SUN'IY INTELLEKTNING KELAJAK UCHUN XAVFSIZLIGINI SAQLASH NAZARIYASI

Choryorqulov G‘iyos Husan o‘g‘li *- assistent*
O‘zbekiston Milliy universitetining Jizzax filiali, O‘zbekiston
choryorqulov@jbnuu.uz
Jo‘raqulova Munisa O‘ktam qizi
O‘zbekiston Milliy universitetining Jizzax filiali, O‘zbekiston
m35209737@gmail.com
Sa‘dullayeva Sabina Rizamat qizi *- talaba*
O‘zbekiston Milliy universitetining Jizzax filiali, O‘zbekiston
sadullayeva9030@gmail.com

Annotatsiya. Sun'iy intellektning to'xtovsiz rivojlanishi natijasida unga katta e'tibor qaratildi. Aqlli tizimlar kelajak insoniyat uchun xavfsiz bo'ladimi? Shu sababli, ko'plab tadqiqotchilar o'zlariga muammolar olib kelishi mumkin bo'lgan muammolarni hal qilish ustida ish boshladilar. Sun'iy intellekt tizimlari nazoratdan tashqarida harakat qilishga imkon beradi yoki odamlar uchun xavfli bo’lgan pozitsiyalarni egallashi kerak. Bunday tadqiqot ishlari hozirda sun'iy intellekt adabiyotlariga kiritilgan.

Sun'iy intellektning xavfsizligi va/yoki kelajagi. Kontekstida tushuntirishlar ushbu tadqiqot hujjati bo'yicha nazariyani taklif qiladi. Uning ishlash muddatini hisobga olgan holda xavfsiz aqlli tizimlarga erishish ba'zilarga ko'ra, sun'iy intellektga asoslangan tizim operatsion o'zgaruvchilar va bartaraf - bir aqlli tugatish yangilarga imkoniyat berish uchun ishlash uchun "yetarlicha eski" tizim xavfsizroq ko'rinadigan tizimlarning avlodlaridir. Nazariyaga qisqacha kirish u bo'yicha qo'shimcha tadqiqotlar eshiklarni ochadi.

Kalit so‘zlar: Sun'iy intellekt, FIT nazariyasi, trening holat, umbrbodlik holati, optimallashtirish muammosi.

I. KIRISH

Ilmiy maydonga ilk qadamlaridan boshlab, Sun'iy intellekt juda yaxshilandi va deyarli hammaga zamonaviy hayot sohalariga ta'sir qildi. Nazariy va amaliy birlashtirib kompyuter fanining aspektlari va ularni boshqarish kompyuter kabi ba'zi ilg'or texnologiyalarni, elektronika va aloqa qo'llab-quvvatlashni o’z ichiga oladi [1]. Sun'iy intellekt hozirda real dunyoning barcha turlarini yengish uchun katta kuchga

asoslangan muammolar, hatto ular turli darajalarga tegishli murakkablikdir. Moslashuvchan va ochiq yechim doirasida sun'iy intellektni takomillashtirishkatta rol o'ynaydi. Bu uchun yechim yondashuvlarining samaradorligi real dunyoga asoslangan muammolar va hayotni odamlar uchun ko'proq qulay qiladi. Ayniqsa, matematik va mantiqiy fondagi yondashuvlar har qanday narsani moslashtirishni osonlashtiradi. Bu yerda sun'iy intellekt o'rtasidagi farq va falsafiy ma'noda yana bir ilmiy soha bo'ladi, bu esa aqlli tizimlarni ishlab chiqish uchun muammo emas [2].

Sun’iy intellekt kelajakning eng kuchli ilmiy sohasidir. Ammo yangi texnologik yaxshilanishlardan xavotir - O'zgarishlar odamlarni har doim har qanday narsani muhokama qilishga majbur qildi Bu uchun xavfli yoki zararli bo'lishi mumkin bo'lgan stsenariylar insoniyatning mavjudligi yoki hech bo'lmaganda barqaror hayotining yerdagi standartlardir. Nihoyat, sun'iy soha bunday tashvishga duch keldi va Yangi sub-tadqiqot sohasi paydo bo'ldi [3].

Xavfsizlik. Bundan tashqari, aqlli mashinalar qilish etikasi bilan bog'liq tizimlari, Sun’iy intellekt xavfsizligini ta’minlashga qaratilgan zararli bo'lmagan xavfsiz aqlli tizimlar insoniyat va ularning muammolarini hal qilishda samarali yechimdir. Qachon tegishli adabiyotlarni ko'rib chiqsak, biz ushbu tadqiqotni ko'rishimiz mumkin xavfsizlik muammolari bo'yicha tadqiqotlar ba'zi tadqiqotlar bilan javob beradi [4].

Bu tadqiqotlarning barchasi tushunchalar bilan bog'liq xavfsiz aqlli tizimlarga erishish bilan va tushunishga harakat qilish umumiy yondashuvlar, qoidalar, strategiyalar va siyosatlar yo'naltirilgan kerakli xavfsiz sun'iy intellektni ishlab chiqish tizimlaridir. Batafsil, Sun'iy intellekt etikasi asarlari haqida aqlli tizimlar yuzaga kelishi mumkin bo'lgan axloqiy dilemmalarni hal qilish bo'yicha uchrashish keltirilgan [5]. Bu yerda "axloq" tushunchasini qanday tushunish haqida munozara, aqlli tizimlar istiqboli va boshqa tushuncha keltirilgan:

Sun'iy intellekt xavfsizligi muhandisligi uchun taklif qilingan [6]. Uning eng muhim bosqichlaridan biri Sun'iy intellekt xavfsizligiga oid o'zgarishlar bo'lishi mumkin. Sun'iy intellekt xavfsizligi bo'yicha tadqiqot dasturining boshlanishi 2015 -yilda asosan Ilon Mask tomonidan moliyalashtirilgan [7].

Sun'iy intellekt mavjud bo'lgan tadqiqot institutlari / markazlari Xavfsizlikka yo'naltirilgan tadqiqotlarni quyidagicha sanab o'tish mumkin:

- Future of Humanity Institute – Oksford universiteti,
- Center for Human-Compatible AI – UC Berkeley,
- Machine Intelligence Research Institute,
- Leverhulme Centre for the Future of Intelligence – Kembrij universiteti,

- Vector Institute for Artificial Intelligence –Toronto universiteti
- Future of Life Institute,
- Open AI,
- Ekzistensial xavfni o'rganish markazi.

Sun'iy intellekt xavfsizligini hisobga olgan holda ilmiy-tadqiqot ishlari odatda mavjudligiga qarab tuzilgan aqlli agentlar mavjud. Ushbu tadqiqotda bunday agentlar ham mavjud aqlli tizimlar deb ataladi. Oddiy agent bittadan iborat bo'lishi mumkin yoki unga erishish uchun ko'proq Sun'iy intellekt texnikasi ekzistensial tuzilmalar bo'lishi mumkin. Ammo bu omil bir necha yildan beri e'tiborga loyiq emas Sun'iy intellekt xavfsizligiga yo'naltirilgan ishlarning asosiy nuqtasi bunday tizimlarni qanchalik yaxshi o'rgatish yoki qanday qilib muammolarni hal qilish uchun, ularni nazorat qilish yaxshidir [8].

E'tiborga molik bo'lganlardan hozirgi kunda tadqiqotchilarni odatda quyidagilar qiziqtiradi:

- Inverse Reinforcement Learning / ReinforcementLearning [9-10],
- Interruptible Agents / Ignorant Agents / Inconsistent Agents / Bounded Agents [11-19],
- Corrigibility [20],
- Rationality [21-23],
- Super Intelligence [22].

Hozirgacha berilgan tushuntirishlar kontekstida buning maqsadi tadqiqot xavfsiz intellektga erishish bo'yicha nazariyani taklif qilishdir. Sun'iy intellektning ishlash muddatini hisobga olgan holda agentlar / tizimlar ba'zi operatsion ko'ra razvedka asoslangan tizim o'zgaruvchilar va yo'q qilish - aqlli agentni tugatish /imkoniyat berish uchun ishlash uchun "etarlicha eski" tizim xavfsizroq ko'rinadigan yangi avlod tizimlari deb ataladi [12].

"Fading Intelligence Theory" bu aqlli deb hisoblanadi agent/tizim o'z darajasiga yetganda uni qo'shimcha o'rgatib bo'lmaydi. Eng yuqori ta'lim qobiliyati, aks holda u o'z maqsadlarini qo'ldan boy beradi, bu endi xavfsiz emasligini anglatadi. Bundan tashqari, ba'zida to'xtash kerak, bilan aql darajasining pasayishiga olib kelishi uchun uni o'rgatish, yaxshi taqsimlanmagan ta'lim ma'lumotlari. Nihoyat, ba'zilari bo'lishi kerak, qaysi aqlli tizimlar ishlashini aniqlash uchun global ko'rsatkichlar yoki yo'q qilish - tugatish. Shunday qilib, har bir kishi uchun hayot vaqti bo'lishi kerak [13].

II. FIT nazariyasi.

Fading Intelligence Theory (FIT) haqida qisqacha aqlli agent / tizimning bandlik holati va bunda yo'l, buning uchun umrbod umumiy tuzilmani shakllantirish. Bu nazariyani ikki jihat ostida tekshirish mumkin (1-rasm):

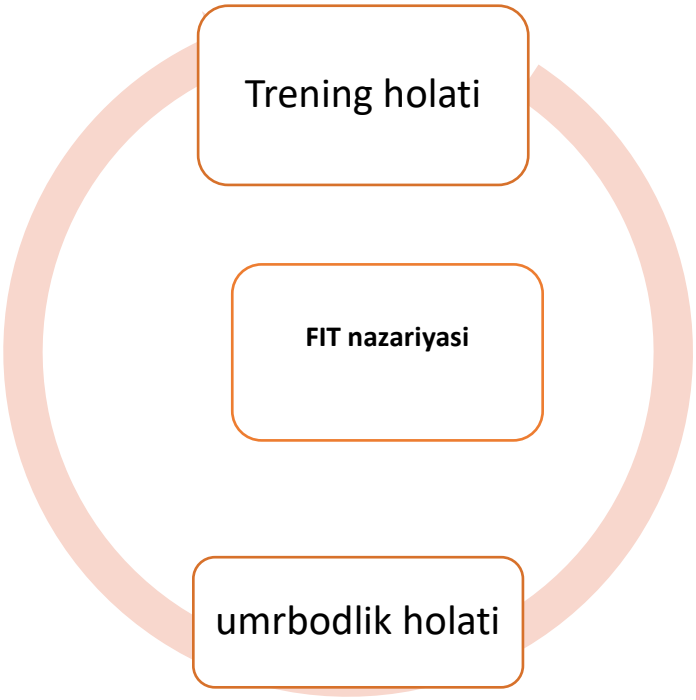
Tizimning trening holati,

Tizimning umr bo'yi holati.

Umuman olganda, nazariya ostida ikkita yondashuvni olishimiz mumkin tegishli jihatlarni hisobga olgan holda:

Mashg‘ulotning davlat yondashuvi **FIT-a nazariyasi**. P bo'lishi mumkin bo'lgan global muammolar to'plami bo'lsin aqlli agent / tizim Ag tomonidan hal qilinadi. Shuningdek, T q p ę P ni hal qilishga erishadigan o'quv ma'lumotlarining global to'plami 100% muvaffaqiyat darajasi bilan Ag bo'lsin. Gipotezaga ko'ra, Ag ko'proq o'qitilgan bo'lishi mumkin emas va u "to'liq aqlli agent" deb ataladi [14].

FIT-a isboti. Qo'shilishi kerak bo'lgan yangi o'quv ma'lumotlari haqida o'ylab ko'ring. T. ga Chunki u muvaffaqiyat darajasiga kafolat bera olmaydi yangi T to'plami bilan o'zgaradi va T ning tarqalishiga ta'sir qiladi ch p ę P ni 100% muvaffaqiyat darajasida echilishi mumkin qilish, the Natijalarni ko'rish uchun joriy agent/tizim qayta o'qitilishi kerak. Shunday qilib, to'liq Ag bir xil agent / tizim emas, chunki to'liq Ag oldingi T bilan bog'langan, t ni o'z ichiga olmaydi. Biri kerak. yangi T to'plami bo'yicha butunlay yangi treningni ko'rib chiqish haqida o'ylang, yangi agent / tizim bilan, ya'ni to'liq Ag ko'proq o'rgatib bo'lmaydi. Bundan tashqari, to'liq mashq qilishni tanlash Ag o'z mavqeini yo'qotishiga, muvaffaqiyat darajasining pasayishiga sabab bo'ladi [15]. Bundan tashqari, sun'iy intellekt xavfsizlik agentning umumiy muvaffaqiyat darajasi bilan bog'liq agent/tizim inson ehtiyojlarini qondirsa jami 100% stavka bilan muammoni hal qilish, keyin har qanday qo'shimcha T ning o'zgarishi "kapalak effektlari" va xavfsizlikka olib keladi.



1-rasm. Fitning ikki jihati

FIT-b nazariyasi. FIT-a nazariyasini ko'rib chiqsak, bu mumkin agent / tizimning umr bo'yi holati haqida o'ylash Ag. Qisqasi, FIT-a nazariyasi Agni ko'proq o'rgatish mumkin emasligini ko'rsatadi agar u "to'liq aqlli agent / tizim" bo'lsa. Shunday qilib, gipoteza bo'yicha, T va / yoki P dagi har qanday o'zgarish Ag ning yo'q qilinishiga olib keladi [16].

FIT-b isboti. To'liq Ag davom etishi mumkinligini ko'rib chiqing .100% muvaffaqiyat darajasini ko'rsatish, hatto T va P o'zgartiriladi. Bunda har qanday yaxshi taqsimlanmagan ta'lim ma'lumotlari qo'shilmasligi kerak mashg'ulot natijalariga ta'sir qiladi. Ammo bu mumkin emasligi sababli, bir xil Ag har doim bo'lishini kafolatlay olmaydi. Yangilangan P.larni yangilangan Ts orqali hal qilish. Shunday qilib, agent / tizim, yangilangan P larni yangilangan T larga nisbatan hal qila oladigan narsa bilan bir xil emas. To'liq Ag, bu yo'q qilish - tugatish degan ma'noni anglatadi. Yangi P va / yoki T uchun to'liq Ag. Umuman olganda, bu yangi P to'plamini anglatadi va/yoki yangi T to'plami yangisini talab qiladi [17].

FIT-c nazariyasi. Gsr tomonidan aytilgan global muvaffaqiyat darajasi bo'lsin organlari P to'plami uchun olingan eng yaxshi yechim natijasi sifatida, bir xil T ustida va bir xil turdagi agent / tizim bilan parametrlarni o'zgartirish. Ag yangi agent / tizim bo'lsin, qaysi hozirda tegishli P uchun tegishli T. tomonidan foydalanilmoqda. Gipoteza, Ag gsr ga bog'liq hayot muddatiga ega bo'lishi kerak va uning ta'limi

bo'yicha odatiy hisob-kitob orqali - qo'llanilishi marta va ba'zi parametrlar, shu jumladan gsr [18].

FIT-c isboti. Ko'p agentlar/tizimlar mavjudligini ko'rib chiqing bir xil P ustidan ariza berish uchun kundan-kunga ortib borayotgan bir xil T ustida. Bundan tashqari, Agb agenti bo'lsin / tizimi ega gsr. Bunday holda, o'lmas agentlar / tizimlardan "ishlash" P ustidan T yaxshiroq hal qilinsa, xavfsizlik buzilishiga olib keladi. Yangi agentlar/tizimlarni loyihalash va ulardan foydalanishni davom ettirish evristik usul va faqat Agb dan foydalanishni o'ylamaydi yechimlar bo'yicha ko'proq takomillashtirish faqat allaqachon qiladi qayta-qayta kashf qilinadigan yechim joylari kuzatildi. Shunday qilib, chunki bunday agentlar/tizimlarni ishga olish ko'pchilikka sabab bo'ladi. xavfsizlik nuqtai nazaridan muammolar, gsrga qarab umr bo'yi borligini inkor etib bo'lmaydi. Shuningdek, oddiy hisob-kitob, ya'ni ta'lim - dastur vaqtlari va parametrlari, shu jumladan gsr umr bo'yi aniq qiymat berishi mumkin. Agentning ishlash muddatini hisoblash bo'yicha Tizimga xos optimallashtirish muammolari shakllantirilishi mumkin qachon yo'q qilishni aniqlang - boshqasidan tashqari uni to'xtating FITda ko'rsatilgan shartlar. Bunga quyidagilar orqali erishish mumkin:

Global optimallashtirishga yo'naltirilgan minimallashtirish orqali gsr va tajribali tomon umumiy xato muammosi ba'zi o'zgaruvchilarni aniqlash uchun hozirgacha ta'lim natijalari shu jumladan agent/tizimning qolgan foydalanish vaqti.

• Global optimallashtirishga yo'naltirilgan maksimallashtirish gsr va vaznli tomon muvaffaqiyat darajasi muammosi gacha bo'lgan o'rtacha mashg'ulotlar miqdori Ba'zi o'zgaruvchilarni, shu jumladan qolgan foydalanishni ham aniqlang agent / tizim vaqti.

• bilan shug'ullanuvchi bir kombinatoryal muammo tuzilishi ustidan o'tgan o'quv mashg'ulotlarining parametrlarini aniqlash gsr yoki undan yuqori natijalarga yaqin natijalarga olib keladigan optimal yo'l global natijalar, shu jumladan qolgan foydalanish vaqti haqidagi ma'lumotlar [20].

Ushbu optimallashtirish muammosiga batafsil e'tibor aniq hayot vaqtini aniqlash keyingi tadqiqot mavzusi bo'lishi mumkin. Belgilangan FITni hisobga olgan holda, umumiy aqlli agent / tizimning ishlash muddati uchun sxema bo'lishi mumkin.

V. XULOSALAR VA KELAJAKDAGI ISHLAR

Ushbu maqola Fading Intelligence nazariyasini taqdim etdi, sun'iy saqlashda e'tiborga olinishi mumkin. Intellektual tizimlarning ishlash muddati bo'yicha razvedka xavfsizligi. Batafsil, nazariya mashg'ulotlarni qachon davom ettirish kerakligi bilan bog'liq, yo'q qilish - tugatish yoki aqlli tizimni o'rgatmaslik tufayli

yuzaga kelishi mumkin bo'lgan har qanday istalmagan vaziyatlardan qoching "eski" aqlli tizim. Buni tushunish mumkin nazariya Sun'iy intellektga asoslangan tizimlarni halokatli qiladi. Garchi barcha aqlli tizimlarni o'lmas deb hisoblash mumkin klonlash mumkin bo'lgan dasturiy ta'minotga yo'naltirilgan jihatlari tufayli tegishli yondashuvlar bilan uzatiladi yoki qayta yaratiladi. Sun'iy intellekt tizimlarini o'lik deb qabul qiladigan vaziyat (bir umrga ega) uchun umumiy xavfsizlikni ta'minlash, chunki kelajakning sun'iy intellekti. Boshqa tomondan, nazariya, shuningdek, umrining umumiy doirasini belgilashga harakat qiladi aqlli agentlar / tizimlar.

Nazariya qo'shimcha kuzatishlarni talab qilishi aniq doiradgia amalga oshirilgan vakillik baholashga qo'shimcha tadqiqot hisoblanadi. Buni butunlay maqsadli deb aytish mumkin. Fading Intelligence nazariyasini isbotlovchi kuzatish o'qitilgan aqlli agent / tizimni isbotlash bilan bog'liq. Hozirgi ilmiy ma'lumot 100% aniqlik bilan, bu bilan imkonsiz ko'rinadi. Biroq, mualliflar tomonidan amalga oshiriladigan ko'proq va ko'proq kuzatishlar kelajakdagi ishlarga asoslanadi.

Foydalanilgan adabiyotlar:

1. Nizomiddin N. et al. TA'LIMDA DASTURLASH JARAYONINI BAHOLASHGA ASOSLANGAN AVTOMATLASHTIRILGAN TIZIMNI TADBIQ ETISH //International Journal of Contemporary Scientific and Technical Research. – 2023. – С. 24-28.
2. Choryorqulov G. H., & Qosimov NS (2023) //ELEKTRON JADVAL MODELINING TAVSIFLANISHI. PEDAGOGS Jurnali. – Т. 30. – №. 3. – С. 67-73.
3. Чорркулов Г., Норматов Н., Мамараимов А. Роль анализа текстовых связей в электронных документах в информационной безопасности //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 1. – С. 67-71.
4. Норматов Н., Мамараимов А. Ta'lim tizimida baholash tizimini avtomatlashtirishni joriy etish jarayonlari va foydalanish metodlari //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 2. – С. 356-359.
5. Мамараимов А., Чорёркулов Г., Норматов Н. Tanib olish modullarini dasturiy amalga oshirish //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 2. – С. 38-44.
6. Sanoqul o'g'li Q. N. et al. ELEKTRON HUJJAT ALMASHINUVINI AVTOMATLASHTIRISH MODELINI ANALITIK TAHLILI //Лучшие интеллектуальные исследования. – 2023. – Т. 10. – №. 5. – С. 89-100.
7. Mamatkulovich B. B., Shuhrat o'g'li M. S., Jasurjonovich B. J. SPECIAL DEEP CNN DESIGN FOR FACIAL EXPRESSION CLASSIFICATION WITH A SMALL AMOUNT OF DATA //Open Access Repository. – 2023. – Т. 4. – №. 3. – С. 472-478.
8. Mamatkulovich B. B. et al. Simplified machine learning for image-based fruit quality assessment //Eurasian Journal of Research, Development and Innovation. – 2023. – Т. 19. – С. 8-12.
9. Mamatkulovich, B. B., Dilshod o'gli, Y. A., & Akmal o'g'li, A. A. (2023). Predicting daily energy production in a blockchain-based P2P energy trading system. Texas Journal of Engineering and Technology, 18, 7-11.

10. Javlon, K., & Erali, M. (2023). STRUCTURE AND PRINCIPLE OF OPERATION OF FULLY CONNECTED NEURAL NETWORKS. International Journal of Contemporary Scientific and Technical Research, 136-141.
11. Мустафоев, Е., & Холматов, Ж. (2023). Brayl matn tasviri sifatini oshirish usullari. Информатика и инженерные технологии, 1(2), 23-27.
12. Obid o‘g A. S. J. et al. Numpy Library Capabilities. Vectorized Calculation In Numpy Va Type Of Information //Eurasian Research Bulletin. – 2022. – Т. 15. – С. 132-137.
13. Javlon X. et al. Классификатор движения рук с использованием биомиметического распознавания образов с помощью сверточных нейронных сетей с методом динамического порога для извлечения движения с использованием датчиков EF //Journal of new century innovations. – 2022. – Т. 19. – №. 6. – С. 352-357.
14. Фитратович В. и др. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ МНОГОФАЗНОЙ ФИЛЬТРАЦИИ В НЕФТЕГАЗОВОМ ПЛАСТЕ ПРИ ЕГО ЗАВОДНЕНИИ //INTERNATIONAL CONFERENCES ON LEARNING AND TEACHING. – 2022. – Т. 1. – №. 4. – С. 520-525.
15. Jamshid S. ENTROPY EVALUATION CRITERION IN DECISION TREE ALGORITHM EVALUATION //International Journal of Contemporary Scientific and Technical Research. – 2023. – С. 236-239.
16. Салимов Ж., Абулаева А. Классификации дерева в машинном обучении и гиперпараметрах //Информатика и инженерные технологии. – 2023. – Т. 1. – №. 1. – С. 71-79.
17. Obid o‘g‘li S. J., Nodir o‘g‘li X. A., Jasurjonovich B. J. SUPERVISED LEARNING REGRESSION ALGORITHM SIMPLE LINEAR REGRESSION //Academia Science Repository. – 2023. – Т. 4. – №. 04. – С. 69-76.
18. Ramazon, Mixliyev, and Babayarov Abdusattor. "MIKROSKOP YORDAMIDA HUJAYRALARDAGI QON VA OQ QON HUJAYRALARI SONI BO‘YICHA BEMORLARNING SOG‘LIG‘INI ANIQLASH." International Journal of Contemporary Scientific and Technical Research (2023): 133-137.
19. Norqo’ziyev, Q. (2023). MOBIL ROBOTLAR UCHUN YO’LNI REJALASHTIRISH ALGORITMI. Research and Implementation. извлечено от https://fer-teach.uz/index.php/rai/article/view/746
20. Норкозиев, К., & Тоджиев, А. (2023). Использование искусственных нейронных сетей при разработке алгоритма поиска оптимального пути мобильных роботов в динамических средах. Информатика и инженерные технологии, 1(1), 25–29. извлечено от https://inlibrary.uz/index.php/computer-engineering/article/view/25025
21. Тоджиев, А., & Норкузиев, К. (2023). The role of artificial intelligence technology in individualized teaching . Информатика и инженерные технологии, 1(2), 153–156. извлечено от https://inlibrary.uz/index.php/computer-engineering/article/view/25014
22. Ramazon, Mixliyev, and Babayarov Abdusattor. "MIKROSKOP YORDAMIDA HUJAYRALARDAGI QON VA OQ QON HUJAYRALARI SONI BO‘YICHA BEMORLARNING SOG‘LIG‘INI ANIQLASH." International Journal of Contemporary Scientific and Technical Research (2023): 133-137.