СТАНОВИЩЕ
за дисертационния труд на Невена Христова
„Изкуства и изкуствен интелект“
от доц. д-р Росен Люцканов
Институт за изследване на обществата и знанието,
Българска академия на науките

Дисертационният труд на Невена Христова е с обем 194 страници. Включва предговор, четири глави, заключение и използвана литература. Основната задача на текста е да изобрази или въобрази линиите на развитие в бъдещето, които водят до точката на „технологичната сингулярност“ и отвъд нея. Веднага бих искал да отбележа някои от силните страни на предложената за разглеждане работа (по нарастване на значимостта им): 1. чудесният език на изложението; 2. изключително строгата структура, която позволява на читателя да проследи предложената аргументация; 3. ясната принадлежност на текста към една специфично българска мисловна традиция в полето на онтологията; 4. безспорно оригиналната постановка на изследването; 5. реално осъществените приноси, изброени в заключението.
Всъщност, целите на текста са двупосочни. От една страна, той трябва да ни осигури „моментна снимка“ на „неясните и неподдаващи се на определяне обстоятелства“, свързани с потенциално значими „социални промени“ (с. 3). От друга страна, да направи въпросните значими социални промени възможни чрез самия акт на тяхното въобразяване (с. 7) Диалектиката на тези два момента – дескриптивен и прескриптивен, преминава през цялото изложение. Изборът на темата – технологичната сингулярност и (силният) изкуствен интелект, е свързан с убедеността на автора, че те са „неизбежни“ (с. 4). Разбира се, целият проблем е в това да се определи какво точно е онова, което сме приели за неизбежно. За да се отговори на този въпрос на помощ са призовани „артистичните проектории“, реализирани в наративните изкуства – основно литературата и киното (с. 4-5).
Първа глава е поглед към онтологията на теорията на изкуствения интелект. Разграничени са четири типа системи с изкуствен интелект – реактивни машини, машини с ограничена памет, машини с теория на ума и машини със самосъзнание, като е посочено, че в момента сме стигнали до втория тип (впрочем, неясно ми е защо представянето на другите като мислещи и на самия себе си като мислещ се отнасят до различни типове – по същество това не е ли една и съща задача?). По-нататък, компютърните програми с интелект от тип 1 и 2 са идентифицирани като „обекти“ (с. 15) и мястото им в онтологичната йерархия е съотнесено съответно с това на химическите реакции и на растенията (с. 16). От своя страна, тези от тип 3 и 4 са идентифицирани като „субекти“ и на тях са съотнесени нивата, присъщи съответно на животните и хората (с. 19). Основание за това е фактът, че „един обект не може да влияе върху околната си среда, не може да помни, да запаметява информация или да притежава какъвто и да било вид знание“ (с. 20). За да бъде прокарана по-ясно границата между обектното и субектното ниво на развитие са използвани индикаторите, предложени от Тонони и Кох и наречени от самите тях „аксиоми на опита“ (с. 28).
Разбирането за субектност, което се съотнася на системите от тип 3 и 4, е свързано с концепцията на професор Андонов, според която „субектът е такава организация на материята, която може да произвежда и възпроизвежда себе си и сама създава предпоставките за собственото си развитие“ (с. 34). Това е възможно благодарение на факта, че субектът се намира в конститутивно отношение със своята собствена история – “субектността е историческа“ (с. 35). Така основният въпрос се оказва: могат ли машините да имат история. Според приведеното мнение на Христо Стоев отговорът е отрицателен – машините нямат история, развитието им във времето няма цел, съответно те са лишени от индивидуалност (с. 20-21). Предложеният по-долу преглед показва, че това възражение е основателно дотолкова, доколкото ограничаваме вниманието си до безплътни компютърни програми – въплътеният компютърен интелект по мнението на Родни Брукс разполага със „способността целенасочено да формира средата си“ (с. 50), което му осигурява тъкмо онзи

тип индивидуалност, телеологичност и историчност, която обикновено приписваме на самите себе си (и изискваме от всяка друга предполагаемо интелигентна система).
Тук бих си позволил един малко по-разгърнат коментар. Интелигентността е аспект на поведението на една система, разглеждано в отношението ѝ към средата. За да имаме основания да определим една система като интелигентна, тя трябва да бъде достатъчно комплексно устроена, което ѝ позволява да взаимодейства по комплексен начин със своята комплексна среда. Виртуалните софтуерни среди са неизмеримо по-елементарни от реалния свят, съответно не предоставят нужните условия за формиране на истински интелигентно поведение. Освен това хуманоидната форма цели да предостави на въпросните системи възможността да функционират в среда, пригодена според потребностите на нашето тяло. Според мен това е основният мотив за избор на хуманоидната форма – не само нуждата от изграждане на „емоционална връзка“ със „собственика“ (с. 52). Всъщност, това е нож с две остриета. Препоръчвам на Невена Христова да се запознае с добре проучения в естетиката феномен на „зловещата долина“ (uncanny valley). Той е свързан с факта, че хората изпитват „неприязън, отвращение и дори страх“ към хуманоидни роботи, които приличат много на хора, но все пак са отличими от тях (за това вероятно има чисто еволюционни причини).
Втората глава е посветена на „онтологическите проектории“ и техните артистични превъплъщения. Според дефиницията на Веселин Дафов, проектория е „метафизичната интенционалност на човешкото съзнание, която е интегрирана в самата реалност“ (с. 64). Съответно, примери за артистични проектории ни дават мащабни саги от типа на „Междузвездни войни“, които „създават съвсем различна вселена със собствени социални и политически неписани норми“ (с. 65), благодарение на свойството, присъщо според Рикьор на всяка символна система - „да допринася за оформянето на реалността“ (с. 104). Съответно, „художествената измислица има трансфигуративен ефект, позволява на читателя да взаимодейства с хора, различни от него самия, да вижда света през очите им ... Разказите са начинът ни да реагираме спрямо и да интерпретираме света, в който живеем“ (с. 105). Според Невена Христова, за да може обществото ни да еволюира отвъд ограниченията на доминиращите днес социални, икономически и политически модели, имаме нужда от точно такива алтернативни начини на мислене, които ни осигуряват артистичните проектории (с. 107). Подобно разбиране намира опора при Хегел, според когото „изкуството е обективен посредник, чрез който една общност колективно рефлектира върху самата себе си“ (с. 167).
Трета и четвърта глава са посветени на социално-политическите последици от евентуалната поява на общ изкуствен интелект. В началото на тази част от текста е поставен въпросът „в кой момент изкуственият интелект спира да бъде творение на даден човек (или група от хора) и кога придобива субектност, превръща се в самостоятелна личност“ (с. 134). Отговорът на Христова гласи, че това е моментът, в който той „започне да разбира понятието за моралност и да прилага възприетите морални норми към хората около себе си и към самия себе си“ (пак там). Иначе казано, в момента, в който изкуственият интелект стане субект на морални задължения, със самото това той става носител на права, съответно на основното право, от което произлизат всички останали - правото на лична неприкосновеност (с. 134-135).
Четвърта глава поставя един въпрос с централно значение за самото изследване – за какво говорят произведенията на киното и на литературата, които привидно се отнасят до бъдещето. Ясно е, че „Ние хората използваме идеята за изкуствен интелект в изкуството и медиите за означаване на въстанието срещу правилата, наложени на човечеството от една по-висша сила“ (с. 169), че „този дебат се е случвал много пъти в историята, във времена, когато не всички хора са били считани за хора, а равните права за всички са били само проблясък на хоризонта“ (с. 178), че „сами се заблуждаваме, че проектираме бъдеще [докато] в най-добрия случай рефлектираме върху настоящето“ (с. 182). Тогава обаче съвсем не е ясно (поне за мен) дали в пиесата „Р.У.Р.“ на Чапек изобщо става дума за роботи. Те ли са онези същества, които полагат тежък, унизителен и дехуманизиращ труд във фабриките и рано или късно ще въстанат срещу своите господари? Не са ли интелигентните роботи просто един дигитален пролетариат? За андроиди ли става дума в художествените произведения, предлагащи ни

едно бъдеще, в което въпросните същества се разглеждат като непълноценни индивиди, които не споделят културата и етическите ни норми, но ще „отнемат работните ни места“ (с. 181)? Съответно, можем ли да разграничаваме ефективно изпълнението на двете задачи (наречени по-горе дескриптивна и прескриптивна), които обслужва текста?

Искам да завърша становището си с още един въпрос. Целта на текста беше да очертае с помощта на метод, изхождащ от концепцията за „артистични проектории“, едно радикално различно бъдеще, включващо радикално различни, нечовешки интелигентни субекти. В хода на анализа се оказа, че въпросните субекти трябва да имат тяло (да бъдат въплътени), да се развиват и еволюират самостоятелно (да бъдат неприкосновени), да бъдат обучавани (също като нашите деца), да са способни да различават добро от лошо (да бъдат морални), да притежават съзнание (понятие за себе си като себе си). Тогава въпросът е: защо определяме тези интелигентни системи като изкуствени? По какво изобщо те се различават от нас? Съществени ли са тези разлики? Не е ли интелектът по начало един съвсем естествен, природен феномен, който може да се реализира в различни носители? Свидетелстват ли разгледаните произведения за нашата способност да си въобразим радикалната другост?

Обсъждането на двете групи от въпроси според мен би могло да бъде от полза при евентуална преработка и публикуване на текста, което убедено препоръчвам. В заключение бих искал да констатирам, че текстът на дисертационния труд свидетелства за изключително добре изпълнена образователна подготовка и съдържа безспорни изследователски приноси. Ето защо ще гласувам за това на Невена Христова да бъде присъдена научно-образователната степен „доктор“ по направление 2.3 (философия).

10.06.2019,
София /Р. Люцканов/