РЕЦЕНЗИЯ

за дисертационния труд
„Arts and Artificial Intelligence“ [Изкуства и изкуствен интелект]
на **Невена Николаева Христова** за придобиване на образователната и научна степен „доктор” в професионално направление 2.3. Философия (Философия с преподаване на английски език)

от доц. д-р **Васил Видински**
СУ „Св. Климент Охридски“, катедра „История на философията“;
Философия на Новото време и Съвременна философия

Дисертационният труд на Невена Христова „*Arts and Artificial Intelligence*“ [Изкуства и изкуствен интелект] е в обем 194 с. и е разделен на увод, четири основни глави и заключение. Написан е на английски език. Библиографията се състои от 59 единици – на английски и на български език. Към нея е представен и филмографски списък с 10 заглавия, които са анализирани в изложението. Авторефератът е на български език и отразява съдържанието на дисертацията – той е в обем 28 стр., като 5 от тях са за библиографията и филмографския списък. Преди библиографията под заглавие „Приносни моменти на дипломната работа“ (22-23 с.) в автореферата са посочени три коректно изведени приноса. Същите приноси присъстват и в дисертацията (186-187 с.).

1. СЪДЪРЖАТЕЛНО ПРЕДСТАВЯНЕ НА ОБЩАТА ИДЕЯ, ОТДЕЛНИТЕ ГЛАВИ И ПРИНОСИТЕ В ДИСЕРТАЦИОННИЯ ТРУД

А. ОБЩАТА ИДЕЯ НА ИЗСЛЕДВАНЕТО

Като начало е важно да се посочи, че на пета страница от дисертацията Невена Христова преформулира заглавието, като заявява, че по-точното наименование на вече свършената изследователска работа е “**Narrative Arts and Artificial Intelligence**” [Наративни изкуства и изкуствен интелект]. Това е видимо стесняване на изследователското поле в сравнение с официалната тема и е пропуск, който е могъл лесно да бъде отстранен чрез навременна заявка за промяна на заглавието. Насочването към *разказвателната, наративна страна в изкуствата* подпомага, разбира се, по-детайлното реконструиране на *идеите* за изкуствен интелект (и съпътстващите ги проблеми), които са разгледани в съответните произведения. От философска гледна такова ограничение е смислено, то намалява абстрактността и двусмислеността, характерна за някои произведения, и е похвално, че е направено, макар и със закъснение.

Що се отнася до конкретната работа, то Невена Христова извежда като основен **изследователски обект** „артистичните проектории, които спомагат за развитието и разглеждането на възможни реалности“ (автореферат, 2 стр.; дисертация, 5 стр.). **Целта** е в рамките на артистичните проектории да се анализират процесите по интеграция на общия изкуствен интелект в едно бъдещо човешко общество, в което е настъпила вече *сингулярността*, като се взимат предвид всички предизвикателства, проблеми и трудности, които тази интеграция представя. Това е същината на дисертацията и може да се каже, че Невена Христова представя *убедителни аргументи в полза на подхода и в полза на смислеността на резултатите от него*. Това справедливо е посочено и като принос на дисертацията, който видимо ще може да бъде надграждан и в бъдеще, т.е. има изследователско-научна стойност. Важно е да се отбележи също, че това е продължение на вече направени философски разработки на български език, които Невена Христова коректно посочва. Подобен континуитет е и общностно значим.

В хода на изследването са разгледани най-вече популярни филми и класически за тази област литературни примери. Но амбицията на дисертацията е повече от анализ на конкретни теми в рамките на „артистичните въображаеми светове“. Невена Христова заявява, че **(1)** *планира да даде жизнени и изпълними [viable] предложения за интеграция на изкуствения интелект в обществото* (Д, 7; АР, 2) като посочва и рамката на това съвместно съществуване: **(2)** *трябва да дадем еднакви човешки права на всички личности [persons], били те биологични или изкуствени* (Д, 8)[1]. Тези **два момента** – първото като цел, второто като нормативно положение – ми се струват едни от най-съществените твърдения, чрез които може да се резюмира замисълът на Невена Христова.

Б. Съдържанието на отделните глави

Първа глава (11-63) *опитва да разгледа понятието за изкуствен интелект*. Като още в началото са отделени четири типа ИИ според степента на автономност и функционалност. Искам обаче веднага да обърна внимание, че има страшно много

[1] „**This dissertation will therefore have for its purpose to propose a method of using artistic projectories to create a snapshot of a possible reality, in this particular case: the advent of the Singularity, and using narrative forms of art such as literature, filmmaking and other media, to explore the process of integrating Artificial General Intelligence (AGI) realities into advanced human society, taking into account the challenges this would present, what approaches may work best, and what possible outcomes we may expect**.“ (Д, 6-7; оригинален получер). Както и „I plan to establish viable proposals for integration, as well as flag any potential problems with the process that may arise, on the basis of established history as well as projected possible events.“ (Д, 7) и „…respectively we must grant all persons, be it biological or artificial, the same rights to a good, just, and fulfilling existence that are afforded to us.“ (Д, 8). Или следното твърдение: „Всяка дадена субектност има право на интериорност, на вътрешен живот, който спира при границата на тялото, т.е. външният слой епидермис. Всяка намеса на ниво мозъчна активност е също толкова неприемлива спрямо една ИИ личност колкото и човешка такава.“ (АР, 18).

различни класификации, които подразделят областта на ИИ. Те в повечето случаи служат на съвсем утилитарни цели (точно както и предложената в дисертацията) и не са понятийно прецизни, нито е необходимо да са философски промислени. В това отношение е могло да бъде направено повече в дисертационното изследване. Полезното в случая обаче е това, че Невена Христова използва четиридялбата, за да *ограничи и посочи темата*, която същински ще я интересува и занимава: единствено този изкуствен интелект, който има съзнание, самосъзнание, емоции и желания – т.е. четвъртият тип в класификацията ѝ (Д, 13-14, 30-36; АР, 4). Именно поради това в тази глава са разгледани понятията за съзнание, самосъзнание, еволюция[2], цялостност, субектност, въплъщаване и др. От друга страна не е разгледано самото понятие за изкуствен интелект в неговото сложно понятийно и историческо развитие през XX и XXI век, а това е неочакван пропуск.

Втора глава (64-112) анализира артистичните проектории и техния социален аспект. Периодично в изложението се изтъква, че въображението и творчеството са едни от фундаменталните човешки характеристики (наред с любопитството) – това е и основата, от която дисертацията се насочва към проектността. Разгледан е начинът за създаване на проектории, авторовата роля, темата за другостта, както и идеята за социалния договор, т.е. *включването на другостта в рамките на всеобщото* (и по-конкретно в рамките на общността). Специално внимание е отделено на темата за езика и ролята на наративността, което е напълно закономерно с оглед на преформулираното заглавие.

Трета глава (113-162) представя по-детайлни анализи върху конкретни произведения (т.е. артистични проектории)[3]. В повечето случаи анализът е чисто тематичен. Наблюденията са извлечени от преразказването на сюжета, от най-типичната фабулна развръзка или от очевидните характеристики на героите. Понякога усещането е, че това е изследване върху популярната култура и спонтанните реакции на хората спрямо тази популярна култура[4]. Не е обаче напълно ясно дали подобни популярни употреби, автори или примери допринасят

[2] Разсъжденията върху еволюцията (Д, 37-38) звучат наивно и на моменти странно. Те изглеждат по-скоро като спонтанни и артистични реакции. Точно в тази връзка не са аргументирани и някои твърде общи твърдения в автореферата, например: „Все още се наблюдава част от ранната ни еволюция в нас, която мисли през противоречия, а не хармония.“ (АР, 1).

[3] Тук са анализирани: „Изгубеният рай“ (1667) на Джон Милтън, „Аз, роботът“ (1950) на Айзък Азимов (и едноименният филм от 2004 г.), „Франкенщайн или новият Прометей“ (1818) на Мери Шели, „Отмъстителите: Ерата на Ултрон“ (2015) на Джос Уидън, „Матрицата“ (1999) на братята Уашовски, „Тя“ (2013) на Спайк Джоунз, „Блейд Рънър“ (1982) на Ридли Скот (и съответно „Сънуват ли андроидите електрически овце?“ на Филип Дик), „Терминатор“ (1984) на Джеймс Камерън (и следващите филми от поредицата), „Дух в броня“ (1995; 2004) на Мамору Ошии, сериалът „Западен свят“ (2016-…) на Джонатан Нолан и Лиса Джой, както и други произведения.

[4] Именно в такъв контекст се появяват внезапно разсъжденията и цитатите от Туитър на журналистката диетолог Мишел Алисън (Michelle Allison, @fatnutritionist) (Д, 153); или размишленията на Ърнест Бекер (Ernest Becker), който е наречен философ, но е писател и антрополог и чието име всъщност е сгрешено (Д, 153); или примерът с Питър Тиъл, основателят на *PayPal* (Д, 153) и т.н.

задължително за по-дълбокото разбиране или пък за обективното представяне на проблема. Изглежда, че те са по-скоро лакмус за някакви *социални очаквания* в конкретен исторически контекст. Вероятно това е била целта на Невена Христова, но тя не е изрично експлицирана в дисертацията.

Четвърта глава (163-184) заявява още чрез заглавието си, че ще се заеме с деконструкцията на артистичните проектории, но по-нататък внезапно се сменя обектът и се обсъжда деконструкцията на нашите лични възгледи и вярвания (Д, 167). За съжаление до края на главата думата „deconstruction“ се използва изключително свободно (както всъщност е и в цялата дисертация) – нито като термин, нито като понятие [5]. Разбира се, част от описаните употреби на „deconstruction“ могат да бъдат напълно правомерни (вж. тук бележка под линия № 5), доколкото евентуално биха могли да са *действителен ефект* от използването й. Но пък е изненадващо, че не е представено ясно и недвусмислено разбиране какво всъщност е деконструкция. Не става и напълно ясно какво се случва в тази по-есеистична и кратка четвърта глава, което да е различно спрямо предходните анализи. Дали в нея се преобръщат или променят изводите, които вече са направени в първите три глави, след като става дума за деконструкция? Не само това, но е смущаващо и необяснимо, че не е нито цитиран, нито споменат, нито дори използван Жак Дерида, на когото дължим идеята за деконструкция. Само ще припомня, че според него тя не е нито критика в смисъла на Кант, нито класически тип анализ (срв. с бел. 5 тук).

В. Приносите и обобщенията

Приносите, които най-вече покриват II и III глава, коректно отразяват съществени моменти в дисертацията (Д, 186-187; АР, 22-23). Смятам, че тъкмо тези посоки са действително най-важни и най-стойностни в изложението. Към приносите имам и най-малко критични бележки. С това искам да посоча, че голяма част от изследването би била далеч по-убедителна, ако се съсредоточаваше именно върху тази страна на проблема, а не върху много обширни, разнопосочни и реално необхванати теми, както е в първа и четвърта глава.

Що се отнася до обобщенията, те понякога звучат свободно, немотивирано и са оставени на интуитивно ниво (т.е. представени са като самоочевидни и естествени).

[5] От време на време деконструкцията се свързва с критическо мислене (Д, 163, 173), поставяне под съмнение на статуквото (Д, 172), разобличаване на предразсъдъци (Д, 9, 67), свързване на различните проектории в едно цяло (Д, 182), хетерогенност, предотвратяване на тоталитаризма, борба с митовете или доминантната идеология (Д, 171, 173, 183), начин за дестилация на реалността (Д, 10), предоставяне на категории за запитвания (АР, 21), осъзнаване на властовите структури (Д, 166), поглеждане към себе си отстрани, себеосъзнаване (Д, 168), поглед към Другия (Д, 166), постструктуралистки автентичен отговор на въпросите свързани с властта на наратива (Д, 163), анализ на съобщението на негови компоненти и преосмисляне на разбирането (Д, 172), промяна на гледната точка (Д, 172; АР, 21), на едно място дори Тери Пратчет е разгледан като деконструктувист (Д, 176) и т.н.

Това не означава, че няма интересни твърдения или асоциации, но често липсва един по-внимателен подход към контекста и детайлите. Например, след всички разнообразни примери, които са използвани преди това в дисертацията, изглежда твърде грубо предложението да се резюмират А и Б, както и тяхното крайно противопоставяне:

> **А**. Всички компютри и машини в проекториите се водят от желанието си за самосъхранение и от стремежа за запазване на човешкия вид[6].
>
> **Б**. Човечеството като цяло е арогантен и алчен вид, същото се отнася и до човешките общества[7].

2. Оценка на отделни теми и казуси в дисертационния труд

Работата с примери

Дисертацията съдържа много ценни, положителни моменти: разгледана е обстойно темата за сингулярността в популярното изкуство и е поставена поне *рамката за правата* на общия изкуствен интелект под формата на ясно заявено нормативно положение. Включена е идеята за *съвместността на творчеството*, което е плодотворна и интересна линия. Освен това е изключително важно разглеждането от философска гледна точка на изкуствения интелект не просто като технологична иновация, а като *проект*. Затова и използването на разнородни примери от изкуството, литературата, философията, науката или всякакви други области е напълно оправдано и това е най-важната и интересна част от изложението. От друга страна, на чисто методологическо ниво дисертацията работи много повече през примери и оприемеряване, отколкото през (понятиен) анализ. Нека припомня, че примерът функционира по-скоро като потвърждение на вече приета теза, отколкото като критическа позиция. Оттам идва и общото усещане за преразказ в голяма част от изложението.

Но има и друга трудност – въпреки че в целия калейдоскоп от примери може да бъде открита *ясна обща тенденция*, при внимателно вглеждане се виждат и несъвместими различия. Не съм убеден, че Жан-Франсоа Лиотар, Георг Хегел, Ърнест Бекер, Тери Пратчет, Асата Шакур и Даниел Денет са толкова лесно съпоставими без допълнителен и по-внимателен сравнителен анализ. Същото се случва с конкретни

[6] "all the computers and machines that we see in various media projectories act out of desire for self-preservation and also the preservation of human life in terms of the species, if not the particular individual." (Д, 164).

[7] "Humanity as a whole is an arrogant, greedy species, thinking we know best, ignoring the checks and balances in search of profit and affluence. We take things that are not ours to possess, and we have no regard for life other than our own. This is frequently not true for individuals, but is almost always true for societies as a whole." (Д, 164-165). Освен че тези обобщения звучат някак тривиално, не става ясно защо общият изкуствен интелект ще запазва живота на вида *Homo sapiens*, при условие, че именно човечеството като цяло е арогантно и алчно. Като цяло от изложението не става ясна принципната функция на това резюме.

теми – например, когато става дума за позициите върху ролята и произхода на езика се изброяват Стивън Пинкър, Мартин Хайдегер, Ханс-Георг Гадамер и историка Ювал Харари (Д, 87-92). Веднага след това е споменат Левинас, хипотезата на Сапир – Уорф и т.н. Не е ли добре да бъдат по-ясно разграничени, групирани, съпоставени тези разнородни позиции и школи? Или друг пример: прословутият и все още доста безинтересен робот София (2016)[8], който е всъщност чатбот с хуманоидно лице и който получи изненадващо граждански „права“ в Саудитска Арабия (Д, 48), е анализиран редом с фикционалния, но самосъзнателен R2-D2 от „Междузвездни войни“ (Д, 51), който пък се появява едновременно със съвременните изследователски програми в *MIT AI Lab* (Д, 50) или социалните експерименти в Станфордския университет (Д, 52-53), и т.н. Тези преходи са по-скоро асоциативни, което не подпомага изграждането на обща аналитична картина около темата за „въплътяването“ (embodiment), която се представя именно на тези страници (Д, 47-55).

ТЕМАТА ЗА ИЗКУСТВОТО

Изкуството се разглежда от Невена Христова единствено като артистична проектория, т.е. не точно като изкуство, а далеч по-абстрактно. Това означава, че се игнорират стиловете, техниките, ролята на образността, историческият контекст и т.н. В някакъв смисъл изкуството е редуцирано до набор от идеи, проекти, наративи, което въображението е породило. Това е напълно оправдана редукция с оглед на непосредствените цели на дисертацията, но ми се струва, че би се *променило и обогатило* самото разбиране за проектория, ако се разгледа изкуството именно като изкуство, а не само като хипотетичен модел или въображаем свят.

Тук има и един допълнителен казус. Няма предложени аргументи в полза на странното твърдение, което се появява в автореферата, че изкуството „работи винаги в насока да придава отражение на истината на нашата реалност“, както и да показва „истинското лице на властта“ (АР, 2). В дисертацията има сходни заявления: „It is the purpose of art to place a mirror to ourselves“ (Д, 165; АР, 21). Ще припомня, че отражателната теория на изкуството (която се различава от теориите за изкуството като рефлексия) има сериозни теоретични и емпирични трудности в това да съотнася фикционално и реално или да описва функцията на изкуството в неговата хилядолетна история. От друга страна, позицията на Невена Христова не изглежда съвсем еднозначна и на моменти твърденията ѝ са по-нюансирани и изкуството се представя през неговата *критическа или „деконструираща“ функция*, а не през отражателната (Д, 6, 165 и т.н.). Подозрението ми е, че се синонимизират неволно

[8] Погрешно е посочено, че София е получила гражданство в Обединени арабски емирства [UAE], държавата е всъщност Саудитска Арабия. Освен това София е погрешно назована „Ерика“ в бележка под линия № 38 (Д, 48). Ерика е съвсем различен „робот“, който просто чете телевизионни новини в Япония.

понятия (репрезентация, отражение, рефлексия), които са доста различни; именно това води до непрецизни или проблеми твърдения[9].

Използването на вторичната литература

Обхватът на дисертацията в нейното конкретно изложение (а не само като формулирано заглавие) е наистина обширен. Това е важно от философска гледна точка, доколкото демонстрира осъзнаване на многостранността на проблема и е опит за поставянето му в широка понятийна и историческа област. На свой ред това води до някои рискове, а именно – недостатъчно критическо задълбочаване в някои от разискваните области. Не виждам достатъчно *първична или вторична литература*, свързана със следните ключови проблеми, които са изрично споменати (искам да подчертая, че това е различен проблем от недостатъчния анализ на конкретни теми):

a. Човешкото в неговата обществената или политическа роля, а оттам и темите за интеграцията на изкуствения интелект в човешката общност.
b. Водещите етическите теории през XX и XXI век, които изграждат до голяма степен рамката, в която се поместват разсъжденията на Невена Христова за ролята на изкуствения интелект.
c. Правните аспекти пред появата на изкуствения интелект. Посочването само на общата рамка ми се струва недостатъчно.
d. Развитието на философия на изкуството през XX и XXI век и дискусиите върху отношението между фикция и реалност.

Има огромно количество допълнителни книги и статии, които биха могли да бъдат прегледани във връзка с посочените теми. Най-натрапчив ми се вижда пропускът, който е свързан с отсъствието на анализите върху изкуствения интелект на основните изследователи Марвин Мински, Джон Макарти и т.н., както и важните за философията критики на Джон Сърл, Хюбърт Драйфус и др. Особено по отношение на първата глава от дисертацията тези липси изглеждат необясними. За странното отсъствие на Жак Дерида по отношение на „деконструкцията“ в IV глава вече споменах.

И една техническа бележка: когато се пише на английски език и се цитира чужда – да кажем българска – литература, то стандартът е или да се остави заглавието в оригинал или да се транскрибира и чак след това (в скоби) то може и да се преведе. Защото, ако всичко е директно дадено на английски, както е в

[9] Именно Жил Дельоз, който Невена Христова използва, разграничава имитацията и миметичността от повторението: „(подражанието е копие, а изкуството е симулакрум, то преобръща копията в симулакруми)“. Това е доста различно твърдение от заявеното в дисертацията (в същия параграф): „It is the purpose of art to place a mirror to ourselves“.

дисертацията, то така се създава погрешното впечатление, че книгата е написана оригинално на английски език.

3. Заключение

Както вече отбелязах, самата дисертация е често наративна и не толкова аналитично-понятийна. От това има три следствия. Първо, изложението се чете с лекота и е на моменти увлекателно. Второ, някои от важните проблеми биват „преодолявани“ чрез разказ, вместо да бъдат пресрещнати фронтално и да бъдат представени внимателно аргументи и контрааргументи. Трето, много различни теми са хем разделени, хем смесени на една и съща плоскост; това разделяне и смесване е на места много евристично, но в други ситуации разграниченията са недостатъчни и предизвикват допълнителни въпроси. Понякога се получава и така, че поставените от Невена Христова проблеми, са много по-мощни и интересни от предложените в дисертацията отговори, които остават на интуитивно ниво, т.е. нещо, което сякаш и преди това сме знаели или предполагали. В дисертацията също има поредица от нормативни твърдения – например, как трябва да се случат нещата от етическа гледна точка. Това е отговорно и похвално, но ми се струва, че и тук трябва да има много повече аргументация.

Въпреки тези трудности изложението съдържа *самостоятелни научни резултати с оригинален принос* и е очевидно, че Невена Христова има *задълбочени познания* във философското дискутиране на поставената тема за артистичните проектории. Разглеждането им е важно и смислено начинание, коментиран е разнообразен художествен материал. Недвусмислено е демонстрирана философската значимост на темите, проблемите и произведенията. Посочените реални приноси и свързаните с тях изводи се вписват във вече съществуваща философска традиция и освен това биха могли допълнително да се надграждат – този континуитет е важен и ценен. С оглед на всичко дотук може да се заключи, че Невена Христова се е справила задоволително с поставените й задачи.

Дисертационният труд „Изкуства и изкуствен интелект“ на Невена Христова **отговаря на необходимите минимални изисквания** за придобиването на образователната и научна степен „доктор” в професионално направление 2.3. Философия. Като член на научното жури **ще гласувам с „да“**. Нямам съвместни публикации с Невена Христова и не съм в конфликт на интереси.

Дата: 11 юни 2019 г.

доц. д-р Васил Видински
Философия на Новото време и
Съвременна философия;
СУ „Св. Климент Охридски“