SUN’IY INTELLEKT MODELLARINING ZAIFLIKLARI VA ULARNI HIMOYALASH USULLARI (AI SECURITY)

Muallif: ***Abdullayev Lochinbek***
Kompyuter muhandisligi yo‘nalishi talabasi,
Xalqaro Nordik Universiteti
Email: abdulayevlochinbek.05@gmail.com,
Tel: +998 93-548-21-08
Ilmiy rahbar: ***Abdullayev Munis Qurbonovich***
Xalqaro Nordic universiteti “Sanoatni boshqarish va raqamli texnologiyalar” kafedrasi mudiri,
PhD, dotsent. TDIU mustaqil izlanuvchisi.
ORCID: https://orcid.org/0000-0003-0290-8453
Email: m.abdullayev@nordicuniversity.uz

Annotatsiya

Ushbu maqolada sun’iy intellekt tizimlarining xavfsizligi bilan bog‘liq dolzarb muammolar va ularning zamonaviy axborot tizimlaridagi ahamiyati tahlil qilinadi. Tadqiqotning asosiy maqsadi sun’iy intellekt asosida ishlovchi tizimlarda uchraydigan xavfsizlik tahdidlarini aniqlash, ularni tasniflash hamda mavjud himoya yondashuvlarini o‘rganishdan iborat. Tadqiqot jarayonida tahliliy va taqqoslash usullaridan foydalanilib, mavjud ilmiy adabiyotlar va amaliy holatlar asosida xavfsizlik muammolari ko‘rib chiqildi. Maqolada chalg‘ituvchi hujumlar, ma’lumotlarning buzilishi, modelni noqonuniy nusxalash hamda shaxsiy ma’lumotlar maxfiyligi bilan bog‘liq xavflar alohida tahlil qilinadi. Shuningdek, sun’iy intellekt tizimlarining barqarorligi va ishonchliligini oshirishga qaratilgan himoya mexanizmlari, jumladan, xavfsiz ma’lumotlar bilan ishlash, modelni tekshirish va monitoring qilish usullari yoritib beriladi. Tadqiqot natijalari shuni ko‘rsatadiki, sun’iy intellekt xavfsizligini ta’minlash faqat texnik choralar bilan cheklanib qolmasdan, kompleks va tizimli yondashuvni talab etadi. Mazkur maqola sun’iy intellekt asosidagi tizimlarni xavfsiz va ishonchli ishlab chiqish bo‘yicha ilmiy va amaliy tavsiyalar berishga qaratilgan.

Kalit so‘zlar: Sun’iy intellekt, AI xavfsizligi, mashinaviy o‘rganish, chalg‘ituvchi hujumlar, ma’lumotlar manipulyatsiyasi, model himoyasi, kiberxavfsizlik, ishonchli AI, ma’lumotlar maxfiyligi.

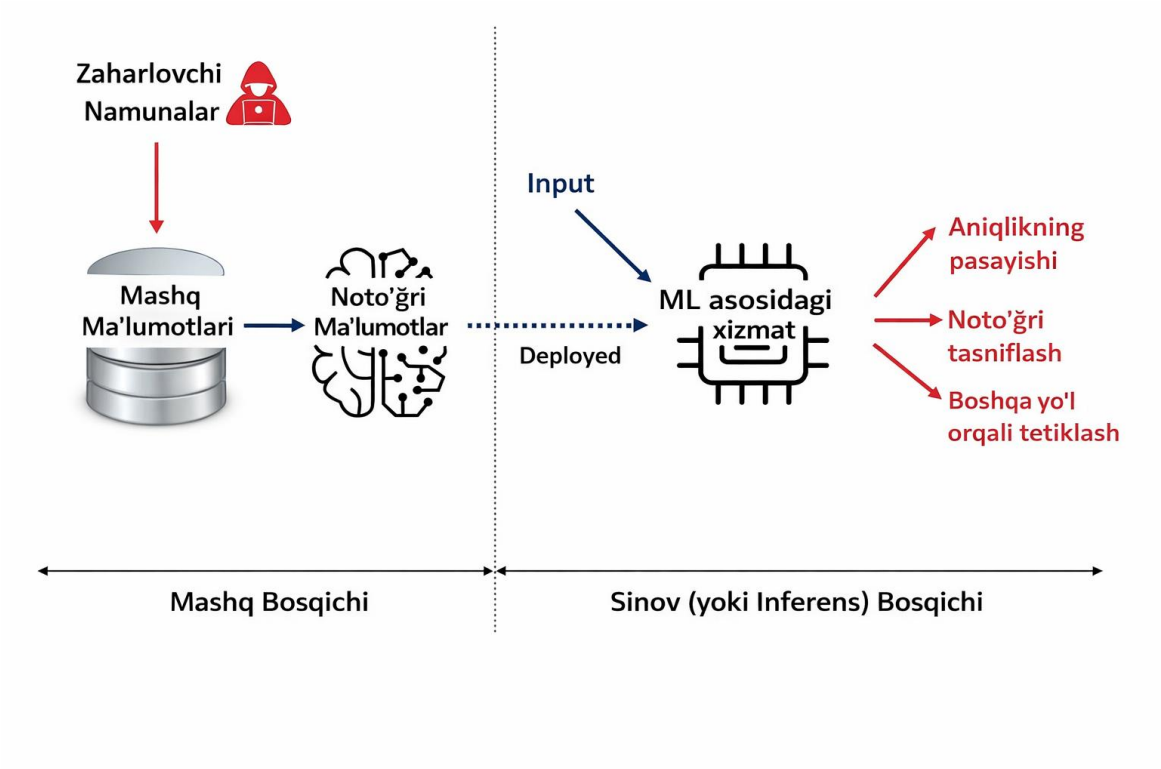
Kirish

So‘nggi yillarda sun’iy intellekt (Artificial Intelligence — AI) texnologiyalarining jadal rivojlanishi ularning turli sohalarda keng qo‘llanilishiga olib keldi. Hisoblash quvvatining ortishi, katta hajmdagi ma’lumotlarning paydo bo‘lishi va chuqur o‘rganish (deep learning) algoritmlarining takomillashuvi natijasida AI tizimlari tibbiyot, moliya, sanoat, transport va davlat boshqaruvi kabi strategik yo‘nalishlarda muhim vositaga aylanmoqda. Bunday tizimlar jarayonlarni tezlashtirish, inson omilidan kelib chiqadigan xatolarni kamaytirish va boshqaruv samaradorligini oshirish imkonini bermoqda.

Shu bilan birga, sun’iy intellektdan foydalanish ko‘lamining kengayishi xavfsizlik bilan bog‘liq yangi muammolarni yuzaga chiqarmoqda. AI tizimlarining ishlashi bevosita ma’lumotlarga bog‘liq bo‘lgani sababli, ularning ishonchliligi ko‘p jihatdan ma’lumotlar sifati va yaxlitligiga taqaladi. Amaliyotda esa turli zararli ta’sirlar natijasida tizimlar noto‘g‘ri qarorlar qabul qilishi mumkin. Xususan, chalg‘ituvchi hujumlar (adversarial attacks) yordamida kirish ma’lumotlari o‘zgartirilib, modelning xulosalari buziladi. Shuningdek, o‘quv ma’lumotlarini manipulyatsiya qilish (data poisoning) orqali tizimni noto‘g‘ri o‘rganishga majbur qilish mumkin. Bundan tashqari, maxfiy ma’lumotlarning sizib chiqishi (privacy leakage) hamda modelni o‘g‘irlash (model stealing) holatlari AI infratuzilmasining zaif tomonlarini ko‘rsatmoqda.

Ko‘plab holatlarda ishlab chiquvchilar modelning aniqligiga katta e’tibor berib, uning hujumlarga nisbatan barqarorligini (robustness) yetarli darajada baholamaydi. Natijada real muhitda yuqori samaradorlik ko‘rsatgan tizimlar favqulodda vaziyatlarda ishdan chiqishi yoki kutilmagan natijalar berishi mumkin. Bu esa nafaqat texnik, balki ijtimoiy va iqtisodiy risklarni ham kuchaytiradi.

Shu sababli sun’iy intellekt xavfsizligini chuqur o‘rganish, mavjud tahdidlarni aniqlash va samarali himoya mexanizmlarini ishlab chiqish dolzarb masalalardan biri hisoblanadi. Mazkur yo‘nalishda olib boriladigan tadqiqotlar AI tizimlarining ishonchliligi va barqarorligini oshirish, ularni real sharoitlarda xavfsiz qo‘llash imkoniyatlarini kengaytirishga xizmat qiladi.



1-rasm. Sun'iy intellekt tizimidagi xatolik misoli

Sun'iy intellekt tizimlariga tahdidlar tahlili

Sun'iy intellekt tizimlarining keng qo'llanilishi ularning turli xil zararli ta'sirlarga nisbatan zaif tomonlarini aniqlash zaruratini yuzaga keltirdi. An'anaviy dasturiy tizimlardan farqli ravishda, AI modellarining xatti-harakati qat'iy yozilgan qoidalarga emas, balki o'rganilgan ma'lumotlarga asoslanadi. Shu sababli, hujumchilar ko'pincha tizim algoritmidan ko'ra ma'lumotlar bilan bog'liq jarayonlarga ta'sir ko'rsatishga harakat qiladi. Quyida amaliyotda eng ko'p uchraydigan tahdid turlari ko'rib chiqiladi.

Chalg'ituvchi hujumlar (adversarial attacks)

Chalg'ituvchi hujumlar AI tizimiga uzatilayotgan kirish ma'lumotiga juda kichik, ko'pincha inson sezmaydigan o'zgarishlar kiritish orqali modelni noto'g'ri qaror qabul qilishga majbur etadi. Bunday hujumlarning xavfliligi shundaki, tashqi tomondan ma'lumot deyarli o'zgarmagandek ko'rinadi, biroq model uni mutlaqo boshqacha talqin qiladi. Goodfellow va hamkorlari neyron tarmoqlarning bunday nozik manipulyatsiyalarga yuqori darajada sezgir ekanligini qayd etadi[1]. Masalan, tasvirni aniqlash tizimlarida obyektga qo'shilgan mayda shovqin (noise) natijasida model avtomobilni boshqa predmet sifatida qabul qilishi mumkin. Avtonom transport vositalarida bunday xatolik jiddiy avariyaviy vaziyatlarga sabab bo'lish ehtimolini oshiradi. Shu bois ushbu turdagi hujumlar AI xavfsizligining eng muhim muammolaridan biri sifatida qaraladi.

O'quv ma'lumotlarini manipulyatsiya qilish (data poisoning)

[1] Goodfellow I., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. ICLR, 2015.

Mazkur hujum turi modelni o‘qitish bosqichida amalga oshiriladi. Hujumchi trening ma’lumotlari tarkibiga ataylab noto‘g‘ri yoki chalg‘ituvchi namunalarni qo‘shadi. Natijada tizim noto‘g‘ri bog‘liqliklarni o‘rganadi va kelajakda xatoli qarorlar chiqaradi. Bunday manipulyatsiya ayniqsa ochiq ma’lumotlardan foydalanadigan yoki foydalanuvchi tomonidan to‘ldiriladigan tizimlarda osonroq amalga oshiriladi. Eng muhim jihati shundaki, muammo ko‘pincha model ishga tushgandan keyin seziladi, ammo uni keltirib chiqargan sababni aniqlash murakkab bo‘ladi.

Maxfiy ma’lumotlarning sizib chiqishi (privacy leakage)

AI modellar katta hajmdagi shaxsiy va sezgir ma’lumotlarni qayta ishlaydi. Maxsus tahlil usullari orqali model javoblaridan foydalanuvchi haqidagi ayrim ma’lumotlarni qayta tiklash mumkinligi aniqlangan. Bu esa tibbiy, moliyaviy yoki shaxsiy axborotlar bilan ishlaydigan tizimlar uchun katta xavf tug‘diradi. Natijada foydalanuvchi ishonchi kamayadi va tashkilotlar huquqiy javobgarlikka duch kelishi mumkin.

Modelni o‘g‘irlash (model stealing)

Ko‘plab AI tizimlari xizmat sifatida (AI-as-a-Service) taqdim etiladi. Hujumchi modelga ko‘plab so‘rovlar yuborish orqali uning ishlash mantiqini taxminan tiklashi va o‘xshash nusxa yaratishi mumkin. Bu esa intellektual mulkni himoya qilish hamda tijorat sirlarini saqlash bilan bog‘liq jiddiy muammolarni keltirib chiqaradi.

Ko‘rib chiqilgan tahdidlar shuni ko‘rsatadiki, sun’iy intellekt tizimlariga qarshi hujumlar ko‘pincha an’anaviy dasturiy xatoliklardan emas, balki modelning o‘rganish mexanizmlaridan foydalanish orqali amalga oshiriladi. Shu sababli xavfsizlikni ta’minlash jarayoni faqat dasturiy himoya bilan cheklanmasdan, ma’lumotlar hayotiy siklining barcha bosqichlarini qamrab olishi zarur.

1-jadval. Sun’iy intellekt tizimlariga qarshi asosiy tahdidlar va ularning oqibatlari

Tahdid turi	Ta’sir bosqichi	Oqibat
Chalg‘ituvchi hujum	Kirish ma’lumotlari	Noto‘g‘ri tasnif
Ma’lumot manipulyatsiyasi	Trening	Xato o‘rganish
Privacy leakage	Ekspluatatsiya	Maxfiylik buzilishi
Model stealing	API	Intellektual yo‘qotish

Sun’iy intellekt tizimlarini himoya qilish strategiyalari

Sun’iy intellekt tizimlariga nisbatan tahdidlarning murakkablashuvi ularni himoya qilish bo‘yicha kompleks va ko‘p qatlamli yondashuvni talab etadi. Amaliyot shuni ko‘rsatadiki, yagona himoya mexanizmi barcha turdagi hujumlarga qarshi yetarli darajada samarali bo‘la olmaydi. Shu sababli zamonaviy tadqiqotlarda ma’lumotlar bilan ishlash bosqichidan tortib modelni real muhitda ekspluatatsiya qilishgacha bo‘lgan jarayonlarning barchasida xavfsizlikni ta’minlash zarurligi ta’kidlanadi. Quyida AI xavfsizligini oshirishda keng qo‘llanilayotgan asosiy strategiyalar ko‘rib chiqiladi.

Barqaror o‘qitish (adversarial training)

Chalg‘ituvchi hujumlarga qarshi eng samarali usullardan biri bu modelni oldindan buzilgan yoki shovqin qo‘shilgan ma’lumotlar bilan o‘qitish hisoblanadi. Ushbu jarayon davomida tizim turli manipulyatsiyalarga duch keladi va ular bilan ishlashni “o‘rganadi”. Natijada model keyinchalik real hujumlarga nisbatan bardoshliligi (robustness) yuqori bo‘lgan holatda ishlaydi. Biroq bu usul katta hisoblash resurslarini talab qiladi hamda modelni o‘qitish vaqtini sezilarli darajada oshirishi mumkin.

Ma’lumotlar yaxlitligini nazorat qilish (data validation)

Trening va kirish ma’lumotlarini tekshirish AI xavfsizligining muhim qismi hisoblanadi. Maxsus filtratsiya mexanizmlari yordamida noodatiy yoki shubhali namunalar aniqlanadi va tizimga kiritilishidan oldin chiqarib tashlanadi. Bu usul o‘quv ma’lumotlarini manipulyatsiya qilish (data poisoning) ehtimolini kamaytirishga yordam beradi. Shuningdek, ma’lumot manbalarining ishonchliligini baholash va ularni sertifikatlash amaliyotda muhim ahamiyat kasb etmoqda.

Shifrlash va maxfiylikni himoya qilish (encryption & privacy protection)

AI tizimlari ko‘pincha sezgir axborot bilan ishlagani sababli ma’lumotlarni himoyalangan holda saqlash va uzatish talab etiladi. Shifrlash (encryption) usullari ma’lumotlardan ruxsatsiz foydalanishni cheklaydi. So‘nggi yillarda federativ o‘rganish (federated learning) va differensial maxfiylik (differential privacy) kabi texnologiyalar ma’lumotni markazlashtirmasdan modelni o‘qitish imkonini berib, maxfiylikni oshirishda muhim vositaga aylanmoqda.

Monitoring va anomaliyalarni aniqlash (monitoring & anomaly detection)

Model real muhitda ishlayotgan paytda uning xatti-harakatlarini doimiy kuzatib borish zarur. Agar tizim odatiy bo‘lmagan natijalar bera boshlasa, bu potentsial hujum belgisi sifatida qaraladi. Zamonaviy monitoring vositalari avtomatik ravishda shubhali faoliyatni aniqlab, administratorlarga ogohlantirish yuboradi.Bu yondashuv ayniqsa moliyaviy operatsiyalar yoki muhim infratuzilma bilan bog‘liq tizimlarda katta ahamiyatga ega.

Modelni himoyalash usullari (model protection)

Intellektual mulkni saqlash maqsadida modellarni noqonuniy nusxalashdan himoya qilish mexanizmlari ishlab chiqilmoqda. Masalan, watermarking texnikasi yordamida model ichiga yashirin identifikatsiya belgisi joylashtiriladi. Bu esa model nusxalangan taqdirda uning haqiqiy egasini aniqlash imkonini beradi.

Keltirilgan strategiyalar shuni ko‘rsatadiki, AI xavfsizligi alohida bitta vosita bilan emas, balki o‘zaro bog‘langan himoya qatlamlari orqali ta’minlanadi. Ma’lumot, algoritm va ekspluatatsiya bosqichlarining har biri nazorat ostida bo‘lishi zarur. Faqat shundagina tizimning ishonchliligini yuqori darajada ushlab turish mumkin.

Sun’iy intellekt tizimlarida qo‘llaniladigan himoya mexanizmlarining haqiqiy qiymatini aniqlash ularning amaliy sharoitlarda qay darajada samarali ishlashini baholash orqali amalga oshiriladi. Zamonaviy tadqiqotlarda xavfsizlik choralarini tekshirish uchun turli eksperimental modellar, test ma’lumotlari va maxsus hujum ssenariylaridan foydalaniladi. Bunday yondashuv tizimning real tahdidlarga nisbatan qanchalik barqaror ekanligini aniqlash imkonini beradi.

Himoya usullarining samaradorligi

AI xavfsizligini baholashda bir nechta asosiy ko‘rsatkichlar qo‘llaniladi. Eng keng tarqalgan mezonlardan biri aniqlik (accuracy) bo‘lib, u modelning to‘g‘ri javob berish darajasini ko‘rsatadi. Biroq xavfsizlik nuqtai nazaridan faqat yuqori aniqlik yetarli emas. Model chalg‘ituvchi hujumlar sharoitida ham o‘z natijalarini saqlab qolishi lozim. Shu sababli bardoshlilik (robustness) darajasi muhim mezon sifatida qaraladi.

Amaliy tajribalar shuni ko‘rsatadiki, oddiy sharoitda yaxshi ishlaydigan modellar adversarial ta’sir ostida sezilarli ravishda samaradorligini yo‘qotadi. Masalan, kirish ma’lumotiga juda kichik shovqin qo‘shilganda tasniflash aniqligi keskin pasayishi mumkin. Biroq adversarial training qo‘llanilgan modellar bunday sharoitlarda ancha barqaror natijalar namoyish etadi. Bu esa oldindan tayyorgarlik ko‘rish tizimni real hujumlarga nisbatan chidamli qilishini tasdiqlaydi.

Ma’lumotlar yaxlitligini tekshirish mexanizmlari ham o‘z samaradorligini ko‘rsatmoqda. Filtrlash va anomaliyalarni aniqlash algoritmlari yordamida zararli namunalar erta bosqichda aniqlanib, modelga yetib bormasdan oldin bartaraf etiladi. Natijada noto‘g‘ri o‘rganish xavfi kamayadi va tizimning umumiy ishonchliligi oshadi.

Maxfiylikni himoya qilishga qaratilgan usullar, jumladan federativ o‘rganish va differensial maxfiylik, foydalanuvchi ma’lumotlaridan to‘g‘ridan-to‘g‘ri foydalanmasdan ham yuqori sifatli modellar yaratish mumkinligini ko‘rsatmoqda. Bunday yondashuvlar ayniqsa tibbiyot va moliyaviy xizmatlar kabi sezgir sohalarda muhim ahamiyatga ega. Shu bilan birga, barcha himoya strategiyalari ma’lum xarajatlar bilan bog‘liq. Qo‘shimcha hisoblash quvvati, murakkab infratuzilma va monitoring tizimlarini joriy qilish zarurati ba’zi tashkilotlar uchun muammoli bo‘lishi

mumkin. Demak, xavfsizlik va samaradorlik o‘rtasida optimal muvozanatni topish dolzarb vazifa bo‘lib qolmoqda.

IBM Security va NIST tadqiqotlariga ko‘ra [6, 8], adversarial training qo‘llanilgan modellar ayrim hujumlarga nisbatan 20–40 foizgacha yuqori bardoshlilik ko‘rsatgan. Shuningdek, differensial maxfiylik va federativ o‘rganish texnologiyalari maxfiy ma'lumotlarning sizib chiqish xavfini sezilarli darajada kamaytirishi aniqlangan.

Kelajakdagi tadqiqot yo‘nalishlari

Olib borilgan tahlillar shuni ko‘rsatadiki, sun’iy intellekt tizimlarida xavfsizlikni ta’minlash ko‘p qirrali va doimiy rivojlanib boruvchi jarayon hisoblanadi. Mavjud himoya mexanizmlari ma’lum darajada samarali natijalar berayotgan bo‘lsa-da, hujum usullarining murakkablashuvi ularni muntazam ravishda takomillashtirib borishni talab qiladi. Ayniqsa, chalg‘ituvchi hujumlarning yangi avlodlari an’anaviy filtratsiya va nazorat mexanizmlarini chetlab o‘tish qobiliyatiga ega bo‘lib bormoqda. Tadqiqotlar shuni ko‘rsatadiki, modelni mustahkamlashga qaratilgan yondashuvlar ko‘pincha aniqlik va tezkorlikning pasayishiga olib keladi. Bu esa ishlab chiquvchilarni xavfsizlik va samaradorlik o‘rtasida tanlov qilishga majbur etadi. Real amaliyotda esa har ikki omil muhim bo‘lgani sababli optimal muvozanatni topish murakkab vazifa bo‘lib qolmoqda. Bundan tashqari, AI tizimlarining tobora ko‘proq avtonom qarorlar qabul qilishi javobgarlik masalasini ham kun tartibiga olib chiqmoqda. Agar tizim xato qilsa, mas’uliyat kim zimmasida bo‘lishi kerakligi hali ham ochiq savol bo‘lib qolmoqda. Shu sababli texnik himoya bilan bir qatorda huquqiy va etik me’yorlarni ham rivojlantirish zarur.

Kelajakdagi tadqiqotlar sun’iy intellektni o‘z-o‘zini himoya qila oladigan, ya’ni hujumni mustaqil ravishda aniqlab, unga mos ravishda javob qaytara oladigan tizimlar yaratishga qaratilmoqda. Bunda avtomatik monitoring, real vaqt rejimida xavfni baholash va moslashuvchan mudofaa mexanizmlari muhim rol o‘ynaydi. Shuningdek, izohlanadigan sun’iy intellekt (explainable AI) konsepsiyasining rivojlanishi model qarorlarini chuqurroq tushunish va potentsial xatolarni erta aniqlash imkonini beradi. Umuman olganda, AI xavfsizligi masalasi yakuniy yechimga ega emas, balki texnologiya rivojlangan sari yangilanib boradigan jarayondir. Shu bois ushbu yo‘nalishda muntazam ilmiy izlanishlar olib borish, yangi tahdidlarni prognoz qilish va ularga mos himoya strategiyalarini ishlab chiqish muhim ahamiyat kasb etadi.

LLM tizimlariga xos xavfsizlik tahdidlari

So‘nggi yillarda ChatGPT, Gemini, Claude va DeepSeek kabi katta til modellari (Large Language Models – LLM) keng qo‘llanila boshladi. Ushbu modellar an’anaviy mashinaviy o‘rganish tizimlaridan farqli ravishda yangi turdagi xavfsizlik muammolariga duch kelmoqda.

Prompt Injection hujumlarida foydalanuvchi modelga maxsus buyruqlar yuborib, uning dastlabki cheklovlarini chetlab o‘tishga harakat qiladi. Jailbreaking usullari esa modelni xavfsizlik siyosatlariga zid javoblar berishga majbur qilishi mumkin.

Bundan tashqari, Model Backdoor Attack hujumlari davomida model o‘qitish bosqichida yashirin zararli xatti-harakatlar bilan ta’minlanadi. Natijada ma’lum shartlar bajarilganda model noto‘g‘ri qaror qabul qiladi. Shu sababli LLM tizimlari uchun maxsus monitoring va himoya mexanizmlarini ishlab chiqish zamonaviy AI xavfsizligining muhim yo‘nalishlaridan biri hisoblanadi.

2-jadval. Sun’iy intellekt tizimlariga qarshi asosiy hujumlar va ularning xususiyatlari

Hujum turi	Ta’sir qiluvchi bosqich	Asosiy maqsad	Xavf darajasi	Himoya usuli
Adversarial Attack	Inference	Noto‘g‘ri tasniflash	Yuqori	Adversarial Training
Data Poisoning	Training	Noto‘g‘ri o‘qitish	Juda yuqori	Data Validation
Privacy Leakage	Deployment	Maxfiy ma’lumotlarni olish	Yuqori	Differential Privacy
Model Stealing	API Access	Model nusxasini yaratish	O‘rta	Watermarking
Prompt Injection	LLM Input	Modelni boshqarish	Yuqori	Input Filtering
Backdoor Attack	Training	Yashirin zararli xatti-harakat	Juda yuqori	Dataset Auditing

Xulosa

Sun’iy intellekt texnologiyalarining jadal rivojlanishi ularni iqtisodiyot va jamiyatning turli sohalarida muhim qaror qabul qiluvchi vositaga aylantirdi. Bunday tizimlar samaradorlikni oshirish, jarayonlarni optimallashtirish va inson omilidan kelib chiqadigan xatolarni kamaytirish imkonini bermoqda. Shu bilan birga, AI infratuzilmasining kengayishi uning ishonchliligi va himoyalanganligini ta’minlash masalasini ustuvor vazifaga aylantirmoqda.

Bayon etilgan tahdidlar shuni ko‘rsatadiki, sun’iy intellekt modellarining zaifligi ko‘pincha ularning ma’lumotlarga bog‘liq tabiati bilan izohlanadi. Kirish signallariga kichik o‘zgartirishlar kiritish, o‘quv jarayoniga aralashish yoki tizimdan bilvosita

ma’lumot ajratib olish imkoniyati real sharoitda jiddiy oqibatlarga olib kelishi mumkin. Ayniqsa sog‘liqni saqlash, moliyaviy xizmatlar va transport tizimlarida bunday xatarlar inson xavfsizligiga bevosita ta’sir qiladi.

Keltirilgan himoya strategiyalari sun’iy intellekt xavfsizligi ko‘p qatlamli yondashuvni talab qilishini tasdiqlaydi. Ma’lumotlarni tekshirish, modelni barqarorlashtirish, maxfiylikni saqlash va tizim faoliyatini uzluksiz monitoring qilish birgalikda qo‘llanilganda yuqori natijaga erishish mumkin. Shu bilan birga, xavfsizlik choralarini joriy etish resurs sarfi va texnik murakkablikni oshirishi sababli amaliyotda optimal yechimni tanlash muhim ahamiyatga ega.

Kelgusida sun’iy intellekt tizimlari yanada murakkablashib, mustaqil qaror qabul qilish darajasi ortib boradi. Bu esa xavfsizlik talablarini yanada kuchaytiradi va yangi himoya mexanizmlarini ishlab chiqishni zarur qiladi. Shu nuqtai nazardan, mazkur yo‘nalishdagi ilmiy izlanishlar va innovatsion yondashuvlar AI texnologiyalarining barqaror, ishonchli va mas’uliyatli rivojlanishini ta’minlashda muhim rol o‘ynaydi.

Foydalanilgan adabiyotlar:

1. Goodfellow I., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples. ICLR, 2015, 1–11-betlar.
2. Papernot N., McDaniel P., Jha S., et al. The Limitations of Deep Learning in Adversarial Settings. IEEE European Symposium on Security and Privacy, 2016, 372–387-betlar.
3. Biggio B., Roli F. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition, 2018, 317–331-betlar.
4. Carlini N., Wagner D. Towards Evaluating the Robustness of Neural Networks. IEEE Symposium on Security and Privacy, 2017, 39–57-betlar.
5. Goldblum M., Tsipras D., Xie C., et al. Data Security for Machine Learning: Challenges and Opportunities. IEEE Security & Privacy, 2020, 20–29-betlar. https://arxiv.org/abs/2007.04177
6. NIST. AI Risk Management Framework (AI RMF 1.0), 2023.
7. OWASP Top 10 for LLM Applications, 2025.
8. IBM AI Security and Trust Report, 2024.
9. Huang L., Joseph A.D., Nelson B., Rubinstein B.I.P., Tygar J.D. Adversarial Machine Learning. ACM Workshop on Security and Artificial Intelligence, 2011.
10. Papernot N., McDaniel P., Goodfellow I. Practical Black-Box Attacks against Machine Learning. ACM AsiaCCS, 2017.