Süni İntellekt Əsaslı Sistemlərin və Texnologiyaların Rəqəmsal Ekspertizası Problemləri

Ramiz Şıxəliyev
İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
shikhramiz61@gmail.com

***Xülasə*— Məqalədə süni intellekt sistemlərinin rəqəmsal ekspertizası sistemli və inteqrativ yanaşma əsasında analiz edilmiş, modellərin "qara qutu" xarakteri, izah edilə bilməmə xüsusiyyəti, qeyri-deterministik davranış və təlim verilənlərindən asılılıq kimi xüsusiyyətləri araşdırılmışdır. Bununla yanaşı, izah edilə bilən süni intellekt yanaşmaları, təhlükəsizlik riskləri, audit mexanizmləri, standartlaşdırma məsələləri, həmçinin hüquqi və etik aspektlər kompleks şəkildə analiz edilmişdir.**

Açar sözlər*— *süni intellekt; rəqəmsal ekspertiza; izah edilə bilən süni intellekt; audit; etik prinsiplər.

I. GİRİŞ

Süni intellekt (Sİ) sistemləri son illərdə səhiyyə, maliyyə, kibertəhlükəsizlik və dövlət idarəçiliyi kimi müxtəlif sahələrdə geniş tətbiq olunmağa başlamışdır [1]. Maşın təlimi, xüsusilə dərin təlim metodları əsasında işləyən bu sistemlər böyük həcmli verilənlər üzərində mürəkkəb nümunələri aşkar edərək qərar qəbul edir [2].

Hazırda Sİ sistemləri yalnız texniki alət deyil, həm də hüquqi və sosial nəticələr doğuran qərarvermə mexanizminə çevrilmişdir [3, 4]. Bu sistemlərin qərar qəbuletmə prosesində artan rolu onların etibarlılığı, şəffaflığı və təhlükəsizliyi ilə bağlı ciddi suallar doğurur [5]. Xüsusilə, mürəkkəb və qeyri-şəffaf modellərin istifadəsi nəticəsində qəbul edilən qərarların izah olunması və yoxlanılması çətinləşir [6]. Bu çətinliklər eyni zamanda Sİ sistemlərinin qəsdən manipulyativ hücumlara (adversarial attacks), məxfilik pozuntularına və təlim verilənlərindəki qərəzlərə qarşı həssaslığını artırır [7, 8]. Bütün bu hallar Sİ sistemlərinin fəaliyyətinin rəqəmsal ekspertiza baxımından araşdırılmasını zəruri edir.

Rəqəmsal ekspertiza kontekstində Sİ sistemlərinin təhlili, onların necə işlədiyini, hansı məlumatlara əsaslandığını və nəticələrin nə dərəcədə etibarlı olduğunu müəyyən etməyə yönəlir [9]. Bu yanaşma ənənəvi rəqəmsal kriminalistikadan fərqli olaraq yalnız verilənlərin (loq fayllar, metaverilənlər) deyil, həm də modelin özünün, onun təlim prosesinin və dinamik qərarvermə mexanizmlərinin analizini tələb edir [9, 10]. Xüsusilə böyük dil modelləri (LLM) tərəfindən yaradılan rəqəmsal sübutların etibarlılığının qiymətləndirilməsi əlavə metodoloji protokollar tələb edir.

Məqalənin məqsədi Sİ sistemlərinin rəqəmsal ekspertizasının nəzəri və praktiki aspektlərinin, onların ekspertizasını çətinləşdirən qeyri-determinizm, izaholunmazlıq, təlim verilənlərindən asılılıq və şəffaflıq kimi xüsusiyyətlərin, həmçinin bu sistemlərə qarşı mövcud təhlükələr və onların ekspertiza nəticələrinə təsirinin araşdırılmasıdır. Məqalədə həmçinin Sİ sistemlərinin audit və standartlaşdırma yanaşmaları araşdırılmışdır.

II. SÜNİ İNTELLEKT SİSTEMLƏRİNİN XÜSUSİYYƏTLƏRİ

Müasir süni intellekt sistemlərinin böyük əksəriyyəti maşın təlimi və xüsusilə dərin təlim yanaşmalarına əsaslanır. Bu modellər böyük həcmli və çoxölçülü verilənlər üzərində təlim keçərək mürəkkəb nümunələri, gizli əlaqələri və statistik asılılıqları aşkar etmə qabiliyyətinə malikdir [11]. Bu baxımdan onlar ənənəvi proqramlaşdırma yanaşmalarından fərqli olaraq, əvvəlcədən müəyyən edilmiş qaydalarla deyil, verilənlər əsasında öyrədilən nümunələr əsasında qərar qəbul edir. Nəticədə, modellərin məhsuldarlığı birbaşa olaraq təlim verilənlərinin keyfiyyəti, müxtəlifliyi və təmsilçilik dərəcəsi ilə bağlı olur.

Bu sistemlərin əsas fərqləndirici xüsusiyyətlərindən biri onların "qara qutu" xarakterli olmasıdır. Bu xüsusiyyət modelin daxili strukturunun insan tərəfindən intuitiv və şəffaf şəkildə izah olunmasının çətinliyi ilə xarakterizə olunur [6, 12]. Xüsusilə dərin neyron şəbəkələrində məlumatların laylar üzrə transformasiyası yüksək səviyyədə abstraksiyaya malik olduğundan, giriş verilənlərinin çıxış qərarlarına necə təsir etdiyini dəqiq izləmək çətinləşir. Bu isə izaholunma problemlərini ön plana çıxarır və izah edilə bilən süni intellekt (XAI) metodlarının inkişafını zəruri edir.

Dərin təlim arxitekturaları, məsələn, konvolyusiyalı neyron şəbəkələr, təkrarlanan neyron şəbəkələr və transformer əsaslı modellər çoxsaylı parametr və qeyri-xətti transformasiyalar vasitəsilə mürəkkəb funksiyaları aproksimasiya edir [2, 13]. Bu cür arxitekturalar yüksək dəqiqlik və ümumiləşdirməni təmin etsə də, onların struktur mürəkkəbliyi analizi, izahı və xüsusilə tərs mühəndisliyi (reverse engineering) çətinləşdirir. Bu xüsusiyyətlər rəqəmsal ekspertiza kontekstində modelin davranışının izahı və hadisələrin bərpası baxımından əlavə çətinliklər yaradır.

Bundan əlavə, müasir Sİ modellərinin təlimində geniş şəkildə stoxastik (ehtimal əsaslı) optimizasiya alqoritmlərindən, məsələn, stoxastik qradiyent enişindən istifadə olunur. Bu isə modellərin qeyri-determinizm xüsusiyyətini formalaşdırır: eyni hiperparametrlər və təlim verilənləri ilə fərqli təlim sessiyalarında model müxtəlif lokal minimumlara yaxınlaşa və fərqli daxili reprezentasiyalar formalaşdıra bilər [7, 14]. Nəticədə, model davranışının təkrarlanması çətinləşir ki, bu da xüsusilə hüquqi və kriminalistik analizlərdə mühüm problemdir.

Digər mühüm xüsusiyyət isə modellərin təlim verilənlərində

DOI: 10.25045/forensic.gov.2026.36

mövcud olan gizli qərəzləri mənimsəməsi və bu qərəzlərin qərarvermə prosesinə ötürülməsidir. Əgər təlim verilənləri müəyyən sosial, iqtisadi və ya demoqrafik qrupları qeyri-adekvat şəkildə təmsil edirsə, model nəticələrində ayrı-seçkilik və ədalətsizlik halları yarana bilər [15]. Bu isə Sİ sistemlərinin etibarlılığına və tətbiq olunduğu sahələrdə, xüsusilə səhiyyə, maliyyə və hüquq kimi kritik sahələrdə qəbul edilən qərarların qanuniliyinə birbaşa təsir göstərir.

Müasir Sİ sistemləri yüksək adaptivlik və dinamik təlim qabiliyyətinə malikdirlər. Bəzi modellər davamlı təlim və ya onlayn təlim mexanizmləri vasitəsilə yeni verilənlərə uyğunlaşa bilir. Bu isə sistemin zamanla dəyişən mühitlərdə effektivliyini artırsa da, eyni zamanda modelin davranışının sabitliyini və audit olunmasını çətinləşdirir. Beləliklə, Sİ sistemlərinin xüsusiyyətləri onların həm texniki analizini, həm də rəqəmsal ekspertizasını mürəkkəb və çoxsəviyyəli problemə çevirir.

III. SÜNİ İNTELLEKT SİSTEMLƏRİNİN RƏQƏMSAL EKSPERTİZASI

Sİ sistemlərinin rəqəmsal ekspertizası bu sistemlərin fəaliyyət mexanizmlərinin araşdırılması, baş vermiş hadisələrin bərpası və modelin qərarvermə prosesindəki səbəb-nəticə əlaqələrinin müəyyən edilməsini əhatə edən kompleks elmi-praktiki sahədir [5]. Bu yanaşma vasitəsilə modelin qərarlarının formalaşma prinsipləri, giriş xüsusiyyətlərinin çıxışa təsiri və müxtəlif şəraitlərdə səhv və ya qərəzli nəticələrin yaranma səbəbləri müəyyən olunur [5, 9]. Beləliklə, Sİ ekspertizası yalnız nəticələrin təsviri ilə kifayətlənməyərək, qərarvermənin daxili mexanizmlərinin izahını da təmin edir.

Ənənəvi rəqəmsal kriminalistika ilə müqayisədə Sİ ekspertizası daha geniş və dinamik analizi əhatə edir. Klassik yanaşmalar əsasən statik verilənlərin (loq faylları, metaverilən və şəbəkə trafiki) bərpası və təhlilinə yönəldiyi halda, Sİ ekspertizası əlavə olaraq modelin daxili vəziyyətlərini, təlim verilənlərinin təsirini və dinamik qərarvermə davranışını araşdırır [9, 10].

Müasir yanaşmaya əsasən, Sİ sistemlərinin rəqəmsal ekspertizası üç əsas mərhələdən ibarətdir: sübutların əldə edilməsi, analizi və qiymətləndirilməsi [9]. Bu mərhələlər çərçivəsində müxtəlif tipli məlumat mənbələri potensial sübut kimi çıxış edir. Bunlara model çıxışları (təsnifat ehtimalları, reqressiya nəticələri və generativ cavablar), sistem loq faylları (çağırış vaxtları, giriş verilənləri, istifadəçi sessiyaları və API sorğuları), metaməlumatlar (model versiyası, hiperparametrlər, təlim tarixi), təlim verilənlərinin nümunələri və modelin daxili təsvirləri (məsələn, neyron şəbəkələrin ara qat aktivasiyaları) daxildir [2, 9, 16].

Sübutların toplanması zamanı onların bütövlüyünün və dəyişməzliyinin qorunması əsas prinsip kimi çıxış edir. Bu məqsədlə zəncirvari saxlanma qaydalarına riayət olunmalı və hər bir sübut üçün kriptoqrafik heş qiymətləri hesablanmalıdır [9]. Toplanmış məlumatlar əsasında hadisələrin səbəb-nəticə əlaqələrinin qurulması mümkün olur. Qərarların izahı üçün izah oluna bilən süni intellekt (XAI) metodları, o cümlədən SHAP və LIME yanaşmaları istifadə edilir [17, 18].

Böyük dil modelləri (LLM) kontekstində ekspertiza prosesi əlavə problemlər yaradır. Bu modellərin qeyri-deterministik davranışı ekspertiza zamanı nəticələrin təkrar əldə edilməsini çətinləşdirir. Buna görə generasiya parametrlərinin qeyd olunması vacibdir [10]. Eyni zamanda, LLM sistemlərində təlim verilənlərinin model çıxışlarında təzahürü ayrıca ekspertiza prosedurları tələb edir [10, 19].

Sübutların etibarlılığının qiymətləndirilməsi Sİ ekspertizasının mühüm mərhələlərindən biridir və bu zaman sübutların dəqiqliyi, bütövlüyü, tamlığı və obyektivliyi əsas meyarlar kimi çıxış edir [9]. Bu məqsədlə çoxmənbəli yoxlama, model çıxışlarının ehtimal paylanmalarının və anomaliyalarının təhlili, təkrar testetmə və müstəqil ekspert qiymətləndirməsi kimi metodlardan istifadə edilir [7, 9, 10]. Bu yanaşmalar sübutların etibarlılığını artırmaqla yanaşı, qeyri-determinizm və potensial manipulyasiya risklərini də aşkar etməyə imkan verir.

Sübutların bütövlüyünün təmin olunması üçün kriptoqrafik üsullarla yanaşı, blokçeyn texnologiyalarının tətbiqi də perspektivli istiqamətlərdən biri hesab olunur. Bu halda hər bir sübut dəyişməz reyestrdə vaxt nişanı ilə birlikdə saxlanılır ki, bu da sonradan hər hansı müdaxilənin olub-olmadığını yoxlamağa imkan verir [9, 20]. Bu yanaşma xüsusilə məhkəmə ekspertizası və tənzimləyici audit baxımından mühüm əhəmiyyət kəsb edir.

Bununla yanaşı, nəzərə alınmalıdır ki, Sİ sistemlərinin özləri də müxtəlif təhlükələrə məruz qala bilər. Qəsdən manipulyativ hücumlar, məxfilik pozulmasın yönəlmiş hücumlar və təlim verilənlərindəki qərəzlər sübutların etibarlılığına mənfi təsir göstərə bilər [7, 8, 15]. Bu səbəbdən ekspertiza prosesində modellərin dayanıqlıq testlərindən keçirilməsi və təlim verilənlərinin sənədləşdirilməsi kimi tədbirlər həyata keçirilməlidir [14, 16].

Beləliklə, Sİ sistemlərinin rəqəmsal ekspertizası yalnız texniki analizlə məhdudlaşmamalı, eyni zamanda hüquqi və etik aspektləri də əhatə etməlidir. Bu kontekstdə tənzimləyici çərçivələr, o cümlədən süni intellekt üzrə normativ aktlar, məlumatların qorunması qaydaları və ədalətlilik prinsipləri nəzərə alınmalıdır [4, 21, 22]. Belə inteqrativ yanaşma Sİ ekspertizasının nəticələrinin etibarlılığını artırar və onların məhkəmə proseslərində qəbul edilməsini təmin edər.

IV. SÜNİ İNTELLEKT SİSTEMLƏRİNİN EKSPERTİZA PROBLEMLƏRİ

Sİ sistemlərinin rəqəmsal ekspertizasını çətinləşdirən əsas problemlər texniki, metodoloji və hüquqi səviyyələrdə qruplaşdırıla bilər.

Sİ modellərinin 'qara qutu' xarakteri ekspertiza zamanı qərarların əsaslandırılmasını çətinləşdirir. Bu problem xüsusilə hüquqi müstəvidə izah hüququnun təmin edilməsi baxımından aktualdır [6, 23]. Ekspert modelin hansı xüsusiyyətlərə əsaslanaraq konkret çıxış nəticəsini verdiyini müəyyən etməkdə çətinlik çəkir. Bu problem xüsusilə "izah hüququ" (right to explanation) tələb edən hüquqi kontekstlərdə (məsələn, GDPR, AI Act (EU)) özünü göstərir [21, 22].

Təlim verilənlərinin keyfiyyəti və adekvatlığı modelin qərarlarına birbaşa təsir göstərir. Xüsusilə təlim verilənlərində mövcud olan qərəzlilik ekspertiza zamanı ayrıca araşdırılmalıdır [16]. Təlim verilənlərindəki qərəzlilik, səhvlər və ya manipulyasiya modelin qərarlarına təsir edir [15]. Ekspertiza zamanı bu cür təsirlərin aşkarlanması çox vaxt yalnız modelin çıxış nəticələrini deyil, həm də təlim verilənləri dəstinin özünün analizini tələb edir [9, 16].

Qeyri-deterministik davranış və həssaslıq problemi eyni giriş verilənləri üçün modelin fərqli nəticələr verməsi və ya girişdəki kiçik dəyişikliklərə həddindən artıq həssas olması ekspertiza prosesini mürəkkəbləşdirir [7, 14]. Belə həssaslıq modelin dayanıqlığını zəiflədir və ekspertizanın təkrar edilə bilməsini çətinləşdirir. Təxribat xarakterli nümunələr həm "ağ qutu", həm də "qara qutu" ssenarilərində modelin çıxış nəticələrinə məqsədyönlü təsir göstərmək qabiliyyətinə malikdir [7, 24].

Sİ modelləri, xüsusilə böyük dil modelləri, təlim verilənlərindəki həssas məlumatları "yadda saxlamaq" və çıxarış zamanı bu məlumatları açmaq qabiliyyətinə malik olduğu üçün məxfilik riskləri yarana bilər [19]. Mənsubluğu müəyyənləşdirmə hücumları konkret bir verilən nümunəsinin təlim verilənlər dəstinə daxil olub-olmadığını müəyyən etməyə [8], tərs mühəndislik hücumları isə modeldən istifadə etməklə ilkin məlumatların qismən bərpasına [25] imkan verir. Bu hallar ekspertiza zamanı sübutların məxfilik və bütövlüyünü təhdid edir və əlavə anonimləşdirmə tədbirlərinin həyata keçirilməsini tələb edir.

Metodoloji standartların olmaması nəticəsində, hazırda Sİ sistemlərinin rəqəmsal ekspertizası üçün beynəlxalq standartlar mövcud deyil [3, 9]. Mövcud çərçivələr (NIST AI RMF, ISO/IEC 42001) daha çox risk idarəetməsinə yönəlmişdir və konkret ekspertiza prosedurlarını (sübutların toplanması, modelin tərs mühəndisliyi, sübutların məhkəmədə təqdimatı) tənzimləmir [26, 27]. Bu vəziyyət ekspert rəylərinin müqayisə edilə bilməsini və etibarlılığını azaldır.

Yuxarıda sadalanan problemlər Sİ sistemlərinin effektiv ekspertizasının yalnız texniki deyil, həm də metodoloji və tənzimləyici vasitələrə ehtiyac olduğunu göstərir.

V. İZAH EDİLƏ BİLƏN SÜNİ İNTELLEKT

İzah edilə bilən süni intellekt (Explainable AI -- XAI) modellərin qərarlarını insanlar tərəfindən başa düşülən, izah edilə bilən və məntiqi cəhətdən əsaslandırıla bilən formada təqdim etməyi hədəfləyir [6]. XAI-nin əsas məqsədi "qara qutu" modellərinin qərarvermə mexanizmlərini şəffaflaşdıraraq onların etibarlılığını, hesabatlılığını və istifadəçi etibarını artırmaqdır [12]. XAI sahəsində qlobal izah və lokal izah kimi iki əsas səviyyə müəyyən edilir.

Qlobal izah modelin bütün davranışının, yəni giriş nümunələri üzrə orta hesabla necə qərarlar qəbul etdiyinin izahıdır. Bu səviyyə modelin ümumi məntiqini, ən vacib xüsusiyyətlərini və potensial qərəzlərini anlamağa imkan verir. Lokal izah isə konkret bir giriş nümunəsi üçün modelin verdiyi tək qərarın səbəblərinin izahıdır [6, 23].

Yanaşma üsuluna görə isə XAI metodları modeldən asılı olmayan və modeldən asılı olan kimi iki kateqoriyaya bölünür. Modeldən asılı olmayan metodlar modelin daxili quruluşundan asılı olmayaraq yalnız giriş-çıxış məlumatlarına əsaslanaraq izah yaradır. Bu metodlar istənilən maşın təlimi modelinə (neyron şəbəkələr, qərar ağacları, SVM və s.) tətbiq edilə bilər [18, 23]. Modeldən asılı olan metodlar isə modelin daxili arxitekturasından (məsələn, neyron şəbəkə arxitekturasından) istifadə edərək izah yaradır. Lakin bu metodlar dəqiq olsa da, adətən yalnız müəyyən model tipləri üçün işləyir [6, 13].

XAI sahəsində ən geniş yayılmış metodlardan biri, oyun nəzəriyyəsinə əsaslanan SHAP (SHapley Additive exPlanations) metodudur [17]. Bu metod hər bir giriş xüsusiyyətinin modelin çıxışına əlavə (artım üzrə) töhfəsinin Şapley qiymətlərinin hesablanmasına əsaslanır. LIME (Local Interpretable Model-agnostic Explanations) metodu [18] isə lokal səviyyədə sadə izah modeli (məsələn, xətti reqressiya) ilə mürəkkəb modelin qərarını yerli səviyyədə yaxınlaşdırır. Hər iki metod modeldən asılı olmayan lokal izah üçün geniş istifadə edilir [17, 18].

Bundan əlavə, model kartları (model cards) [28] və verilənlər toplusu üzrə sənədləşmə [16] kimi sənədləşdirmə standartları modelin qurulma prinsipləri, istifadə olunmuş verilənlər, mövcud məhdudiyyətlər və qərəzlər, eləcə də gözlənilən məhsuldarlıq xüsusiyyətləri barədə strukturlaşdırılmış metainformasiya təqdim edir. Bu sənədlər ekspertiza prosesi zamanı modelin kontekstinin düzgün və hərtərəfli şəkildə anlaşılması baxımından mühüm əhəmiyyət kəsb edir.

Real tətbiqlərdə, həmçinin diqqət xəritələri (attention maps) və Grad-CAM kimi modeldən asılı olan metodlar, xüsusilə dərin neyron şəbəkələrinin vizual izahının təmin edilməsi məqsədilə geniş şəkildə istifadə olunur [13].

XAI yanaşmaları Sİ sistemlərinin rəqəmsal ekspertizasında qərarların əsaslandırılması, sübutların izahı və audit prosesində izlənilə bilənlik (traceability) kimi bir sıra mühüm funksiyaları yerinə yetirir. XAI metodları vasitəsilə ekspert modelin konkret çıxışa hansı giriş xüsusiyyətlərinin təsir etdiyini müəyyən edərək qərarların əsaslandırılmasını təmin edə bilər. Bu da sübutların məntiqi əsaslandırılmasını gücləndirir [6, 23]. Mürəkkəb model çıxışlarının (məsələn, ehtimal vektorları) XAI vasitəsi ilə sadə insan dili və ya vizual izaha çevrilməsi sübutların izahını asanlaşdırır və nəticədə ekspertizanın keyfiyyəti və şəffaflığı artır [6]. Audit prosesində izlənilə bilənlik eyni giriş üçün eyni izahın təkrar verilməsinə və modelin davranışının ardıcıllığını yoxlamağa imkan verir, bu da, xüsusilə məhkəmə ekspertizasında çox vacibdir [9, 23].

Bununla belə, qeyd etmək lazımdır ki, XAI metodları tam şəffaflığı təmin etmir. Birincisi, izahların özləri bəzən modelin işləmə mexanizmini sadələşdirir və ya təhrif edir. İkincisi, müxtəlif XAI metodları eyni model üçün fərqli, hətta ziddiyyətli izahlar verə bilər. Üçüncüsü, bəzi hallarda XAI metodları özləri də yeni izah problemləri yarada bilər, məsələn, izahın dəqiqliyi ilə sadəliyi arasında kompromisin yaradılması çətin olur [13, 25]. Həmçinin, XAI metodları qəsdən manipulyativ hücumlara qarşı həssas ola bilər, yəni kiçik dəyişikliklər izahı əhəmiyyətli dərəcədə dəyişə bilər [7].

Ekspertiza sahəsində yalnız XAI metodlarının istifadəsi kifayət etmir və onlar model kartları, verilənlər toplusu üzrə sənədləşmə, dayanıqlıq testləri və hüquqi qiymətləndirmə ilə birlikdə istifadə edilməlidir [16, 28, 29]. Bu kompleks yanaşma Sİ sistemlərinin həm texniki, həm də hüquqi cəhətdən etibarlı ekspertizasını təmin edə bilər.

VI. AUDİT VƏ STANDARTLAŞDIRMA YANAŞMALARİ

Sİ sistemlərinin kritik sahələrdə geniş tətbiqi onların sistemli şəkildə yoxlanılması, qiymətləndirilməsi və

sertifikatlaşdırılmasını zəruri edir. Bu kontekstdə Sİ auditi texniki, hüquqi, etik və təşkilati aspektləri birləşdirən multidissiplinar yanaşma kimi formalaşmışdır. Audit prosesi yalnız modelin texniki göstəricilərinin yoxlanılması ilə məhdudlaşmır, eyni zamanda onun normativ tələblərə, etik prinsiplərə və institusional siyasətlərə uyğunluğunu da qiymətləndirir [30]. Bu yanaşma Sİ sistemlərinin həyat dövrü boyunca risklərin sistemli idarə olunmasına və etibarlı tətbiqinə xidmət edir.

Sİ auditi sistemlərin uyğunluğunu (compliance), etibarlılığını (reliability), təhlükəsizliyini (security) və ədalətliliyini (fairness) qiymətləndirən strukturlaşdırılmış proses kimi xarakterizə olunur [30, 31]. Bu prosesin əsas məqsədi Sİ sisteminin layihələndirmə, təlim, tətbiq və istismar mərhələlərində yaranan potensial riskləri aşkar etmək və minimallaşdırmaqdır.

Auditin subyekti və icra mexanizminə əsasən üç əsas növü fərqləndirilir. Daxili audit təşkilat daxilində müstəqil audit bölmələri və ya keyfiyyət təminatı qrupları tərəfindən həyata keçirilir və adətən tez-tez icra olunur. Bu audit növü modelin məhsuldarlığının monitorinqi, əməliyyat risklərinin idarə edilməsi və daxili siyasətlərə uyğunluğun təmin olunmasına yönəldilmişdir [32]. Xarici audit isə müstəqil üçüncü tərəflər, məsələn, sertifikatlaşdırma qurumları və ya tənzimləyici orqanlar tərəfindən aparılır və daha geniş, dərin qiymətləndirməni əhatə edir. Xüsusilə yüksək riskli Sİ sistemləri üçün bu audit növü normativ tələblər çərçivəsində məcburi xarakter daşıya bilər [22, 30]. Davamlı audit yanaşması isə real-vaxt rejimində və ya qısa intervallarla avtomatlaşdırılmış monitorinqi nəzərdə tutur və model davranışında baş verən anomaliyaların, o cümlədən konsept dəyişmələrinin və qəsdən manipulyativ hücumların operativ aşkarlanmasına imkan yaradır [31, 32]. Bu audit növləri bir-birini tamamlayaraq çoxsəviyyəli nəzarət mexanizmi formalaşdırır.

Sİ sistemlərinin texniki auditində müxtəlif metodoloji yanaşmalardan istifadə olunur. Modelin formal yoxlanılması riyazi sübutlara əsaslanaraq modelin əvvəlcədən müəyyən edilmiş spesifikasiyalara uyğunluğunu qiymətləndirir. Bu məqsədlə SMT (Satisfiability Modulo Theories) əsaslı alətlərdən, məsələn, Reluplex alqoritmindən istifadə edilir [20]. Bu yanaşma xüsusilə kritik sistemlər üçün əhəmiyyətlidir, lakin böyük miqyaslı modellər üçün yüksək hesablama resursları tələb olunur.

Dayanıqlıq testləri modelin giriş verilənlərindəki kiçik dəyişikliklərə və qəsdən manipulyativ hücumlara qarşı davamlılığını qiymətləndirir. Bu testlər çərçivəsində modelə müxtəlif kiçik təsadüfi dəyişmələr tətbiq edilir və çıxışın sabitliyi ölçülür [7, 14]. Nəticədə modelin ən pis hal ssenarilərində davranışı təhlil olunur. Təhlükəsizlik analizləri isə modelin məxfilik hücumlarına qarşı həssaslığını qiymətləndirməyə yönəlmişdir. Bu mərhələdə modelin həssas məlumatların məxfiliyinə potensial təhdid təlim nümunəsinə aidiyyətin müəyyənləşdirilməsi, modeldən ilkin verilənlərin bərpası və təlim verilənlərinin çıxarılması kimi hücum ssenariləri vasitəsilə qiymətləndirilir [19, 25].

Sİ sistemlərinin audit və idarəetmə proseslərinin standartlaşdırılması məqsədilə bir sıra beynəlxalq çərçivələr və standartlar hazırlanmışdır. Bu sənədlər vahid terminologiya, metodoloji yanaşmalar və ən yaxşı təcrübələr təqdim etməklə müxtəlif maraqlı tərəflər arasında uyğunluğu təmin edir.

NIST tərəfindən hazırlanmış AI RMF 1.0 (AI Risk Management Framework) risk əsaslı idarəetmə prinsiplərinə əsaslanır [26]. Bu çərçivə Sİ sistemlərinin şəffaflıq, izah olunma və etibarlılıq kimi xüsusiyyətlərinin artırılmasına yönəlmiş praktik tövsiyələr təqdim edir. ISO və IEC tərəfindən qəbul edilmiş ISO/IEC 42001:2024 standartı isə Sİ idarəetmə sistemlərinin yaradılması və tətbiqi üçün tələbləri müəyyən edir və digər idarəetmə standartları ilə inteqrasiya oluna bilir [27].

IEEE tərəfindən işlənmiş standartlar, xüsusilə P7001, avtonom sistemlərin şəffaflığını və izah olunmasını təmin etməyə yönəlmişdir. Eyni zamanda, ISO/IEC 24027:2021 standartı Sİ sistemlərində qərəzliliyin idarə olunması məsələlərini əhatə edir [15]. Avropa Komissiyasının Etibarlı Sİ üzrə etik çərçivələri və AI Act (EU) kimi normativ sənədlər isə audit proseslərinin hüquqi əsas kimi çıxış edir və yüksək riskli sistemlər üçün məcburi uyğunluq qiymətləndirməsini tələb edir [4, 22].

Mövcud standartlar və çərçivələr Sİ sistemlərinin idarə olunması üçün mühüm baza yaratsa da, onların bir çoxu hələ tam yetkinlik səviyyəsinə çatmamışdır və rəqəmsal ekspertiza prosesləri ilə tam inteqrasiya olunmamışdır [3, 30]. Gələcək tədqiqat istiqamətləri bu standartların daha praktik və avtomatlaşdırılmış audit alətləri ilə zənginləşdirilməsini, eləcə də ekspertiza metodologiyaları ilə daha sıx inteqrasiyasını əhatə etməlidir [32].

VII. HÜQUQİ VƏ ETİK ASPEKTLƏR

Avtomatlaşdırılmış qərarvermə sistemləri ilə bağlı mühüm hüquqi məsələlərdən biri istifadəçilərin qərarların izahını tələb etmək hüququdur. Bu hüquq GDPR çərçivəsində qismən təmin edilir [21]. GDPR məlumat subyektlərinə qərarvermə prosesləri haqqında "məntiqli məlumat" əldə etmək imkanı verir və tam avtomatlaşdırılmış qərarların hüquqi təsirlər doğurduğu hallarda əlavə məhdudiyyətlər müəyyən edir.

Məsuliyyət məsələsi də Sİ sistemləri kontekstində açıq problemlərdən biri olaraq qalır. Sİ tərəfindən verilən qərarlara görə məsuliyyətin kimə aid olması, məsələn, tərtibatçı, istifadəçi, verilənlərin təminatçısı və ya digər tərəflər bir çox hüquq sistemlərində tam müəyyən edilməmişdir [21, 22]. Bu səbəbdən ekspertiza zamanı məsuliyyətin müəyyənləşdirilməsi üçün modelin arxitekturası, təlim verilənləri, tətbiq konteksti və nəzarət mexanizmləri kompleks şəkildə təhlil olunmalıdır.

Sİ sistemlərinin etibarlı və məsuliyyətli istifadəsi yalnız hüquqi uyğunluqla deyil, həm də etik prinsiplərə riayət olunması ilə təmin edilir. Avropa Komissiyasının Etibarlı Sİ üzrə Etik Qaydaları bu sahədə əsas istiqamətverici çərçivələrdən biridir və üç fundamental komponentdən ibarətdir: qanunauyğunluq, etik uyğunluq və texniki etibarlılıq [4].

Bu çərçivə daxilində bir sıra əsas etik prinsiplər müəyyən edilmişdir. İnsanın fəal rolu və nəzarəti prinsipi Sİ sistemlərinin insan muxtariyyətini nəzərə almasını və onların fəaliyyətinin effektiv insan nəzarəti altında saxlanılmasını zəruri edir. Texniki etibarlılıq və təhlükəsizlik prinsipi sistemlərin səhvlərə və hücumlara qarşı dayanıqlığını nəzərdə tutur. Məxfilik və verilənlərin idarə edilməsi məlumatların qorunmasını və keyfiyyətinin təmin olunmasını əhatə edir. Şəffaflıq prinsipi

qərarların izah olunmasını və istifadəçilərin məlumatlandırılmasını tələb edir. Ədalətlilik və ayrı-seçkilik etməmə prinsipi isə müxtəlif istifadəçi qrupları arasında bərabər münasibətin təmin olunmasına yönəlmişdir [4, 15]. Bundan əlavə, cəmiyyət və ətraf mühitin rifahı, eləcə də hesabatlılıq prinsipləri Sİ sistemlərinin sosial məsuliyyətini müəyyən edir.

Bu prinsiplərin praktiki tətbiqi üçün etik audit mexanizmlərinin inkişaf etdirilməsi xüsusi əhəmiyyət daşıyır. Etik audit texniki uyğunluqla yanaşı təşkilati prosesləri və maraqlı tərəflərin iştirakını da qiymətləndirir. Bu proses adətən etik prinsiplərin konkret tələblərə uyğun çevrilməsi, uyğunluq meyarlarının müəyyən edilməsi, məlumatların analizi və nəticələrin qiymətləndirilməsi mərhələlərini əhatə edir [10]. Sİ sistemlərinin rəqəmsal ekspertizası hüquqi tələblər və etik prinsiplərlə inteqrasiya olunmuş şəkildə həyata keçirilməlidir. Bu yanaşma yalnız texniki düzgünlüyü deyil, həm də sosial ədalət, şəffaflıq və hüquqi legitimlik kimi fundamental meyarların təmin olunmasına imkan verir [4, 10, 22].

NƏTİCƏ

Aparılmış təhlil göstərirr ki, Sİ sistemlərinin "qara qutu" xarakteri, qeyri-determinizm xüsusiyyəti və təlim verilənlərindən yüksək asılılığı onların ekspertizasını ənənəvi rəqəmsal kriminalistikadan əhəmiyyətli dərəcədə fərqləndirir. Xüsusilə model davranışının izah olunması, nəticələrin təkrar generasiyası və sübutların etibarlılığının qiymətləndirilməsi problemləri yaradır.

XAI metodları qərarların izahı baxımından mühüm imkanlar təqdim etsə də, izahların təxmini xarakter daşıması, metodlar arasında ziddiyyətlər və manipulyativ hücumlara həssaslıq kimi məhdudiyyətlərə malikdirlər. Eyni zamanda, qəsdən manipulyativ və məxfilik hücumları Sİ sistemlərinin təhlükəsizliyini və ekspertiza nəticələrinin etibarlılığını təhdid edir.

Mövcud audit yanaşmaları və NIST AI RMF və ISO/IEC 42001 kimi beynəlxalq standartlar mühüm çərçivə təqdim etsə də, vahid və universal ekspertiza protokollarının olmaması praktiki tətbiqlərdə çətinliklər yaradır.

Sİ sistemlərinin etibarlı ekspertizası yalnız multidissiplinar yanaşma əsasında mümkündür. Gələcək tədqiqatlar avtomatlaşdırılmış audit sistemlərinin yaradılması, izah edilə bilən və təhlükəsiz modellərin yaradılması, həmçinin hüquqi və etik tələblərin texniki alətlərlə inteqrasiyası istiqamətində aparılmalıdır. Bu yanaşma Sİ sistemlərinin şəffaflığını, hesabatlılığını və cəmiyyət tərəfindən qəbul olunmasını təmin etməyə imkan verir.

ƏDƏBİYYAT

[1] J. P. Khatiwala et al., "Evaluating the Reliability of Digital Forensic Evidence Discovered by Large Language Model," in 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), 2025, doi: 10.1109/COMPSAC65507.2025.00138.

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[3] B. D. Lund et al., "Standards, frameworks, and legislation for artificial intelligence (AI) transparency," AI and Ethics, vol. 5, pp. 3639–3655, 2025, doi: 10.1007/s43681-024-00456-1.

[4] European Commission, "Ethics Guidelines for Trustworthy AI," 2019.://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[5] J. Schneider and F. Breitinger, "Towards AI Forensics: Did the Artificial Intelligence System Do It?," arXiv preprint arXiv:2005.13635, 2020.

[6] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

[7] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39–57.

[8] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in IEEE Symposium on Security and Privacy, 2017, pp. 3–18.

[9] A. A. Solanke and M. A. Biasiotti, "Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques," KI - Künstliche Intelligenz, vol. 36, pp. 143–161, 2022, doi: 10.1007/s13218-022-00763-9.

[10] J. Laine, M. Minkkinen, and M. Mäntymäki, "Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders," Information & Management, vol. 61, no. 5, art. 103969, 2024, doi: 10.1016/j.im.2024.103969.

[11] M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," Science, vol. 349, no. 6245, pp. 255–260, 2015.

[12] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[13] U. Bhatt et al., "Explainable machine learning in deployment," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 648–657.

[14] T.-W. Weng et al., "Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach," in Sixth International Conference on Learning Representations (ICLR), 2018.

[15] ISO/IEC 24027:2021, "Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making," International Organization for Standardization, 2021.

[16] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021.

[17] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[19] N. Carlini et al., "Extracting Training Data from Large Language Models," in USENIX Security Symposium, 2021.

[20] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks," in Computer Aided Verification: 29th International Conference (CAV), 2019, pp. 97–117.

[21] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," AI Magazine, vol. 38, no. 3, pp. 50–57, 2017.

[22] European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence (AI Act)," 2021.

[23] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub, 2022. https://leanpub.com/interpretable-machine-learning

[24] N. Papernot et al., "Practical Black-Box Attacks against Machine Learning," in ASIACCS, 2017.

[25] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proceedings of CCS, 2015, pp. 1322–1333.

[26] NIST (National Institute of Standards and Technology), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST, USA, 2023.

[27] ISO/IEC 42001:2024, "Information technology — Artificial intelligence — Management system," International Organization for Standardization, 2024.

[28] M. Mitchell et al., "Model cards for model reporting," in Proceedings of FAT, 2019, pp. 220–229.

[29] L. Waltersdorfer et al., "AuditMAI: Towards an Infrastructure for Continuous AI Auditing," 2024.

[30] J. Mökander, "Auditing of AI: Legal, Ethical and Technical Approaches," DISO, vol. 2, art. 49, 2023, doi: 10.1007/s44206-023-00074-y.

[31] I. D. Raji, A. Smart, R. N. White et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), 2020, pp. 33–44.

[32] IEEE P7001, "Standard for Transparency of Autonomous Systems," IEEE, 2021.

Digital Forensics Challenges of Artificial Intelligence-Based Systems and Technologies

Ramiz Shikhaliyev

Institute of Information Technology, Baku, Azerbaijan
shikhramiz61@gmail.com

Abstract – The article analyzes the digital forensics of artificial intelligence systems based on a systematic and integrative approach, and examines the characteristics of models such as the "black box" nature, inexplicability, non-deterministic behavior, and dependence on training data. In addition, explainable artificial intelligence approaches, security risks, audit mechanisms, standardization issues, as well as legal and ethical aspects are analyzed in a comprehensive manner.

Keywords – artificial intelligence; digital forensics; explainable artificial intelligence; audit; ethical principles.