РЕЦЕНЗИЯ

за дисертацията на Невена Христова

на тема

ИЗКУСТВА И ИЗКУСТВЕН ИНТЕЛЕКТ

предложена за присъждане на образователната и научна степен
„Доктор по философия“,

изготвена от Доц. Д-р Борис Грозданов
Секция *Логически модели и системи*,
ИИОЗ, БАН

Дисертацията е в обем от 194 страници и се състои от Увод (9 стр.), 4 глави (176 стр.), Заключение (5 стр.) и Библиография (4 стр.). Текстът е написан на много добър английски език, който дава практическа възможност да бъде публикуван като монография в западно издателство. Този факт, сам по себе си, заслужава адмирации, защото е изключително трудно докторант да напише подобен текст на език, различен от родния си. Но най-вече, защото проправя пътека в българската академична практика докторски дисертации да бъдат писани на английски език, което отваря автоматично едно огромно международно поле на академична достъпност за научните постижения на младите български учени.

Темата на дисертацията е формулирана върху връзката между изкуството и изкуствения интелект (AI), но реалният й фокус, както експлицитно твърди авторката (стр. 3), е върху проблема на т.н. сингулярност (както присъства в дебатите около природата на Изкуствения Интелект и не в областта на физиката и ОТО). Сингулярността е хипотетичен сценарий за развитието на Изкуствения Интелект, при който изкуствената интелигентност се изравнява и започва да надхвърля, според някои автори експоненциално, тази на homo sapiens. Единичността, кодирана в термина, често се приема в дебатите като *единствената* интелигентност, която ще остане и която ще има смисъл да бъде

наричана интелигентност (моя интерпретация на употреба на термина в дебатите).

Невена Христова подчертава контраста между двете различни и противопоставени природи на интелекта: човешкият, който е природен, и изкуственият, който не е непосредствен продукт на природата и еволюцията, а е създаден като артефакт от човешкия интелект. Христова приема, че събитието на сингулярността (стр. 4) е "далеч в бъдещето", и всъщност се "надява" да е "далеч в бъдещето", защото "човекът изобщо не е поготвен за съществуване в среда, в която функционира (силен) Изкуствен Интелект като Общия Изкуствен Интелект или AGI. Напълно споделям тази надежда и най-вече нейната аргументация.

Водещото предложение в текста, което да адресира, ако не да разреши хипотетичната конфронтация между човека и изкуствения интелект в сценария на "сингулярността", е за използването на т.н. Проекции или проектории (projectories в текста), които графично и пространствено биха могли да ни позволят да разберем по-добре на рационално ниво комплексната реалност на "възможните бъдеща". Тук бъдещето се рисува от хора на изкуствата, които онагледяват негови множество варианти като нарисувани, или пространствено-визуално онагледени реалности. Христова твърди, че в основата на артистичните проекции, разглеждани в дисертацията е "вградено" съществуването на интелект, който е много *различен* от нашия.

Първа глава разглежда понятието на AI, историята му, предлага онтологическа перспектива и въвежда една от водещите тази в дисертацията, а именно, че AGI или изкуственият интелект от тип IV, следва да се разглежда като нов вид субективност, която има свои собствени онтологически основания с тежестта на тези, които има човекът.

Втора глава въвежда и анализира арт-проекциите, като разгръща тезата за тяхна инструментална роля в развитието на текста. Главата разглежда интегрирането на AI като пълноправна субективност в обществото на хората. Въвежда понятието на семиотично кодиране, и разглежда проблемите за

автентичността, автономността, личната отговорност и контрола в случая с AGI.

Трета глава третира представянето на природата на AGI и неговите характерни (хипотетични) свойства чрез различните форми на изкуството, като наративни разкази, текстови арт форми и визуални арт форми. Разглежда проблемите за личността и отговорността. Примерите илюстрират използваните социални теории за интегрирането и приемането.

Четвърта глава "деконструира" предразсъдъците и пред-концепциите зад арт-проекциите. Дава историчност на перспективата на анализа.

Един от ключовите фокуси на работата, макар и на места неексплицитен, но подлежащ анализа и примерите, е нормативният фокус:

Инструментът на арт-проекциите носи със себе си *особена нормативна функция*, а именно, "да информира (в смисъла на регулира и ръководи) нашите действия в бъдеще" (стр. 3). Доколкото океанът от дебати по проблемите на Изкуствения Интелект може да се редуцира до два основни проблема: (1) техническия проблем дали изобщо е възможно и ако да – как точно, да се програмира Общ Изкуствен Интелект и (2), дали човечеството ще може и ако да – как, да контролира (технически и етически) такъв Интелект, и особено в крайния сценарий на Сингулярността, Христова ориентира анализа си етически по втория проблем; защото въпросът, на който съвсем скоро човечеството ще трябва да отговори по начин, който ще му позволи да съществува и евентуално, да съществува по етически позитивен модел, касае именно мотивацията зад действията в средата на силен изкуствен интелект. От една страна, това са действия на хората, но от друга това са и етически норми, които трябва да бъдат или вградени в Общия Изкуствен Интелект (ако това е възможно) или, да се надяваме, че ще се формират от самия него, като решение на самопоставената или самовъзникналата в него задача за оптимална стратегия за поведение.

Този подход, за достигане на нормативни и етически заключения по рационален път, основно контрафактически, показва силата на контра-

фактическия метод и неговата ползотворност при ненормативни и нормативни приложения. Приветствам ориентирането на текста към нормативна перспектива и напълно споделям, че ако постигнем успешни резултати, които в случая с етическите норми и нормите за човешко поведение в условията на Сингулярността, то те биха могли да бъдат постигнати само чрез много внимателното разумно контрафактическо изследване на възможните сценарии и нормативни комбинации.

Заключенията на текста, 3 на брой, са дадени изчистено в края на дисертацията и касаят "теорията на онтологическите проекции (ТОП)", като водещият заявен принос тук отново е формулиран в стандартен философски контрафактически стил:

(1) "(Текстът и ТОП) дадоха възможност да се изследват различни случаи на решения, които са изпълнени в реалност, която е *подобна* на нашата, като бяха анализирани потенциалните изходи от тези избори." Приемам този принос без забележки. Бих добавил само, че моделът на възможните бъдеща има поне два формално разгърнати подхода, един от физиката, и по-специално в Интерпретация на Относителното състояние на Квантовата Механика, където Хю Евърет показва, че сериозното вземане на формализма на КМ води необходимо логически до заключението на огромно множество от състояния, които реализират всички физически възможни резултати от едно квантово взаимодействие, като например измерването на даден физически параметър на елементарна частица като електрона. В този модел множеството състояния не са възможни, както е в дисертацията на Невена Христова, а актуални. Вторият подход е, разбира се, от областта на философската логика и по-специално семантиката на възможните светове, която следва линията Лайбниц-Крипке-Луис. Там имаме възможно съществуващи "светове", подредени според своята отдалеченост от актуалния @ свят. Може би е интересно да се изследва в една бъдеща, доразвиваща тематиката работа, тройния паралел между арт-проекциите, множеството актуални светове и множеството възможни светове.

(2) Във втория заявен принос текстът привежда аргументи в подкрепа на тезата, че “общия изкуствен интелект AGI трябва да се приеме като онтологически самостоятелна субективност и затова следва да получи аналогична позиция в обществото на тази, която имат хората.” Приемам този принос *само и в чисто модален аспект*: хипотетично е възможно това да е един от сценариите за развитието на AI и за постигането на AGI, точно както е възможно и никога да не постигнем AGI, в който, както и в други възможни сценарии, твърдението на втория принос няма да се окаже актуално. Това не отслабва, разбира се, неговата евристична научна роля на изследване на възможна и, според редица специалисти, много вероятна възможност за развитието на AI. Частта от твърдението за онтологическата тежест на субективността приемам отново само в горния модален и така евристичен смисъл.

(3) Текстът прокарва паралел между разработването и развитието на общия Изкуствен Интелект, от една страна, и арт-проекциите, от друга. Като принос се заявява конструирането на “нови светове, с техни собствени понятия и реалности, които могат да станат основа за по-нататъшни резултати.” Приемам този принос без забележки, отново само бъдещето ще покаже, как актуално ще се развият нещата.

Бих искал да ограничa въпросите си към текста до следните:

П*ървият ми въпрос* към дисертантката е: какви са основанията да очакваме бъдещият Изкуствен Интелект да бъде *различен* (ако разчитам акцента в различен като *качествен*, и не като квантитативен, примерно като измерен в брой аритметични операции с плаваща запетая в секунда) от човешкия, след като *всички* успешни подходи, познати до момента, за разработването на AGI (Artificial General Intelligence), се опират по един или друг начин именно на човешкия интелект, за да създадат Изкуствения? Може би най-добрата илюстрация за този феномен е именно водещия в момента инструмент на т.н. невронни мрежи (neural networks), които претърпяват експлозивно развитие, и който най-вероятно ще бъде основата на един

скорошен AGI. В невронните мрежи основният аналог е именно невронът на човешкия мозък, като (съвсем нефилософски) създателите на модела идентифицират човешкия разум и разумни способности само и изключително с мозъка и с неговата структурна функционалност. Без да коментирам това пълно игнориране на реално цялата история на проблема за изкуствения интелект от областта на Philosophy of Mind, искам да подчертая, че невронните мрежи работят изключително успешно в множество области като разпознаване на образи и звуци, генериране на реч и текст, имат победи в много игри, като шах и Го, който дълго време бяха поставяни на пиедестал, за който се смяташе, че може да се контролира само и единствено от човешкия разум. В този смисъл *идеята за моделиране на изкуствен интелект върху човешкия* е най-успешния проект до момента за AI изобщо.

Вторият ми въпрос към текста касае твърдението, което представлява обосноваването на избора на подхода в темата на дисертацията:

Стр. 5/6 “Причината за този избор е, че тези форми на изкуството дават възможността за по-детайлно изследване на една въобразена реалност, която се осъществява чрез възможностите на наративния и визуалния език и семиотичната практика, които зависят от споделеното културно значени, etc.”

Приемам, че изборът в този подход е определено оригинален и много интересен, вероятно не само от философска гледна точка, но дори и от математическа гледна точка на специалистите по AI, които създават моделите му, използвайки смесено формални (алгебрични) – графични (геометрични и алгоритмични) техники. Но придаването на приоритет на метафорично-илюстративния подход с арт-проекции над мощта и креативността на методите на съвременната математика, която стои не само зад всички съвременни модели на AI (невронни мрежи, познавателни графи и др.), може би е прибързано и всъщност не е необходимо. Ползата от един подобен подход, включително съвсем измеримата практическа полза за разработчиците на AI, би била по-скоро в идейното им и креативно вдъхновение за създаването на нов модел, или промяната на стар в нова перспектива, появила се като резултат от артистично

и философско вдъхновение. Този критичен коментар има за цел не да оспори подхода, а по-скоро да го позиционира малко по-точно, може би, в смесеното пространство на AI, което заема и от абстрактните философски идеи, от идеите на алгебрата, геометрията, математическия анализ и статистиката, и от съвсем конкретните полета на биологията и компютърната индустрия.

Приемам водещите тези на авторката като, оригинални, интересни и обогатяващи дебата в областта на съвременната философия на изкуството и особено в интердисциплинарната й комуникация със съвременната философия на технологията, където специфично попада и (техническия) дебат за изкуствения интелект, и особено Общия Изкуствен Интелект в сценария на Сингулярността, като игнорираме неговото присъствие във аналитичната философия на науката и аналитичната философия на съзнанието. Намирам, че заявените, налични и защитени приноси отговарят на академичните изисквания за докторска степен по смисъла на Българския закон и академичната практика в България за защита на докторски дисертации в областта на философията. Нямам общи публикации с авторката. Гласувам положително за защитата и препоръчвам на Комисията да приеме настоящия дисертационен труд с висока оценка.

Доц. Д-р. Борис Грозданов
ИИОЗ, БАН
София 10 Юни 2019