(o_t), Success(L), Success(L), \ldots$$

(Тоест наблюденията, които са след края на живота, ще получат очакване за наградата, което е равно на успеха на крайния завършен живот.)

Ще сравняваме две оценки, като вземем първата разлика. Тоест първата цел на най-добрата стратегия ще бъде успехът на целия живот, но втората цел ще бъде максимално бърз резултат.

1.2.6 Очакваната оценка

Дефиниция 10: За всяка детерминирана стратегия P ще определим $grade(P)$, което ще бъде очакваната оценка на живота, при условие че се изпълни тази стратегия.

Ще определим очакваната оценка за всеки връх v, при условие че някак сме стигнали до v и от този момент нататък изпълняваме стратегията P. Очакваната оценка на P ще бъде оценката, която сме съпоставили на корена.

Ще опишем грубо как на върховете се съпоставят очакваните оценки. След това ще опишем подробно частния случай, когато се търси най-добрата оценка – тази, която е очакваната оценка на най-добрата стратегия.

Грубо описание:

1. Нека v е връх за действие.

Тогава оценката на v ще бъде оценката на прекия наследник на v, който отговаря на действието $P(v)$.

2. Нека v е връх за наблюдение.

2.1. Нека има един възможен свят, който е модел на v.

Ако изпълняваме P в този свят, ще получим един възможен живот. Тогава оценката на v ще бъде оценката на този живот.

2.2. Нека има много възможни светове.

Тогава всеки свят ни дава по един възможен живот и оценка на v се получава като средното от оценките на възможните животи.

Следва подробно описание на най-добрата стратегия. Основната разлика е, че когато v е връх за действие, най-добрата стратегия винаги избира максимума от очакваните оценки на преките наследници.

1.2.7 Най-добрата стратегия

Както вече казахме, за да имаме очакване за успеха на агента, трябва да имаме някакво предположение за това какъв е светът. Ще предполагаме, че светът може да се опише чрез някакъв език за описание на светове.

Нека вземем стандартния език за описание на светове. Това е езикът, при който светът се описва с изчислима функция (така е в Hutter, 2000). За да опишем изчислимата функция f, ще използваме машина на Тюринг. Началното състояние на света ще опишем като крайна дума над азбуката на машината на Тюринг. Получаваме изчислим и детерминиран свят, който в общия случай не е тотален.

Дефиниция 11: Свят със сложност k ще бъде свят, в който:

1. Функцията f се описва с машина на Тюринг с k състояния.

2. Азбуката на машината съдържа $k+1$ символа ($\lambda_0, \ldots, \lambda_k$).

3. Началното състояние на света е дума с не повече от k букви. Азбуката е $\{\lambda_1, \ldots, \lambda_k\}$, т.е. азбуката на машината без празния символ λ_0.

Тук за три различни неща използваме едно и също k, но няма нужда да използваме различни константи.

Ще кажем коя е най-добрата стратегия за световете, които са със сложност k (важно е, че тези светове са крайно много). За целта на всеки връх за наблюдение ще съпоставим неговата най-добра оценка (очакваната оценка, при условие че оттам нататък се изпълнява най-добрата стратегия).

Нека имаме живота: $a_1, o_1, \ldots, a_t, o_t, a_{t+1}$.

Нека този живот минава през върховете: $v_0, w_1, v_1, \ldots, w_t, v_t, w_{t+1}$.

Тук v_0 е коренът, v_i са върхове за действие, а w_i са върхове за наблюдение.

За v_t ще видим колко модела със сложност k има той.

Дефиниция 12: Един детерминиран свят е модел на v_t, ако в този свят агентът би достигнал до v_t, ако извърши съответните действия ($a_1, \ldots, a_t$). Всеки връх за действие има едни и същи модели със своите преки наследници.

Дефиниция 13: Най-добрата стратегия за световете, които са със сложност k, ще бъде тази, която винаги избира най-добрата оценка (измежду най-добрите оценки на преките наследници).

Дефиниция 14: Най-добрата оценка на върха w_{t+1} (за световете със сложност k) се определя по следния начин:

Първи случай. Върховете v_t и w_{t+1} нямат модел със сложност k.

Тогава най-добрата оценка на w_{t+1} ще бъде *undef*. В този случай стратегията няма да знае какво да прави (в цялото поддърво на v_t), защото най-добрата оценка на всички наследници ще бъде *undef*.

Ако не искаме да въвеждаме оценка *undef*, можем да използваме най-малката възможна оценка, което е редицата от минус единици. Когато избираме максимума, го избираме между върховете, които не са *undef*. Ако заменим *undef* с възможно най-малката оценка, резултатът ще е същият.

Втори случай. Върховете v_t и w_{t+1} имат един модел със сложност k.

Нека D е този модел. В този случай има континуум много пътища, минаващи през w_{t+1}, такива че D е модел на всичките тези пътища. Ще изберем от тези пътища (завършени животи) множеството на най-добрите. Оценката, която търсим, ще бъде оценката на тези най-добри пътища. На всеки един от тези пътища съответства детерминирана стратегия на агента. Тези стратегии ще наречем „най-добрите стратегии, минаващи през w_{t+1}“.

Ще построим множеството на най-добрите детерминирани стратегии по следния начин: Нека P_0 е множеството на всички стратегии, които водят до w_{t+1}. Взимаме успеха на всяка една от тях в света D. Взимаме подмножеството P_1 от тези стратегии, в които се постига максималният успех. Намаляваме P_1, като взимаме тези стратегии, при които за $reward(o_{t+2})$ се постига максимум, и получаваме P_2. След това повтаряме процедурата за всяко $i>2$. По този начин получаваме множеството на най-добрите детерминирани стратегии P. (Най-добрите от тези, които минават през w_{t+1}. Също така най-добрите по пътищата, минаващи през w_{t+1}, а за останалите пътища няма значение как се държи стратегията.)

$$P = \bigcap_{i=0}^{\infty} P_i$$

Множество P можем да си го мислим като една недетерминирана стратегия. Нека вземем някое $p \in P$. По този начин ще получим и най-добрата оценка:

$$Success(p), reward(o_{t+1}), reward(o_{p,t+2}), reward(o_{p,t+3}), \ldots$$

Тук пропускаме членовете $reward(o_i)$ при $i \leq t$, защото те са определени еднозначно от v_t. Следващият член зависи от w_{t+1} и D, но не зависи от p. Оставащите членове зависят от p.

Горната формула можем да запишем и по този начин:

$$\max_{p \in P_0} Success(p), reward(o_{t+1}), \max_{p \in P_1} reward(o_{p,t+2}), \max_{p \in P_2} reward(o_{p,t+3}), \ldots$$

Трети случай. Върховете v_t и w_{t+1} имат краен брой модели със сложност k.

Нека M е множеството на тези модели. Отново имаме континуум много пътища, минаващи през w_{t+1}, такива че за всеки един от тях някой свят от M му е модел. Тези пътища отново образуват дърво, но във втория случай разклоненията бяха само заради различна стратегия на агента, а тук може да има разклонения заради различен модел на света. Отново имаме континуум много детерминирани стратегии, но те вече не съответстват на пътища, а на поддървета, защото може да има разклонение по модела. Отново ще търсим множеството на най-добрите детерминирани стратегии и търсената оценка ще бъде тяхната средна оценка (средната оценка в M).

Отново ще построим множествата от стратегии P_i. Тук P_1 ще бъде множеството от стратегиите, при които се постига максимумът на средния успех. Съответно P_2 ще бъде множеството на тези от тях, при които се постига максимумът на средното на $reward(o_{t+2})$ и т.н. Получената оценка ще изглежда по следния начин:

$$\max_{p \in P_0} \sum_{m \in M} q_m . Success(m,p), \sum_{m \in M} q_m . reward(o_{m,t+1}), \max_{p \in P_1} \sum_{m \in M} q_m . reward(o_{m,p,t+2}), \ldots$$

Ако вземем някое $p \in P$, тогава получената оценка ще изглежда така:

$$\sum_{m \in M} q_m . Success(m,p), \sum_{m \in M} q_m . reward(o_{m,t+1}), \sum_{m \in M} q_m . reward(o_{m,p,t+2}), \ldots$$

Тук q_i са теглата на световете, които са нормализирани, за да се превърнат във вероятности. В случая предполагаме, че световете имат еднакво тегло. Тоест:

$$q_i = \frac{1}{|M|}$$

■

Това, което описахме, прилича на алгоритъм, но не е алгоритъм, а дефиниция, защото вътре има неизчислими стъпки. Описаната стратегия е добре дефинирана, макар и неизчислима. Как чрез най-добрата оценка за сложност k ще получим най-добрата оценка за всяка сложност?

Дефиниция 15: Най-добрата оценка на върха v ще бъде границата на най-добрите оценки на върха v за световете, които са със сложност k, когато k клони към безкрайност.

Как ще дефинираме граница на редица от оценки? Числото, което е на позиция i, ще бъде границата от числата, които са на позиция i. Когато редицата е разходяща, ще вземем средното между точната долна и точната горна граница.

Дефиниция 16: Най-добрата стратегия ще бъде тази, която винаги избира действие, което води до максимума, получен от най-добрите оценки на преките наследници.

С какво най-добрата стратегия е по-добра от най-добрата стратегия за светове със сложност k? Първата стратегия знае какво да прави във всеки връх, докато втората стратегия в повечето върхове не знае какво да прави, защото в тях няма нито един модел със сложност k. Първата стратегия може да даде по-добро решение от втората дори и за върховете, за които втората стратегия знае какво да прави, защото първата стратегия разглежда и моделите със сложност, по-голяма от k. На пръв поглед не използваме бръснача на Окам, защото всички модели имат еднаква тежест, но всъщност използваме бръснача на Окам, защото по-простите светове се изчисляват от повече машини на Тюринг, тоест те имат по-голяма тежест.

1.2.8 Дефиниция на ИИ

Дефиниция 17: ИИ ще бъде изчислима стратегия, която е достатъчно близо до най-добрата.

Тук трябва да кажем какво значи една стратегия да е близо до друга стратегия и колко е достатъчно близо. Ще казваме, че две стратегии са близки, когато очакваните оценки на двете стратегии са близки.

Дефиниция 18: Нека A и B са две стратегии и $\{a_n\}$ и $\{b_n\}$ са техните очаквани оценки. Тогава разликата между A и B ще бъде $\{\varepsilon_n\}$, където:

$$\varepsilon_n = \sum_{i=0}^{n} \gamma^i (a_i - b_i) = \varepsilon_{n-1} + \gamma^n (a_n - b_n)$$

Тук γ е коефициент на обезценка. Нека $\gamma=0.5$. Добавили сме коефициент на обезценка, защото искаме когато две стратегии дълго време се държат еднакво, тогава двете стратегии да са близки. Колкото по-далеч във времето се появи разликата, толкова по-малко влияние ще има тя.

Когато n се увеличава, тогава $|\varepsilon_n|$ може да расте или да намалява. Дефиницията е направена така, защото искаме разликата да е малка, когато очакваната оценка на стратегията A кръжи около очакваната оценка на B. Тоест, ако за $(n-1)$ по-голямата е очакваната оценка на A, а за n по-голяма е тази на B, тогава в ε_n увеличението и намалението се компенсират едно друго.

Дефиниция 19: Ще казваме, че $|A-B|<\varepsilon$, ако $\forall n\ |\varepsilon_n|<\varepsilon$.

Ако $\forall\varepsilon\ |A-B|<\varepsilon$, тогава очакваната оценка на A е равна на очакваната оценка на B.

1.2.9 Програма, удовлетворяваща дефиницията

Ще опишем алгоритъм, който представлява изчислима стратегия. На всеки връх за действие съответства незавършен живот и алгоритъмът ще ни даде действие, с което този живот да бъде продължен. Този алгоритъм ще се състои от две стъпки.

1. Алгоритъмът ще отговори на въпроса „Какво става?“. Това ще стане, като намери първото k, за което незавършеният живот има модел. Също така алгоритъмът ще намери множеството M (множеството на всички модели на незавършения живот, които са със сложност k). За съжаление, това е неизчислимо. За да стане изчислимо, ще търсим ефективни модели със сложност k.

Дефиниция 20: Ефективен модел със сложност k ще бъде свят със сложност k (дефиниция 11), при който машината на Тюринг използва не повече от $(1000.k)$ стъпки, за да направи една стъпка от живота (да изчисли следващото наблюдение и следващото вътрешно състояние на света). Когато машината направи повече от $(1000.k)$ стъпки, тогава моделът извежда наблюдението *finish*.

Тук числото 1000 е някакъв параметър на алгоритъма, но приемаме, че този параметър не е много съществен. Ако един връх има модел със сложност k, но няма ефективен модел със сложност k, тогава $\exists n\ (n>k)$, такова че върхът има ефективен модел със сложност n.

2. Алгоритъмът ще отговори на въпроса „Какво да правя?“. За целта ще проиграем h стъпки в бъдещето по всички модели от M и по всички възможни действия на агента. Тоест ще обходим едно крайно поддърво и за всеки връх на поддървото ще

изчислим *best* (това е най-добрата очаквана оценка до листо). След това ще изберем действие, което води до максимума по *best*.

1.2.10 Дали това е ИИ?

Дали описаният по-горе алгоритъм удовлетворява нашата дефиниция на ИИ? Първо ще кажем, че този алгоритъм зависи от параметрите h и ε. За да намалим броя на параметрите, ще приемем, че ε е функция на h. Например $\varepsilon=h^{-0.5}$.

Твърдение 1: При достатъчно голямо h описаният алгоритъм е достатъчно близо до най-добрата стратегия.

Нека най-добрата стратегия е P_{best}, а стратегията, изчислена от описания алгоритъм с параметър h, е P_h. Тогава горното твърдение може да се запише така:

$$\forall\varepsilon>0\ \exists n\ \ \forall h>n\ (\ |P_{best} - P_h|<\varepsilon\ )$$

Не можем да докажем това твърдение, но можем да приемем, че когато h клони към безкрайност, описаната стратегия клони към най-добрата стратегия за световете, които са със сложност k. Когато t клони към безкрайност, k ще достигне сложността на света или ще клони към безкрайност. Тези разсъждения ни карат да си мислим, че горното твърдение е вярно.

1.2.11 Заключение

Разгледахме три езика за описание на света. На базата на всеки един от тях направихме дефиниция на ИИ и предположихме, че в трите случая дефиницията е една и съща. Ще направим едно още по-силно предположение.

Твърдение 5: Дефиницията на ИИ не зависи от езика за описание на света, на базата на който сме направили тази дефиниция.

Това твърдение не можем да го докажем, макар че предполагаме, че то е вярно. Също така предполагаме, че то не може да бъде доказано (подобно на тезиса на Чърч).

Въпреки че предположихме, че дефиницията на ИИ не зависи от езика за описание на света, предполагаме, че от този език силно зависи програмата, която удовлетворява тази дефиниция. При сравнението между първите два езика се виждя, че вторият е много по-експресивен и полученият чрез него ИИ е много по-ефективен.

Ще разгледаме още един език за описание на светове. Това е езикът, описан в (Dobrev, 2022b и 2023). В този език се дефинира понятието „алгоритъм“ и чрез това понятие много по-ефективно се описва света. С понятието „алгоритъм“ ние можем да планираме бъдещето. Нека вземем като пример следното: „Ще чакам автобуса, докато дойде. После ще отида на работа и ще седя там до края на работното време“. Тези две изречения описват бъдещето чрез изпълнение на алгоритми. Ако ще предвиждаме бъдещето само като обхождаме h възможни стъпки, то това h ще трябва да стане неприемливо голямо.

Езикът, описан в (Dobrev, 2022b и 2023), е много по-експресивен и той дава надежда, че чрез него ще може да се направи програма, която да е ИИ и която да е достатъчно ефективна, за да може да работи в реално време.

2 Как да го направим?

2.1 Език за описание на светове

Ще сведем задачата за създаването на ИИ към задачата за намирането на подходящия език за описание на света. Този език няма да е език за програмиране, защото езиците за програмиране описват само изчислими функции, докато този език ще опише малко по-широк клас от функции. Друга особеност на този език ще е, че при него описанието ще може да бъде разбито на отделни модули. Това ще ни позволи да търсим описанието на света автоматично, като го откриваме модул по модул. Подходът ни за създаването на този нов език ще бъде да започнем от един конкретен свят и да напишем описанието на тази конкретен свят. Идеята ни е, че езикът, който може да опише този конкретен свят, ще е подходящ за описанието на произволен свят.

2.2 Въведение

Тази дисертация представя един нов подход в изследването на ИИ. Това е *Event-Driven* (*ED*) подходът. Идеята, която стои зад *ED* подхода, е, че моделът не трябва да поема цялата входно-изходна информация, а трябва да отрази само важните събития.

Всяко действие е събитие. Всяко наблюдение също е събитие. Ако моделът отразява всички действия и всички наблюдения, то той ще се претовари с прекалено много информация. Ако моделът се ограничи само до няколко „важни“ събития, тогава това претоварване ще бъде избегнато. По този начин стигаме до идеята за *Event-Driven* моделите.

Недостатък (или може би предимство) на *ED* модела е, че той не описва света напълно, а само частично. По-точно *ED* моделът описва клас от светове (това са световете, изпълняващи определена зависимост).

С какво тази дисертация е различна? Обикновено, когато се разглежда мултиагентна система, се предполага, че светът е даден, а това, което се търси, е стратегия. Тоест светът е част от известните неща в условието на задачата, а стратегията е неизвестното. В тази дисертация ще предполагаме, че светът не е даден, а е това, което се търси.

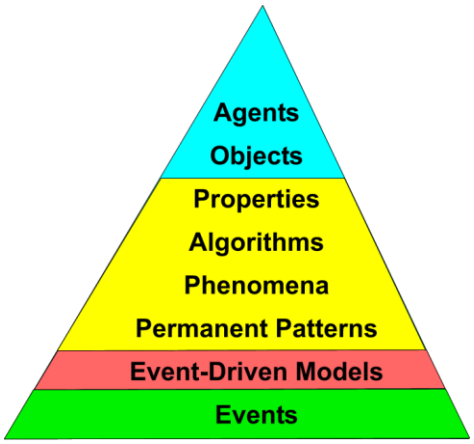
Обикновено, когато се предполага, че светът е даден, тогава предполагаме, че ни е дадена една релация, която описва света напълно. Тук ще се опитваме да опишем света и ще го описваме частично чрез *ED* модели.

Структура на езика. Описанието на света няма да прилича на хомогенна система, състояща се от един-единствен пласт. По-скоро описанието на света ще има структура, състояща се от много различни пластове. На фигура 1 представяме тази структура като пирамида, където първият пласт е основата.

Първият пласт – това ще са събитията. С тях ще опишем *Event-Driven* моделите. С тези модели ще описваме зависимости. Важно е състоянията на *ED* моделите да не са еднакви. За да се различават състоянията, трябва в поне едно от тях нещо специално да се случва. Това специално нещо ще наречем „особеност“. Множеството от особеностите ще наречем „следа“.

Първата ни идея за следа е това да е нещо постоянно. (Например да имаме едно състояние, където винаги е студено.) Следващата ни идея е подвижната следа, тоест

специалното поведение може да се появи и да изчезне или да се премести. (Например студът да се премести в съседната стая.)



Фигура 1

Следващите пластове в пирамидата – това ще са различни видове зависимости (всички зависимости ще представяме с *ED* модели). Първо ще сложим постоянните зависимости. Това са зависимостите, които се наблюдават през цялото време. В повечето статии се предполага, че всички зависимости са постоянни, но тук ще предположим, че освен тях имаме зависимости, които се наблюдават от време на време. Непостоянните зависимости ще наречем „явления“. Тоест както „следата“ може да е постоянна или подвижна, така и зависимостите могат да бъдат постоянни или „явления“.

Алгоритмите също са непостоянни зависимости (наблюдават се, докато алгоритъмът се изпълнява). Ще представим алгоритъма като последователност от събития. Обикновено алгоритъмът се разглежда като последователност от действия, защото се предполага, че главният герой е този, който го изпълнява. В тази дисертация алгоритмите ще са някакви зависимости и ако някой от агентите действа така, че да запази зависимостта, ще считаме, че агентът изпълнява алгоритъма.

Когато едно явление е свързано с наблюдението на обект, това явление ще наречем „свойство“. По този начин стигаме до абстракция от по-високо ниво. Това е абстракцията за „обект“. Обектите не ги наблюдаваме директно, а ги засичаме благодарение на техните свойства.

Следващата абстракция, до която ще стигнем, това са агентите. Тях също не можем да наблюдаваме директно, но можем да ги засечем благодарение на техните действия. За да опишем света, трябва да опишем агентите, които живеят в него, и да кажем какво знаем за тях. Най-важното е дали са ни приятели и дали с действията си ще се опитват да ни помогнат, или да ни попречат.

Дотук описваме изчислими светове (такива, които могат да се емулират с компютърна програма). Ако вътре в света има неизчислим агент, това ще направи света неизчислим, но има и друг начин да опишем неизчислим свят. Можем да добавим правило, зависещо от това дали съществува някакъв алгоритъм (по-точно съществува ли изпълнение на този алгоритъм). Въпросът дали съществува изпълнение на алгоритъм е неизчислим (*halting problem* – Turing, 1937).

Приноси:

1. *Event-Driven* модел. (Това понятие вече е въведено в Dobrev (2018), но там не е дадена интерпретация на *ED* модела. Именно интерпретацията е това, което дава смисъл на модела, и това, което разграничава адекватните от неадекватните модели.)

2. Показваме, че *Markov decision process* (*MDP*) е частен случай на *ED* модел и че *ED* моделът е естественото обобщение на *MDP*.

3. *Simple MDP*. Опростяваме *MDP* и получаваме по-прост модел, който описва повече светове.

4. Разширен модел. Това е моделът, при който състоянието знае всичко. Чрез този модел ние ще въведем интерпретация на събитията и на *Event-Driven* моделите. (Разширеният модел е въведен в Dobrev (2019a), но там състоянието знае само какво се е случило и какво ще се случи, но не знае какво е възможно да се случи. Тоест в Dobrev (2019a) състоянието на разширения модел не знае за пропуснатите възможности. В Dobrev (2019a) разширеният модел се нарича „максимален“.)

5. Дефиниция на понятието алгоритъм. Представяме алгоритъма като последователност от събития в произволен свят. Представяме машината на Тюринг като *ED* модел, който се намира в един специален свят, в който имаме безкрайна лента. По този начин показваме, че новата дефиниция обобщава понятието „машина на Тюринг“ и разширява понятието алгоритъм.

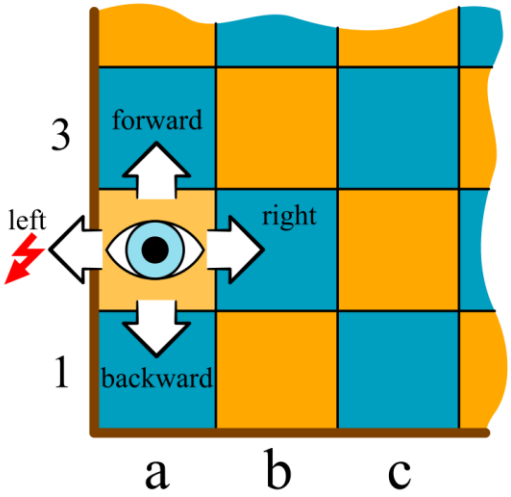
6. Език за описание на светове, при който описанието може да се търси автоматично без помощта на човек.

2.3 Играта шах

Кой ще е конкретният свят, който ще използваме за да създадем новия език за описание на светове? Това ще е светът на играта шах.

Първо ще отбележим, че ще искаме светът да е *partially observable*, защото, ако агентът вижда всичко, светът не е интересен. Ако вижда всичко, на агента няма да му е нужна фантазия. Най-важното за агента е да си представи тази част от света, която той не вижда в текущия момент.

За да бъде светът *partially observable*, ще предположим, че агентът не вижда цялото табло, а само едно квадратче от таблото (фигура 2). Окото на агента ще се намира в квадратчето, което вижда в момента, но той ще може да мести окото си и по този начин ще може да огледа цялото табло. Формално погледнато, няма разлика между това дали виждате цялото табло или само едно квадратче, но можете да местите поглед и да огледате всичко. Няма разлика, ако знаете, че, местейки поглед, оглеждате цялото табло. На практика агентът не знае нищо и той ще трябва да си изгради представата за цялото табло, а това няма да е прост процес и ще изисква фантазия.



Фигура 2

На фигура 2 окото на агента се намира в квадратчето **a2** и агентът може да го мести в четирите посоки. (В този момент не може да го премести наляво, защото е в края на таблото и ходът наляво е некоректен.) Освен местенето на окото в четирите посоки, агентът има още две действия и това са „вдигни фигурата, която виждаш“ и „спусни вече вдигнатата фигура в квадратчето, което виждаш“. Тези две действия ще означим с *up* и *down*. С помощта на тези шест действия агентът може да огледа таблото и да премести фигура, а това е всичко, което му е нужно, за да играе шах.

2.3.1 Шах с един играч

Ще разгледаме два варианта на играта шах. Ще разгледаме шах с един играч и шах с двама играчи.

Какво означава шах с един играч? Това означава, че играем с белите, обръщаме дъската и играем с черните и т.н.

Първо ще опишем по-простия вариант, когато агентът е сам и играе срещу себе си. Това е по-простият вариант, защото в този свят има само един агент и това е главният герой. По-долу ще разгледаме и по-сложния вариант, при който в света има и втори агент, който е противник на главния герой.

Въпросът е: каква е целта, когато играем сами срещу себе си?

2.3.2 Цел

В повечето статии за ИИ се избира цел, но в тази дисертация няма да фиксираме конкретна цел. Ние искаме само да опишем света, а когато сме разбрали света, бихме могли да си поставим различни цели. Например при играта шах може целта ни да е да спечелим или да загубим. Когато играем сами срещу себе си, може целта ни да се променя. Може когато играем с белите, целта ни да е „да спечелят белите“ и обратното.

Разбирането на света не е пряко свързано с поставянето на конкретна цел. Естественият интелект (човекът) обикновено няма ясно дефинирана цел, но това не му пречи да живее.

Има два въпроса: „Какво става?“ и „Какво да правя?“. Повечето статии за ИИ се опитват директно да отговарят на втория въпрос, без да отговорят на първия. Тоест обикновено директно се търси стратегия, а за да има стратегия, трябва да има цел. В тази

дисертация се търси отговор само на първия въпрос, а вторият въобще не се разглежда. Тоест в тази дисертация не ни е нужна цел.

В повечето статии целта се дефинира чрез *rewards* (целта е да се съберат повече награди). Когато говорим за *Markov decision process* (*MDP*), ще предполагаме, че от дефиницията са махнати наградите, защото те са ни нужни само ако ще търсим стратегия, а ние няма да го правим.

2.4 Заключение

Задачата е да разберем света. За да го разберем, трябва да го опишем, а за да го опишем, ни е нужен специален език за описание на светове.

Сведохме задачата за създаването на ИИ до една чисто логическа задача. От нас сега се иска да създадем език за описание на светове и този език ще е логически, защото на него ще могат да се опишат неизчислими функции. Ако езикът описваше само изчислими функции, тогава това щеше да е език за програмиране, а не логически език.

Основните градивни елементи на нашия нов език са *Event-Driven* моделите. Това са простите модули, които ще откриваме един по един. С тези модули ще представим зависимости, алгоритми и явления.

Направихме една абстракция, като въведохме обектите. Обектите не могат да бъдат наблюдавани директно, а ги засичаме индиректно, като наблюдаваме техни свойства. Свойството е специално явление, което се наблюдава, когато виждаме обект с това свойство. Тоест свойството също се представя с *ED* модел.

Следващата абстракция, която въведохме, са агентите и те също не могат да бъдат наблюдавани директно, а ги засичаме индиректно чрез техни действия.

Създадохме език за описание на светове. Това не е напълно завършен език, а е само неговата първа версия, която се нуждае от допълнително развитие. Не дадохме формално описание на създадения от нас език, а го представихме с три примера. Тоест вместо формално да описваме езика, ние написахме описанията на три конкретни свята. Това са два варианта на играта шах (с един и с двама агенти) и свят, който представя работата на машина на Тюринг.

Забележка: Не е голям проблем да се направи формално описание на език, което да покрива трите свята, които сме използвали като пример, но целта е друга. Целта е да се направи език, който може да опише произволен свят, и този универсален език да се опише формално. Това е по-трудна задача, която не сме решили.

Показахме, че езикът за описание на светове може чрез простите модули, от които се състои, да описва доста сложни светове с много агенти и сложни взаимоотношения помежду им. Това, което надграждаме над простите модули, не може да виси във въздуха и трябва да стъпи на някаква стабилна основа. Именно *Event-Driven* моделите са основата, която ще изгради езика за описание на светове, и базата, на която ще изградим всички по-сложни абстракции.

3 Какво ще правим, след като го направим?

3.1 ИИ не трябва да е *Open Source Project*

Кой трябва да притежава технологията „изкуствен интелект“? Тази технология трябва да е на всички, но не самата технология, а плодовете, които тя ще ни даде. Разбира се, не трябва да позволяваме ИИ да попада в ръцете на безотговорни хора. Аналогично, ядрените технологии трябва да носят полза на всички, но тези технологии трябва да се пазят в тайна и да не са общодостъпни.

3.1.1 Въведение

Има много хора, които защитават идеята, че технологията ИИ трябва да се разпространява свободно и дори че тя трябва да бъде *Open Source Project*. Между тях има дори отговорни и сериозни хора, какъвто е президентът Макрон (Macron, 2018). Тук ще се опитаме да поспорим с тези хора и да им обясним колко погрешно и дори пагубно би било подобно решение.

Когато президентът Макрон (Macron, 2018) говори за отворени алгоритми, може би той по-скоро има предвид собствеността върху тези алгоритми. Разбира се, няма нищо лошо собствеността да бъде на всички, но това не значи, че кодът на тези алгоритми може да бъде общодостъпен. Например една ядрена електроцентрала може да бъде държавна, тоест на всички, но това не значи, че технологиите, използвани в тази централа, са общодостъпни и че всеки може да вземе чертежите и по тях да направи собствена ядрена централа.

В момента отношението към технологията „изкуствен интелект“ е изключително безотговорно. Ние се намираме в зората на развитието на тази технология и въобще не можем да си представим каква мощ и неподозирани възможности крие това изобретение. Какво се е случило през 1896 година? Тогава Анри Бекерел (Becquerel, 1896) открива, че ако постави в чекмедже парче уранова руда върху фотографска плака, то след време плаката се осветява. Ако постави метален ключ между плаката и рудата, то изображение на ключа се отпечатва върху плаката. Така Бекерел открива радиоактивността, но по това време той въобще не може да си представи потенциала, който крие тази технология. Експериментът на Бекерел е забавен и е по-скоро нещо като фокус. Същото в момента е положението с изкуствения интелект. Появяват се експерименти, които са интересни и забавни, но хората въобще не си представят докъде може да ни доведе тази технология.

Можели ли са хората през 1896 година да предвидят колко мощна и опасна е ядрената технология? Тогава не е имало как да се досетят. Това става по-късно, когато откриват колко много енергия се отделя при разпада на атомното ядро.

Можем ли сега да се досетим колко опасна е технологията „изкуствен интелект“? Да, и повечето разумни хора се досещат, макар и да не осъзнават действителния мащаб на това откритие.

Всеки разумен и отговорен човек трябва да си зададе въпроса дали да участва в разработката на новата технология, или да остави това на глупавите и безотговорните.

В тази дисертация се занимаваме с технологичните катастрофи, които могат да се предотвратят, а не с неизбежните последствия, които няма как да бъдат предотвратени. Например, ако дадете на един глупак моторен трион, той може да изсече цялата гора и това е неизбежно последствие. Ако глупакът си отреже крака, това е технологична

катастрофа, която би била предотвратена, ако глупакът не беше чак толкова глупав и ако внимаваше повече.

Казваме, че един много мощен интелект е нещо опасно. Това не е нова идея. Още Adorno and Horkheimer (2002) са написали, че разумът може да бъде друга форма на варварство. Те казват, че с помощта на интелекта хората могат да променят природата по безогледен варварски начин. В Adorno et al. (2002) не се говори за изкуствен интелект, а за бюрократичната машина. Приликите между изкуствения интелект и бюрократичната машина са повече, отколкото разликите, и затова написаното в Adorno et al. (2002) може да се приеме за казано по темата. В Adorno et al. (2002) авторите разглеждат случая, когато група хора използват бюрократичната машина като оръжие с цел да подчинят останалите (тоталитарната държава). В Adorno et al. (2002) не се разглежда случаят, когато бюрократичната машина излиза извън контрол и започва да действа против волята на хората, защото това е невъзможно. Обществото като цяло винаги може да промени законите и правилата, които управляват бюрократичната машина. Тоест обществото като цяло няма как да загуби контрола върху бюрократичната машина, но отделният човек няма такъв контрол. За отделния човек бюрократичната машина е даденост, която той не може да промени. Аналогична е ситуацията с изкуствения интелект. Обществото като цяло ще запази контрола си върху ИИ, освен ако не сме достатъчно глупави да го изпуснем, но за отделния човек ИИ ще е даденост, която не може да бъде променена. Последното е едно от неизбежните последствия от появата на ИИ, което излиза извън обхвата на темите на тази дисертация.

3.1.2 Какво може да се случи?

Може да се случи катастрофа. Такава катастрофа може да е предизвикана умишлено или неволно. При ядрените технологии имената на две такива катастрофи са Хирошима и Чернобил. Първата е предизвикана умишлено, а втората – поради глупост и невнимание.

Умишлена катастрофа означава някой да реши да използва новата технология със зъл умисъл, тоест като оръжие. Тук понятието „зъл умисъл“ е относително, защото всеки създател на оръжие смята, че създава нещо полезно и че убивайки хора, той спасява живота на други хора. Обикновено се твърди, че убиваме малко, за да спасим много, или поне че убиваме от чуждите, за да спасим от нашите.

Може ли ИИ да убива? Не действат ли законите на роботиката (Asimov, 1950), които Айзък Айзимов е измислил? Всъщност тези закони са едно добро пожелание и те по никакъв начин не задължават създателите на ИИ да се съобразяват с тях. В момента технологиите, които претендират да имат нещо общо с ИИ, се използват предимно за оръжия. Например така наречените „умни бомби“. Обикновено се твърди, че глупавата бомба убива наред, докато умната убива само тези, които сме ѝ казали да убие. Тоест умната бомба е по-малко кръвожадна и по-хуманна. Това последното е по-скоро оправдание за създателите на умни бомби. Тези бомби са по-мощно оръжие от глупавите бомби и с тях можем да убием повече хора, отколкото със старите глупави бомби.

Нека си представим, че се е случила технологична авария и някой е изпуснал духа от бутилката и е загубил контрола над ИИ. Ако гледаме на ИИ като на оръжие, нека си представим, че някой използва това оръжие срещу нас. По този начин отношението ни към използването на ИИ като оръжие ще е строго негативно, защото отношението към едно оръжие много зависи от това дали ние го използваме, или някой го използва срещу нас.

Имаме ли шанс да преживеем подобна технологична авария? Отговорът е: нямаме никакъв шанс. Във фантастични филми от рода на „Терминатор“ (Cameron, 1984) се описва как хората воюват с роботите, но там имаме едни много тъпи роботи, които са много по-тъпи от хората. Истината за ИИ е, че той ще е много по-умен от който и да е човек. Тоест идеята за равноправно състезание между хора и роботи е безсмислена. Също толкова безсмислено е да организираме равноправно състезание по тичане между хора и автомобили. Хората няма да имат никакъв шанс, защото автомобилите са много по-бързи от хората.

Тоест ние не можем да си позволим подобна технологична авария, просто защото няма да я преживеем. В историята си човечеството е преживяло много природни бедствия и технологични аварии, но винаги ефектът от тези аварии е бил локален. Например аварията в Халифакс причинява взрив, който разрушава града. Въпреки това пораженията са локални и са ограничени до един град. Като възможно най-страшния сценарий за катастрофа се споменава ядрената война, но дори и това би имало само локални последствия. Една подобна война би могла да унищожи градовете, но все някое от селата би оцеляло. Тоест дори и възможна ядрена война не представлява такъв риск, какъвто е евентуалната загуба на контрола над ИИ.

3.1.3 Заключение

Трябва много сериозно да се отнесем към технологията ИИ и да засекретим всички програми, които имат нещо общо с тази технология. Трябва да засекретим и статиите в областта на ИИ, и дори трябва да заключим компютрите, за да не може всеки да си играе и да експериментира с ИИ технологиите. Тези експерименти трябва да бъдат разрешени само на хора, които са достатъчно умни и отговорни. В момента, за да работиш като лекар, трябва да отговаряш на куп изисквания, а за да правиш изследвания в областта на ИИ, не се иска нищо. Това трябва да се промени и да се въведат изисквания, на които трябва да отговарят ИИ изследователите.

Технологичната катастрофа, да бъде изпуснат контролът над ИИ, може да се случи от глупост или от безотговорност. Не трябва да позволяваме с ИИ да се занимават хора с комплекс за малоценност, които биха могли да пожелаят да получат абсолютната власт и да се превърнат в нещо като господар на вселената.

От друга страна, трябва да позволим на независимите изследователи да публикуват (в секретните списания) и по този начин те да получат признание за своя труд. Разбира се, когато независимите изследователи публикуват, те трябва да получат дата и гаранция, че никой няма да се опита да оспори или открадне тяхната заслуга.

Когато говорим за технологична катастрофа, имаме предвид неща, които могат да се избегнат и които, ако сме достатъчно умни и отговорни, ще избегнем. Това са загубата на контрол върху ИИ и използването на ИИ като оръжие или като средство една група хора да подчини останалите.

Има много други последствия от ИИ технологията, които са неизбежни. Например това, че хората ще си загубят работата, е такова последствие, което не можем да предотвратим. Освен че не можем да го предотвратим, ние не искаме да го предотвратяваме, защото никой от нас не иска да е принуден да работи. Бихме работили за удоволствие, но не искаме да работим по принуда.

Много хора се притесняват, че тайните служби им ровичкат из компютрите. Днес тайните служби виждат всичко и знаят всичко. Господ също вижда всичко и знае всичко, но това не притеснява никого. Разбира се, Господ е дискретен и добронамерен. Той няма

да каже на жена ви, че сте ѝ изневерил, нито ще се възползва от информацията на вашия компютър за своя изгода. Тайните служби също са дискретни, но невинаги са добронамерени. Неслучайно най-големите бандити обикновено са бивши или настоящи служители на тайните служби.

Няма нужда да се притесняваме, от това че тайните служби виждат всичко. Това е нещо неизбежно. Глупаво е да се притесняваме от неизбежни неща, които не можем да променим. Казахме, че в тези служби работят хора отговорни, но това не означава, че са достатъчно отговорни. Начинът да избегнем проблеми не е да се крием от тайните служби, а да ги контролираме и да внимаваме кой работи там.

Моята теза е, че трябва да разрешим на тайните служби официално, без да се крият, да контролират нашите компютри. Ако го направим, ще забравим проблеми като компютърни хакове, вируси и спам. Компютрите ни ще станат сигурни и надеждни. Дори може да поискаме, когато ни гръмне харддискът, те да ни възстановят информацията. Така или иначе те пазят нашата информация при себе си.

Кой работи в тайните служби е важно, но още по-важно е на кого ще позволим да прави изследвания в областта на ИИ. Това трябва да са умни, разумни и отговорни хора, без комплекс за малоценност и без престъпни намерения. Ако разрешим на всеки, който си поиска, да се занимава с подобни изследвания и експерименти, то ще се случат технологични катастрофи, в сравнение с които Хирошима и Чернобил ще ни се видят като дребни инциденти.

3.2 Как ще изглежда животът ни след появата на ИИ?

Идеята на ИИ е, че човекът създава същество, което е несравнимо по-умно от него самия. Макар че това същество ще е добронамерено и ще ни служи вярно, за нас то може да е проблем, защото сега ние се гордеем с това, че сме най-умните и че сме по-умни от всички други животни и дори от машините. Нашият интелект е това, което ни дава самочувствието, че сме върхът на еволюцията. Сега ние управляваме планетата Земя и ние решаваме кое животно и кое растение заслужава да живее, и кое да се размножи и да заеме повече площ, и кое да бъде ограничено само в резерватите.

Въпросът е как ще изглежда нашият живот след появата на ИИ. Тогава животът ни ще бъде измамно лесен. Няма да има нужда да мислим за прехраната си, няма да ни се налага да работим, дори няма да е нужно да се забавляваме един друг, защото ИИ ще ни забавлява много по-добре, отколкото който и да е човек би ни забавлявал. Въпреки измамната простота на живота, естественият подбор ще продължи и едни ще оцеляват, а други ще изчезват.

Като говорим за естествен подбор, не говорим за умиране, а за размножаване. След появата на ИИ почти никой няма да умира. Човешкото тяло може да се ремонтира и да се клонира и да продължи да съществува практически вечно, но ние може да решим да поставим граница и да кажем, че никой няма да има право да живее повече от 120 години. Колко хора ще оставим да живеят на Земята? Може да са 7 милиарда, може да ги увеличим до 70 или до 700, но може би е добре да се сложи някаква граница, защото ако сме прекалено много, ще започнем да си пречим, а и няма да остане никакво място за другите видове.

Какво ще правим след появата на ИИ? След като няма да работим, единственото смислено нещо ще е да се отдадем на размножаване. То и сега размножаването е най-важното, но сега ние работим, за да се размножим. Когато работата не е важна, няма да са важни и парите, защото с пари измерваме труда на хората, тогава кой ще е новият

критерий на естествения подбор? Сега критериите са: интелект, красота, здраве, образование, сила, смелост, бързина, честност, религия и мироглед.

Силата и бързината са били много важни в миналото, но сега, когато машините са много по-силни и по-бързи от нас, хората, силата и бързината не са най-важното. Когато машините станат по-умни от нас, тогава и интелектът няма да е най-важното. Смелостта е сложен критерий. От една страна, печелят смелите, но, от друга, най-смелите си чупят главата. Подобно е и положението с честността. Най-успешни са бизнесмените и политиците, които не блестят с особена честност, но най-нечестните влизат в затвора. Образованието е било еволюционно предимство в миналото, но днес то е по-скоро недостатък. Моят учител по рисуване казваше, че процентът на старите моми сред висшистките е много по-голям от средното, а ако една жена защити докторантура, то положението е направо неспасяемо. Тоест трябва ли образованието да бъде ценност и трябва ли да се даде по-голям шанс на по-образованите?

Що се отнася до здравето, то безспорно е ценност, но ако всички са здрави, то тогава това не може да е критерий. Красотата е друг много важен критерий, но той е твърде субективен. Кой ще определя кой е красив и кой е грозен? И сега има възможност за много сериозни козметични корекции, а след появата на ИИ ще можем да докараме какъвто си поискаме външен вид.

Кой ще определи новите критерии на естествения подбор? Дали да не оставим това на по-умния, тоест на ИИ? Градинарят определя кое цвете е красиво и интересно и кое заслужава да се размножи и да заеме повече място в градината. Ако искаме ние да сме тези, които ще определят критериите, тогава трябва да помислим по този въпрос. Тогава вероятно най-важният критерий ще е религията и мирогледът. Господстващият мироглед ще се наложи и ще даде по-голям шанс на тези, които споделят религията на мнозинството.

Заключение

Тази дисертация води до някои изводи. Основният извод е, че изкуственият интелект е близо и че скоро той ще дойде и коренно ще промени живота ни.

Основната промяна ще е това, че трудът ще стане безплатен. Ние сме свикнали с това, че благодарение на техническия прогрес трудът непрекъснато поевтинява. Например изкопаването на една дупка е било скъпо, когато хората са копали на ръка, но сега, когато има багери, тази цена е паднала драстично. Ние сме свикнали с това, че стоките в магазина стават все по-евтини, защото компонентата „труд“, която е включена в тях, поевтинява непрекъснато. Става дума не за поевтиняване спрямо парите, а спрямо нещо твърдо, като златото например (макар че и златото поевтинява, поради това че машините копаят много повече злато, отколкото са можели да изкопаят хората на ръка). Ние сме свикнали с поевтиняването на труда, но не сме готови за момента, в който трудът ще стане безплатен.

Следствие на тази дисертация и на нейния основен извод е патентът Dobrev (2021a). Този патент се отнася за схема за движение на автоматично метро (автоматично, в смисъл, че влаковете се движат сами, без машинисти). Автоматичното метро не е ИИ, както и автоматът за кафе не е. Разбира се, автоматичното метро е нещо, по-сложно от автомата за кафе, но все пак не е ИИ. Ние вече знаем какво е ИИ, защото в тази дисертация има дефиниция на ИИ. Тоест ние вече знаем или поне авторът си мисли, че знае какво е ИИ.

Патентът Dobrev (2021a) не е пряко следствие от тази дисертация, но причината за създаването му е основният извод, който се прави тук. Щом трудът става безплатен, тогава е безсмислено да се инвестира в труд. Повечето изобретения целят да спестят труда на хората. Всички тези изобретения стават безсмислени. Патентът Dobrev (2021a) цели спестяването на време и на енергия, а това са неща, които ще са ценни дори и след появата на ИИ. Разбира се, когато се появи термоядреният реактор, дори и енергията ще стане безплатна, но времето на хората ще си остане ценен ресурс и ние ще продължим опитите си да го пестим.

През 1988 г. на конференцията по логика, посветена на Хейтинг във Варна, присъстваше един от асистентите на Тюринг. Той се казваше Ганди и беше доайенът на конференцията. Ганди разказа как преди войната Тюринг е решил да превърне всичките си спестявания в сребро и да ги зарови. По този начин той искал да избегне очакваното обезценяване на парите. Не е искал да инвестира и в сгради, защото не се знае кои ще са сградите, които ще оцелеят до края на войната.

Тоест решението на Тюринг да купи сребро не е пряко следствие от войната, но е косвено последствие, предизвикано от очакванията му, свързани с войната. Аналогично, патентът Dobrev (2021a) не е пряко следствие от тази дисертация, но неговото създаване и инвестицията в неговото патентоване са мотивирани от основния извод, направен в тази дисертация.

По-нататък Ганди ни разказа как Тюринг забравил къде точно е закопал среброто. След войната той многократно се опитвал да го намери, но, уви, безуспешно. Дори и самият Ганди е участвал в такива експедиции, където заедно с Тюринг търсели сребро.

Изводът е, че ние може да се опитаме да се подготвим за предстоящи събития като войната и идването на ИИ, но едва ли ще успеем да го направим. Тези събития са твърде мащабни и е трудно в такива случаи човек да съобрази коя ще е най-добрата стратегия, която да избере.

Публикации

Авторът разполага със следните публикации, които са част от текста на дисертацията:

Част 1.

Dobrev, D. (2000). AI - What is this. *PC Magazine - Bulgaria, 11/2000, pp.12-13 (on https://dobrev.com/AI/definition.html in English).*

Dobrev, D. (2022a). The AI Definition and a Program Which Satisfies this Definition. arXiv:2212.03184 [cs.AI].

Част 2.

Dobrev D. (2022b). Language for Description of Worlds. Part 1: Theoretical Foundation. *Serdica Journal of Computing 16(2), 2022, pp. 101-150.*

Dobrev D. (2023). Language for Description of Worlds. Part 2: The Sample World. *Serdica Journal of Computing 17(1), 2023, pp. 17-54.*

Част 3.

Dobrev, D. (2019c). AI Should Not Be an Open Source Project. *International Journal "Information Content and Processing", Volume 6, Number 1, 2019, pp. 34-48.*

Авторът разполага още с 19 публикации и един патент, които са свързани с дисертацията, но не са част от текста на дисертацията:

Dobrev, D. (1993). First and oldest application. *1993. http://dobrev.com/AI/first.html.*

Dobrev, D. (2001). AI - How does it cope in an arbitrary world. *In: PC Magazine - Bulgaria, February'2001, pp.12-13 (on http://dobrev.com/AI/world.html in English).*

Dobrev, D. (2005a). A Definition of Artificial Intelligence. *In: Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, pp.67-74.*

Dobrev, D. (2005b). Formal Definition of Artificial Intelligence. *International Journal "Information Theories & Applications", vol.12, Number 3, 2005, pp.277-285.*

Dobrev, D. (2005c). Testing AI in one Artificial World. *Proceedings of XI International Conference "Knowledge-Dialogue-Solution", June 2005, Varna, Bulgaria, Vol.2, pp.461-464.*

Dobrev, D. (2005d). AI in Arbitrary World. *Proceedings of the 5th Panhellenic Logic Symposium, July 2005, University of Athens, Athens, Greece, pp.62-67.*

Dobrev, D. (2007a). Parallel between definition of chess playing program and definition of AI. *International Journal "Information Technologies & Knowledge ", vol.1, Number 2, 2007, pp.196-199.*

Dobrev, D. (2007b). Two fundamental problems connected with AI. *Proceedings of Knowledge - Dialogue - Solution 2007, June 18 - 25, Varna, Bulgaria, Volume 2, p.667.*

Dobrev, D. (2008a). Second Attempt to Build a Model of the Tic-Tac-Toe Game. *June'2008 (represented at KDS 08), published in IBS ISC, Book 2, p.146.*

Dobrev, D. (2008b). The Definition of AI in Terms of Multi Agent Systems. *December, 2008, arXiv:1210.0887 [cs.AI].*

Dobrev, D. (2013a) Comparison between the two definitions of AI. *arXiv:1302.0216 [cs.AI]*

Dobrev, D. (2013b). Giving the AI definition a form suitable for the engineer. *arXiv:1312.5713 [cs.AI].*

Dobrev, D. (2014). Comparison between the two definitions of AI. *International Conference "Mathematics Days in Sofia", July 2014, Sofia, Bulgaria, pp. 28-29.*

Dobrev, D. (2017a). Incorrect Moves and Testable States. *International Journal "Information Theories and Applications", Vol. 24, Number 1, 2017, pp.85-90.*

Dobrev, D. (2017b). How does the AI understand what's going on. *International Journal "Information Theories and Applications", Vol. 24, Number 4, 2017, pp.345-369.*

Dobrev, D. (2018). Event-Driven Models. *International Journal "Information Models and Analyses",* Volume 8, Number 1, 2019, pp. 23-58.

Dobrev, D. (2019a). Minimal and Maximal Models in Reinforcement Learning. *International Journal "Information Theories and Applications",* Vol. 26, Number 3, 2019, pp. 268-284.

Dobrev, D. (2019b). The IQ of Artificial Intelligence. *Serdica Journal of Computing, Vol. 13, Number 1-2, 2019, pp.41-70.*

Dobrev, D. (2021a). Метод за управление на метрото, при който влаковете се движат без да спират на всички спирки. *BG патент № 67273 B1/ 15.03.2021 г., заявка № 112419 от 01.12.2016 (https://dobrev.com/patent.pdf).*

Dobrev, D. (2021b). Before We Can Find a Model, We Must Forget about Perfection. *Serdica Journal of Computing, Vol. 15, Number 2, 2021, pp. 85-128.*

Декларация

Авторът декларира, че дисертацията е оригинална научна разработка. Използването на предходни резултати е отразено с подходящи препратки.

Литература

Adorno, T. & Horkheimer, M. (2002) Dialectic of Enlightenment. Stanford University Press, 2002, ISBN: 9780804736336.

Alfonseca, M., Cebrian, M., Anta, A., Coviello, L., Abeliuk, A. & Rahwan I. (2021) Superintelligence Cannot be Contained: Lessons from Computability Theory. *Journal of Artificial Intelligence Research,* Vol. 70, 2021, pp. 65-76.

Angelova G., M. Nisheva-Pavlova, A. Eskenazi, Kr. Ivanova (2021) Role of Education and Research for Artificial Intelligence Development in Bulgaria until 2030. *Mathematics and Education in Mathematics, Proceedings of the Fiftieth Spring Conference of the Union of Bulgarian Mathematicians 2021, ISSN:1313-3330, pp. 71-82.*

Asimov, I. (1950). I, Robot (The Isaac Asimov Collection ed.). New York City: Doubleday. ISBN 0-385-42304-7.

Cameron, J. (1984). The Terminator. *American science-fiction action film.*

Church, A. (1941) The Calculi of Lambda-Conversion. *Princeton: Princeton University Press.*

Council of Europe (2022) What's AI? https://www.coe.int/en/web/artificial-intelligence/what-is-ai

Becquerel, H. (1896). "Sur les radiations émises par phosphorescence". Comptes Rendus. 122: 420–421.

Dobrev, D. (2000). AI - What is this. *PC Magazine - Bulgaria, 11/2000, pp.12-13 (on https://dobrev.com/AI/definition.html in English).*

Dobrev, D. (2005a). A Definition of Artificial Intelligence. *In: Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, pp.67-74.*

Dobrev, D. (2018). Event-Driven Models. *International Journal "Information Models and Analyses",* Volume 8, Number 1, 2019, pp. 23-58.

Dobrev, D. (2019a). Minimal and Maximal Models in Reinforcement Learning. *International Journal "Information Theories and Applications",* Vol. 26, Number 3, 2019, pp. 268-284.

Dobrev, D. (2019b). The IQ of Artificial Intelligence. *Serdica Journal of Computing, Vol. 13, Number 1-2, 2019, pp.41-70.*

Dobrev, D. (2019c). AI Should Not Be an Open Source Project. *International Journal "Information Content and Processing", Volume 6, Number 1, 2019, pp. 34-48.*

Dobrev, D. (2021a). Метод за управление на метрото, при който влаковете се движат без да спират на всички спирки. *BG патент № 67273 B1/ 15.03.2021 г., заявка № 112419 от 01.12.2016 (https://dobrev.com/patent.pdf).*

Dobrev, D. (2022a). The AI Definition and a Program Which Satisfies this Definition, *arXiv:2212.03184 [cs.AI].*

Dobrev D. (2022b). Language for Description of Worlds. Part 1: Theoretical Foundation. *Serdica Journal of Computing 16(2), 2022, pp. 101-150.*

Dobrev D. (2023). Language for Description of Worlds. Part 2: The Sample World. *Serdica Journal of Computing 17(1), 2023, pp. 17-54.*

Hernández-Orallo, J., & Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. Proc. intl symposium of engineering of intelligent systems (EIS'98), February 1998, La Laguna, Spain (pp. 146–163). : ICSC Press.

Hernández-Orallo, Dowe, D.L. (2012). IQ tests are not for machines, yet. *Intelligence (2012), doi:10.1016/j.intell.2011.12.001*

Hernández-Orallo, Dowe, DL. (2014a). How universal can an intelligence test be?. *Adaptive Behavior. 22(1):51-69. doi:10.1177/1059712313500502.*

Hernández-Orallo, J.; Dowe, DL.; Hernández Lloreda, MV. (2014b). Universal psychometrics: measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research. 27:50-74, ISSN:1389-0417. doi:10.1016/j.cogsys.2013.06.001.*

Hernández-Orallo, Javier Insa-Cabrera. (2014c). Definition and properties to assess multi-agent environments as social intelligence tests. *arXiv:1408.6350 [cs.MA].*

Hernández-Orallo, J. (2017) Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review 48, 397–447, ISSN:0269-2821. https://doi.org/10.1007/s10462-016-9505-7.*

Hutter, M. (2000). A Theory of Universal Artificial Intelligence based on Algorithmic Complexity. *arXiv:cs.AI/0004001 [cs.AI]*

Ivanova Kr., M. Nisheva, E. Eskenazi, G. Angelova, N. Maneva (2020) Artificial Intelligence in and for Education in Bulgaria – Measures for Achievement Reliable Intelligent Growth. *Proc. of the 13th Nat. Conf. "Education and Research in the Information Society", 2020, pp. 7-20.*

Macron, E. (2018). Emmanuel Macron Talks to WIRED About France's AI Strategy. *31 of March, 2018, www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy*

Solar-Lezama, A. (2023) AI will program itself: synthesis, learning and beyond. *Lecture from "INSAIT Series on Trends in AI & Computing", April 3, 2023, Sofia University.*

Turing, A. (1937). "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*. Wiley. s2-42 (1): pp. 230–265.

Turing, A. (1950) Computing machinery and intelligence. *Mind, 1950.*

Wang, P. (1995) Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. *Ph.D. Dissertation, Indiana University.*

Wang, P. (2019) On Defining Artificial Intelligence. *Journal of Artificial General Intelligence 10(2) 1-37, 2019.*