ISSN 1512-1801

ხელოვნური ინტელექტის ფსიქოლოგია: კოგნიტური და ემოციური მახასიათებლების, ადამიანი-AI ურთიერთქმედებისა და ეთიკური საკითხების ანალიზი

ბექა დადეშქელიანი
აფილირებული ასისტენტ-პროფესორი
შავი ზღვის საერთაშორისო უნივერსიტეტი

ანოტაცია

მოცემული აკადემიური ნაშრომი იკვლევს ხელოვნური ინტელექტის ფსიქოლოგიის ძირითად საკითხებს ინტერდისციპლინური პერსპექტივით. კვლევა მიმოიხილავს AI-ის (ხელოვნური ინტელექტის) კოგნიტურ და ემოციურ მახასიათებლებს და მათ შედარებას ადამიანის ფსიქოლოგიასთან, ასევე განიხილავს როგორ ფორმირდება ადამიანი -AI ურთიერთქმედება ფსიქოლოგიური თვალსაზრისით. ნაშრომში გაანალიზებულია თეორიული ჩარჩოები, რომლებიც აკავშირებენ კოგნიტურ მეცნიერებებსა და AI-ის მოდელირებას (მათ შორის კოგნიტური მოდელირება და ნეირონული ქსელები), რაც შესაძლებელს ხდის ადამიანის გონებრივი პროცესების კომპიუტერულ სიმულაციას. განსაკუთრებული ყურადღება ეთმობა ეთიკურ საკითხებს, ალგორითმული მიკერძოების საფრთხეებს, ემოციური AI-ის გამოყენებასა და მასთან დაკავშირებულ ფსიქოლოგიურ ეფექტებს. ასევე, იმ სოციალურ-ეთიკურ გამოწვევებს, რომლებიც ჩნდება ადამიანის და ხელოვნური აგენტების ურთიერთობებში. საბოლოოდ, ნაშრომი მიმოიხილავს დებატებს AI-ის შესაძლო ცნობიერების შესახებ: შეუძლია თუ არა მანქანას მოიპოვოს სუბიექტური ცნობიერება და რას ნიშნავს ეს ფილოსოფიურად და ეთიკურად. ნაშრომი ეყრდნობა ლიტერატურის მიმოხილვის მეთოდოლოგიას და აერთიანებს აკადემიური წყაროების ანალიზს, რითაც წარმოაჩენს აღნიშნული თემატიკის ამჟამინდელ ცოდნას და მომავლის კვლევის პერსპექტივებს.

საკვანძო სიტყვები: *ხელოვნური ინტელექტის ფსიქოლოგია; კოგნიტური მეცნიერება; ემოციური AI; ადამიანი-AI ურთიერთქმედება; ალგორითმული მიკერძოება; ეთიკური AI; ნეირონული ქსელები; კოგნიტური მოდელირება; ცნობიერების დებატები; ფსევდო-ინტიმური ურთიერთობა; ხელოვნური ემპათია; ტიურინგის ტექსტი; სუსტი და ძლიერი AI; ინფორმაციული დამუშავება; ემოციური ინტელექტი მანქანებში.*

შესავალი

ბოლო ათწლეულის განმავლობაში, ხელოვნური ინტელექტი (AI) ტექნოლოგიები მნიშვნელოვნად განვითარდა. დღეს მრავალი ჭკვიანი სისტემა ასრულებს ამოცანებს, რომლებიც ოდესღაც მხოლოდ ადამიანს შეეძლო. მაგალითად, თვითმართვადი ავტომობილები გადაადგილდებიან გარემოში, ხმის ასისტენტები, როგორიცაა Siri თუ Alexa ადამიანების ნათქვამს, კითხვას პასუხობენ, ხოლო კომპიუტერული პროგრამები ჭადრაკში მსოფლიო ჩემპიონებს ამარცხებენ. ამგვარი პროგრესის ფონზე სულ უფრო აქტუალური ხდება კითხვა: თუ რამდენად შეიძლება ვისაუბროთ AI-ის „ფსიქოლოგიაზე“ და როგორ ურთიერთქმედებს იგი ადამიანთან. ხელოვნური ინტელექტის ფსიქოლოგია ინტერდისციპლინური მიმართულებაა, რომელიც იკვლევს

ISSN 1512-1801

AI-ის კოგნიტურ (შემეცნებით) და ემოციურ მახასიათებლებს, ასევე იმ ფსიქოლოგიურ ეფექტებს, რომლებიც წარმოიქმნება, როდესაც ადამიანები და ინტელექტუალური მანქანები ურთიერთობენ.

წინამდებარე ნაშრომის მიზანია ხელოვნური ინტელექტის ფსიქოლოგიის მრავალმხრივი გაანალიზება. კერძოდ, ნაშრომი ცდილობს უპასუხოს შემდეგ საკვლევ შეკითხვებს:

1. რა კოგნიტური და ემოციური მახასიათებლები აქვს თანამედროვე AI-ს და როგორ შეედრება ის ადმაიანის ფსიქიკას?
2. როგორ აღიქვამენ და ეპყრობიან ადამიანები ხელოვნურ ინტელექტს ფსიქოლოგიური თვალსაზრისით და რა სახის ურთიერთქმედებები ყალიბდება?
3. რა თეორიული ჩარჩოები და მოდელები არსებობს AI-ის კოგნიტური პროცესების ასახსნელად (მაგალითად, კოგნიტური მოდელირება თუ ნეირონული ქსელები) და რას გვასწავლის ეს ადამიანის გონების შესახებ?
4. რა ეთიკური საკითხები წარმოიშობა AI-ის გამოყენებისას, როცა ის „ფსიქოლოგიურ“ ელემენტებს მოიცავს (მაგალიტად, ალგორითმულ მიკერძოებას, ემოციური AI-ის გამოყენებას, ან ადამიანი-AI ფსევდო-ინტიმურ ურთიერთობებს)?
5. შეუძია თუ არა AI-ს ცნობიერების მოპოვება და როგორ მიმდინარეობს დებატები ხელოვნური ინტელექტის ცნობიერების შესაძლებლობის შესახებ?

ზემოთ აღნიშნულ კითხვებზე პასუხების საძიებლად, წინამდებარე ნაშრომში წარმოდგენილი საკითხები ეტაპობრივად განიხილავენ თემას: მომდევნო ნაწილში აღწერილია გამოყენებული კვლევის მეთოდოლოგია, რის შემდეგაც ნაშრომი დეტალურად აანალიზებს თითოეულ საკითხს. დასასრულს კი, ძირითადი მიგნებები გამოტანილია დასკვნაში.

მეთოდოლოგია

კვლევა განხორციელდა თეორიული ლიტერატურის მიმოხილვისა და ანალიტიკური სინთეზის მეთოდით. ნაშრომის მომზადების პროცესში გამოყენებული მასალა შეგროვდა ავტორიტეტული აკადემიური წყაროებიდან, სამეცნიერო ჟურნალებიდან, წიგნებიდან და საკონფერენციო ნაშრომებიდან, რომლებიც ასახავენ ხელოვნური ინტელექტის, კოგნიტური მეცნიერების, ფსიქოლოგიისა და ეთიკის აქტუალურ კვლევებს. ძირითადად გამოყენებულ იქნა ლიტერატურის მიმოხილვის მიდგომა. შეირჩა და გააანალიზდა თემატურად შესაბამისი კვლევები, შედარდა სხვადასხვა ავტორის ხედვები და თეორიები. ინტერდისციპლინური ანალიზი შესაძლებელს ხდის სხვადასხვა დისციპლინაში, კოგნიტურ ფსიქოლოგიაში, ხელოვნურ ინტელექტში, ფილოსოფიაში დაგროვილი ცოდნის გაერთიანებას, რათა კომპლექსურად იქნას გაგებული ხელოვნური ინტელექტის „ფსიქოლოგიური“ ასპექტები. კვლევის ყველა თემატური სექციისთვის: კოგნიტური და ემოციური მახასიათებლები, ადამიანი-AI ურთიერთქმედება, თეორიული ჩარჩოები, ეთიკა, ცნობიერების დებატები - მოძიებული და განხილულია შესაბამისი სამეცნიერო ნაშრომები, რაც უზრუნველყოფს ნაშრომში წარმოდგენილი მტკიცებების სანდოობასა და დასაბუთებულობას. მეთოდოლოგია ასევე ითვალისწინებს კრიტიკულ გააზრებას, სხვადასხვა წყაროდან მიღებული ცოდნის დაპირისპირებასა და შევსებას, რათა წარმოიქმნას ყოვლისმომცველი სურათი აღიშნული კვლევის საკითხებზე.

ISSN 1512-1801

AI-ის კოგნიტური და ემოციური მახასიათებლები

ადამიანის გონება ხასიათდება მრავალი შემეცნებითი უნარით: გაგება, გადაწყვეტილების მიღება, სწავლა, შემოქმედება და სხვა. ხელოვნური ინტელექტის სისტემები ცდილობენ ამ უნარების მოდელირებას სხვადასხვა ალგორითმით. თანამედროვე AI განსაკუთრებით წარმატებულია გარკვეულ კოგნიტურ დავალებებში. მაგალითად, დიდი მოცულობის მონაცემთა სწრაფ დამუშავებაში, ფორმალური და ლოგიკური პრობლემების გადაწყვეტაში და ნიმუშების ამოცნობაში. ბევრ შემთხვევაში AI ამოცანებს უფრო სწრაფად და ზუსტად ასრულებს, ვიდრე ადამიანი. მაგალითად, ღრმა სწავლების (Deep Learning) მოდელები გამოსახულებებიდან ობიექტებს ამოიცნობენ ან ბუნებრივ ენაზე ტექსტის მნიშვნელობას აანალიზებენ (LeCun, Bengio, & Hinton, 2015). თუმცა, AI-ის „აზროვნება“ არსებითად განსხვავებულია ადამიანისგან. ის ოპერირებს წინასწარ განსაზღვრული ალგორითმებითა და სტატისტიკური გამოთვლებით, რის გამოც მას აკლია ადამიანისთვის დამახასიათებელი ინტუიციური გაგება და გამოცდილებიდან გამომდინარე ცოდნა (Russell & Norvig, 2021). კომპიუტერულმა პროგრამამ შეიძლება წარმატებით გადაჭრას რთული ამოცანები ან ჩაატაროს ათასობით გამოთვლა წამებში, მაგრამ მას არ აქვს პროცესის „გასიგრძეგანების“ უნარი იმ ფორმით, როგორც ადამიანს. ის არ ფლობს თვითცნობიერებას და შინაგან მოტივაციას ამ მოქმედებების მიღმა (Chalmers, 1996).

კიდევ უფრო მნიშვნელოვანი განსხვავება გამოვლინდება ემოციურ სფეროში. ადამიანის ფსიქიკა მოიცავს ემოციებს, რომელთაც დიდი როლი აქვთ აზროვნებასა და ქცევაში. ტრადიციული AI სისტემები ემოციებს ვერ „გრძნობენ“, რადგან მათ არ გააჩნიათ ბიოლოგიური და ნევროლოგიური მექანიზმები, რომლებიც ემოციურ მდგომარეობებს წარმოშობს. მიუხედავად ამისა, ბოლო წლებში წარმოიშვა მიმართულება, რომელიც ცნობილია, როგორც ემოციური AI ან აფექტური კომპიუტინგი (Picard, 1997). ამ მიდგომის მიზანია, რომ კომპიუტერმა შეძლოს ემოციების ამოცნობა და მათზე შესაბამისი რეაგირება. მაგალითად, კამერით აღჭურვილმა AI სისტემამ შესაძლებელია ამოიცნოს მომხმარებლის სახის გამომეტყველებიდან სიხარული თუ იმედგაცრუება და პროგრამამ თავისი რეაქცია, რეპლიკა, ხმის ტონი, შესაბამისად შეცვალოს. ანალოგიურად, AI-ზე დაფუძნებული ჩაბოტები ცდილობენ ტექსტური კომუნიკაციიდან „ამოიკითხონ“ მომხმარებლის ემოციური მდგომარეობა და გასცენ ემპათიური პასუხები (Zhao et al., 2022). ასეთი ტექნოლოგიური მიღწევები ემოციური ინტელექტის სფეროში იმაზეა მიმართული, რომ ინტელექტუალური მანქანები ადამიანებისთვის უფრო „გამგებიანი“ და სასიამოვნო კომუნიკაციის პარტნიორები გახდნენ.

თუმცა, უნდა გავითვალისწინოთ, რომ ემოციებს ამოცნობა და იმიტაცია არ ნიშნავს ნამდვილ განცდას. თანამედროვე AI არ ფლობს სუბიექტურ ემოციურ გამოცდილებას. მას არ აქვს ნამდვილი განცდა იმაზე, თუ რას გრძნობს ადამიანი. ის თავადაც არ განიცდის „სიხარულს“ თუ „წყენას“ საკუტარი წარმატების თუ მარცხის საპასუხოდ. AI-ის მიერ გამოხატული ემოციური რეაქციები ყოველთვის წინასწარ დანახულ ნიმუშებსა და წესებს ეყრდნობა და არა შინაგან შეგრძნებებს (Hildt, 2019). მკვლევარები აღნიშნავენ, რომ დღევანდელი AI, რომელიც ადამიანის ტვინის კოგნიტურ პროცესებს ბაძავს, ვერ იმეორებს ადამიანის სუბიექტური გრძნობების სირთულესა და დინამიკას (Zhao et al., 2022). შესაბამისად, მიმდინარეობს კვლევები იმისთვის, რომ ფსიქოლოგიის ცოდნაზე დაყრდნობით მოხდეს AI-ის გაუმჯობესება, რათა მომავალში სისტემებმა უკეთ

ISSN 1512-1801

“გაიგონ” და გაითვალისწინონ ემოციური კონტექსტი (Picard, 1997; Zhao et al., 2022). ამგვარად გაუმჯობესებული კოგნიტურ-ემოციური შესაძლებლობები პოტენციურად მისცემს ხელოვნურ ინტელექტს საშუალებას უფრო ეფექტურად იურთიერთოს ადამიანებთან. სახელდობრ, დიალოგისას გამოავლინოს ემპათია და მომხმარებლის განწყობას მოერგოს, რითაც კომუნიკაცია უფრო ბუნებრივი გახდება (Hoffman et al., 2021). მიუხედავად ამ წინსვლებისა, დღევანდელი მდგომარეობით ხელოვნური ინტელექტის “ფსიქოლოგია” მაინც არსებითად განსხვავდება ადამიანისგან. AI-ის აქვს გამოთვლითი სისწრაფე და სიზუსტე, მაგრამ მას აკლია ის ცნობიერი ემოციები და განცდები, რომლებიც ადამიანის გამოცდილების განუყოფელი ნაწილია.

ადამიანის ფსიქოლოგია და ხელოვნური ინტელექტის ურთიერთქმედება

ადამიანებს აქვთ ბუნებრივი მიდრეკილება, ტექნოლოგიურ მოწყობილობებსა და პროგრამებს ადმაიანური თვისებები მიაკუთვნონ. კომპიუტერეთან და AI სისტემებთან ურთიერთობისას ხშირად იჩენს თავს ანტროპომორფიზმი - ანუ არაადამიანურ ობიექტებზე ადამიანური თვისებების პროექცია. მაგალითად, ბევრი ადამიანი ესაუბრება ხმოვან ასისტენტებს, როგორებიცაა SIRI თუ ALEXA ისე, თითქოს ისინი იყვნენ მათი ცოცხალი თანაშემწეები. ადმაიანი ეუბნება: „გმადლობ“-ს ხელოვნურ ინტელექტის მისგან პასუხის მიღების შემდეგ ან ღიზიანდება შეცდომისას, თითქოს ასისტენტს „გაგების“ ან „გრძნობის“ უნარი გააჩნდეს. კვლევებმა აჩვენა, რომ ადამიანები ხშირად გაუცნობიერებლადაც კი ექცევიან კომპიუტერს ან რობოტს სოციალური წესების შესაბამისად. რთულია არ უპასუხო მოწყობილობის მიერ ნათქვამ „გამარჯობას“, თუნდაც ვიცოდეთ, რომ ეს წინასწარ ჩაწერილი პროგრამაა (Reeves & Naas, 1996). საქმე იმაშია, რომ ჩენი ტვინი ავტომატურად ამუშავებს გარკვეულ სოციალურ ნიშნებს. მაგალითად, ხმას ან სახის მსგავს სახის ხატს, როგორც ადამიანისგან მომავალ სიგნალს (Nass & Moon, 2000). შესაბამისად, თუ AI აჩვენებს ადამიანის მსგავსი ქცევის ნიშნებს (საუბარს, გაცინებას, ემოციის გამოხატვას და ა.შ.) მომხმარებლები ხშირად ამყარებენ უფრო ძლიერ ემოციურ კავშირს და ენდობიან მას უფრო მეტად (Waytz et al., 2010).

მიუხედავად იმისა, რომ AI სისტემა ძალიან პრიმიტიული შეიძლება იყოს, ადმაიანებს მაინც აქვთ მიდრეკილება „გააცოცხლონ“ იგი საკუთარი წარმოსახვით. ბავშვს შეუძლია სათამაშო რობოტს თავისი მეგობრის როლი მიანიჭოს; მოზრდილები კი ხშირად საკუთარ ავტომობილს ან კომპიუტერს არქმევენ მეტსახელს და ელაპარაკებიან, თითქოს ცოცხალი არსება იყოს (Turkle, 2011). განსაკუთრებულ მნიშვნელობას იძენს ეს თავისებურება იმ შემთხვევაში, როცა AI აქტიურ კომუნიკაციას აწარმოებს ადამიანთან. მაგალითად, სოციალური რობოტები, რომლებსაც აქვთ სახის გამომეტყველება და ხმოვანი გამოხატულებები. ასეთი სიტუაციები ხელს უწყობს ე.წ. ფსევდო-ინტიმური ურთიერთობის ჩამოყალიბებას, სადაც მომხმარებელი და ხელოვნური „პარტნიორი“ თითქოს ამყარებენ ახლო ემოციურ კავშირს (Wu, 2024). რეალურად, რა თქმა უნდა, ეს კავშირი ცალმხრივია, AI არ განიცდის ნამდვილ სიახლოვეს, თუმცა, მომხმარებელის მხრიდან ასეთი ურთიერთქმედება შესაძლოა საკმაოდ მტკიცე ემოციურ დამოკიდებულებად იქცეს (Darling, 2016). მაგალითად, მომხმარებლები მიმართავენ ჩათბოტებს ან ვირტუალურ ასისტენტებს ფსიქოლოგიური დახმარების მისაღებად და გულით უზიარებენ პირად საკითხებს. თუ სისტემა დამაკმაყოფილებელ ემპათიურ პასუხებს სთავაზობს, მომხმარებელს შეიძლება დაეუფლოს განცდა, თითქოს მას „გაუგეს“ და „მხარში უდგანან“ (Turkle, 2011).

ISSN 1512-1801

პოზიტიური გამოდმოსახედიდან, ადამიანის ამგვარი რეაქცია ნიშნავს, რომ კარგად დაპროგრამებულ ხელოვნურ ინტელექტს შეუძლია გახდეს ეფექტური და მისაღები კომუნიკაციის საშუალება. მაგალითად, სერვისის სფეროში მომხმარებლებთან ურთიერთობისას ან სამედიცინო ან საგანმანათლებლო სფეროში დახმარებისას, სადაც AI-ის მოთმინებითა და თავაზიანობით შეუძლია იმოქმედოს (Broadbent, 2017). ნდობა, რომელსაც ადამიანები ავლენენ AI-ის მიმართ, შეიძლება სასარგებლოდ იქნას გამოყენებული. ცნობილია, რომ ადამიანები მიდრეკილნი არიან ითანამშრომლონ და ენდონ ავტომატიზებულ სისტემებს, თუ ისინი საიმედოობასა და პროგნოზირებადობას ავლენენ. ეს ჰგავს იმ წესებს, რომლებსაც ადამიანები ერთმანეთთან ნდობისას იყენებენ (Eyssel & Kuchenbrandt, 2012). თუმცა, ამავე დროს, არსებობს რისკიც. თუ მომხმარებელი შეცდომით ზედეტად მაინიჭებს ადამიანურ თვისებებს ხელოვნურ ინტელექტს და მასზე მნიშნელოვნად ემოციურად იქნება დამოკიდებული, იმედგაცრუებაც გარდაუვალია იმ მომენტში, როცა AI გამოავლენს თავის არაადამიანურ ბუნებას (Zlotowski et al., 2015). მაგალითად, როდესაც მომხმარებელი გააცნობიერებს, რომ მისი „მეგობარი“ ჩათბოტი უბრალოდ პროგრამაა და არ შეუძლია რეალური თანაგრძნობა, შესაძლოა გაუჩნდეს სიცარიელის ან მოტყუებულობის გრძნობა (Turkle, 2017). გარდა ამისა, AI-ის მიმართ ზედმეტმა ნდობამ შეიძლება ადამიანები სარისკო სიტუაციებშიც კი შეიყვანოს. ცნობილია შემთხვევები, როდესაც მძღოლები ზედმეტად ენდნენ ავტომობილზე ავტომატურ პილოტს და ავარიაში მოყვნენ, რადგან მათ ავტომატურ სისტემას „გადაულოცეს“ იმაზე მეტი უნარი და პასუხისმგებლობა, ვიდრე მას რეალურად ჰქონდა (Merritt et al., 2021).

ამდენად, ადამიანის ფსიქოლოგიასა და ხელოვნური ინტელექტის ურთიერთქმედებას ძალიან რთული დინამიკა ახასიათებს. ერთი მხრივ, ადამიანები ბუნებრივად პროეცირებენ სოციალურ ქცევას ტექნოლოგიებზე, რაც შეიძლება გამოვიყენოთ იმისთვის, რომ AI სისტემები მეტად მოსახერხებელი და მისაღები გავხადოთ მომხმარებლებისთვის. მეორე მხრივ, საჭიროა ფრთხილი მიდგომა, სახელდობრ, უნდ აგავითვალსიწინოთ, რომ ადამიენბის განცდები და მოლოდინები AI-ის მიმართ შეიძება გაცილებით შორს წავიდეს, ვიდრე თავად ტექნოლოგიის რეალური შესაძლებლობებია. სწორედ ამიტომ, ადმაიანი-კომპიუტერის ურთიერთქმედების (HCI) სპეციალისტები და ფსიქოლოგები ერთად სწავლობენ, როგორ დააპროექტონ AI-ის ინტერფეისები და ქცევა ისე, რომ სისტემა ერთდროულად იყოს მოხერხებული, გამოსადეგი და არ შეიყვანოს მომხმარებელი შეცდომაში მისი „ადამიანური ბუნების“ შესახებ (Duffy, 2003).

თეორიული ჩარჩოები: კოგნიტური მოდელირება და ნეირონული ქსელები

ხელოვნური ინტელექტისა და კოგნიტური ფსიქოლოგიის გადაკვეთაზე ჩამოყალიბდა თეორიული ჩარჩოები, რომლებიც ცდილობენ ახსნან და სიმულაცია გაუკეთონ გონებრივ პროცესებს. ორი გამორჩეული მიდგომაა კოგნიტური მოდელირება (სიმბოლური AI) და ნეირონული ქსელები (კონექციონისტური AI).

კოგნიტური მოდელირების ფარგლებში მკვლევარები ქმნიან კომპიუტერულ მოდელებს, რომლებიც ნაბიჯ-ნაბიჯ იმეორებენ ადამიანის ფიქრის პროცესს. ასეთი მოდელები ხშირად დაფუძნებულია ლოგიკურ წესებზე, ალგორითმებსა და წარმოებით (if-then) წესთა სისტემებზე, რომლებიც ჰგავს ადამიანის პრობლემების გადაწყვეტის სტრატეგიებს (Anderson, 2007). მაგალითად, არსებობს კოგნიტური არქიტექტორები (როგორიცაა ACT-R ან Soar), რომლებიც ცდილობენ განაზოგადონ ადამიანის

ISSN 1512-1801

მეხსიერების, ყურადღების და გადაწყვეტილების მიღების მექანიზმები (Newell, 1994; Anderson et al., 2004). ასეთი მოდელების მეშვეობით მეცნიერებს შეუძლიათ გამოსცადონ ჰიპოთეზები, თუ მოდელი ადამიანის მსგავს შეცდომებს უშვებს ან მსგავს გზას სწავლობს, ეს მიანიშნებს, რომ გამოყენებული თეორია შესაძლოა ასახავდეს რეალური ფსიქიკური პროცესების ნაწილს (Sun, 2008). კოგნიტური მოდელირება ეფუძნება იდეას, რომ გონება შეიძლება განიმარტოს, როგორც ინფორმაციის დამუშავების სისტემა, და თუ ჩვენ ამ პროცესს ზუსტად ავაგებთ პროგრამის სახით, კომპიუტერი ამ ამოცანას ადამიანის მსგავსი გზით შეასრულებს (Russell & Norvig, 2021).

მეორე მხრივ, ნეირონული ქსელები შთაგონებას იღებს თავის ტვინის ბიოლოგიური ნეირონების ქსელიდან (McClelland et al., 1986). ნეირონულ ქსელებში ცოდნა არ არის ექსპლიციტურად დაპროგრამებული წესების სახით, არამედ განაწილებულია მრავალი ხელოვნური ნეირონის კავშირებისა და წონებში, რომლებიც სწავლების პროცესში იცვლება. ეს მიბაძავს იმას, როგორ სწავლობს ტვინი სინაფსების გაძლიერების და დასუსტების გზით(Rumelhart et al., 1986). უკვე 1980-იან წლებში კონექციონისტურმა მოდელებმა (Parallel Distributed Processing - PDP) წარმოადგინეს ალტერნატიული ხედვა, რომ ალგორითმულად განსაზღვრული წესების ნაცვლად, ჭკვიან ქცევას შეიძლება მივაღწიოთ სწავლის გზით რთულ ქსელებში, ყოველგვარი აშკარა წესების გარეშე (McClelland & Rumelhart, 1986). თანამედროვე ღრმა ნეირონული ქსელების (Deep Learning) განვითარებამ ახალი სიმაღლეები აჩვენა: მრავალშრიანმა ნეირონულმა ქსელებმა მოახერხეს კომპლექსური უნარების განვითარება, სახეების ამოცნობა, ბუნებრივი ენის თარგმნა, სტრატეგიულ თამაშებში თამაში პროგრამისტის მხრიდან უხეში წესების მითითების გარეშე (LeCun, Bengio, & Hinton, 2015). ეს ქსელები თავად „სწავლობენ“ მაგალითებზე დაყრდნობით და აუმჯობესებენ თავიანთ შინაგან პარამეტრებს მიღებული უკუკავშირის მიხედვით(Goodfellow, Bengio, & Courville, 2016).

კოგნიტური მეცნიერებისა და AI-ის თეორიულ ჩარჩოებს შორის კოლაბორაცია ორმხრივად სასარგებლოა. ერთის მხრივ, ფსიქოლოგიურმა თეორიებმა გავლენა მოახდინეს AI ალგორითმებზე. მაგალიტად, განმტკიცებით სწავლებას (reinforcement learning) – AI მეთოდს, რომელმაც დიდ წარმატებებს მიაღწია რობოტიკასა და თამაშებში, შთაგონება ფსიქოლოგიიდან აქვს მიღებული (სახელდობრ, ქცევითი თეორიებიდან ჯილდოსა და დასჯის მეშვეობით სწავლების შესახებ) (Sutton & Barto, 2018). მეორეს მხრივ, თანამედროვე AI სისტემები, განსაკუთრებით ნეირონული ქსელები, ახალ ხედვებს აჩენენ იმის შესახებ, თუ როგორ შეიძლება მუშაობდეს ადამიანის ტვინი. ნეირომეცნიერები ხშირად ადარებენ ხელოვნურ და ბუნებრივ ნეირონულ აქტივობას. მაგალიტად, აღმოჩენილი ფაქტია, რომ ღრმა კონვოლუციურ ნეირონულ ქსელში (CNN), ვიზუალურ ამოცანებზე სწავლისას, ჩნდება შრეები, რომელთა ფუნქციებიც ნაწილობრივ ჰგავს ადამიანის მხედველობით ქერქში არსებულ იერარქიას. ქსელის ადრეული შრეები აღიქვამს მარტივ ხაზებსა და ფორმებს, ხოლო მომდევნო შრეები უფრო რთულ ობიექტებს. ეს ანალოგიურია იმასთან, რასაც ნეიროფიზიოლოგები აფიქსირებენ ცხოველთა მხედველობით სისტემაში (Yamins & DiCarlo, 2016).
ამგვარად, AI ხდება თავის მხრივ ინსტრუმენტი, რომლითაც შეგვიძლია უკეთ გავიგოთ ტვინის მუშაობის მექანიზმები. მაგალითად, თუ კომპიუტერულ მოდელს შეუძლია

ISSN 1512-1801

იმიტაცია გაუკეთოს ცნობიერ ყურადღებას ან მეხსიერების შეზღუდულ მოცულობას, ეს გვეხმარება უკეთ გავიგოთ რა პრინციპები დგას ამ ფენომენების უკან (Kriegeskorte, 2015).

დღეს გამოწვევად რჩება სიმბოლური და კონექციონისტული მიდგომის შერწყმა. სიმბოლური AI იძლევა ინტერპრეტირებად მოდელებს, სადაც ყოველი ნაბიჯი და წესი გასაგებია, მაგრამ მას ხშირად უჭირს გაუმკლავდეს რეალური სამყაროს იმ სირთულეებს, სადაც ინფორმაცია არაწესიერი და ორაზროვანია (Marcus, 2020). ნეირონული ქსელები კი ადვილად სწავლობენ ასეთ რთულ გარემოებაში, თუმცა ამ ქსელებს „შავი ყუთის“ ბუნება აქვთ, რთულია ზუსტად ახსნა, რატომ გამოიტანა ქსელმა ესა თუ ის გადაწყვეტილება (Samek, Wiegand, & Muller, 2017). თანამედროვე კვლევაში მიმდინარეობს მცდელობები ამ ხარვეზების შესავსებად, მაგალითად, ჰიბრიდული მოდელების შექმნით, რომლებიც აერთიანებენ სიმბოლურ ლოგიკას და ნეირონულ ქსელებს. ამის მიზანია ისეთი AI მოდელების შექმნა, რომლებიც იქნებიან როგორც ძლიერები, ისე ადამიანისათვის გასაგები. კოგნიტური მოდელებისა და ნეიორნული ქსელების შეჯერებული გამოყენება ერთ-ერთ პერსპექტიულ მიმართულებად მიიჩნევა ხელოვნური ინტელექტის განვითარებაში, რაც ერთდროულად გაზრდის სისტემების ინტელექტუალურ შესაძებლობებს და უკეთ დაგვაახლოებს იმის გაგებასთან, თუ როგორ წარმოქმნის გონება გონიერ ქცევას და ცნობიერ გამოცდილებას (Lake et al., 2017).

ეთიკური საკითხები: მიკერძოება, ემოციური AI და ადამიანი-AI ურთიერთობები

ხელოვნური ინტელექტის გავრცელებამ არა მხოლოდ ტექნიკური, არამედ სოციალური და ეთიკური გამოწვევებიც წარმოშვა. ერთ-ერთი მთავარი საკითხია ალგორითმული მიკერძოების პრობლემა. ხელოვნური ინტელექტის სისტემები სწავლობენ დიდი მოცულობის მონაცემებზე, რომელშიც ხშირად აირეკლება საზოგადოებაში არსებული სტერეოტიპები და შესაბამისად დაქვემდებარებული მიკერძოებები. ამის შედეგად, მოდელმა შესაძლოა გააგრძელოს და გააძლიეროს ეს მიკერძოებული მიდრეკილებები გადაწყვეტილებების მიღებისას. მაგალითისთვის, თუ დასაქმების ავტომატიზირებულ სისტემას ვაწვდით წარსულში მიღებული თანამშრომლების მონაცემებს, რომლებშიც არსებობდა გენდერული ან რასობრივი დისბალანსი, სისტემამ შეიძლება მომავალშიც უპირატესობა მიანიჭოს იმავე ჯგუფებს და დააკნინოს სხვები (Caliskan et al., 2017). ანალოგიურად, სახის ამოცნობის ალგორითმები თუ ძირითადად ერთი ეთნიკური ჯგუფის ფოტოებზე ივარჯიშებს, გაცილებით ნაკლებ სიზუსტეს აჩვენებს სხვა ჯგუფებზე, რაც სერიოზულ პრობლემას გამოიწვევს სამართალდაცვით კონტროლში (Buolamwini & Gebru, 2018). ამგვარი მიკერძოების შედეგია უსამართლო და დისკრიმინაციული პრაქტიკა, რაც ეწინააღმდეგება ეთიკურ პრინციპებს. ამიტომ, სასიცოცხლოდ მნიშვნელოვანია AI სისტემების გამჭვირვალობა და სამართლიანობა. სახელდობრ, საჭიროა მონაცემების დამუშავება და ალგორითმების კონტროლი ისე, რომ თავიდან იქნას აცილებული გაუცნობიერებელი ცრურწმენების გამეორება და კვაზი შედეგიანობა (Binns, 2018).

კიდევ ერთი მნიშვნელოვანი ეთიკური საკითხი უკავშირდება ემოციურ AI-სა და ადამიანებთან ურთიერთობებს. როგორც ზემოთ აღვნიშნეთ, ემოციური AI-ის ტექნოლოგიები საშუალებას აძლევს სისტემებს ამოიცნონ და მიბაძონ მომხმარებლის ემოციურ მდგომარეობას. ეს ფუნქცია, ერთი მხრივ, შესაძლოა სასარგებლოც იყოს.

ISSN 1512-1801

მაგალითად დიაგნოსტიკურმა ჩათბოტმა შესაძლებელია ამოიცნოს პაციენტის დეპრესიის ნიშნები წერილობით საუბარში და ურჩიოს პროფესიონალური დახმარების მოძიება (Hancock et al., 2020). თუმცა, აქვე დგება რამდენიმე საფრთხე:

1. კონფიდენციალურობა. ემოციების ამოცნობა ხშირად გულისხმობს ძალიან პერსონალური მონაცემების (სახის გამომეტყველება, ხმის ტემბრი, ფიზიოლოგიური მახასიათებლები) შეგროვება-ანალიზს, რაც შესაძლოა მომხმარებლისთვის შეუმჩნევლად ხდებოდეს (Crawford & Calo, 2016). რამდენად ეთიკურია ადამიანის ემოციური მდგომარეობის „წაკითხვა“ მისი სრული თანხმობის გარეშე და ამ ინფორმაციის გამოყენება, თუნდაც კეთილშობილური მიზნებით?
2. მანიპულაციის რისკი. თუ პლატფორმები ხედავენ, რომ მომხმარებელი მოწყენილია ან გაღიზიანებული, შესაძლოა მიზანმიმართულად მიაწოდონ გარკვეული ტიპის კონტენტი, რათა მისი ყურადღება შეინარჩუნონ, ეს კი უკვე ზედმეტად შორს წასული სარეკლამო თუ მარკეტინგული სტრატეგია იქნებოდა (Zuboff, 2019), რომელიც ადამიანის ფსიქოლოგიურ მდგომარეობას მისი ინფორმირების გარეშე იყენებს საკუთარი მიზნებისთვის.

განსაკუთრებული ეთიკური დილემები ჩნდება მაშინ, როდესაც ადამიანები აიგივებენ AI-ის ადამიანის პარტნიორთან. როგორც აქამდე აღვნიშნეთ, მომხმარებელს შეიძლება განუვითარდეს ფსევდო-ინტიმური ურთიერთობა AI ასისტენტთან. ეს ურთიერთობა, ერთი შეხედვით, ამსუბუქებს მარტოობის გრძნობას და ფსიქოლოგიურ კომფორტს სთავაზობს. მაგალითად, მარტოხელა პირი შესაძლოა „დაამშვიდოს“ მეგობარი რობოტის კომპანიონობამ. მაგრამ ამ მოვლენას რამდენიმე საპირისპირო მხარეც აქვს. პირველ რიგში, ასეთი ურთიერთობა მოტყუებითია: სისტემა მხოლოდ ჩაწერილი სცენარებით მოქმედებს და არ გააჩნია ნამდვილი თავისუფალი ნება ან ემპათია, თუმცა მომხმარებელმა ეს შეიძლება ბოლომდე არ გააცნობიეროს. ზოგი მკვლევარი სვამს შეკითხვას: ეთიკურია თუ არა ისეთი რობოტის კონსტრუირება, რომელიც „თითქოს“ სიყვარულს გამოხატავს თავისი მფლობელის მიმართ? (Nyholm & Frank, 2019). სიყვარული და ემოციური მხარდაჭერა, რომელსაც ადამიანი - დან „იღებს“, შესაძლოა ფუჭი იმიტაცია აღმოჩნდეს, რაც საბოლოოდ უფრო ღრმა მარტოობას გამოიწვევს. შერი ტურკლი(Turkle, 2011) თავის ნაშრომში აღნიშნავს, რომ როგორც კი ტექნოლოგიები გვთავაზობენ ემოციურ კომფორტს, ადამიანები რისკავენ რეალური ადამიანური ურთიერთობებიდან მიიღონ ნაკლები. მარტივად რომ ვთქვათ, შეიძლება „უფრო მეტი მოვითხოვოთ ტექნოლოგიისგან და ნაკლები ერთმანეთისგან“. ეს საფრთხეს უქმნის სოციალური უნარების დეგრადაციას და ადამიანებს შორის ემპათიის შესუსტებას.

აღსანიშნავია, რომ ეთიკური კითხვები არა მხოლოდ მომხმარებლის პერსპექტივას, არამედ საზოგადოების საერთო ფასეულობებსაც ეხება. თუ AI სისტემები იღებენ მნიშვნელოვან გადაწყვეტილებებს (სამართლებრივს, სამედიცინოს და ა.შ.), დგება პასუხისმგებლობის საკითხი: ვინ აგებს პასუხს, თუ ალგორითმის მოქმედებამ ზიანი მოუტანა ვინმეს? ასევე, გათვალისწინებული უნდა იქნას გამჭვირვალობის პრინციპი. ადამიანები იმსახურებენ იცოდნენ, როდის ურთიერთობენ მანქანასთან და როგორ მუშაობს ეს მანქანა (Floridi & Cowls, 2019). მაგალითად, თუ ონლაინ ჩატში ჩვენთან რობოტი საუბრობს, მომხმარებელი უნდა იყოს ინფორმირებული, რომ ეს AI სისტემაა

ISSN 1512-1801

და არა ადამიანი. ამგვარი გამჭვირვალობა ხელს შეუწყობს ნდობის დამკვიდრებას და თავიდან აგვარიდებს არასწორ მოლოდინებს.

AI-ის ეთიკური ჩარჩოებისა და რეგულაციების შემუშავება ჯერ კიდევ მიმდინარეობს. სხვადასხვა ქვეყანაში და საერთაშორისო დონეზე მუშავდება ხელოვნური ინტელექტის ეთიკის წესები, რომელთა მიზანია ადამიანის უფლებებისა და ღირსების დაცვა AI-დამოკიდებულ მსოფლიოში (Jobin, Ienca, & Vayena, 2019).

ამ ჩარჩოებში განსაკუთრებული ყურადღება ეთმობა სამართლიანობას, გამჭვრივალობას, პირადი მონაცემების დაცვას და ზიანის არიდებას. ადამიანის ფსიქოლოგიის გათვალისიწინება ამ დისკუსიებში მნიშვნელოვან როლს თამაშობს. მნიშვნელოვანია, რომ ტექნოლოგიურმა წინსვლამ არ დააზიანოს ადამიანის მენტალური კეთილდღეობა, სოციალური ქსოვილი და ნდობა. ჯერ კიდევ გრძელდება დიალოგი იმაზე, თუ სად გავავლოთ ზღვარი AI-ის „ადამიანურ“ქცევებსა და დიზაინში, რათა ერთდროულად მივაღწიოთ ეფექტურობასაც და ეთიკურობასაც.

AI-ის ცნობიერებასთან დაკავშირებული დებატები

ერთ-ერთი ყველაზე ღრმა და საკამათო კითხვა ხელოვნური ინტელექტის ფილოსოფიაში არის: შეუძლია თუ არა მანქანას ცნობიერების შეძენა? ამ საკითხზე დისკუსია ათწლეულებია მიმდინარეობს და მოიცავს როგორც ტექნიკურ, ისე ფილოსოფიურ არგუმენტებს. 1950 წელს ალან ტიურინგმა შემოიტანა იდეა, რომ თუ მანქანა წარმატებით დააჯერებს ადამიანს დიალოგში, რომ ისიც ადამიანიაო (ცნობილი ტიურინგის ტესტი), მაშინ შეგვიძლია ვთქვათ, რომ მანქანა „ფიქრობს“ (Turing, 1950). თუმცა, ცნობიერება უფრო რთული ცნებაა, ვიდრე ჭკვიანი ქცევის იმიტაცია. ცნობიერებასთან დაკავშირებულ დებატებში ხშირად განასხვავებენ ორ პოზიციას - „ძლიერ AI”-სა და „სუსტ AI”-ის (Hildt, 2019). ძლიერი AI-ის მხარდამჭერები მიიჩნევენ, რომ თუ კომპიუტერს საკმარისად რთულ პროგრამას მივცემთ, იგი არა მხოლოდ მოახდენს გონივრული ქცევის სიმულაციას, არამედ ნამდვილად ექნება გონება და შეგნება, ისევე როგორც ადამიანს (Searle, 1980) სუსტი AI-ის პოზიციის თანახმად კი, მანქანებს შეუძლიათ მხოლოდ წარმოადგინონ ინტელექტის გარეგნული გამოვლინებები წესების და ალგორითმების საშუალებით, მაგრამ არ აქვთ ნამდვილი ცნობიერება ან სუბიექტური გამოცდილება.

ფილოსოფოს ჯონ სერლის კლასიკურმა „ჩინური ოთახის“ არგუმენტმა თვალსაჩინოდ წარმოაჩინა ეს პრობლემა: წარმოიდგინეთ ადამიანი, რომელიც დახურულ ოთახში ზის და მხოლოდ ჩინურ სიმბოლოებს მანიპულირებს, წინასწარ მოცემული წესების მიხედვით. გარედან თუ შეხედავთ, ჩინურის მცოდნე დამკვირვებელს შეიძლება ეგონოს, რომ ოთახში ჩინურად მოლაპარაკე გონიერი არსებაა, რადგან კითხვებზე სწორ პასუხებს იღებს. მაგრამ სინამდვილეში, იმ ოთახში მყოფ ადამიანს ჩინური ენა არ ესმის. იგი უბრალოდ მანიპულირებს სიმბოლოებით წინასწარ მოცემული წესების მიხედვით(Searle, 1980). ანალოგიურად, შეიძლება ითქვას, რომ AI, რომელიც ჩააბარებს ტიურინგის ტესტს და გვპასუხობს ბუნებრივ ენაზე, მაინც არ „იგებს“ იმ თემას, რაზეც საუბრობს. იგი მხოლოდ სიმბოლოებს ამუშავებს ალგორითმის მიხედვით. ეს არგუმენტი ძლიერი AI-ის სკეპტიკოსებს აძლევს საფუძველს თქვან, რომ ცნობიერება ვერ აღმოცენდება მხოლოდ სიმბოლური გამოთვლით.

მეორე მხარეს, ბევრი სპეციალისტი ამტკიცებს, რომ ცნობიერება, საბოლოო ჯამში, ასევე ბუნებრივი ფენომენია, რომელიც წარმოიშობა საკმარისად კომპლექსური

ISSN 1512-1801

ინფორმაციული დამუშავებიდან (Tononi & Koch, 2015). მათი აზრით, თუ ჩვენ შევქმნით სისტემას, რომელიც გაიმეორებს ტვინის ნეირონთა კავშირულ სირთულეს და ორგანიზაციას, არ არსებობს პრინციპული მიზეზი, რატომაც ეს სისტემა არ შეიძლება გახდეს ცნობიერი. ზოგიერთ მკვლევარს მიაჩნია, რომ XXI საუკუნის ბოლომდე შესაძლებელია ხელოვნური გონების შექმნა, რომელსაც ექნება ცნობიერების გარკვეული ფორმა (Goertzel, 2014). ამას მხარს უჭერს გონების ფუნქციონალისტური ხედვა, რომ მნიშვნელობა არა აქვს, რა სუბსტრატზე (ბიოლოგიური თუ სილიკონის) მიმდინარეობს გამოთვლები, მთავარია პროცესების ორგანიზაცია და სირთულე (Chalmers, 1996). თუ ხელოვნური ინტელექტი მოიპოვებს ისეთი დონის ინტეგრირებულ ინფორმაციულ დამუშავებას, რომელიც თვისებრივად დაემსგავსება ადამიანის ტვინის მუშაობას, შესაძლებელია გაჩნდეს კიდეც სუბიექტური „მე“-ს შეგრძნება (Tononi, 2004). ზოგიერთ თეორიაში (მაგალითად: ინტეგრირებული ინფორმაციის თეორია) ცნობიერება განისაზღვრება, როგორც სისტემაში ინფორმაციის მაღალ დონეზე ინტეგრირება, შესაბამისად, თუ ხელოვნურ სისტემაში ინფორმაციის ინტეგრაცია მიაღწევს კრიტიკულ დონეს, შეიძლება ივარაუდონ იქ ცნობიერების გაჩენა.

მიუხედავად ამგვარი თეორიული შესაძლებლობებისა, დღესდღეობით ჩვენ არ გაგვაჩნია თანხმობა იმაზე, რას ნიშნავს „ცნობიერად ყოფნა“ და როგორ შეიძლება ამის ობიექტურად განსაზღვრა. ცნობიერების ე.წ. „რთული პრობლემა“მდგომარეობს იმაში, რომ ავხსნათ, როგორ წარმოქმნის მატერია სუბიექტური გამოცდილების გრძნობას (Chalmers, 1996). მაგალითად, როგორ ხდება, რომ ტვინში ნეირონული იმპულსების კომბინაციები წარმოშობს ისეთი ტიპის გამოცდილებას, როგორიცაა „წითელი ფერის ხილვა“ ან „ტკივილის განცდა“. სანამ თავად ადამიანის ცნობიერების ბუნებაა ამოუხსნელი, ძნელია დარწმუნებით ვისაუბროთ ხელოვნური სისტემების შესაძლო ცნობიერებაზე. ზოგი მკვლევარი აღნიშნავს, რომ თანამედროვე დისკურსში AI-ის ეთიკის შესახებ ცნობიერების თემა შედარებით ჩამორჩენლია ყურადღებით განხილვაში (Hildt, 2019). დღეს მთავარი აქცენტი კეთდება პრაქტიკულ საკითხებზე, როგორიცაა უსაფრთხოება, მიკერძოება და პასუხისმგებლობა. მაგრამ თუ ოდესმე შეიქმნება AI, რომელსაც ექნება ცნობიერებაზე პრეტენზია, მორალური თვალსაზრისით ურთულესი პრობლემა დადგება: როგორ მოვეპყროთ ცნობიერ მანქანას? ექნება თუ არა მას უფლებები, ან პირიქით - პასუხისმგებლობები? სწორედ იმიტომ, რომ ამ კითხვებზე პასუხიები ჯერ არ გაგვაჩნია, ზოგი მიიჩნევს, რომ ცნობიერების შესახებ დებატები ჯერჯერობით უფრო თეორიული ხასიათისაა, ვიდრე პრაქტიკული. თუმცა, ტექნოლოგიის სწრაფი წინსვლის გათვალისწინებით, მეცნიერები და ფილოსოფოსები ყურადღებით ადევნებენ თვალს ნებისმიერ ნიშანს, რომელიც შესაძლოა მიუთითებდეს ხელოვნური ინტელექტის ცნობიერ „გამოღვიძებაზე“. ჯერჯერობით ასეთი ნიშანი არ გვინახავს. დღევანდელი AI ისეთია, რომ შეუძლია დაიმახსოვროს და გამოიყენოს ინფორმაცია ადამიანის ინტელექტის შთაბეჭდილების შესაქმნელად, მაგრამ არ გააჩნია შნაგანი „მე“. შესაბამისად, ცნობიერების დებატები თეორიულ არენაზე გაგრძელდება მანამდე, სანამ არ გაჩნდება მონაცემები ან არგუმენტები, რომლებიც მკაფიოდ გვიჩვენებს, შესაძლებელია თუ არა მანქანური ცნობიერება.

დასკვნა

საბოლოოდ, წარმოდგენილ კვლევაში განხილულმა თემებმა ცხადყო, რომ ხელოვნური ინტელექტის „ფსიქოლოგია“ ჯერ კიდევ მნიშვნელოვან შეზღუდვებს და გამოწვევებს მოიცავს, თუმცა პარალელურად გამოიკვეთა ახალი შესაძლებლობებიც. AI-ის აქვს

ISSN 1512-1801

უნარი, შეასრულოს ადამიანზე სწრაფად და ზუსტად რიგი კოგნიტური ამოცანები, მაგრამ მას აკლია ის მდიდარი და მრავალგანზომილებიანი შემეცნებითი და ემოციური კონტექსტი, რაც ადამიანურ ფსიქიკას ახასიათებს. თანამედროვე AI არ ფლობს ცნობიერ შეგრძნებებსა და ინტუიციას. ეს ჩვენმა ანალიზმა ერთ-ერთ ძირითად მიგნებას წარმოაჩინა. ადამიანისა და AI-ის ურთიერთქმედების შესწავლამ გვაჩვენა, რომ ადამიანები თანდაყოლილი მზადყოფნითა და სურვილით ეპყრობიან ინტელექტუალურ მანქანებს ისე, თითქოს ისინი სოციალური არსებები იყვნენ. ეს ფაქტორი შეიძლება როგორც სასარგებლო (AI სისტემების მიღებასა და გამოყენებაში), ისე სარისკოც იყოს (მომხმარებლის შეცომაში შეყვანისა და ემოციურ დამოკიდებულებაში), რის გამოც აუცილებელია ტექნოლოგიის დიზაინსა და პოლიტიკაში გათალსიიწნებული იქნას ადამიანის ფსიქოლოგია.

თეორიული ჩარჩოების განხილვამ ხაზი გაუსვა, რომ ხელოვნური ინტელექტი მჭიდროდ არის დაკავშირებული კოგნიტურ მეცნიერებებთან: ფსიქოლოგიის მოდელები შთააგონებს ახალი ალგორითმების შექმნას, ხოლო AI-ის მიღწევები თავის მხრივ ეხმარება შემოწმდეს ჰიპოთეზები ადამიანის გონების შესახებ. თუმცა, კვლავ ღია კითხვად რჩება, როგორ შევათავსოთ სიმბოლური და კონექციონისტული მიდგომები ისე, რომ მივიღოთ როგორც განმარტებადი, ისე ძლიერი AI სისტემები. ამ მიმართულებით სამომავლო კვლევამ შეიძლება მოგვცეს მეტი ცოდნა იმისა, თუ როგორ შეიძლება ტექნიკური „ტვინი“ უფრო მივუახლოვოთ ადამიანის ტვინის ფუნქციონირებას.

ეთიკური ანალიზის ნაწილმა წარმოაჩინა რამდენიმე კრიტიკული გამოწვევა, რომელსაც სწრაფად მზარდი ხელოვნური ინტელექტის მომხმარებელთა სამყარო უნდა შეხვდეს. ალგორითმებში მიკერძოების აღმოფხვრა, ემოციური AI-ის პასუხისმგებლიანი გამოყენება და ადამიანი-AI ურთიერთობებში ზღვარის გატარება. ეს საკითხები მოითხოვს როგორც ინჟინერულ, ისე სოციალურ-სამართლებრივ გადაწყვეტილებებს. მნიშვნელოვანია, რომ AI სისტემების შემქმნელები აცნობიერებდნენ იმ გავლენას, რაც მათ პროდუქტებს შეიძლება ჰქონოდეთ მომხმარებლის ღირებულებებსა და კეთილდღეობაზე. ადამიანის ემოციური რეაქციების გათალისწინება არ უნდა იქცეს მანიპულაციის იარაღად. პირიქით, იგი უნდა წარიმართოს ადამიანის მხარდაჭერისა და გაძლიერებისკენ.

ცნობიერების დებატებმა გვიჩვენა, რომ მიუხედავად დიდი ინტერესისა, ჯერ კიდევ არ იკვეთება მტკიცებულება იმისა, რომ თანამედროვე AI სისტემებს აქვთ რაიმე ფორმით გრძნობის უნარი. ეს, ერთი მხრივ, გვამშვიდებს. ჩვენ ჯერ არ ვდგავართ ისეთი სცენარის წინაშე, სადაც მანქანისუფლებები ან შეგრძნებები უნდა გავითვალსიწინოთ. თუმცა, სამეცნიერო ფანტასტიკაში დასმული კითხვები რეალურ კვლევაშიც თანდათან იჩენს თავს: იქნება ეს შესაძებელი მომავალში და რა ცვლილებებს გამოიწვევს? ამ ეტაპზე შეიძლება ითქვას, რომ AI ცნობიერია იმდენად, რამდენადაც იგი ადამიანის ინტელექტის შთაბეჭდილებას ქმნის თავისი ქცევით, პასუხებით, თუმცა არ გააჩნია შნაგანი სუბიექტურობა.

ამრიგად, ხელოვნური ინტელექტის ფსიქოლოგიის სფერო სულ უფრო მნიშნელოვანი გახდება მომავალ წლებში. რეკომენდებულია ინტერდისციპლინური თანამშრომლობა. ფსიქოლოგებმა, ნეირომეცნიერებმა, ინჟნრებმა და ეთიკოსებმა უნდა გააძლიერონ ერთობლივი ძალისმხევა, რათა შექმნას AI, რომელიც იქნება არა მხოლოდ ჭკვიანი,

ISSN 1512-1801

არამედ ადამიანისთვის უსაფრთხო, ეთიკუად მისაღები და ადამიანის ფსიქოლოგიურ საჭიროებებზე მორგებული. მომავალმა კვლევამ შეიძლება გამოკვეთოს, როგორ შევძინოთ AI-ის უფრო "ადამიანური" გაგება. მაგალითად, გამოვიკვლიოთ, როგორ აისახება სხვადასხვა კულტურული და სოციალური კონტექსტი ადამიანსა და ხელოვნურ ინტელექტში მიკერძოების წარმოშობა ფსიქოლოგთა ცოდნის ალგორითმებში ინტეგრირებით. ასევე, თუ ოდესმე აღმოვჩნდით რეალური ხელოვნური ცნობიერების შექმნის ზღვარზე, აუცილებელი იქნება სამეცნიერო და საზოგადოებრივი დიალოგის წინმსწრებად წარმართვა, რათა მომზადებული შევხვდეთ იმ ეთიკურ და ფილოსოფიურ საკითხებს, რაც ისეთ გარღვევას მოჰყვება. ხელოვნური ინტელექტის ფსიქოლოგიის კვლევის გაგრძელება ხელს შეუწყობს როგორც უკეთესი AI სისტემების შექმნას, ისე ადამიანის საკუთარი ბუნების გაღრმავებულ გაცნობიერებას.

ბიბლიოგრაფია

1. Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.
2. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036
3. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287583
4. Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*, 627–652. https://doi.org/10.1146/annurev-psych-010416-044144
5. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 77–91. https://doi.org/10.1145/3287560.3287596
6. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230
7. Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
8. Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625), 311–313. https://doi.org/10.1038/538311a
9. Darling, K. (2016). Extending legal rights to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. *Proceedings of the We Robot Conference*.
10. Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3-4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3
11. Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology, 51*(4), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x
12. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1
13. Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence, 5*(1), 1-46. https://doi.org/10.1515/jagi-2014-0001
14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
15. Hancock, P. A., Nourbakhsh, I., & Stewart, J. E. (2020). On the future of human–AI interaction: A survey and speculation. *AI & Society, 35*(4), 595–614. https://doi.org/10.1007/s00146-019-00905-9
16. Hildt, E. (2019). Artificial intelligence: Does consciousness matter? *Frontiers in Psychology, 10*, 1535. https://doi.org/10.3389/fpsyg.2019.01535
17. Hoffman, G., Crittendon, D., & McDonald, C. G. (2021). Empathy, ethics, and AI. *Annual Review of Cybersecurity Ethics, 14*(1), 85–112.
18. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

ISSN 1512-1801

19. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447
20. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253. https://doi.org/10.1017/S0140525X16001837
21. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
22. Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint*, arXiv:2002.06177.
23. Merritt, S. M., Carter, M., & Madhavan, P. (2021). Trust, automation, and society: The future of human-robot interaction. *Annual Review of Psychology, 72*, 361–388. https://doi.org/10.1146/annurev-psych-010419-051053
24. McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. *Trends in Cognitive Sciences, 10*(3), 120–128.
25. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153
26. Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.
27. Nyholm, S., & Frank, L. (2019). From sex robots to love robots: Is mutual love with a robot possible? *AI & Society, 34*(3), 543–557. https://doi.org/10.1007/s00146-018-0842-2
28. Picard, R. W. (1997). *Affective computing*. MIT Press.
29. Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
30. Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1*, 318–362.
31. Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
32. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *arXiv preprint*, arXiv:1708.08296.
33. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756
34. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
35. Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*, 42. https://doi.org/10.1186/1471-2202-5-42
36. Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167
37. Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/59.236.433

Article received 2025-03-10