SUN'IY INTELLEKT TIZIMLARIDA ISHONCHLILIK, ETIK MUAMMOLAR VA XAVFSIZLIK: ZAMONAVIY YONDASHUVLAR VA RISKLARNI BOSHQARISH MEXANIZMLARI

RELIABILITY, ETHICAL ISSUES, AND SECURITY IN ARTIFICIAL INTELLIGENCE SYSTEMS: MODERN APPROACHES AND RISK MANAGEMENT MECHANISMS

НАДЁЖНОСТЬ, ЭТИЧЕСКИЕ ПРОБЛЕМЫ И БЕЗОПАСНОСТЬ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: СОВРЕМЕННЫЕ ПОДХОДЫ И МЕХАНИЗМЫ УПРАВЛЕНИЯ РИСКАМИ

FAN TARMOG'I
13.00.00 – Pedagogika

QABUL QILINDI
06.08.2026

CHOP ETILDI
09.08.2026

Ilhom Dustnazarovich G'aniyev
Chirchiq davlat pedagogika universiteti dotsenti v.b., pedagogika fanlari bo'yicha falsafa doktori (PhD)

ANNOTATSIYA

Mazkur maqolada sun'iy intellekt tizimlarining ishonchliligi, etik muammolari va xavfsizligini ta'minlashning zamonaviy yondashuvlari tahlil qilingan. Algoritmik tarafkashlik, tushuntiriladigan sun'iy intellekt (XAI), qaror qabul qilishning shaffofligi hamda adversarial tahdidlarga qarshi himoya mexanizmlari yoritilgan. Shuningdek, ma'lumotlar sifati, risklarni boshqarish, xalqaro etik tamoyillar va normativ-huquqiy yondashuvlarning sun'iy intellekt tizimlari ishonchliligini oshirishdagi o'rni asoslab berilgan. Tadqiqot natijalari inson markazli, xavfsiz va mas'uliyatli sun'iy intellekt tizimlarini yaratishda kompleks yondashuv muhim ekanligini ko'rsatadi.

Kalit so'zlar: sun'iy intellekt, ishonchli sun'iy intellekt (Trustworthy AI), algoritmik tarafkashlik, tushuntiriladigan sun'iy intellekt (Explainable AI, XAI), sun'iy intellekt xavfsizligi, etik tamoyillar, ma'lumotlar boshqaruvi, qaror qabul qilishning shaffofligi, adversarial hujumlar, risklarni boshqarish.

ABSTRACT

This article analyzes contemporary approaches to ensuring the reliability, ethics, and security of artificial intelligence (AI) systems. It examines issues related to algorithmic bias, Explainable Artificial Intelligence (XAI), transparency in decision-making, and protection mechanisms against adversarial attacks. In addition, the study highlights the role of data quality, risk management, international ethical principles, and regulatory frameworks in improving the reliability of AI systems. The research

findings demonstrate that a comprehensive approach is essential for developing human-centered, secure, and responsible artificial intelligence systems.

Keywords: artificial intelligence, Trustworthy AI, algorithmic bias, Explainable Artificial Intelligence (XAI), AI security, ethical principles, data governance, decision-making transparency, adversarial attacks, risk management.

АННОТАЦИЯ

В данной статье проанализированы современные подходы к обеспечению надёжности, этики и безопасности систем искусственного интеллекта. Рассмотрены вопросы алгоритмической предвзятости, объяснимого искусственного интеллекта (Explainable AI, XAI), прозрачности принятия решений, а также механизмов защиты от состязательных (adversarial) атак. Кроме того, обоснована роль качества данных, управления рисками, международных этических принципов и нормативно-правовых подходов в повышении надёжности систем искусственного интеллекта. Результаты исследования показывают, что комплексный подход является важным условием создания человекоориентированных, безопасных и ответственных систем искусственного интеллекта.

Ключевые слова: искусственный интеллект, доверенный искусственный интеллект (Trustworthy AI), алгоритмическая предвзятость, объяснимый искусственный интеллект (Explainable AI, XAI), безопасность искусственного интеллекта, этические принципы, управление данными, прозрачность принятия решений, состязательные атаки (adversarial attacks), управление рисками.

KIRISH

Ushbu maqolada sun’iy intellekt (SI) tizimlarining ishonchliligi, etik muammolari va xavfsizlik masalalari zamonaviy raqamli muhit sharoitida tahlil qilinadi. Algoritmik tarafkashlik, tushuntiriladigan sun’iy intellekt (Explainable AI), qaror qabul qilish jarayonining shaffofligi hamda adversarial hujumlarga nisbatan zaifliklar asosiy e’tibor markazida bo‘ladi. Ma’lumotlar sifati, model arxitekturasi va tibbiyot, moliya, kiberxavfsizlik kabi yuqori xavfli sohalarda avtomatlashtirilgan qarorlar bilan bog‘liq risklar o‘rtasidagi bog‘liqlik yoritiladi. Shuningdek, zamonaviy etik tamoyillar, normativ-huquqiy yondashuvlar va texnik risklarni kamaytirish strategiyalari tahlil qilinadi. Natijalar inson markazli va ishonchli SI tizimlarini ishlab chiqishda kompleks va fanlararo yondashuv muhimligini ko‘rsatadi.

ASOSIY QISM

Sun’iy intellekt (SI) texnologiyalari XXI asrning eng tez rivojlanayotgan va jamiyat taraqqiyotiga bevosita ta’sir ko‘rsatayotgan innovatsion yo‘nalishlaridan biri

hisoblanadi. Katta hajmdagi ma’lumotlarni qayta ishlash, murakkab naqsh va qonuniyatlarni aniqlash hamda inson aralashuvisiz yoki minimal nazorat ostida qaror qabul qilish imkoniyatlari tufayli SI tizimlari bugungi kunda tibbiyot, moliya, sanoat, ta’lim, kiberxavfsizlik va davlat boshqaruvi kabi strategik ahamiyatga ega sohalarda keng qo‘llanilmoqda. Xususan, chuqur o‘rganish (deep learning), katta til modellari (Large Language Models) va generativ sun’iy intellekt yechimlari diagnostika aniqligini oshirish, moliyaviy risklarni prognozlash, ishlab chiqarish jarayonlarini optimallashtirish hamda murakkab analitik vazifalarni bajarishda inson imkoniyatlaridan ustun natijalarni namoyon etmoqda. Shu bilan birga, SI texnologiyalarining jadal rivojlanishi va keng miqyosda joriy etilishi ularning ishonchliligi, xavfsizligi va etik oqibatlari bilan bog‘liq muammolarni yanada dolzarb qilib qo‘ymoqda. Avtomatlashtirilgan qaror qabul qilish tizimlari inson hayoti, sog‘ligi, ijtimoiy mavqei yoki huquqiy holatiga bevosita ta’sir ko‘rsatadigan sohalarda qo‘llanilganda, noto‘g‘ri, tushuntirilmaydigan yoki tarafkash qarorlar jiddiy ijtimoiy va huquqiy salbiy oqibatlarga olib kelishi mumkin. Masalan, tibbiy diagnostikadagi xatolar, kredit berishdagi adolatsizlik yoki sud tizimidagi noto‘g‘ri prognozlar inson huquqlarining buzilishiga sabab bo‘lishi ehtimoli mavjud.

So‘nggi yillarda ilmiy adabiyotlarda va xalqaro siyosiy-huquqiy hujjatlarda “Trustworthy AI” – ishonchga loyiq sun’iy intellekt konsepsiyasi markaziy o‘ringa chiqdi. Ushbu yondashuv SI tizimlarining nafaqat yuqori aniqlik va samaradorlikka ega bo‘lishini, balki shaffoflik, adolat, xavfsizlik, tushuntiriluvchanlik va inson nazoratiga ochiqlik kabi tamoyillarga mos kelishini talab qiladi. Yevropa Ittifoqining AI Act hujjati, UNESCO va OECD tomonidan ishlab chiqilgan etik tavsiyalar ushbu konsepsiyaning global miqyosda normativ asosga ega bo‘layotganini ko‘rsatadi.

Algoritmik tarafkashlik (bias), “black-box” muammosi, adversarial hujumlarga nisbatan zaiflik hamda shaxsiy ma’lumotlar xavfsizligi bilan bog‘liq xatarlar SI tizimlariga bo‘lgan ijtimoiy ishonchni sezilarli darajada pasaytirishi mumkin. Bundan tashqari, yetarli nazorat va audit mexanizmlarining mavjud emasligi AI texnologiyalarining noto‘g‘ri yoki zararli qo‘llanilishiga olib kelish ehtimolini oshiradi. Shu nuqtai nazardan, SI tizimlarining ishonchliligini ta’minlash masalasi faqat texnik vazifa emas, balki kompleks etik, huquqiy va ijtimoiy muammo sifatida qaralmoqda. Mazkur maqola sun’iy intellekt tizimlarining ishonchliligini belgilovchi asosiy texnik, etik va xavfsizlik omillarini zamonaviy ilmiy tadqiqotlar va amaliy faktlar asosida kompleks tarzda tahlil qilishga qaratilgan. Tadqiqot natijalari inson markazli, xavfsiz va ishonchga loyiq SI tizimlarini ishlab chiqish uchun muhim nazariy va amaliy xulosalarni shakllantirishga xizmat qiladi.

Sun’iy intellekt tizimlarining qaror qabul qilish jarayoni bevosita o‘qitish (trening) ma’lumotlarining sifati, tarkibi va tuzilishiga bog‘liq. Agar trening datasetlari tarixiy notenglik, ijtimoiy stereotiplar yoki muayyan guruhlarning yetarli

darajada vakillik qilinmasligini aks ettirsa, SI modeli ushbu tarafkashliklarni o‘zlashtiradi va amaliy qo‘llanilish jarayonida ularni takrorlashi yoki yanada kuchaytirishi mumkin. Kredit skoringi, ishga qabul qilish jarayonlari hamda sud-huquq tizimlarida qo‘llanilayotgan algoritmlarda ayrim ijtimoiy guruhlar uchun adolatsiz qarorlar qayd etilgani bu muammoning dolzarbligini tasdiqlaydi.

So‘nggi ilmiy tadqiqotlar algoritmik tarafkashlik faqat ma’lumotlar bilan cheklanib qolmasligini, balki model arxitekturasi, xususiyatlar (feature) tanlovi, optimallashtirish strategiyalari va baholash metrikalarida ham shakllanishi mumkinligini ko‘rsatmoqda. Shu sababli zamonaviy SI tizimlarini ishlab chiqishda fairness-aware learning, balanslangan va diversifikatsiyalangan datasetlar, ma’lumotlar boshqaruvi siyosatlari (data governance), shuningdek, sotsiotechnik yondashuv va doimiy audit mexanizmlarini joriy etish muhim ahamiyat kasb etadi. Ushbu choralar SI tizimlarining adolatli, ishonchli va ijtimoiy jihatdan maqbul bo‘lishini ta’minlashga xizmat qiladi.

Tushuntiriladigan sun’iy intellekt va qarorlarning shaffofligi. Chuqur neyron tarmoqlar va murakkab mashinali o‘rganish modellarining yuqori aniqligi ko‘pincha ularning ichki ishlash mexanizmlarining inson uchun tushunarsizligi bilan birga keladi. Ushbu “black-box” muammosi tibbiy diagnostika, moliyaviy risklarni baholash va davlat boshqaruvi kabi yuqori mas’uliyatli sohalarda jiddiy xavf tug‘diradi, chunki qarorlarning sabablarini izohlash imkoniyati cheklangan bo‘ladi.

Explainable AI (XAI) yondashuvlari ushbu muammoni bartaraf etishga qaratilgan bo‘lib, LIME, SHAP, attention mexanizmlari va kontrfaktual tushuntirishlar kabi usullar model qarorlariga qaysi omillar ta’sir ko‘rsatganini aniqlash imkonini beradi. XAI texnologiyalari nafaqat foydalanuvchi va mutaxassislar ishonchini oshiradi, balki modellarni tekshirish, xatoliklarni aniqlash, huquqiy javobgarlikni ta’minlash hamda etik me’yorlarga rioya etilishini nazorat qilishda muhim vosita hisoblanadi. Shu bois ko‘plab regulyatorlar SI tizimlarida tushuntiriluvchanlikni majburiy talab sifatida ilgari surmoqda.

Xavfsizlik muammolari va adversarial tahdidlar. Zamonaviy sun’iy intellekt tizimlari turli xil kiberxavfsizlik tahdidlariga, xususan adversarial hujumlarga nisbatan zaif bo‘lishi mumkin. Bunday hujumlar kiritma ma’lumotlariga juda kichik va inson ko‘ziga sezilmas o‘zgarishlar kiritish orqali modelni noto‘g‘ri yoki zararli qarorlar qabul qilishga undaydi. Avtonom transport vositalari, biometrik identifikatsiya tizimlari va kiberxavfsizlikda qo‘llaniladigan SI modellarida bunday hujumlar real xavf sifatida baholanmoqda.

Xavfsiz SI tizimlarini yaratish maqsadida adversarial training, model robustligini baholash, stress-testlar, doimiy monitoring va real muhitda sinovdan o‘tkazish strategiyalari keng qo‘llanilmoqda. So‘nggi xalqaro hisobotlar va ekspert

xulosalari ushbu choralarni SI tizimlarini ishlab chiqish va joriy etish jarayonida majburiy xavfsizlik standarti sifatida joriy etish zarurligini ta’kidlamoqda.

Etik va normativ-huquqiy yondashuvlar. Sun’iy intellekt texnologiyalarining jamiyatga ko‘rsatadigan ta’siri global miqyosda e’tirof etilib, UNESCO, OECD va Yevropa Ittifoqi kabi xalqaro tashkilotlar tomonidan etik va normativ-huquqiy hujjatlar ishlab chiqilmoqda. Xususan, Yevropa Ittifoqining AI Act hujjati SI tizimlarini xavf darajasiga ko‘ra tasniflash, shaffoflikni ta’minlash, inson nazoratini joriy etish va ishlab chiquvchilar hamda foydalanuvchilarning javobgarligini belgilashni nazarda tutadi.

Mazkur yondashuvlar texnologik innovatsiyalarni rag‘batlantirish bilan bir qatorda, ijtimoiy mas’uliyat va inson huquqlarini himoya qilish muhimligini ko‘rsatadi. Shu bilan birga, ko‘plab mamlakatlarda milliy darajada ham SI tizimlarini huquqiy tartibga solish, etik kodekslar ishlab chiqish va davlat nazoratini kuchaytirishga qaratilgan ilmiy-amaliy tadqiqotlar faollashmoqda.

XULOSA

Sun’iy intellekt texnologiyalari zamonaviy jamiyatning raqamli transformatsiyasida muhim va hal qiluvchi rol o‘ynamoqda. Biroq ushbu texnologiyalarning samaradorligi va keng ijtimoiy qabul qilinishi sun’iy intellekt tizimlarining ishonchliligi, xavfsizligi hamda etik me’yorlarga muvofiqligi bilan bevosita bog‘liq. Algoritmik tarafkashlik, qarorlarning tushuntirilmasligi va adversarial xavfsizlik tahdidlari yetarlicha nazorat qilinmasa, SI tizimlari ijtimoiy tengsizlikni kuchaytirishi, inson huquqlariga zarar yetkazishi va texnologiyalarga bo‘lgan jamoatchilik ishonchini pasaytirishi mumkin.

Mazkur tadqiqot natijalari shuni ko‘rsatadiki, sun’iy intellekt tizimlarining ishonchliligini ta’minlash faqatgina yuqori aniqlikka ega algoritmlarni ishlab chiqish bilan cheklanib qolmasligi kerak. Bunda ma’lumotlar sifati va boshqaruvi, algoritmik adolat, tushuntiriladigan sun’iy intellekt yondashuvlari, xavfsizlikni mustahkamlovchi texnik choralar hamda doimiy audit va monitoring mexanizmlari o‘zaro uyg‘un holda joriy etilishi lozim. Ayniqsa, yuqori xavfli sohalarda inson nazorati (human-in-the-loop) tamoyilini qo‘llash va qaror qabul qilish jarayonida javobgarlikni aniq belgilash muhim ahamiyat kasb etadi.

Shu sababli kelajakdagi ilmiy tadqiqotlar va amaliy ishlanmalar texnik aniqlik, etik mas’uliyat va huquqiy boshqaruvni yagona tizim sifatida ko‘rib chiqishi zarur. Fanlararo hamkorlikka asoslangan yondashuv – ya’ni axborot texnologiyalari, huquqshunoslik, sotsiologiya va etikani integratsiyalash - ishonchli va mas’uliyatli sun’iy intellekt tizimlarini yaratishda muhim omil bo‘lib xizmat qiladi.

Xulosa qilib aytganda, inson markazli, shaffof va xavfsiz sun’iy intellekt tizimlarini ishlab chiqish nafaqat texnologik taraqqiyotning muhim sharti, balki

barqaror, adolatli va xavfsiz raqamli jamiyatni shakllantirishning asosiy omillaridan biridir. Ushbu yo‘nalishda olib boriladigan izchil ilmiy tadqiqotlar va puxta boshqaruv mexanizmlari sun’iy intellekt texnologiyalarining jamiyat manfaatlariga xizmat qilishini ta’minlaydi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. Russell S., Norvig P. *Artificial Intelligence: A Modern Approach.* Pearson, 2021.
2. Floridi L., Cowls J. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 2019.
3. European Commission. *Ethics Guidelines for Trustworthy AI*. Brussels, 2020.
4. Goodfellow I., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *ICLR*, 2015.
5. Barocas S., Hardt M., Narayanan A. *Fairness and Machine Learning.* MIT Press, 2023.
6. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017.
7. Bostrom N. *Artificial Intelligence: Stages, Threats, Strategies.* Moscow, 2022.
8. UNESCO. *Recommendation on the Ethics of Artificial Intelligence.* Paris, 2021.
9. OECD. *Artificial Intelligence and Trust.* OECD Publishing, 2022.