Јанко Нешић[1]
Универзитет у Београду
Филозфски факултет

УДК 123:159.91; 004:17
Оригинални научни рад
Примљено: 25. 03. 2014.

КОМПЈУТЕРСКИ СИСТЕМИ КАО ДЕЛАТНИЦИ

Резиме: У овом раду испитујем да ли компјутерски системи као вештачки ентитети могу постати делатници. Показујем да компјутерски системи не задовољавају услове стандардног модела делања. Они не поседују ментална стања чиме се отварају проблеми менталне каузалности и слободне воље. Могући су реалистички и антиреалистички приступ. Аргументујем да је потребно заузети одређени реалистички приступ проблему моралности машина.

Кључне речи: компјутерски систем, делатник, ментална стања, слободна воља.

1. Увод

Да ли неки вирус чини зло дело када уништава важне фајлове и да ли робот може повредити људско биће, те због тога бити „кажњен“ искључивањем, зависи од тога да ли се они могу сматрати *делатницима* (енг. *moral agent*). Да ли се може говорити о „поступцима“ робота? Ово је питање о томе да ли се машине могу подвести под царство моралности, које је, до сада, искључиво било насељено људским бићима. Компјутерски артефакти постају све независнији. Неки их већ сматрају *вештачким делатницима* (енг. *artificial moral agent*). Многи од дизајнера таквих система сматрају да су на трагу одређеног облика вештачке интелигенције и да постојећи системи већ показују особине свесних и аутономних ентитета.

Сврха наше полемике биће у томе да ове ставове доведемо у питање и одредимо морални статус компјутерских система. Дакле, да би се могло говорити о компјутерској етици и да бисмо решавали конкретне проблеме компјутерске етике, потребно је прво одговорити на питање: да ли се компјутерски системи могу сматрати делатницима? Компјутерски системи су продукт људске делатности и они ће бити онакви каквим их људи начине. Дакле ако одлучимо да од њих не начинимо моралне ентитете, они то неће бити. Но, у природи човека је да, када има могућности да нешто учини или створи, то највероватније и учини. Али, не постоји неки нужан процес, еволуција компјутерских система, која ће довести до стварања таквих ентитета. Оно о чему можемо одлучивати јесте да ли ћемо створене вештачке ентитете укључити у заједницу делатника.

[1] eshatolog@gmail.com

Постоје два приступа проблему вештачке моралности: *реалистички* и *антиреалистички*. Први узима овај проблем као метафизички проблем, и тражи „истину“ о моралности машина. Друга стратегија „решава“ проблем релативистички. Ја приступам овом питању реалистички и покушаћу да покажем који се *метафизички* проблеми налазе на путу конструкције вештачких делатника. Ово ћу учинити преко стандардног модела делања (2), који постулира поседовање менталних стања као услов за добијање статуса делатника (3). Размотрићу проблем менталне каузалности и додаћу услов да мора постојати слободна воља да би било делања (3). Такође, критиковаћу антиреалистички приступ и његов бихејвиоризам и релативизам (4). Потребно је заузети реалистички приступ (5).

2. Стандардни модел делања

Прво што морамо учинити јесте да одредимо шта значи бити *делатник*. Позајмићу одређење Дебре Џонсон. Узећу у обзир оно што се може сматрати *стандардним моделом* делања (Johnson 2006: 198). Људи се разликују од других бића по томе што делају слободно, а не из нужности (или могу да делају и слободно, а не само из нужности). Одговорност постоји само ако има слободне воље. Ако је неко починио злочин зато што је под претњом натеран на то, ово ће умањити његову кривицу, у моралном и правном смислу. Ако смо нешто учинили под утицајем неког биолошког узрока и морални суд о нашем поступку мора бити другачији. Случајност, такође, може умањити нашу одговорност.

Колоквијални концепт слободне воље је дуалистички. Данашња наука и филозофија имају другачије поимање слободне воље, које је материјалистичко. Неки филозофи сматрају да слободна воља не постоји, или да не поседујемо онолико слободе колико мислимо да имамо. Овим проблемима бавићу се касније у раду, што ће нас приближити одређењу моралног статуса компјутерских система. Ипак, треба увидети да у одређеном смислу ни морални статус људи није више тако сигуран.

Посматрајмо стандардни модел делања. Када човек дела својом вољом, он то чини зато што поседује унутрашња стања, оно што називамо *менталним стањима*, која су узроци спољашњег понашања. Само захваљујући менталним стањима за човека се може рећи да је нешто учинио зато што је то хтео и зато што је имао неке *разлоге* да то учини. По стандардом моделу постоје следећи потребни услови да би се неко људско понашање назвало *делањем*:

1. Постоји неки делатник и он има унутрашња стања (ментална стања);
2. Постоји спољашњи догађај који изазива тај делатник;
3. Унутрашње стање је *узрок* тог спољашњег догађаја;
4. Спољашње понашање има неки спољашњи ефекат, последицу;
5. Тај ефекат има свог примаоца, оног који трпи последицу (Johnson 2006: 198).

Потребно је да постоји делатник који поседује унутрашња, ментална стања, попут жеља, циљева, веровања и најважније, намера да се дела. Међу свим менталним стањима за делање најважнија је интенција да се дела и разлози, који су узроци тог делања. Нека промена је потребна, промена коју у спољашњости изазива делатник. Прави узроци тих спољашњих промена су унутрашња, ментална стања. Спољашње понашање делатника ствара одређене последице, ефекте који утичу на друге људе, нпр. ја имам намеру да неког физички повредим и то чиним, те постоји неко ко трпи ту радњу. На овај начин се затвара троугао делања. Морална евалуација се односи на делање.

Стандардни модел је погодан за даље промишљање моралног статуса компјутерских система. Нећу покушавати да га изменим да би га прилагодио некој унапред постављеној замисли компјутерских система, са намером да их што лакше учиним „моралним“. Намера ми је да, упоређујући компјутерске системе са стандардним моделом, истражим да ли ти системи могу бити делатници.

Слажем се са Џонсоновом да се за компјутерске системе може рећи, без много проблема, да задовољавају услове 2–5. Унутрашње промене у систему изазивају спољашње промене, неку врсту понашања које може бити подложно моралној евалуацији. Постоји, дакле, отеловљено понашање, као код људи, које се морално оцењује. Међутим, постоји проблем у говору о унутрашњем и спољашњем код компјутерских система. Ово може бити само метафоричка подела код компјутера. Ми говоримо о менталном животу човека, као његовом унутрашњем свету који узрокује спољашње понашање. Да ли се код компјутерских система може говорити о нечем сличном, о софтверу као о менталном компјутеру. Коначно, проблем је у томе да ли се за понашање машина може рећи да је делање за које је потребна слободна одлука.

Дакле, можемо претпоставити да услови од 2 до 5 могу бити задовољени код компјутерских система. Највећи проблем са приписивањем моралности машинама је што оне немају ментална стања као људи, дакле у првом услову. Такође, недостатак менталних стања повлачи са собом и проблеме 2. и 3. услова, јер је граница између унутрашњег и спољашњег нејасна. Потребно је одговорити на друга питања, попут оног о односу духа и тела (енг. *mind-body problem*), да би се уопште расправљало о могућности да машине буду делатници. Проблем односа духа и тела је неразрешен проблем. Ипак, од нашег одређења према овом питању зависиће како ћемо посматрати вештачке артефакте. Зато ћу скицирати нека решења и правце који нас очекују, зависно од тога коју теорију прихватимо.

3. Ментална стања

Да бих показао како од метафизичких полазишта у дебати о односу духа и тела зависи и наш став према моралном статусу компјутерских система, потребно је да себи поставимо питање које Грахек поставља у књизи *Материја,*

свест и сазнање: „Може ли се говорити о менталном животу робота?“ Грахек га поставља да би заоштрио полемику између функционално-физикалних теорија и антиредукционистичких теорија. Питање гласи: да ли савршено конструисан робот (што у нашем случају можемо заменити термином „компјутерски систем“) заиста поседује ментални живот попут људи или се о њему може говорити само метафорички? Дакле, да ли се и људско понашање може описати као један функционално-физикални систем, тј. да ли се ментална стања могу описати само физикалистичким језиком и појмовима. Помоћу овог примера Грахек испитује важење функционално-физикалистичких теорија.

Потребно је узети у обзир феноменална обележја менталних стања, јер највећи проблем неке физикалистичке теорије представља осећај који се јавља у одређеном менталном стању, нпр. осећај бола. Чак и када би се мој бол могао до танчина описати неурофизиолошким терминима, тиме се не би ни дотакао квалитативни *осећај бола*. Ова *sui generis* својства менталних стања називају се *квалије* (енг. *qualia*). Грахек указује да се неће бавити практичним питањем да ли се може и како конструисати савршено програмирани робот који би подражавао људско понашање, већ теоријским питањем појмовног карактера, да ли роботи могу имати свест и ментална стања попут људи. А ово је питање о различитим теоријама свести и статуса менталних стања. Од питања о „аутентичности менталног живота“ зависе функционалне-физикалистичке теорије.

Три су главна разлога, по Грахеку зашто је редукционистима „потребан“ такав модел робота. Први разлог је што би тако савршено конструисани робот, као чисто физички систем био парадигматичан пример функционално-физикалистичке теорије о менталном животу човека. Он би био доказ таквих теорија. Други разлог је епистемолошки, јер савршени робот посматран из перспективе трећег лица изгледаће изоморфан људима, којима се без проблема приписују ментални појмови. У том случају и роботу би могла бити приписана ментална стања. Коначно, постојање овако конструисаног робота пружило би емпирички доказ да и људски организам представља само један функционално-физикалистички механизам.

Физикалисти тврде да могу постојати случајеви феноменално или квалитативно необојених стања, која би ипак имала функционалне и бихејвиоралне карактеристике менталног стања, нпр. да неко може бити у стању бола, али да не *осећа* бол (*ersatz* бол). С друге стране, ови примери антиредукционистима говоре да једно биће изоморфно људском бићу, попут робота или зомбија, у функционалном и бихејвиоралном смислу једнако човеку, ипак не би имало феноменални карактер менталних стања, а самим тим ни свест. По њима, функционално-физикалистички модел свести једноставно није довољан опис људског стања, он је редуктиван. Питање о моралности робота и компјутерских система јесте *метафизичко* питање. И као што Грахек показује, и редукционисти и антиредукционисти, без обзира на опречне одговоре о статусу менталних стања робота, сматрају да је ипак могуће дати коначан, истинит или лажан, одговор на то питање. Може се, дакле, јасно одредити да ли роботи осећају и мисле или не. Ово ће онда коначно разрешити питање да ли се они могу сматрати делатницима у заједници са људима или не.

Суштина питања о делатништву робота је у менталним стањима, нарочито у интенционалности, која ствара могућност за делање. Џонсонова сматра да се компјутерски системи могу узети као морални ентитети зато што имају интенционалност, али вештачку, уграђену од стране њихових дизајнера. Дакле, они делују интенционално у одређеном смислу па се могу сматрати бар моралним *ентитетима*; ипак, нагласак је на интенционалности њихових твораца. Али та интенционалност није аутентична, зато што машине не поседују ментална стања; ово је „сурогат" интенционалност. Џонсонова говори о непостојању *воље* код машина, компјутерски системи не могу се сматрати *вољним* бићима, као што су то људи. Они вољу и „избор" добијају од својих дизајнера.

3.1 Ментална каузалност

Компјутери, захваљујући великом броју процеса које могу да обављају истовремено, боље од људи могу „прорачунати" и предвидети последице својих поступака као и број опција које имају при неком избору. Ово их, наизглед, чини „компетентнијим" за похвално делање од људи. Но, они немају могућност избора. Њихов избор детерминисан је циљевима дизајнера. Компјутерски системи не чине слободан и свестан избор међу тим могућностима, већ то чине у складу са циљевима, задатим од стране дизајнера. Такође, с обзиром да не познају *смисао* избора, он постаје потпуно илузоран.

Услов да делатник мора имати слободну вољу, не постоји међу 5 услова стандардног модела, али је имплицитно садржан у 1. и 3. услову. Овај проблем је у директној вези са проблемом менталних стања робота. Ако роботи немају ментална стања, па самим ти ни интенционалност, не могу имати ни слободну вољу. Али физикалисти и не сматрају да постоји „картезијанска" слободна воља. Долазимо до проблема менталне узрочности (3. услов). Физикалисти тврде да ни ментална стања нису ништа више од својих неурофизиолошких (физичких) основа. Такво је становиште редуктивног физикализма. Ако се прихвати функционално-физикалистичко становиште, није тешко замислити један чисто физички ентитет који би могао бити делатник. Ако ни људи немају неку врсту слободне воље, зашто би онда роботи и компјутерски системи, били искључени из моралности?

Ким (1993) сматра да се не може бранити силазна узрочност. Силазна узрочност је способност емергентних својстава виших слојева стварности, да независно утичу на догађаје и процесе у нижим слојевима. Када неки систем дође до одређеног степена комплексности своје организације, појављују се нова, емергентна својства, каквих није било у нижим слојевима и чију појаву нисмо могли предвидети. Живот, свест и рефлексивно мишљење класични су примери јако емергентних својстава. Ово нам омогућава и појаву слободне воље. Ким тврди да емергентизам пропада са силазном узрочношћу. Емергентисти морају показати како силазна узрочност избегава двоструку детерминацију (енг. *overdetermination*): постојање два или више довољних, а различитих узрока исте последице. То је Кимов аргумент каузалног искључења.

Ким жели да покаже како је ментално сувишно и да сав каузални „посао“ обавља физичко (неурофизиолошко). Овде се подразумева *принцип физичке реализације*. По том принципу исто ментално стање могу реализовати различите физичке основе. Реализација обично претпоставља да својство које је реализовано наслеђује каузалне моћи основе која је реализује. Варијације се одвијају на нижем нивоу. Насупрот овоме, у емергенцији се варијације одвијају на вишем нивоу и каузалне моћи нису наследне. Ово је и став дуализма.

Когнитивна и интенционална ментална стања повезана су са понашањем и по Киму могу се функционално окарактерисати; квалије, вероватно, не могу (Kim 2005: 165). Иако функционализам није до сада успео да пружи, а можда никада и неће успети да пружи потпуне дефиниције сложених менталних стања, Ким сматра да су делимичне дефиниције довољне да би омогућиле напредовање и константну допуну у научном истраживању подлежућих физичких и неуралних механизама. Ким сматра да је Чалмерсова хипотеза „зомбија“ (Chalmers 1996) сувише контроверзна, па се пре одлучује за хипотезу инверзије квалија. Без обзира колико ове хипотезе деловале контроверзно Киму, оне су валидне и аргументоване критике функционализма. Критика из инверзија квалија указује да је функционализам „слеп“ за разлике у квалитативном карактеру менталних стања: опажај боја има исту „каузалну улогу“ и у случају када неко опажа предмет као црвен, док други посматрач тај исти предмет опажа као зелен.

Особина функционализма да посматра само *релациона* својстава, без осврта на *инстрисична* својства менталних стања, ствара додатне проблеме (Lowe 2000: 44-67). Ово се односи пре свега на занемаривање квалитативних својстава, чија се каузална улога негира. Из инверзије квалија не можемо закључити да квалије не постоје, ни да су каузално инертне. Може се на овај аргуемнт одговорити на следећи начин: да замена боја управо утиче на последице које те боје могу оставити на нас (црвена изазива страх, зелена умирује). Али овај одговор не може дати функционалист ако сматра да ове разлике долазе из нашег *веровања* у утицај боја на расположење. И да имају исти каузални утицај две квалије, па их не можемо разлучити, ово не значи да никаквог каузалног утицаја и нема, већ да је критеријум непрецизан. Нелегитимно је закључити из аргумента инверзног спектра да су квалије епифеноменалне (Lowe 2000: 58). Лоув указује на једну битну особину функционализма: он је онтолошки неодређена позиција. Функционализам може бити компатибилан са разним теоријама: физикализмом, нередуктивним физикализмом (ако укључује реализацију), али и са интеракционим дуализмом. Слобода у примењивости функционализма није предност и не увећава привлачност ове позиције.

Такође, сматра се да нема менталног утицаја на физичко јер по *принципу каузалне затворености физичког домена*, не може бити нефизичког узрока физичког догађаја. Приговара се и да је ментална силазна узрочност врста *механичке* узрочности и да укључује *трансференцију*, пренос енергије или момента са узрока на последицу. Овим принципима покушава се оправдање забране утицаја ичега нефизичког на физички домен. Међутим, каузална затвореност је

претпоставка која још увек нема научну потврду. Да нема сложног става по питању каузалне затворености показује и велики број дефиниција овог принципа. Општа дефиниција могла би да гласи: „Сваки физички догађај који има узрок у тренутку *m*, има физички узрок у тренутку *m*“ (Kim 1993: 280). Али могуће је замислити и такав облик дуализма који би ипак био конзистентан са каузалном затвореношћу физичког домена.[2]

У теорији интерактивног дуализма, ментално не започиње физичке каузалне ланце, ментални догађај само *омогућава* да се нека физичка узрочност изврши. Нема никаквог проблема у томе да ментални догађај узрокује одређену физичку *каузалну чињеницу* (Lowe 2008). У том случају кохерентно је да ментални догађај M узрокује „да буде случај да одређени физички догађаји P1, P2... Pn имају одређену физичку последицу P“. Ментални догађај M није директни узрок P1, већ читавог каузалног ланца, као што Бог може узроковати актуелност света, а да није ниједан од физичких узрока у бесконачном каузалном ланцу света.

Менталној узрочности приговара се да укључује трансференцију. Ово би значило да ментална узрочност претпоставља пренос одређеног квантитета енергије или момента са узрока на последицу. У том случају ментална силазна узрочност била би врста *механичке узрочности*. С обзиром да целина не може предати енергију својим деловима, јер је апстрактни конструкт, силазна узрочност се одбацује. Интерактивни дуализам не прихвата теорију узрока као механизма (Gibb 2010). Дуализам не претпоставља менталну узрочност која је пренос енергије или момента са узрока на последицу. Самим тим, не прихвата да је једини начин за неки нефизички систем да утиче на одређени физички систем, додатак нове енергије или редистрибуција енергије или момента у њему.

Видели смо због чега физикалисти одбацују постојање или каузалну ефикасност менталних стања. Они не морају одбацити слободну вољу, али сматрају да постоји само у једном ограниченом облику. Ако је физикализам у праву, онда није немогуће замислити да ће у будућности бити конструисани роботи и компјутерски системи способни за делање.

3.2 Слободна воља и неуронауке

Проблем менталне каузалности је питање да ли воља као дистинктна ментална диспозиција (моћ) може имати каузални утицај слободан од физичких својстава. За слободну вољу потребна је ментална каузалност. Да ли решење проблема слободне воље може доћи из истраживања у неуронаукама?

Експерименти на које се најчешће позива у филозофској литератури о слободној вољи су Либетови експерименти са „слободним волунтарним акцијама“. Ови су показали, по неким интерпретацијама, да несвесни мождани процеси покрећу вољне акције. Вољном акту претходи електрофизиолошки *потенцијал спремности* (енг. *readiness potential*). Овај потенцијал је негативни

[2] За дискусију вид.: Lowe (2000).

помак у електричном потенцијалу мозга који се може забележити пре самопокренутог вољног моторног акта (Libet 1985: 529). Експериментална открића су довела Либета да закључка да вољни акти бивају иницирани несвесним можданим процесима пре него што се јави свесна намера да се дела. Ипак, свесна намера и даље има контролу над вољним актом. Свесна контрола нема могућност да покрене вољни процес, већ да само омогући или забрани (стави „вето“ на) коначни покрет. Ова открића изазвала су сумњу у либертаријанску слободну вољу.

Меле (2007) критикује Либетове закључке о слободној вољи који долазе из истраживања (Libet 1985, 2001). Подаци добијени тим експериментима су корисни, али су интерпретације проблематичне. Хагард примећује да би „концептуална анализа била од помоћи“ (Haggard, Libet 2001). Либет не успева да увиди да се мора начинити концептуална дистинкција између *одлуке* и *намере* (*deciding and intending*) с једне стране и мотивационих стања као што је *жеља* (*urge*), *хтење* (*wanting*) с друге стране (Mele 2007: 241). Либет користи термине *интенција*, *жеља*, *одлука* синонимно. Ако нешто хоћемо, то не значи да смо се одлучили да то и учинимо. Када имамо намеру да нешто урадимо одлука је већ донесена, мада се може променити. Иде се од жеље, хтења ка намери и самом делању. Ови имају различите функционалне улоге у интенционалном понашању (Mele 1992). Одлучивање (*deciding*) јесте тренутно и разликује се од намеравања (*deliberating*). Интенције и одлуке могу бити *дисталне* и *проксималне* (Меле 1992: 143–144, 158).

Да је Либет разликовао интенцију од жеље (*intending and wanting/urge*) он би пре повезао потенцијал спремности са овим другим (Mele 2007: 245). Аутори (Libet, Gleason, et al. 1983: 640) нису показали да мозак започиње неку акцију пре него што постоји субјективна свест о тој акцији. За ово време, од појаве грубо проксималне жеље ((*roughly*) *proximal urge**) у тренутку -550 msec до времена W, субјект може свесно покренути проксималну интенцију да се покрене рука, када то жели да учини намерно. Либет тврди да се ова интенција или одлука јавља у -550 msec. Либет сматра да до волунтарног акта доводи финални „делати сада“ процес (Libet 2001: 61). Овај процес се започиње несвесно. Међутим, када су процеси у питању, тешко је рећи када тачно почињу, где је тачно почетак неког процеса. Слободна воља и не мора да почиње (производи) ту прву жељу (*roughly proximal urge*) за покретом.

Слободна воља се обично посматра као одлучивање или бирање. Интенција се може *касније* укључити у волунтарни процес и ставити „вето“ на почетну жељу. И сам Либет каже: „свесна вољна контрола може радити тако . . . да селектује и контролише [„вољни процес“], или тако што дозволи (permiting) или активира (triggering) коначни моторни исход (final motor outcome) несвесно иницираног процеса или стављајући вето на прогресију ка актуелној моторној активацији“ (Libet 1985: 529). Меле примећује да је окидање врста иницирања, тако да има времена за свесну одлуку да се умеша у процес који води до покрета. Код Мелеа „проксималне жеље за делањем помажу да се покрене делање тако што помажу стварање важних проксималних интенција, а формирање или

стицање ових директно покреће делање“ (Mele 2007: 254, вид.: Меле 1992: 71–77, 143–144, 168–170, 176–177, 190–191). У дуалистичком моделу те проксималне интенције биле би независне од проксималних жеља.

Не може се избећи осећај да сва неурофизолошка истраживања одговарају на погрешно питање; баве се питањем слободне воље на погрешан начин. Њихова истраживања су корисна, али не за дебату о слободној вољи. Иако истраживачи, али и филозофи који интерпретирају такве експерименте тврде да они мењају наш поглед на слободну вољу или да одговарају на питање: да ли слободна воља постоји, чини се да се они уопште и не баве слободном вољом, већ питањем моторне контроле. На ово је већ скренута пажња раније; тешко је повезати термине које ти истраживачи користе када говоре о резултатима својих експеримената са слободном вољом. Под слободном вољом се обично подразумева нешто друго, што није предмет таквих испитивања.

Истраживачи и филозофи сматрају да могу објаснити слободну вољу тако што проучавају моторну контролу (Gallagher 2006: 115). Истраживања моторне контроле показују да се већина тих контролних процеса дешава на субперсоналном, несвесном нивоу и нема ничег чудног у томе. Зашто и како бисмо уопште били у стању да пратимо све своје покрете? Због тога моторна контрола и постоји на несвесном нивоу, да наша свест не би била „пренатрпана“ непотребним задацима. Ако бисмо морали да имамо сталну репрезентацију другог реда својих телесних покрета, да свесно рефлектујемо о тим покретима, морали бисмо да обављамо покрете много спорије и не бисмо успели да на прави начин (и у правом тренутку) у многим ситуацијама реагујемо. Ако овако посматрамо Либетове и сличне резултате истраживања увидећемо да они не указују на нешто необично, ништа што изискује да на посебан и нов начин буде објашњено. Они нам не могу рећи ништа о слободној вољи (Gallagher 2006: 116).

Међутим, слични проблеми постоје и у оквиру филозофских расправа о слободној вољи. Аутори често своје теорије слободне воље поткрепљују примерима телесних покрета. Постављање питања о слободној вољи у оквирима моторних контролних процеса је погрешно јер је слободна воља дуготрајни феномен (не мора се тицати временских размака од неколико стотина милисекунди) и слобода се примарно не примењује на телесне покрете који *чине* интенционално делање, већ на само интенционално делање које се описује на највишем прагматичком нивоу (Gallagher 2005). У овоме „преломну“ улогу има свест. Временски оквир за испољавање слободне воље је онолики колико је минимално потребно за свесну рефлексију процеса. Такође, разлоге за делање не треба одбацивати на основу примитивних схватања каузалности као сударања билијарских кугли. Ако каузалност у расправи о слободној вољи схватимо тако уско (као детерминистички механизам), онда „питање о слободној вољи и није о каузалности“ (Gallagher 2006: 119).

4. Антиреалистички приступ

У поглављу о менталном животу робота, Грахек анализира становиште Хилари Патнама, који одбацује питање „да ли роботи имају ментална стања“, као погрешно постављено. По Патнаму, нема никаквог „открића“ које би нам указало да роботи немају ментална стања, ради се само о томе „да ли ми желимо да роботе укључимо у нашу језичку заједницу“ (Грахек 1990: 179). Моралност постоји једино у заједници. До сада се подразумевало да је у питању заједница људских бића. Ова заједница ће можда бити проширена. Но, чини се да је потребно испитати да ли су ти нови чланови „на истом нивоу“ са људима да би били укључени у заједницу. Тако се враћамо на метафизичко питање о статусу менталних стања робота.

Патнам заузима становиште *перспективе трећег лица*, спољне опсервације која индиректно долази до закључака о унутрашњем менталном животу. Овакво становиште увек ће бити редуктивно. Међутим, Патнам није редукциониста. Његово становиште је *антиреалистичко*. Патнам сматра да не постоје интринсична својства која нам говоре о томе какав је статус менталних стања робота, већ да ми њима *приписујемо* та стања тако што описујемо њихово понашање менталним предикатима и тиме их укључујемо у нашу језичку заједницу. Грахек је исправно приметио да би се таква стратегија могла лако пренети на говор о људској свести, те не постоји ништа што нас спречава да кажемо како другим људима само приписујемо ментална стања. По Крипкеу ово је стратегија свих антиреалистичких доктрина. У њој се врши инверзија кондиционалног става. Ово врло добро објашњава неке одговоре на које наилазимо у етичкој расправи о моралном статусу компјутерских система. Антиреалист тако може тврдити „да не осуђује неки поступак зато што је неморалан, већ је тај поступак неморалан зато што га ми осуђујемо“ (Грахек 1990: 182).

4.1 „Компјутери у друштву“

Џонсонова себе сврстава у групу аутора који се баве проблемом *компјутера у друштву*, дакле на који начин независни компјутерски системи мењају наше друштво и какав морални утицај они могу имати на људски живот. Такав приступ разматра и улогу творца, дизајнера аутономних машина, узима у обзир њихове циљеве и њихово делање преко ових компјутерских система. Компјутерски системи су *сурогат* делатници, по Џонсоновој, њихово делање одређују дизајнери. Ова група не жели да припише делање машинама, јер би се тада изгубили из вида дизајнери и њихова моралност. Они хоће да укажу на то да притисак да се машине што пре учине делатницима, скреће пажњу са озбиљнијих етичких проблема, и истичу да се моралност машина мора посматрати преко моралноцти њихових дизајнера.

Насупрот њима је група аутора, које Џонсонова назива *Компјутерским моделарима* и ту спадају Флориди и Сандерс. Ова група аутора, жели да учини и компјутерске системе делатницима, због тога што сматрају да је компјутерски

модел универзална матрица природе. Због тога за њих нема много проблема да и роботе и компјутерске системе уврсте у делатнике, јер су парадигматични примери таквог модела. Аутори ове групе већ посматрају неке компјутерске системе као делатнике, као да њих не стварају људи са одређеним циљевима. Џонсонова критикује становиште компјутерских моделара. Она исправно указује на то да Флориди и Сандерс „мешају" различите нивое апстрактности да би постигли моралност компјутерских система. На једном нивоу апстрактности они можда делују аутономно, али није тако и на другим нивоима; Флориди и Сандерс чине тај незаконити скок међу нивоима (Floridi, Sanders 2004). Такође, она са правом примећује да Флориди и Сандерс мешају научне моделе и операционалне системе (Johnson, Miller 2008: 128). Научни модел се проверава упоређивањем са стварношћу. Машине су операционални системи, створени да би вршили одређени задатак (операцију) уместо људи; они не морају испољавати понашање изоморфно људском, јер то није сврха њиховог дизајна.

Џонсонова скреће пажњу на чињеницу да само свесном људском одлуком, за неки операционални систем можемо рећи да се понаша „попут" људи, и то у врло ограниченом смислу; да замењује човека у неким задацима. Али из овога се не може извући закључак да се машине понашају попут људи. Примедба Џонсонове Флоридију и Сандерсу је у томе да се машине не налазе у истом односу са људима, као научни модели са стварношћу. Машине су операционални системи који врше одређене задатке уместо људи и то чине на другачији начин од људи. Приступ друге групе као да остаје на површини, не разумевајући делање компјутерских система. Коначно, по Џонсоновој, „моделари" не увиђају да је одлука о моралном статусу робота конвенционална. Они, такође стоје на позицији технолошког детерминизма, да се технологија развија нужном прогресијом, што је такође, врло проблематично.

5. Моралне машине

Савремена оцена компјутерских и роботских система као аутономних је претерана и неумесна, и, најчешће, филозофски непромишљена. Могуће су три групе одговора на наше питање. (1) Елиминативни физикалисти не виде никакав проблем у постојању свесних, аутономних робота, све док се они конструишу тако да буду изоморфни људским бићима. (2) Филозофи другачијих метафизичких становишта, емергентисти и нередукционисти, имају проблема са могућношћу конструисања оваквих машина. Ово су реалистички одговори. (3) Антиреалистичке одговоре дају многи савремени дизајнери робота; они сматрају да ће људи одлучити о томе да ли да прихвате машине као делатнике.

Џонсонова сматра да ово није „чињенично-појмовно" питање. Ја мислим да оно то свакако мора бити, али док се коначно не одговори на ово питање, упутно је прихватити решења Џонсонове. Ово значи да компјутерске системе не смемо посматрати као делатнике. Они јесу морални ентитети, али не и делатници. Роботи се могу програмирати да чине добро, могу бити деонтолош-

ки програмирани да не повређују људе,[3] али то није делање. Они не одлучују својом вољом.

Роботи су информатичке машине. Али они само обрађују податке које им задају људи, не разумеју смисао тих информација. Неки аутори (Stahl 2004) указујући на то да компјутерски системи процесирају податке, али не разумеју значење информација које процесирају. Из овог разлога компјутерски системи не пролазе „Морални Туринговтест" и не могу бити делатници. Ја сам проблему моралности машина приступио из менталних стања. Овај приступ јасније показује да се ни у *принципу* не могу изградити вештачки делатници. Сем тога, потребно је задовољити још нужних услова, сем разумевања значења информација, да би машине постале делатници.

Творци аутономних робота морају преиспитати своје филозофске позиције. Да употребим израз Инмана Харвија они „се баве филозофијом духа помоћу шрафцигера" (Harvey 2002: 205). Овај истраживач вештачке интелигенције сматра да је Чалмерсов „тешки" проблем свести, који кочи конструкцију аутономних робота и вештачке интелигенције, непостојећи проблем, који нестаје у лингвистици; да се решава језички. Усвајањем бихејвиористичке перспективе трећег лица, по Харвију проблеми људске (и роботске) свести нестају. Роботи ће бити делатници уколико их ми укључимо у језичку заједницу. Овакав релативистички и антиреалистички приступ чест је у литератури о конструисању аутономних робота и најчешће га заузимају творци нових технологија. Научници имају највише проблема са субјективним карактером свести који је присутан у дефиницијама свести (Harvey 2002: 200). Квалије се не могу програмирати, и не уклапају се у „компјутерски" модел свести.

Стратегија групе *Компјутери-у-друштву* добро критикује програм супротне групе. Указује на њихово преурањено и филозофски нелегитимно приписивање моралне одговорности компјутерским системима. Указују и на проблем моралности дизајнера компјутерских система, статус машина као моралних *ентитета*, који не могу бити искључени из моралне игре у људском друштву, али који се за сада, не могу сматрати делатницима попут људских бића. Људи се могу договорити да компјутерске системе посматрају као делатнике. Ово је ствар *практичне* одлуке и проблематично је колико нам је заиста потребна оваква моралност машина. Сматрам релевантним филозофско питање да ли компјутерски системи *теоријски* могу постати делатници и залажем се за реалистички приступ проблему. Проблеми менталне каузалности и слободне воље су тешки метафизички проблеми за које немамо решења ни када је људско делање у питању. У сваком случају, питање о моралности компјутерских система јесте чињеничко-појмовно питање и на њега се мора дати реалистички одговор.

[3] Најчешћи је пример „подизања руке".

Литература

Chalmers, David J. (1996) *The Conscious Mind: In Search of a Fundamental Theory,* Oxford University Press.

Floridi L., Sanders, J. W. (2004) „*On the Morality of Artificial Agents*“, Minds and Machines 14 (3): 349–379.

Gallagher, S. (2005) *How the Body Shapes the Mind*. Oxford University Press.

Gallagher, S. (2006) „*Where's the Action? Epiphenomenalism and the Problem of Free Will*“ u Susan Pockett, William P. Banks & Shaun Gallagher (eds.), *Does Consciousness Cause Behavior?* MIT Press.

Gibb, S. C., (2010) „*Closure Principles and the Laws of Conservation of Energy and Momentum*“, Dialectica 64 (3): 363–384.

Grahek, N., (1990) *Materija, svest i saznanje*, FDS, Beograd.

Grodzinsky, F. S., Miller, K. W., Wolf, M. J. (2008) „*The ethics of designing artificial agents*“, Ethics and Information Technology 10: 115–121.

Haggard, P., and B. Libet. (2001) ‘‘*Conscious Intention and Brain Activity*’’ Journal of Consciousness Studies 8: 47–63.

Harvey, I., (2002) *„Evolving robot consciousness: The easy problems and the rest“* u James H. Fetzer (ed.), *Consciousness Evolving*, John Benjamins Publishing Company, Amsterdam/Philadelphia.

Johnson, D. G. (2006) „*Computer Systems: Moral Entities but Not Moral Agents*“, Ethics and Information Technology 8 (4).

Johnson, D. G., Miller, K. W. (2008) „*Un-making artificial moral agents*“, Ethics and Information Technology, September 2008, Volume 10, 2–3, pp 123–133.

Kim, J. (1993) *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press.

Kim, J. (2005) *Physicalism, or Something Near Enough*. Princeton University Press.

Libet, B., C. Gleason, E. Wright, and D. Pearl. (1983) ‘‘*Time of Unconscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential)*’’ Brain 106: 623–642.

Libet, B. (1985) ‘‘*Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action*’’ Behavioral and Brain Sciences 8: 529–566.

Libet, B. (2001) ‘‘*Consciousness, Free Action, and the Brain*’’ Journal of Consciousness Studies 8: 59–65.

Lowe, E. J. (2000) „*Causal Closure Principles and Emergentism*“ Philosophy 75 (294): 571–586.

Lowe, E. J. (2008) *Personal Agency, The Metaphysics of Mind and Action*, Oxfor: Oxford University Press.

Mele, A. (1992) *Springs of Action: Understanding Intentional Behavior*. New York: Oxford University Press.

Mele, A. (2007) *„Decisions, Intentions, Urges, and Free Will: Why Libet Has Not Shown What He Says He Has“* u J. K. Campbell, M. O'Rourke & H. S. Silverstein (eds.) *Causation and Explanation.* MIT Press.

Stahl, B. C. (2004) „*Information, Ethics, and Computers: The Problem of Autonomous Moral Agents*“, Minds and Machines 14 (1): 67–83.

Janko Nešić, Beograd

COMPUTER SYSTEMS AS MORAL AGENTS

Abstract. Are computer systems, as artificial entities, moral agents and if not, could they ever become moral agents in the future? Behavior of computer systems doesn't meet the requirements of the standard account of moral agency. Machines don't have mental states and problems of mental causality and free will arise. Realistic and antirealistic accounts are possible. I will argue for a realistic account of moral agency of machines.

Key words: computer system, moral agent, mental states, free will.