AZƏRBAYCAN RESPUBLİKASININ DÖVLƏT STANDARTI (Texniki hesabat)

AZS ISO/IEC TR 24028:2024

İlkin nəşr
2024

Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

İnformasiya texnologiyaları — Süni intellekt — Süni intellektin etimadı doğrultmasına dair icmal

AZƏRBAYCAN STANDARTLAŞDIRMA İNSTİTUTU

İstinad nömrəsi
AZS ISO/IEC TR 24028:2024

Elçin İsaqzadə küç., 7-ci köndələn
Qaynar xətt: +994125149308
Email: office@azstand.gov.az

MÜQƏDDİMƏ

1. Bu standart Azərbaycan Elm Fondunun qrant layihəsi (Qrant №AEF-MQM-QA-1-2021-4(41)-8/03/1) çərçivəsində işlənib hazırlanıb və “İnformasiya-kommunikasiya texnologiyaları” standartlaşdırma üzrə Texniki Komitə (AZSTAND/TK 05) tərəfindən təqdim edilib.
2. Azərbaycan Standartlaşdırma İnstitutunun ..... 2024-cü il tarixli ..... saylı qərarı ilə təsdiq edilib.
3. Bu standart Beynəlxalq Standart ISO/IEC TR 24028 nəşr 1.0 (2020-05) ilə eyndir (IDT). This standart is identical (IDT) to the International Standard ISO/IEC TR 24028, (2020-05).
4. İlk dəfə tətbiq edilir.
5. Dövlət standartında müəyyən edilən tələblərin beynəlxalq standartlara, norma, qayda və tövsiyələrə və digər dövlətlərin müvafiq mütərəqqi milli standartlarına, elm, texnika və texnologiyanın müasir nailiyyətlərinə əsaslanmasını müəyyən etmək üçün standartın dövri yoxlama müddəti ildə 1 dəfədir.

MÜNDƏRİCAT

ÖN SÖZ

ISO (Beynəlxalq Standartlaşdırma Təşkilatı) və IEC (Beynəlxalq Elektrotexniki Komissiya) dünya üzrə standartlaşdırma sahəsində ixtisaslaşmış sistemi formalaşdırırlar. ISO və ya IEC üzvü olan milli orqanlar texniki fəaliyyətin konkret sahələri ilə məşğul olmaq üçün müvafiq təşkilat tərəfindən yaradılmış texniki komitələr vasitəsilə beynəlxalq standartların hazırlanmasında iştirak edirlər. ISO və IEC texniki komitələri qarşılıqlı maraq doğuran sahələrdə əməkdaşlıq edirlər. ISO və IEC ilə əməkdaşlıq edən digər beynəlxalq təşkilatlar, dövlət və qeyri-hökumət təşkilatları da bu işdə iştirak edirlər.

Bu standartın hazırlanması üçün istifadə olunan və həmçinin sonrakı texniki xidmət üçün nəzərdə tutulan prosedurlar ISO/IEC Direktivlərinin 1-ci hissəsində təsvir edilmişdir. Müxtəlif növ sənədlər üçün tələb olunan fərqli təsdiq meyarlarına xüsusilə diqqət yetirilməlidir. Bu sənəd ISO/IEC Direktivlərinin 2-ci hissəsində (www.iso.org/directives və ya www.iec.ch/members_experts/refdocs) verilmiş qaydalara uyğun olaraq hazırlanmışdır.

Bu standartın bəzi elementlərinin patent hüquqlarının predmeti ola bilməsi diqqət çəkir. ISO və IEC bu patent hüquqlarının hər hansı birinin və ya hamısının müəyyən edilməsinə görə məsuliyyət daşımır. Sənədin hazırlanması zamanı müəyyən edilmiş patent hüquqlarının təfərrüatları Girişdə və/və ya ISO və IEC (www.iso.org/patents və patents.iec.ch) alınmış patent bəyannamələrinin siyahısında təqdim olunur.

Bu standartdakı ticarət adları (“trade name”) haqqında məlumatlar istifadəçilərin rahat istifadəsi üçün təqdim olunur və bu təqdimat tövsiyə xarakteri daşımır.

Standartların könüllü xarakter daşıması, uyğunluğun qiymətləndirilməsi üzrə ISO-nun xüsusi termin və ifadələrinin mənası ilə bağlı izahlar, eləcə də Ticarətdə Texniki Maneələrin (Technical Barriers to Trade, TBT) aradan qaldırılması ilə əlaqədar ISO-nun Ümumdünya Ticarət Təşkilatının (ÜTT) prinsiplərinə sadiqliyi haqqında məlumat “www.iso.org/iso/foreword.html” internet informasiya ehtiyatından əldə edilə bilər.

Bu sənəd ISO/IEC JTC 1 “İnformasiya texnologiyaları” Birgə Texniki Komitəsinin “SC 42, Süni intellekt” Altkomitəsi tərəfindən hazırlanmışdır.

Bu sənədlə bağlı istənilən rəy və suallar milli standartlar üzrə quruma yönəldilməlidir. Bu qurumların tam siyahısı ilə “www.iso.org/members.html” internet informasiya ehtiyatında tanış olmaq olar.

GİRİŞ

Bu standartın məqsədi (bundan sonra süni intellekt (Sİ) sistemləri adlandırılacaq) Sİ ilə təmin edən və ya Sİ-dən istifadə edən sistemlərin etimadı doğrultması xüsusiyyətlərinə təsir edə biləcək amilləri təhlil etməkdir. Sənəddə texniki sistemlərin etimadı doğrultmasını dəstəkləyən və ya təkmilləşdirə bilən mövcud yanaşmalar qısaca araşdırılır və onların Sİ sistemlərinə potensial tətbiqi müzakirə edilir. Sİ sistemlərinin etimadı doğrultması ilə bağlı zəifliklərinin aradan qaldırılması üçün standartda mümkün yanaşmalara baxılır. Sənəddə, həmçinin Sİ sistemlərinin etimadı doğrultması səviyyəsinin yüksəldilməsinə dair yanaşmalar müzakirə edilir.

AZƏRBAYCAN RESPUBLİKASININ DÖVLƏT STANDARTI

İnformasiya texnologiyaları - Süni intellekt - Süni intellektin etimadı doğrultmasına dair icmal

AZE ISO/IEC TR 24028:2024

Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

Tətbiq edilmə tarixi ... . 2024-cü il

1 TƏTBİQ SAHƏSİ

Bu standartda Sİ sistemlərinin etimadı doğrultması ilə bağlı mövzulara, o cümlədən aşağıdakılara baxılır:

- Şəffaflığı, izahlılığı, idarəolunanlığı və s. təmin etməklə Sİ sistemlərinə inamın yaradılması üçün yanaşmalar;
- Sİ sitemləri üçün mühəndislik "tələləri" (layihələndirmə mərhələsindəki səhvlər), əlaqəli tipik təhdidlər və risklər, həmçinin, onların mümkün təsirlərinin azaldılması texnikaları və üsulları;
- Sİ sistemlərinin əlçatanlığı, dözümlülüyü, etibarlılığı, dəqiqliyi, mühafizəliliyi, təhlükəsizliyi və məxfiliyinin qiymətləndirilməsi və təmin edilməsi üçün yanaşmalar.

Sİ sistemləri üçün etimadı doğrultma səviyyələrinin spesifikasiyası bu standartın əhatə dairəsinə daxil deyil.

2 NORMATİV İSTİNADLAR

Bu sənəddə normativ istinad yoxdur.

3 TERMİN VƏ ANLAYIŞLAR

Bu sənədin məqsədləri üçün aşağıdakı terminlər və anlayışlar tətbiq edilir.

ISO и IEC-in standartlaşdırma sahəsində istifadə olunan terminoloji məlumat bazaları aşağıdakı ünvanlarda saxlanılır:

— ISO Onlayn baxış platforması: https://www.iso.org/obp;
— IEC Electropedia: http://www.electropedia.org/.

3.1
hesabatlılıq
Obyektin (3.16) fəaliyyətinin yalnız həmin obyektə aid edilməsini izləməyə imkan verən xüsusiyyət

[Mənbə: ISO/IEC 2382:2015, 2126250, düzəliş - Qeydlər silinib.]

3.2
aktor
ünsiyyət quran və qarşılıqlı əlaqədə olan *obyekt* (3.16)

[Mənbə: ISO/IEC TR 22417:2017, 3.1]

3.3
alqoritm
verilənlərin (3.11) məntiqi təsvirinin çevrilməsi (transformasiyası) qaydaları çoxluğu;

[Mənbə: ISO/IEC 11557:1992, 4.3]

3.4
süni intellekt
Sİ
mühəndislik sisteminin (3.38) bilik və bacarıqlar əldə etmək, onları emal və tətbiq etmək qabiliyyəti

Qeyd 1: Bilik təcrübə və ya təhsil nəticəsində əldə edilmiş faktlar, *informasiya* (3.20) və bacarıqlardır.

3.5
aktiv
maraqlı tərəf (3.37) üçün dəyəri olan hər şey (3.46)

Qeyd 1: Aktivlərin bir çox növləri var, o cümlədən:
a) *informasiya* (3.20);
b) proqram təminatı (məsələn, kompüter proqramı);
c) fiziki obyektlər (məsələn, kompüter);
d) xidmətlər;
e) insanlar və onların ixtisasları, bacarıqları və təcrübəsi;
f) reputasiya və imic kimi qeyri-maddi aktivlər.

[MƏNBƏ: ISO/IEC 21827:2008, 3.4, düzəliş - Tərifdə “təşkilat” sözü "maraqlı tərəf" ifadəsi ilə əvəz edilib. Qeyd 1 silinib.]

3.6
atribut
insan və ya avtomatlaşdırılmış vasitələrin köməyi ilə obyektin kəmiyyət və ya keyfiyyətcə fərqləndirilə bilən xassəsi və ya xüsusiyyəti

[MƏNBƏ: ISO/IEC/IEEE 15939:2017, 3.2]

3.7
avtonomluq
avtonom
öz-özünə öyrənmə nəticəsində öz qaydaları ilə idarə olunan *sistemin* (3.38) xüsusiyyəti

Qeyd 1: Belə sistemlər kənar *idarəetməyə* (3.10) və ya nəzarətə tabe deyildir.

3.8
qərəz
digərləri ilə müqayisədə bəzi əşyaların, insanların və ya qrupların şəxsi lütfə görə üstün tutulması (favoritizm)

3.9
məntiqi ardıcıllıq
sənədlər və ya *sistemin* (3.38) hissələri və ya komponentləri arasında vahidlik, standartlaşdırılma və ziddiyyətsizlik dərəcəsi

[MƏNBƏ: ISO/IEC 21827:2008, 3.14]

3.10
idarəetmə
hər hansı *proses* (3.29) üzərindən və ya *proses* (3.29) daxilində müəyyən hədəfə çatmaq üçün məqsədyönlü fəaliyyət

[MƏNBƏ: IEC 61800-7-1:2015, 3.2.6]

3.11
verilənlər
informasiyanın (3.20) ünsiyyət, interpretasiya və ya emal üçün uyğun formatda yenidən interpretasiya eidlə bilən təqdimatı

Qeyd 1: *Verilənlər* (3.11) insan tərəfindən və ya avtomatik olaraq emal oluna bilər.

[MƏNBƏ: ISO/IEC 2382:2015, 2121272, düzəliş - Qeyd 2 və Qeyd 3 silinib.]

3.12
verilənlər subyekti
fərdi məlumatları (3.27) qeydə alınan fiziki şəxs

[MƏNBƏ: ISO 5127:2017, 3.13.4.01, düzəliş - Qeyd 1 silinib.]

3.13
qərarlar ağacı
nəticənin bir və ya daha çox ağacvari struktur üzrə təqdim edildiyi müəllimlə (supervayzerlə) öyrənmə modeli

3.14
effektivlik
planlaşdırılan fəaliyyətlərin həyata keçirilməsi və planlaşdırılan nəticələrin əldə edilməsi dərəcəsi

[MƏNBƏ: ISO 9000:2015, 3.7.11, düzəliş - Qeyd 1 silinib.]

3.15
səmərəlilik
əldə edilən nəticələrlə istifadə olunan resurslar arasında əlaqə

[MƏNBƏ: ISO 9000:2015, 3.7.10]

3.16
obyekt
maraq doğuran hər hansı konkret və ya mücərrəd element

[MƏNBƏ: ISO/IEC 10746-2:2009, 6.1]

3.17
zərər
insanların sağlamlığına yetirilən xəsarət və ya onlara vurulan ziyan, əmlaka və ya ətraf mühitə vurulan ziyan

[MƏNBƏ: ISO/IEC Təlimatı 51:2014, 3.1]

3.18
təhlükə
zərərin (3.17) potensial mənbəyi

MƏNBƏ: ISO/IEC Təlimatı 51:2014, 3.2]

3.19
insan faktorları
İnsanların davranışına və ya təşkilatların işinə təsir edən, koqnitiv insan xüsusiyyətləri ilə birgə ətraf mühitlə, iş fəaliyyəti ilə əlaqədar, təşkilati amillər

3.20
informasiya
məna kəsb edən *verilənlər* (3.11)
[MƏNBƏ: ISO 9000:2015, 3.8.2]

3.21
tamlıq
aktivlərin (3.5) dəqiqliyini və bütövlüyünü qorumaq xüsusiyyəti

[MƏNBƏ: ISO/IEC 27000:2018, 3.36, düzəliş - Tərifdə “dəqiqlik” sözündən əvvəl “aktivlərin”, “bütövlük" sözündən sonra isə “qorumaq” əlavə edilib.]

3.22
təyinatlı istifadə
məhsul və ya *sistemlə* (3.38) birlikdə təqdim edilən *informasiyaya* (3.20) uyğun və ya bu informasiya olmadıqda isə nümunələrə (*obrazlara*) (3.26) uyğun ümumişlək mənada istifadə

[MƏNBƏ: ISO/IEC Təlimatı 51:2014, 3.6]

3.23
maşın öyrənməsi
ML

yeni bilik və bacarıqlar əldə etməklə və ya mövcud bilik və bacarıqların yenidən təşkili ilə funksional vahidin öz perfomansını təkmilləşdirməsi prosesi (3.29)

[MƏNBƏ: ISO/IEC 2382:2015, 2123789]

3.24
maşın öyrənməsi modeli
giriş verilənlərinə əsasən nəticə və ya proqnoz generasiya edən riyazi konstruksiya (3.11)

3.25
neyron şəbəkəsi
paylanmış paralel lokal emaldan istifadə edərək, süni neyronlar adlanan sadə emal elementləri şəbəkəsindən ibarət olub mürəkkəb qlobal davranış nümayiş etdirə bilən hesablama modeli
[MƏNBƏ: ISO 18115-1:2013, 8.1]

3.26
obraz
nümunə
verilmiş kontekstdə *obyekti* (3.16) tanımaq üçün istifadə edilən əlamətlər və onların əlaqələri çoxluğu

[MƏNBƏ: ISO/IEC 2382:2015, 2123798]

3.27
fərdi məlumat
identifikasiya edilmiş və ya identifikasiya edilə bilən şəxsə aid *informasiya* (3.11)

[MƏNBƏ: ISO 5127:2017, 3.1.10.14, düzəliş - Qeyd1, Qeyd2 və qəbul edilmiş tərif silinib.]

3.28
məxfilik
fərd haqqında *verilənlərin* (3.11) əsassız və ya qeyri-qanuni toplanması və istifadəsi nəticəsində onun şəxsi həyatına və ya işlərinə müdaxilənin olmaması

[MƏNBƏ: ISO/IEC 2382:2015, 2126263, düzəliş - Qeyd1 və Qeyd2 silinib.]

3.29
proses
nəzərdə tutulan nəticəni əldə etmək üçün giriş verilənlərindən istifadə edən bir-biri ilə əlaqəli və qarşılıqlı təsir göstərən fəaliyyətlər

[MƏNBƏ: ISO 9000:2015, 3.4.1, düzəliş - Qeydlər nəzərə alınmayıb.]

3.30
etibarlılıq
gözlənilən davranış və nəticələr arasında uyğunluq xassəsi

[MƏNBƏ: ISO/IEC 27000:2018, 3.55]

3.31
risk
qeyri-müəyyənliyin məqsədlərə təsiri

Qeyd 1: Təsir - gözlənilən sapmadır. O, ya müsbət, ya mənfi və ya hər ikisi ola bilər, imkan və *təhdidləri* (3.39) aradan qaldıra, yarada və ya doğura bilər.
Qeyd 2: Məqsədlər müxtəlif aspektlərə və kateqoriyalara malik ola və müxtəlif səviyyələrdə tətbiq oluna bilər.
Qeyd 3: Risk adətən riskin mənbələri, potensial mümkün hadisələr, onların nəticələri və ehtimalları kontekstində ifadə edilir.

[MƏNBƏ: ISO 31000:2018, 3.1]

3.32
robot
nəzərdə tutulmuş tapşırıqları yerinə yetirmək üçün öz mühitində fəaliyyət göstərən, müəyyən *avtonomluq* (3.7) dərəcəsinə malik proqramlaşdırılmış icra mexanizmi

Qeyd 1: Robot *idarəetmə* (3.10) sistemini və idarəetmə *sisteminin* (3.38) interfeysini ehtiva edir.
Qeyd 2: Robotun sənaye robotu və ya xidmət robotu kimi klassifikasiyası nəzərdə tutulan tətbiqindən asılı olaraq aparılır.

[MƏNBƏ: ISO 18646-2:2019, 3.1]

3.33
robototexnika
robotların (3.32) elmi-praktiki layihələndirilməsi, istehsalı və tətbiqi

[MƏNBƏ: ISO 8373:2012, 2.16]

3.34
mühafizəlilik
yolverilməz *riskin* (3.31) olmaması

[MƏNBƏ: ISO/IEC Təlimatı 51:2014, 3.14]

3.35
təhlükəsizlik
məhsulun və ya *sistemin* (3.38) *informasiyanı* (3.20) və *verilənləri* (3.11) mühafizə dərəcəsidir, beləki insanlar, digər məhsullar və ya sistemlər onların növlərinə və avtorizasiya səviyyələrinə uyğun olaraq verilənlərə əlçatanlıq dərəcəsinə malikdir

[MƏNBƏ: ISO/IEC 25010:2011, 4.2.6]

3.36
həssas verilənlər
açıqlandıqda və ya sui-istifadəsi nəticəsində potensial zərərli təsirlərə malik olan *verilənlər* (3.11)

[MƏNBƏ: ISO/IEC 38500:2015, 2.24]

3.37
maraqlı tərəf
hər hansı qərar və ya fəaliyyətə təsir edə bilən, onların təsirinə məruz qalan və ya özlərini onların təsirinə məruz qalan hesab edən fərd, qrup və ya təşkilat
[MƏNBƏ: ISO/IEC 38500:2015, 2.24]

3.38
sistem
müəyyən edilmiş bir və ya daha çox məqsədə çatmaq üçün təşkil edilmiş qarşılıqlı əlaqəli elementlərin kombinasiyası

Qeyd 1: Sistem bəzən məhsul və ya onun təqdim etdiyi xidmətlər kimi nəzərdə tutulur.

[MƏNBƏ: ISO/IEC/İEEE 15288:2015, 3.38]

3.39
təhdid
sistemlərə (3.38), təşkilatlara və ya fərdlərə *zərər* (3.17) vurmaqla nəticələnə bilən arzuolunmaz insidentin potensial səbəbi

3.40
təlim
təlim *verilənlərindən* (3.11) istifadə edərək maşın öyrənməsi *alqoritmi* (3.3) əsasında *maşın öyrənməsi modelinin* (3.24) parametrlərinin müəyyənləşdirilməsi və ya təkmilləşdirilməsi *prosesi* (3.29)

3.41
inam
güvən
istifadəçinin (3.43) və ya *maraqlı tərəfin* (3.37) məhsul və ya *sistemin* (3.38) gözlənilən davranışına əminlik dərəcəsi

[MƏNBƏ: ISO/IEC 25010:2011, 4.1.3.2]

3.42
etimadı doğrultma
maraqlı tərəflərin (3.37) gözləntilərini yoxlanıla bilən şəkildə qarşılamaq qabiliyyəti

Qeyd 1: Kontekstdən və ya tətbiq sahəsindən, həmçinin konkret məhsul və ya xidmətdən, istifadə olunan verilənlərdən (3.11) və texnologiyalardan asılı olaraq, maraqlı tərəflərin gözləntilərinin qarşılanmasını təmin etmək üçün *verifikasiyanın* (3.47) tələb olunduğu müxtəlif xüsusiyyətlər tətbiq edilir.

Qeyd 2: Etimadı doğrultma xüsusiyyətləri məsələn, *etibarlılıq* (3.30), əlçatanlıq, dözümlülük, *təhlükəsizlik* (3.35), *məxfilik* (3.28), *mühafizəlilik* (3.34), *hesabatlılıq* (3.1), şəffaflıq, *tamlıq* (3.21), orijinallıq, keyfiyyət, istifadə rahatlığını ehtiva edir.

Qeyd 3: Etimadı doğrultma xidmətlərə, məhsullara, texnologiyalara, verilənlərə və *informasiyaya* (3.20), eləcə də idarəetmə kontekstində təşkilatlara tətbiq oluna bilən *atributdur* (3.6).

3.43
istifadəçi
sistemlə (3.38) interaktiv əlaqədə olan və ya sistemdən istifadə zamanı faydalanan fərd və ya qrup

[MƏNBƏ: ISO/IEC/IEEE 15288:2015, 4.1.52, düzəliş - Qeyd 1 silinib.]

3.44
həqiqiliyin yoxlanılması
yoxlanılma
konkret *təyinatlı istifadə* (3.22) və ya tətbiq üçün müəyyən edilmiş tələblərin yerinə yetirildiyinin obyektiv sübutların təqdim edilməsi ilə təsdiqi

Qeyd 1: *Sistem* (3.38) düzgün qurulub.

[MƏNBƏ: ISO/IEC TR 29110-1:2016, 3.73, düzəliş – Qeyd1-in yalnız sonuncu cümləsi saxlanılıb və Qeyd2 silinib.]

3.45
qiymət
< verilənlər konteksti> *verilənlər* (3.11) vahidi

[MƏNBƏ: ISO/IEC/IEEE 15939:2017, 3.41]

3.46
dəyər
<sosial kontekst> təşkilatın sadiq qaldığı əqidə(lər) və riayət etməyə çalışdığı standartlar

[MƏNBƏ: ISO 10303-11:2004, 3.3.22]

3.47
verifikasiya
obyektiv sübutların təqdim edilməsi yolu ilə müəyyən edilmiş tələblərin yerinə yetirildiyinin təsdiq edilməsi

Qeyd 1: *Sistem* (3.38) düzgün qurulub.

[MƏNBƏ: ISO/IEC TR 29110-1:2016, 3.74, düzəliş - Qeyd1-in yalnız sonuncu cümləsi saxlanılıb.]

3.48
zəiflik
aktivin (3.5) və ya *idarəetmənin* (3.10) bir və ya bir neçə *təhdid* (3.38) tərəfindən istifadə (istismar) oluna bilən boşluğu

[MƏNBƏ: ISO/IEC 27000:2018, 3.77]

3.49
iş yükü
adətən müəyyən bir kompüter *sistemində* (3.38) yerinə yetirilən tapşırıqlar coxluğu

[MƏNBƏ: ISO/IEC/IEEE 24765:2017, 3.4618, düzəliş - Qeyd1 silinib.]

4 İCMAL

Bu standartda Sİ sistemlərinin etimadı doğrultmasının təmin oluması ilə bağlı mövzuların icmalı təqdim edilir. Bu sənədin məqsədlərindən biri standartların işlənilməsi ilə məşğul olan qurumlara, ümumiyyətlə, standartlaşdırma icmasına Sİ sahəsində mövcud spesifik standartlaşdırma boşluqlarının müəyyən edilməsində köməyin göstərilməsindən ibarətdir.

Standartın 5-ci bəndində texniki sistemlərdə etimadı doğrultma xüsusiyyətlərinin yaradılması (formalaşdırılması) üçün istifadə edilən mövcud yanaşmalar qısa şəkildə araşdırılır və onların Sİ sistemlərinə potensial tətbiq imkanları müzakirə edilir. Standartın 6-cı bəndində maraqlı tərəflər haqqında məlumat təqdim olunur. 7-ci bənddə Sİ sistemlərinin cavabdehliyi, hesabatlılığı, idarəçiliyi və mühafizəliliyi ilə bağlı məsələlər müzakirə olunur. Standartın 8-ci bəndində Sİ sistemlərinin etimadı doğrultması səviyyəsini azalda bilən zəifliklər araşdırılır. 9-cu bənd Sİ sisteminin həyat dövrü ərzində zəiflikləri aradan qaldırmaqla onun etimadı doğrultma səviyyəsini yüksəldən mümkün tədbirləri müəyyən edir. Tədbirlər Sİ sisteminin şəffaflığının, idarəolunanlığının, verilənlərin emal olunmasının, dayanıqlığının, möhkəmliyinin, test olunmasının, qiymətləndirməsinin və istifadəsinin təkmilləşdirilməsini ehtiva edir. Nəticələr 10-cu bənddə təqdim olunur.

5 ETİMADI DOĞRULTMA XÜSUSİYYƏTİ İLƏ BAĞLI MÖVCUD ÇƏRÇİVƏLƏR

5.1 Anlayışların mənşəyi

Bu standartın məqsədləri üçün Sİ sistemlərinin və etimadı doğrultmasının işlək tərifərinin təmin edilməsi vacibdir.

Burada Sİ sistemi dedikdə Sİ istifadə edilən istənilən sistem (məhsul və ya xidmət) nəzərdə tutulur. Sİ sistemlərinin müxtəlif növləri var. Onların bəziləri proqram təminatı kimi, digərləri isə əsasən aparat təminatı (məsələn, robotlarda) kimi realizə olunur.

Etimadı doğrultma xüsusiyyətinin işlək tərifi kimi yoxlanıla bilən şəkildə maraqlı tərəflərin gözləntilərini qarşılamaq qabiliyyəti nəzərdə tutulur. Bu tərif Sİ sistemlərinin, texnologiyalarının və tətbiq sahələrinin geniş spektrinə şamil edilə bilər.

Təhlükəsizlik anlayışı kimi, etimadı doğrultma da istifadə keyfiyyəti kontekstində sistemin fövqəladə xassələrini (yəni atributları ilə birlikdə xüsusiyyətlər çoxluğunu) təyin edən qeyri-funksional tələb kimi başa düşülür və nəzərə alınır. Bu, ISO/IEC 25010 [20] standartında göstərilmişdir.

Bununla yanaşı, təhlükəsizlik kimi, etimadı doğrultma səviyyəsi də xüsusi ölçülə bilən nəticələrin və əsas performans göstəricilərinin (KPI) istifadə olunduğu təşkilati proseslər vasitəsilə yaxşılaşdırıla bilər.

Beləliklə, etimadı doğrultma həm davamlı təşkilati proses, həm də (qeyri-funksional) tələb kimi başa düşülür və nəzərdə tutulur.

UNEP [26] əsasən “ehtiyatlılıq prinsipi” o deməkdir ki, ciddi və ya bərpa olunmaz zərərlə nəticələnə biləcək təhdid olduqda, tam elmi əsaslandırmanın olmaması zərərin qarşısının alınması məqsədilə effektiv tədbirlərin təxirə salınması üçün səbəb kimi istifadə edilə bilməz. Təhlükəsizlik texnikasında maraqlı tərəflərin “qiymətləndirmə” tələblərinin müəyyənləşdirilməsi və sonra ölçülməsi prosesi sistemin istifadəsi kontekstinin, zərərlə nəticələnəcək risklərinin başa düşülməsini və tətbiqi mümkün olan müvafiq hallarda, fiziki şəxslərin hüquq və azadlıqlarına, bütün canlılara, ətraf mühitə, bioloji növlərə və ya populyasiyalara zərər vurmaq kimi potensial gözlənilməz nəticələrə səbəb olacaq risklərin azaldılması texnikası kimi “ehtiyatlılıq prinsipi”nin tətbiqini əhatə edir.

Süni intellekt sistemləri çox zaman mövcud sistemlərə süni intellekt imkanları əlavə edilən təkmilləşdirilmiş sistemlərdir. Bu zaman, sistemin köhnə versiyasına etimadı doğrultma xüsusiyyəti ilə bağlı tətbiq edilən bütün yanaşmalar və mülahizələr təkmilləşdirilmiş sistemə də tətbiq olunur. Bu yanaşma və mülahizələrə keyfiyyətlə (həm metrikalar, həm də ölçmə metodologiyaları), mühafizəliliklə və zərər vurulması ilə nəticələnən risklərlə bağlı yanaşmalar, həmçinin riskləri idarəetmə sistemləri (məsələn, təhlükəsizlik və məxfilik sahəsində mövcud olan) daxildir. 5.2-5.5 yarımbəndləri süni intellekt sistemlərinin etimadı doğrultması xüsusiyyətlərinin kontekstləşdirilməsi üçün müxtəlif çərçivələri təqdim edir.

5.2 Güvən səviyyələrinin tanınması

Sİ sisteminə konseptual olaraq funksional səviyyələrdən ibarət bir ekosistem kimi baxıla bilər. Sİ sisteminə onun mühitində güvənmək üçün hər bir funksional səviyyədə güvən yaradılır və dəstəklənir. Məsələn, “Güvənin təmin edilməsi” (Trust Provisioning) üzrə BTİ-T hesabatında [27] üç güvən səviyyəsi təqdim edilir: verilənlərin toplanması üçün fiziki infrastruktur (məsələn, sensorlar və aktuatorlar), verilənlərin saxlanması və emalı üçün İT-infrastruktur (məsələn, bulud) və tətbiqi proqramlar (ML alqoritmləri, ekspert sistemləri və son istifadəçilər üçün tətbiqlər) nəzərə alınmaqla, müvafiq olaraq fiziki güvən, kibergüvən və sosial güvən səviyyələri.

Fiziki güvən səviyyəsinə gəldikdə, burada göstəricilər (metrikalar) fiziki ölçmələrə və ya testlərə əsaslandığı üçün fiziki güvən çox vaxt etibarlılıq və mühafizəlilik anlayışlarını ehtiva edən ifadənin sinonimi kimi nəzərdə tutulur. Məsələn, avtomobilin texniki nəzarəti avtomobilin və onun daxili mexanizmlərinin etibarlı olmasını təmin edir. Bu kontekstdə güvən səviyyəsi yoxlama siyahısındakı bəndlərin yerinə yetirilmə səviyyəsi ilə müəyyən edilə bilər. Bununla yanaşı, sensorun kalibrlənməsi kimi bəzi proseslər ölçmələrin və nəticə etibarı liə, əldə olunmuş verilənlərin düzgünlüyünü təmin edə bilər.

Kibergüvən səviyyəsində problemlər Sİ sisteminin tamlığının qorunması və verilənlərin təhlükəsiz saxlanılması üçün girişə nəzarət və digər tədbirlər kimi daha çox İT infrastrukturunun təhlükəsizlik tələblərinin müəyyənləşdirilməsinə yönəlir.

Tətbiqlər səviyyəsində Sİ sisteminə güvənin yaradılması, həmçinin proqram təminatının etibarlı və təhlükəsiz olmasını tələb edilir. Kritik sistemlər kontekstində proqram təminatının yaradılması “məhsulu”n verifikasiyası və yoxlanılması prosesləri çoxluğu ilə müəyyənləşdirilir [28]. Bu, Sİ sistemləri və digərləri üçün də doğrudur. Maşın öyrənməsinə əsaslanan Sİ sistemlərinin stoxastik təbiətini nəzərə alaraq, etimadın dogruldulması

xüsusiyyəti həm də sistemin (yersiz qərəzin olmamasına uyğun olan) davranışının ədalətli olmasını nəzərdə tutur.

Sosial güvən insanın həyat tərzinə, inancına, xarakterinə və s. əsaslanır. Daxili funksiyası dəqiq başa düşülməyən sistemin fəaliyyət prinsipləri əhalinin texniki biliklərə malik olmayan təbəqəsi üçün şəffaf olmur. Bu halda, güvənin yaradılması Sİ sisteminin fəaliyyətinin obyektiv verifikasiyasından asılı olmasından daha çox Sİ sisteminin müşahidə olunan davranışının subyektiv pedaqoji izahına əsaslanır.

5.3 Proqram təminatı və verilənlərin keyfiyyət standartlarının tətbiqi

Proqram təminatı tipik Sİ sisteminin etimadın doğruldulması xüsusiyyətinə mühüm təsir göstərir. Nəticədə, proqram təminatının keyfiyyət atributlarının müəyyənləşdirilməsi və təsvir edilməsi bütün sistemin etimadın doğruldulması xüsusiyyətinin təkmilləşdirilməsinə kömək edə bilər [29]. Bu atributlar həm kibergüvənin, həm də sosial güvənin formalaşmasına töhfə verə bilər. Məsələn, ictimai nöqteyi-nəzərdən etimadın doğruldulması xüsusiyyəti imkan, dürüstlük və xeyirxahlıq baxımından təsvir edilə bilər [30]. Aşağıda bu terminlərin hazırda Sİ sistemləri kontekstində necə interpretasiya olunduğuna dair nümunələr verilmişdir.

- İmkan – Sİ sisteminin müəyyən bir işi yerinə yetirmək qabiliyyətidir (məsələn, tibbi təsvirdə şişin aşkarlanması və ya video monitorinq sistemində üzün tanınmasından istifadə etməklə şəxsin identifikasiyası). İmkanlarla bağlı atributlara dayanıqlıq, təhlükəsizlik, etibarlılıq və s. daxildir.

- Tamlıq – Sİ sisteminin sağlam əxlaqi və etik prinsiplərə riayət etməsinə və ya Sİ sisteminin məlumatları qəsdən manipulyasiya etməyəcəyinə əminlikdir. Beləliklə, tamlığın atributlarına bütövlük, dəqiqlik, müəyyənlik, ziddiyyətsizlik və s. daxildir.

- Xoşniyyətlilik – Sİ sisteminin “yaxşılıq etməsinə” inam və ya başqa sözlə, sistemin “zərər verməmək” prinsipinə riayət etməsi dərəcəsidir.

ISO/IEC SQuaRE seriyası modellər və ölçmələr (modellər üzrə ISO/IEC 2501x, ölçmələr üzrə ISO/IEC 2502x) vasitəsilə proqram təminatının keyfiyyətinin təmin olunması məsələlərinə həsr olunub, nəticədə proqram təminatının və verilənlərin keyfiyyət xüsusiyyətlərinin siyahısı təqdim olunur.

SQuaRE seriyasında aşağıdakı modellər fərqləndirilir:

- proqram məhsulunun keyfiyyəti (nəticə 8 xüsusiyyətdən ibarət siyahı şəklində);
- proqram məhsulu, verilənlər və İT xidmətlərindən istifadə keyfiyyəti (nəticə kibergüvənlə sosial güvəni fərqləndirməyə imkan verən, həmçinin azaldılması mümkün olan riskləri göstərən 5 xüsusiyyətdən ibarət siyahı şəklində);
- verilənlərin keyfiyyəti (nəticə 15 xüsusiyyətdən ibarət siyahı şəklində);
- İT xidmətinin keyfiyyəti (nəticə 8 xüsusiyyətdən ibarət siyahı şəklində).

Məsələn, ISO/IEC 25010 [20] standartında yaranan sosial tələblər “riskdən azad” kateqoriyasına aid edilir. [20] standartına əsasən, riskdən azad - “məhsulun və ya sistemin iqtisadi vəziyyəti, insan həyatı, sağlamlığı və ya ətraf mühit üçün potensial riskləri azaltma dərəcəsidir”.

ISO/IEC 25010 Beynəlxalq Standartların SQuaRE seriyasının tərkib hissəsidir və proqram məhsulunun keyfiyyəti və istifadə olunan proqram təminatının keyfiyyəti üçün xüsusiyyətlərdən və altxüsusiyyətlərdən ibarət modeli təsvir edir. ISO/IEC 25012 [19] həmin seriyanın tərkib hissəsidir və öz növbəsində, kompüter sistemlərində strukturlaşdırılmış formatda emal olunan verilənlər üçün verilənlərin ümumi keyfiyyət modelini müəyyən edir. ISO/IEC 25012 standartı kompüter sisteminin bir hissəsi olan verilənlərin keyfiyyətinə diqqət yetirir, insanlar və sistemlər tərəfindən istifadə olunan hədəf verilənlərin keyfiyyət xüsusiyyətlərini müəyyənləşdirir.

SQuaRE seriyası verilənlərini strukturlaşdırılmış şəkildə saxlayan və aşkar məntiqdən ("explicit logic") istifadə edərək emal edən ənənəvi proqram təminatı sistemləri üçün işlənib hazırlanmışdır. ISO/IEC 25012 dəqiqlik, tamlıq, əlçatanlıq, izlənilənlik və daşınanlıq kimi 15 fərqli xüsusiyyətdən istifadə etməklə verilənlərin keyfiyyət modelini təsvir edir.

Sİ sistemlərində həm sistemlərin, həm də verilənlərin keyfiyyət xüsusiyyətlərini ölçmək daha çətin ola bilər. ISO/IEC 25012 standartında verilənlərin keyfiyyət modeli Sİ sistemlərinin verilənlərə əsaslanan təbiətinin bütün xüsusiyyətlərini kifayət qədər əhatə etmir. Məsələn, dərin öyrənmə böyük həcmli verilənlər üzərində çoxsaylı gizli təbəqələrə malik neyron şəbəkələrinin öyrədilməsi vasitəsilə zəngin iyerarxik təsvirlərin yaradılması üçün yanaşmadır [31]. Bundan əlavə, Sİ sistemləri üçün verilənlərin keyfiyyəi modeli ISO/IEC 25012 standartında hazırda təsvir olunmayan, Sİ sistemlərinin işlənməsi üçün istifadə edilən verlənlərdə qərəzlilik kimi digər xüsusiyyətləri nəzərə almalıdır.

Sİ sistemlərini və onların asılı olduqları verilənləri daha adekvat əhatə etmək məqsədilə ISO/IEC 25010 standartında təsvir olunan ənənəvi sistemlərin və proqram təminatının işlənilməsinin və ISO/IEC 25012 standartında verilənlərin keyfiyyət modelinin xüsusiyyətlərindən və tələblərindən kənara çıxmaq üçün mövcud standartların genişləndirilməsi və ya dəyişdirilməsi zəruri ola bilər.

5.4 Riskləri idarəetmə tətbiqləri

Risklərin idarə olunması konkret Sİ məhsulunun və ya xidmətinin bütün ömrü boyu etimadı doğruldulmasının təmin edilməsinə kömək edən preventiv prosesdir. Risklərin idarə edilməsinin ümumi prosesi ISO 31000:2018 [14] standartında müəyyən edilmişdir və maraqlı tərəflərin və onların zəif aktivlərinin və dəyərlərinin müəyyənləşdirilməsini, nəticələrinə və ya təsirinə əsaslanaraq risklərin qiymətləndirilməsini, təşkilatın məqsədlərinə və risklərə dözümlülüyünə əsaslanaraq risklərin aradan qaldırılması üzrə qərarların qəbul edilməsini əhatə edir. ISO 31000 [14] standartına uyğun olaraq, risk "qeyri-müəyyənliyin məqsədlərə təsiri" kimi təyin edilir, burada nəticə gözlənilən dən fərqlənir və gözlənilməz hadisənin baş vermə ehtimalı və maraqlı tərəflərə təsir səviyyəsi baxımından ölçülür və ya qiymətləndirilir. Risklərin idarə edilməsi gözlənilməzliklərin daha çox olduğu yeni texnologiyalar üçün xüsusilə önəmlidir. O, həmçinin mahiyyət etibarı ilə insan səhvləri və bədniyyətli hücumlar kimi risklərlə əlaqəli situasiyalarla işləmək üçün də ən uyğun yanaşmadır. Bundan əlavə, risklərin idarə edilməsi ümumi qəbul edilmiş keyfiyyət göstəricilərinin hələ də müəyyən edilmədiyi sahələrdə qeyri-müəyyənliyin qarşısının alınmasına şərait yaradır. Sadalanan xüsusiyyətlər Sİ sistemləri üçün xarakterik olduğu üçün bu sistemlər risklərin idarə edilməsi nöqteyi-nəzərindən xüsusi maraq kəsb edir.

Burada ISO 31000 standartının əsas konsepsiyaları onların Sİ sistemlərinə necə tətbiq oluna biləcəyi baxımından təqdim olunur. Bu konsepsiyalara maraqlı tərəflərin müəyyən edilməsi, təşkilatın məqsədləri, idarəetmə məsələləri (məqsədləri), nəzarət vasitələri və onlarla əlaqəli tədbirlər daxildir. Sİ sistemlərinə gəldikdə, maraqlı tərəflərin dairəsi xüsusilə geniş ola bilər və yalnız təşkilatın özünü, tərəfdaşlarını və müştərilərini deyil, həm də bütövlükdə cəmiyyəti və ətraf mühiti əhatə edir. Sİ sistemlərinin tərtibatçısı, distribyutoru və ya istifadəçisi olmağından asılı olmayaraq, təşkilatın ən yüksək səviyyəli məqsədlərinə reputasiyanı qorumaq, təhlükəsizlik və məxfiliyi, ədalətlilik və mühafizəliliyi təmin etmək aiddir. Sistemin etimadı doğrultmasına nail olmaq – daha konkret idarəetmə məsələlərinə (və ya risk mənbələrinə) çevrilən bütün təşkilati məqsədlərin təmin olunmasına əsaslanır. İdarəetmə məsələləri adətən zəifliklər, "tələlər" və ya gözlənilən təhdidlərlə əlaqəli olur.[1] Sİ sistemləri üçün bunlara hesabatlılıq, yeni təhlükəsizlik təhdidləri, yeni məxfilik təhdidləri, düzgün olmayan spesifikasiya, natamam tətbiq, yanlış istifadə və müxtəlif qərəzlilik mənbələri ilə bağlı (lakin bunlarla məhdudlaşmayan) məsələlər aiddir. Müəyyən edilmiş idarəetmə məsələlərinin hər biri üçün mümkün nəzarət vasitələri (və ya təsirlərin azaldılması tədbirləri) müəyyən edilə bilər. Sİ sistemləri üçün bunlara aşağıdakılarla məhdudlaşmayan:

- şəffaflığın təmininə dair yanaşmalar;
- təhlükəsizlik üzrə yeni nəzarət vasitələri;
- məxfiliyin təmini üzrə yeni nəzarət vasitələri;
- dayanıqlılıq və dözümlülüklə əlaqəli mülahizələr;
- maşın örənmə alqoritmlərinin seçimi və konfiqurasiyası ilə bağlı mülahizələr;
- verilənlərlə bağlı mülahizələr;
- sistemin idarəolunanlığı ilə bağlı mülahizələr aiddir.

Riskləri idarəetmə prosesi hər bir nəzarət vasitəsi ilə bağlı əlavə tədbirlər görərək, təşkilatın siyasəti və iş şəraitindən asılı olaraq seçim məqsədilə onları müvafiq təlimatlar (və ya tədbirlər) çoxluğuna yönəldir. Riskləri idarəetmə prosesinin strukturu yaradıldıqdan sonra onun düzgün tətbiqi və korrekt quraşdırılması alqoritmik performans metrikaları və sahə üzrə sınaqlar daxil olmaqla (yalnız bunlarla məhdudlaşmayaraq), müxtəlif qiymətləndirmə və ölçmə yanaşmalarından istifadə edərək mütəmadi testlərdən, təhlillərdən və təkmilləşdirmələrdən keçir.

5.5 Texniki təminat dəstəkli yanaşmalar

Tipik maşın öyrənməsi sistemləri (həm təlim, həm də istifadə üçün) sistemin icra prosesinin korrektliyinə təsir göstərə bilən ümumi və hazır istifadə olunan etibarsız platformalarda quraşdırılır. Məsələn, maşın öyrənməsi proqramları çox vaxt çox-istidadəçili (çox-kirayəçili) buludda quraşdırılır. Texniki təminat mexanizmləri verilənlərin və hesablamaların həm təlim, həm istifadə zamanı konfidensiallığını və tamlığını qoruyan "güvənli icra mühitləri" (Trusted execution environments, TEE) təmin etməklə hücum sahəsini azaldılır.

TEE-lər proqramların və ya qorunan icra sahələrinin texniki (aparat) təminat tərəfindən məcburi təcrid edilməsini təmin etməklə seçilmiş kodu və verilənləri açıqlanmadan və ya dəyişdirilmədən qorumaq üçün istifadə olunur ki, bu da hətta etibardan düşmüş (hücuma

[1] Qeyd: Təşkilatın məqsədləri ilə idarəetmə məsələləri (məqsədləri) arasındakı inikas birqiymətli olmaya bilər. Mürəkkəb sistemlərdə (məsələn, Sİ sistemləri) təşkilatın hər bir məqsədinə çatmaq adətən müəyyən edilmiş bir çox idarəetmə məqsədlərinə nail olmağı tələb edir.

məruz qalmış) platformalarda belə təhlükəsizlik səviyyəsini artırır. Tərtibatçılar “güvənli icra mühitləri”ndən istifadə edərək, zərurət yarandıqda model ilə mühafizə olunan verilənlər və ya intellektual mülkiyyət kimi effektiv şəkildə davranmaqla, maşın öyrənməsi modelini onun həyat dövrü ərzində (məsələn, onun təlimi və istifadəsi dövründə) qoruya bilərlər. TEE-lər hətta sistem proqram təminatı səviyyələrində imtiyazlı zərərli proqramların mövcudluğu şəraitində belə maşın öyrənməsinin iş yükünü daşıyan yaddaşın (adətən yaddaşın şifrlənməsi mexanizimindən istifadə etməklə) konfidensiallığını və tamlığını təmin edir.

6 MARAQLI TƏRƏFLƏR

6.1 Ümumi anlayışlar

Bu sənəddə “maraqlı tərəflər”in ISO/IEC 38500 [17] standartında geniş mənada verilmiş tərifindən istifadə olunur, beləki bu tərif fərdləri və təşkilatları maraqlı tərəf kimi tanımaqla yanaşı:

- bir qrup insanı da maraqlı tərəf növü hesab edir; bu, təşkilat olmayan qrupun (yəni qrup onu təmsil edən müştərək idarəetmədən faydalanmır) bölüşdüyü kollektiv mövqelərin başa düşülməsi və nəzərə alınması üçün vacibdir;
- sistemin təsir edə biləcəyi maraqlı tərəf növlərini əhatə edir; bu cür yanaşma məqsədəuyğun olsa da, bəzən sistemlərlə bağlı hər hansı ehtiyac və ya gözləntiləri olmayan tərəfləri də ehtiva edə bilər. Bu da öz növbəsində, sistem barədə qabaqcadan məlumatlı olmağı tələb edir.

Bu daha geniş tərif (bəziləri sistemin mövcudluğu, məqsəd və ya imkanlarından xəbərsiz ola bilən) maraqlı tərəflərin müəyyənləşdirilməsi nəzərdən keçirilərkən əhəmiyyət kəsb edir.

ISO/IEC 27000:2018 [1] standartında istifadə olunan “aktiv” termininin tərifinə gəlincə, bu tərif yuxarıda müzakirə olunan “maraqlı tərəflər”lə bağlı nəzərdən keçirildiyi zaman yetərli olmur. Bunun əvəzinə, bu standartda “aktiv” terminindən istifadə olunduqda “maraqlı tərəf üçün dəyəri olan hər şeyə” istinad olunması nəzərdə tutulur. Bu, [1] istinadında yalnız təşkilatların dəyərli aktivlərə sahib olacağı ilə bağlı qeyd edilən fərziyyəni genişləndirir, çünki bunu fərdlərə və insan qruplarına da şamil etmək olar.

Bu standart çərçivəsində müəyyən olunan dəyərlər təşkilatlarla məhdudlaşmır ([23] istinadında qeyd olunduğu kimi), əksinə, istənilən maraqlı tərəfin “əməl etdiyi” inancları və “riayət etməyə çalışdığı standartları” ehtiva edir.
Sİ sistemləri çevik və dinamik şəkildə idarə edilə bilir, bunu nəzərə alaraq Sİ-nin etimadı doğrultma xüsusiyyətinə dair yanaşmalar həm etimadın qazanılmasına, həm də onun qorunmasına yönəlməlidir. Etimadı doğruldan Sİ-nin konkret xüsusiyyətləri üçün dəqiq kontekst təmin edən təriflər vasitəsilə buna nail olmaq mümkündür. Belə ki, kontekstdə baş verən dəyişiklik bəyan edilən xüsusiyyətin kritik şəkildə yenidən qiymətləndirilməsinə səbəb ola bilər [32]. Bu mənada, sadəcə, “etimadı doğruldan Sİ”yə istinad etmək yetərli deyil, bununla yanaşı, Sİ-nin işlənməsi və istifadəsi zamanı hansı aspektlərdə kimin kimə güvəndiyinin dəqiqləşdirilməsi lazımdır. Beləliklə, maraqlı tərəflərin bu cür kontekstuallaşdırılması etimadı doğruldan Sİ-nin şəffaflıq (bax: 9.2), izahlılıq (bax: 9.3) və idarəolunanlıq (bax: 9.4) kimi xüsusiyyətlərinin nəzərdən keçirilməsinə tətbiq oluna bilər. Kontekstuallaşdırma üçün maraqlı tərəflər dəqiqliklə müəyyən edilməli, onların Sİ sisteminin həyat dövrü və dəyər zəncirlərinin müxtəlif mərhələlərində iştirakı aydın şəkildə başa düşülməlidir.

Müxtəlif maraqlı tərəflər etimadı doğruldan Sİ üçün təklif olunan müxtəlif xüsusiyyətlərin nisbi əhəmiyyətinə dair fərqli fikirlərə malik ola bilər. Ona görə də etimadı doğruldan Sİ üçün şərtlər və konseptual çərçivənin standartlaşdırılması müxtəlif maraqlı tərəflər arasında ünsiyyətin aydın və birmənalı olmasını təmin edəcək. Beləcə, nəzərdə tutulan fikir ayrılığını başa düşmək və onların aradan qaldırılması yollarını axtarmaq mümkün olacaq. Maraqlı tərəflərlə bu cür ünsiyyət çərçivəsində aşağıdakı məqamlara diqqət yetirilməlidir:

- məhsul və ya xidmətdə istifadə edilən Sİ texnologiyası müxtəlif maraqlı tərəflərə necə təsir göstərə bilər;
- müxtəlif maraqlı tərəflərin dəyər verdiyi istənilən aktivdən necə istifadə edilir və ya məhsul və ya xidmətdə Sİ-dən istifadə onlara necə təsir göstərir;
- məhsul və ya xidmətdə Sİ-dən istifadə müxtəlif maraqlı tərəflərin dəyərləri ilə necə əlaqələnir.

6.2 Növlər

Məhsul və ya xidmətlərdə Sİ-dən istifadə ilə əlaqədar maraqlı tərəflərin tipologiyası ilə bağlı dəqiq konsensus hələ mövcud deyil. Biznesdə maraqlı tərəflər nəzəriyyəsi [33] qərar qəbuletməyə olan yanaşmanın üstünlüklərini vurğulayır ki, burada rəhbərliyin səhmdarlar üçün mənfəət formalaşdırmaq məqsədilə fidusiar öhdəliklərindən kənara çıxması nəzərdə tutulur. Bu zaman təşkilat daxilində digər maraqlı tərəflər, o cümlədən işçilər, müştərilər, rəhbərlik, təchizatçılar, kreditorlar, hökumət və tənzimləyici orqanlar, ümumilikdə, cəmiyyət və təbii mühit üçün (gələcək nəsilləri təmsil edən nümayəndə kimi) faydalar (üstünlüklər) nəzərə alınır.

Sİ kontekstində baxılan növ maraqlı tərəfləri Sİ-nin dəyər zəncirində aşağıdakı fərqli rollara münasibətdə (bir maraqlı tərəf bir neçə belə rol icra edə bilər) nəzərdən keçirə bilərik:

- verilənlər mənbəyi: Sİ sistemini öyrətmək üçün istifadə olunan verilənləri təmin edən təşkilat və ya fiziki şəxs;
- Sİ sistemi tərtibatçısı: Sİ sistemini layihələndirən, işləyib hazırlayan və öyrədən (təlim keçən) təşkilat və ya fiziki şəxs;
- Sİ istehsalçısı: ən azı, bir Sİ sistemindən istifadə edən məhsul və ya xidməti layihələndirən, işləyib hazırlayan, test edən və quraşdıraraq tətbiq edən təşkilat və ya fiziki şəxs;
- Sİ istifadəçisi: ən azı, bir Sİ sistemindən istifadə edən məhsul və ya xidmət istehlakçısı olan təşkilat və ya fiziki şəxs;
- Sİ alətləri və aralıq proqram təminatı tərtibatçısı: Sİ alətlərini və ilkin təlim keçmiş Sİ üzrə hazır kod bloklarını (“tikinti blokları”) layihələndirən və işləyib hazırlayan təşkilat və ya fiziki şəxs;
- test və qiymətləndirmə üzrə nümayəndəlik (orqan): müstəqil testləşdirmə və mümkün olduğu halda, sertifikatlaşdırma təklif edən təşkilat və ya fiziki şəxs;
- Si sisteminin tətbiq edildiyi daha geniş cəmiyyət (çünki hətta dəqiq funksionallığa malik Sİ sistemi mövcud bərabərsizliklərin təsdiqlənməsinə gətirib çıxara bilər);
- ayrı-ayrı şəxslərin mövqeyini təmsil edən birliklər;

- Sİ-dən istifadəni monitorinq və tədqiq edən idarəetmə təşkilatları, o cümlədən milli hökumətlər və Beynəlxalq Valyuta Fondu (BVF) kimi beynəlxalq təşkilatlar.

6.3 Aktivlər

Maraqlı tərəfləri onların dəyər verdiyi aktivlərlə xarakterizə etmək olar. Bu zaman məhsul və ya xidmətdə Sİ-dən istifadə zamanı müəyyən rola malik, yaxud Sİ-dən istifadənin təsirinə məruz qalan aktivlər nəzərdə tutulur.

Sİ ilə bağlı maddi aktivlər aşağıdakılardan ibarət ola bilər:

- Sİ sisteminə təlim keçmək üçün istifadə olunan verilənlər;
- təlim keçmiş Sİ sistemi;
- bir və ya bir neçə Sİ sistemindən istifadə edən məhsul və ya xidmət;
- məhsul və ya xidmətin Sİ ilə əlaqəli davranışını yoxlamaq üçün istifadə edilən verilənlər;
- məhsul və ya xidmətə işlənilməsi üçün ötürülən və əsasında Sİ-əsaslı qərarlar qəbul edilən verilənlər;
- Sİ sistemlərinin təlimi, test edilməsi və istismarı üçün istifadə olunan hesablama resursları və proqram təminatı;
- aşağıdakı bacarıqlara malik insan resursları:
    - Sİ sistemlərinə təlim keçmək, onları test və idarə etmək;
    - həmin tapşırıqların tərkibində və ya həmin tapşırıqların həlli üçün istifadə olunan proqram təminatını işləyib hazırlamaq;
    - Sİ təlimi üçün lazımi verilənləri generasiya etmək, seçmək və ya verilənlərə annotasiya vermək.

Daha az maddi aktivlərə aşağıdakılar aiddir:

- Sİ sisteminin və ya ondan istifadə edən xidmət və ya məhsulun işlənməsi, test olunması və ya istismar edilməsi ilə məşğul olan maraqlı tərəfin reputasiyası və ona olan güvən;
- vaxt, məsələn, məhsul və ya xidmət istifadəçisinin Sİ-nin istifadəsi sayəsində qənaət etdiyi vaxt və ya Sİ sisteminin qeyri-münasib (yersiz) tövsiyəsinə reaksiya vermək üçün boşuna sərf olunan vaxt;
- Sİ-nin təmin etdiyi avtomatlaşdırma sayəsində dəyərini itirə biləcək bacarıqlar;
- tapşırıqla əlaqədar informasiyanın Sİ sistemi tərəfindən daha səmərəli şəkildə təqdim edilməsi sayəsində gücləndirilə və ya zəiflədilə bilən avtonomluq (məsələn, Sİ vasitəsilə profilləşdirmədən istifadə etməklə bir fərd və ya bir qrup fərdlər üçün nəzərdə tutulmuş reklam və ya siyasi mesajlar kimi inandırıcı məzmunla).

6.4 Dəyərlər

Maraqlı tərəflərin etimadı doğruldan Sİ sisteminin müvafiq xüsusiyyətlərinə dair fikirləri fərqlənə bilər. Bu onların sadiq qaldığı və ya riayət etməyə çalışdığı müxtəlif dəyərlərdən asılıdır. Etimadı doğruldan Sİ-nin yaradılması ilə bağlı bəzi təkliflər (məsələn, Fundamental

hüquqlara dair Avropa Xartiyasına əsaslanan prinsipləri təklif edən Avropa Komissiyasının Etimadı doğruldan Sİ üzrə Yüksəksəviyyəli Ekspert Qrupunun işçi sənədi [34]) konkret siyasətdə müəyyən edilmiş xüsusi dəyərlərə əsaslanır. Etimadı doğruldan Sİ ilə əlaqədar fərqli baxışlar müxtəlifliyi ilə seçilən mənəvi dünyagörüşü və ya dəyərlər sistemindən də irəli gələ bilər. Qərb etikası, buddizm, ubuntu, sintoizm kimi müxtəlif dünyagörüşlərinin aktuallığı və Sİ-yə təsiri hələ də kifayət qədər araşdırılmayıb [35]. Ümumiyyətlə, "dəyərə həssas" layihələndirməyə [36] gəlincə, etimadı doğruldan Sİ xüsusiyyətlərinin qlobal miqyasda yayılması üçün bu dəyər fərqlərinin aydın başa düşülməsi vacibdir.

7 ƏSAS PROBLEMLƏR

7.1 Məsuliyyət, hesabatlılıq və idarəçilik

Sİ sistemlərinin işlənməsi və tətbiqi elə bir mühiti əks etdirir ki, orada informasiya texnologiyaları çoxsaylı maraqlı tərəflərin iştirak etdiyi mühitdə tətbiq olunur. Bu cür mühitdə etimad formalaşdırmaq və onu qoruyub saxlamaq üçün maraqlı tərəflər arasında məsuliyyət və hesabatlılığın müəyyənləşdirilməsi vacibdir. Sİ sistemləri həm mürəkkəb beynəlxalq kommersiya dəyər zəncirlərində, həm də transmilli sosial strukturlarda tətbiq olunduqlarına görə vacibdir ki, bütün maraqlı tərəflər digər maraqlı tərəflər qarşısında daşıdıqları məsuliyyəti və həmin məsuliyyətə görə hesabatlılığı eyni cür anlasın. Belə bir yanaşmaya (çərçivəyə) dair razılığa gəlməyin əsas səbəblərindən biri Sİ sisteminin həyat dövrü ərzində qərar qəbuletmə məqamlarını müəyyən etməyin mümkün olmasıdır.

Təşkilat daxilində qərarvermə məsuliyyəti və qərarların nəticələrinə görə hesabatlılıq, adətən, idarəetmə sistemi çərçivəsində əhatə olunur. ISO/IEC 38500 [17] standartı təşkilat daxilində yüksəksəviyyəli qərarlar qəbul edən şəxslərə onların İT-dən istifadə ilə əlaqədar öz hüquqi, normativ və etik öhdəliklərini başa düşməyə və yerinə yetirməyə kömək edir. Bu standart məsuliyyət, strategiya, verilənlərin əldə edilməsi, performans, uyğunluq və insan davranışı prinsiplərinin həyata keçirilməsi çərçivəsində İT aspektlərinin qiymətləndirilməsi, idarə edilməsi və monitorinqi ilə əlaqədar tapşırıqları müəyyənləşdirir. ISO/IEC 38505[37] standartı bu İT idarəetmə modelini verilənlərin idarə edilməsinə tətbiq edir. Həmin standartda Sİ-nin zəifliklərindən xüsusi olaraq bəhs edilmir, lakin verilənlərə əsaslanan Sİ ilə əlaqəli bəzi məsələlər, məsələn, maşın öyrənməsi nəzərdən keçirilir. ISO/IEC 38500 standartından götürülmüş İT idarəetmə modelini etimadı doğruldan Sİ sistemlərinə olan ehtiyaca daha da uyğunlaşdırmaq üçün (xüsusən də onların digər sosial qərar qəbuletmə strukturları ilə qarşılıqlı əlaqəsinə və avtonom sistemlərdə istifadəsinə münasibətdə) imkanlar yarana bilər. Bu sənəddə "avtonom sistem" terminindən istənilən növ sistemlərə istinad üçün istifadə olunur, bu zaman nisbətən müstəqil fəaliyyət göstərən və avtomobil və ya "maşınlar" ilə məhdudlaşmayan sistemlər nəzərdə tutulur.

Məsələn, həkim öz diaqnozunu dəqiqləşdirmək üçün Sİ-dən istifadə edə bilər. Həkim qoyduğu diaqnoza görə məsuliyyət daşıyır, çünki həmin konkret sahə üzrə ixtisaslaşmış ekspertdir. Elə bu səbəbdən həkim diaqnoz barədə qərar qəbul edərkən Sİ sisteminin nəticələrinin nəzərə alınmasına görə cavabdehlik daşıyır. Digər tərəfdən, işə qəbul üçün müraciət edən son istifadəçi konkret sahə üzrə ekspert deyilsə, ərizəsinin qəbul edilməməsinin səbəbini anlamaqda daha çox çətinlik çəkə bilər. Sonuncu nümunədə cavabdehlik zənciri aydın şəkildə müəyyən edilməlidir.

Sİ ilə idarə olunan avtonom sistemlərdə etimadı doğrultma xüsusiyyətini təmin etmək üçün avtonom sistemdə nasazlığın baş verdiyi halda məsuliyyət və hesabatlılığın

müəyyənləşdirilməsi vacibdir [38][40]. Bu zaman, avtonom sistemin işi zərərə səbəb ololduğu təqdirdə, müvafiq maraqlı tərəflərin hüquqi məsuliyyətə cəlb olunmasına imkan yaranır. Elm və yeni texnologiyalar sahəsində etika məsələləri üzrə Avropa Qrupu [41] özünün 2018-ci il bəyanatında vurğulayır ki, Sİ ilə idarə olunan sistem hüquqi anlamda avtonom ola bilməz. Odur ki, istənilən avtonom sistemin işi hər hansı zərərə səbəb olarsa, həmin zərərin əvəzinin ödənilməsini təmin etmək üçün dəqiq cavabdehlik çərçivəsi və hüquqi məsuliyyət müəyyən edilməlidir.

7.2 Mühafizəlilik

Mühafizəlilik – etimadı doğrultma xüsusiyyətinin ən vacib aspektidir. Buna görə də mühafizəlilik aspektlərinin nəzərə alınması yüksək prioritetə malikdir. Adətən, sistem tərəfindən zərərin vurulması riskinin ehtimalı nə qədər yüksək olarsa, etimadı doğrultma xüsusiyyətinə dair tələblər də bir o qədər yüksək olur.

İstənilən digər sistemlər kimi, Sİ sistemlərindən də gözlənilən onların heç bir zərərə qəsdən səbəb olmamasıdır. Qeyd olunanlara yalnız maddi (məsələn, canlı varlıqların sağlamlığına, əmlaka və fiziki ətraf mühitə vurulan) zərər deyil, həm də qeyri-maddi (məsələn, sosial və mədəni mühitə vurulan) zərər də aiddir.

ISO/IEC 51 Təlimatında [10] qeyd olunur ki, “Bütün məhsul və sistemlər təhlükə və buna görə də müəyyən dərəcədə qalıq risk ehtiva edir. Bununla belə, həmin təhlükələrlə əlaqədar risk yol verilən səviyyəyə endirilməlidir. Mühafizəlilik riskin yol verilən səviyyəyə endirilməsi ilə əldə olunur ki, bu da hazırkı Təlimatda məqbul risk kimi müəyyən edilir”. Müəyyən səviyyədə avtonomluq təmin edən Sİ sistemləri çox vaxt mühafizəlilik baxımından daha kritik hesab olunur. Bununla belə, təhlükələr və risklər tətbiqdən asılıdır və sistemin avtonomluq səviyyəsi ilə bilavasitə mütləq şəkildə əlaqəli deyil (məsələn, avtonom yol nəqliyyat vasitəsi və ya avtonom məişət tozsoranı).

Sİ sisteminin layihələndirilməsindən tutmuş utilizasiyasına qədər tam həyat dövrü mühafizəlilik aspektləri baxımından nəzərə alınmalı məsələyə çevrilir. Sİ sisteminin tətbiqləri, onun təyinatlı istifadəsi və rasional şəkildə proqnozlaşdırıla bilən sui-istifadə halları, eləcə də istifadə olunduğu mühit və istifadə olunan texnologiyalar diqqətlə nəzərdən keçirilməli olan predmetə çevrilir. ISO/IEC 51 Təlimatında [10] “rasional şəkildə proqnozlaşdırıla bilən sui-istifadə” məhsul və ya sistemin təchizatçı tərəfindən nəzərdə tutulmayan, lakin asanlıqla proqnozlaşdırıla bilən insan davranışından irəli gələ bilən istifadəsi anlamında işlədilir. Sİ-yə məxsus spesifik zəifliklərə görə Sİ sistemləri üçün yeni risklər meydana çıxa bilər. Bu, qeyri-şəffaf və ya qeyri-deterministik qərar qəbuletmə prosesləri kimi Sİ-yə xas xüsusi davranış sayəsində riskin məqbul səviyyəyə qədər azaldılması üçün yeni tədbirlərin həyata keçirilməsinə səbəb olacaq. Konkret bir sistem üçün yalnız Sİ-nin hissəsi deyil, istifadə olunan bütün texnologiyalar və onların qarşılıqlı əlaqəsi də diqqətlə nəzərdən keçirilməlidir. ISO/IEC 51 Təlimatında [10] məqbul risklərə nail olmaq üçün ümumi tövsiyələr təqdim olunur.

8 ZƏİFLİKLƏR, TƏHDİDLƏR VƏ DİGƏR PROBLEMLƏR

8.1 Ümumi müddəalar

Burada Sİ sistemlərinin potensial zəifliklər və onlarla əlaqəli təhdidlər təsvir edilir. Zəifliklər ISO/IEC 27000 [1] standartı tərəfindən bir və ya bir neçə təhdid tərəfindən istifadə edilə bilən aktiv və ya idarəetmə ilə bağlı boşluq kimi müəyyən edilir. Təhdidlər ISO/IEC 27000 [1] standartı tərəfindən sistemə və ya təşkilata zərər verə biləcək arzuolunmaz insidentin potensial səbəbi kimi müəyyən edilir.

Müxtəlif maraqlı tərəflər "zəiflik" anlayışını təsvir etmək üçün müxtəlif terminlərdən istifadə edirlər. Bunlara risk mənbələri, tələlər, uğursuzluq mənbələri, "kök" səbəblər və çətinliklər daxildir.

Zəifliklərin çox hissəsi Sİ sistemlərində maşın öyrənməsinin istifadəsi ilə bağlıdır. Bunlara verilənlərdən asılılıq, maşın öyrənmə modellərinin qeyri-şəffaflığı və proqnozlaşdırılmazlıq aiddir. Xüsusi halda, verilənlərin istifadəsi yeni təhlükəsizlik təhdidlərinə və qərəzli nəticələrə səbəb ola bilər.

Sİ sistemlərinin layihələndirilməsi, işlənməsi və quraşdırılması üçün "ən yaxşı təcrübələrin" olmaması ilə bağlı problemlər əlavə zəifliklərin və təhdidlərin yaranmasına səbəb ola və ya mövcud zəiflikləri və təhdidləri artıra bilər.

Müəyyən təhdidlər Sİ sistemlərinin texnoloji imkanlarının yetərincə başa düşülməməsi və müxtəlif maraqlı tərəflərin onları məqsədəuyğun şəkildə istifadə etməməsi səbəbindən yaranır.

8.2-8.10-cu bəndləri Sİ sistemlərinin potensial zəifliklərini, təhdidlərini və digər problemlərini daha ətraflı təsvir edir.

8.2 Sİ yönələn xarakterik təhlükəsizlik təhdidləri

8.2.1 Ümumi müddəalar

Sİ-nin inkişafı rəqəmsal təhlükəsizlik baxımından həm üstünlüklərə, həm də çatışmazlıqlara malikdir. Bir tərəfdən, Sİ texnologiyaları bədniyyətlər və onların zərərli fəaliyyətlərinin profilinin yaradılması, daha sonra onlarla mübarizə üçün təhlükəsizlik həllərinin işlənməsi üçün istifadə edilə bilər. Digər tərəfdən, Sİ və maşın öyrənməsindən istifadə etməklə hazırlanmış qabaqcıl texnologiya zərərli məqsədlər üçün sui-istifadə edilə bilər. Məsələn, Sİ parolları tapmaq və rəqəmsal hesablara müdaxilə üçün istifadə edilə bilər.

Əksər sistemlərə olan ümumi İT təhlükəsizliyi təhdidlərinə əlavə olaraq (məsələn, proqram təminatı səhvləri, aparat təminatında "arxa qapılar", verilənlərin təhlükəsizliyi pozuntuları), maşın öyrənmə sistemləri kimi müəyyən SI sistemləri xüsusi və ya məqsədli təhlükəsizlik təhdidlərinə qarşı həssas ola bilər. Belə təhdidlərə aşağıdakılar daxildir [43]:

- Sİ sisteminin nasazlığına səbəb olan verilənlərin yoluxması;
- əlverişli Sİ sistemindən sui-istifadə edən rəqabətli hücumlar;
- model oğurluğu.

8.2.2 Verilənlərin yoluxdurulması

Verilənlərin yoluxdurulması (“zəhərləndirilməsi”) hücumları zamanı bədniyyətlər proqnoz modelinin nəticələrini manipulyasiya etmək məqsədilə təlim verilənlərinə qəsdən təsir edirlər. Verilənlərin yoluxdurulması hücumlarında məqsəd serverə bədniyyətli hücumları aşkar edə bilməyən pis bir model öyrədilməsinə imkan verməkdən ibarətdir. Model internetdə onlayn rejimdə əlçatan olduqda bu növ hücum xüsusilə aktualdır, yəni model yeni verilənlər əsasında öyrənmə ilə davamlı şəkildə yenilənir. Təlim verilənlərinin lazımi qaydada filtrlənməsi vasitəsilə “qeyri-normal” verilənlər aşkar olunaraq filtrlənə bilər və beləliklə, mümkün ziyan minimallaşdırıla bilər.

8.2.3 Rəqabətli hücumlar

Sİ sistemləri ilə əlaqəli təhlükəsizlik təhdidlərindən biri maşın öyrənmə sistemlərinə olan rəqabətli hücumlardır. Rəqabətli hücum zamanı faktiki modelə bir qədər təhrif edilmiş giriş verilənləri daxil edilir (məsələn, avtonom nəqliyyat sistemini aldatmaq məqsədilə giriş verilənlərini səhv təsnifatlandırmaq üçün sistemə daxil edilən bir qədər dəyişdirilmiş yol nişanı).

Maşın öyrənməsi sahəsində, xüsusilə də dərin öyrənmə sahəsində əldə olunmuş son əhəmiyyətli irəliləyişlər buna gətirib çıxarıb ki, şirkətlər arasında mühafizəlilik və təhlükəsizlik baxımından kritik olan tətbiqi proqram kontekstlərində maşın öyrənməsi alqoritmlərinin tətbiqinə maraq artıb. Bu nümunələrə konvolyusiyalı (bürünmə) neyron şəbəkəsinə (CNN) əsaslanan semantik seqmentasiyanın avtonom avtomobillərə və ya tibbi diaqnostika üçün neyron şəbəkələrinə inteqrasiyası daxildir. Aydındır ki, bu yeni tətbiq kontekstlərində keyfiyyətə və keyfiyyətin təminatına dair qoyulan tələblər son dərəcə yüksəkdir. Əksər hallarda, yüksək dəqiqliyi yaxşı hazırlanmış təlim, test və yoxlanılma verilənləri çoxluğu üzərində sübuta yetirmək kifayət etmir. Təlim keçmiş maşın öyrənməsi modulu konkret verilənlərin paylanmasının ümumi hallarını emal etməklə yanaşı, baş verən nadir halların və ya hətta pis niyyətlə hazırlanmış giriş nöqtələrinin də öhdəsindən gələ bilməlidir [44].

Szegedy et al.[45] və Goodfellow et al.[46] nəşrləri göstərir ki, kompüter görməsi üçün maşın öyrənməsi alqoritmləri, xüsusən də dərin neyron şəbəkələri rəqabətli nümunələrə (və ya rəqabətli perturbasiyalara) əsaslanan hücumlara məruz qala bilər.

Bu rəqabətli perturbasiyalar rəqabətli hücumdan istifadə edən bədniyyətlər tərəfindən yaradılır və adətən, onlar normal giriş verilənlərinə əlavə edildikdə sözügedən maşın öyrənməsi modulunun səhv davranışını nəzərdə tutur. Əlavə olaraq, bu rəqabətli perturbasiyaları aşkar etmək çox vaxt çətin olur, belə ki, onlar hətta insan gözü üçün görünməz olur. Bu görünməzlik mühafizəlilik və təhlükəsizlik baxımından kritik olan mühüm sənaye sahələrində sistemin arzuolunan şəkildə quraşdırılmasına maneə törədir və bununla yanaşı, həm də insanlarda və süni neyron şəbəkələrində sensor informasiya emalı arasındakı mühüm fərqə işarə edir [47]. Bu zəiflik aşkar edildikdən indiyədək xeyli sayda fərqli rəqabətli hücumlar və müdafiə vasitələri (müdafiə strategiyaları) barədə məlumat nəşr edilmişdir [48][49]. Bu, hücum edənlər və müdafiə olumamlar arasında silahlanma yarışına çevrilmişdir. Bir sıra mövcud hücumlara qarşı yeni dərc edilmiş müdafiə vasitələri çox vaxt bir neçə həftə ərzində lazımsız hala gəlir [50], bunun səbəbi daha güclü hücumların yaradılmasıdır. Bu cür hücumları ümumiləşdirmək çətin olsa da və onlar istehsal sistemlərinin (adətən, gizli saxlanılan) daxili şəbəkə topologiyası haqqında ətraflı biliklərə əsaslansa da, onlar etimadı doğrultma baxımından ciddi şəkildə nəzərdən keçirilməlidir.

8.2.4 Model oğurluğu

Həm təhlükəsizlik, həm də məxfilik aspektlərinə təsir göstərən model oğurluğu hücumlarından modellərin daxili funksiyalarını kopiyalayaraq (replikasiya) onları "oğurlamaq" üçün istifadə olunur. Bu məqsədlə hədəf modelə çoxlu sayda proqnoz sorğuları göndərilir. Nəticədə, əldə olunan cavabdan (proqnozdan) başqa modeli öyrətmək üçün istifadə edilir.

8.2.5 Konfidensiallığa və tamlığa aparat yönlü təhdidlər

Maşın öyrənməsi tətbiqləri digər həssas tətbiqlərlə eyni hücumlara tez-tez məruz qalır. Maşın öyrənməsi tətbiqlərinə tipik proqram və aparat hücumları verilənlərin konfidensiallığına, verilənlərin və hesablamaların tamlığına mənfi təsir edən rəqəmsal hücumlardır. Xidmətdən imtina (əlçatanlığın itirilməsi), informasiya sızması və ya səhv hesablamalara səbəb olan hücumların başqa formaları da mövcuddur.

Yaddaşın tamlığı və güvənli platforma modulları (TEE – “güvənli icra mühitləri”nə daxildir) kimi ənənəvi mexanizmlərdən istifadə etməklə verilənlərin və kodun konfidensiallığını və tamlığını təmin etmək zəruridir, lakin bu, maşın öyrənməsi mexanizmlərinin kod və verilənlərinin konfidensiallığını və tamlığını təmin etmək üçün yetərli deyil. Odur ki, maşın öyrənməsi (ML) proqramlarının icrasının proqramlaşdırılmış məntiqə uyğunluğunu təmin etmək eyni dərəcədə vacibdir. Məsələn, ML tətbiqinin idarəetmə axınına hücum ML modelinin nəticəçıxarma prosesini poza/dağıda bilər və ya təlimin yararsız olmasına gətirib çıxara bilər. İcra mühitinin tamlığını təmin etmək üçün vacib olan ikinci kateqoriya - yaddaş təhlükəsizliyi xətalarının qarşısını almaq üçün nəzərdə tutulan mexanizmlərdir. Proqramlardakı məntiq xətaları buferin həddən artıq dolması, boşaldılmış yaddaşdan istifadə, yol verilən diapazondan kənaraçıxmalar üçün istifadə oluna bilər ki, bu da ML tətbiqlərinin işində uğursuzluqlara səbəb ola bilər.

Maşın öyrənməsi tətbiqlərinin istifadə edə biləcəyi mürəkkəb qurğu modelləri, o cümlədən aparat sürətləndiriciləri üçün təhdidlər nəzərə alınmalıdır. Bir sıra hallarda, bu sürətləndirici və ya qurğular paravirtuallaşdırıla və ya emulyasiya edilə bilər. Bəzən də bulud əsaslı tətbiqlərdə birbaşa onlara təhkim edilmiş qurğulardan istifadə faydalı ola bilər. Bununla belə, bu cür qurğuların verifikasiyası (attestasiya yolu ilə) qurğunun ML tətbiqlərinin məxfilik və təhlükəsizliyə dair tələblərə uyğunluğunun təmin olunmasına kömək edəcək. Qurğuları iş yükləri ilə təhlükəsiz şəkildə əlaqələndirmək üçün aparat vasitələrinin giriş/çıxış (IO) yaddaşının idarə edilməsi imkanlarından (o cümlədən mühafizəli yaddaşa DMA da daxil olmaqla) istifadə etmək olar. Diqqət yetirilməli olan gələcək hücum vektorlarına qurğuların spufinqi, işçi yaddaşın yenidən paylaşılması hücumları və "ortada insan" (“man-in-the-middle”) hücumları aiddir.

8.3 Sİ yönələn xüsusi məxfilik təhdidləri

8.3.1 Ümumi müddəalar

Sİ alqoritmlərinin təkamülü və böyük verilənlərdən istifadə əksər sahələrdə mürəkkəb həllər tapmağa imkan yaratmışdır. Bir sıra Sİ üsulları (məsələn, dərin öyrənmə) böyük verilənlərə əsaslanır, çünki onların dəqiqliyi qismən istifadə olunan verilənlərin miqdarından asılıdır. Bəzi verilənlərin, xüsusilə fərdi və həssas verilənlərin (məsələn, tibbi qeydlərin) qanunsuz

istifadəsi və ya açıqlanması verilənlər subyektləri üçün zərərli nəticələrə gətirib çıxara bilər. Beləliklə, məxfiliyin qorunması böyük verilənlərin təhlili və Sİ sahəsində əsas problemə çevrilmişdir. Problem verilənlərin həyat dövrünün ilkin mərhələsindən (yəni verilənlərin müxtəlif qurumlar arasında toplanılması və paylaşılmasından) başlayır. Bu, verilənlərin təhlili və Sİ alqoritmlərinin tətbiqindən ibarət son mərhələsinə qədər (məsələn, çoxsaylı verilənlər mənbələrindən əldə olunmuş verilənlərin təhlilindən sonra təkrar identifikasiya riski) davam edir.

Məxfilik nöqteyi-nəzərindən bu təhdidlər fərdlərin öz müqəddəratını təyinetmə, ləyaqət, azadlıq və fundamental hüquqlarına mənfi təsir göstərməsi ilə nəticələnə bilər.

8.3.2 Verilənlərin toplanılması

Verilənlərin toplanılması prosesində məxfilik üçün təhdid, əsasən, müəyyən məqsəd üçün əldə edilməsi lazım olan verilənlərin miqdarı ilə bağlıdır. ISO/IEC 29100 [51] standartında təqdim olunan məxfilik prinsiplərindən biri verilənlərin minimuma endirilməsi prinsipidir. Maşın öyrənməsi modellərinin böyük həcmdə yüksək keyfiyyətli verilənlərin əlçatanlığından asılı olduğunu nəzərə alaraq (bu cür məlumatların müxtəlif mənbələrdən olmasına üstünlük verilir), verilənlərin əldə edilməsini məhdudlaşdırmaq çətindir.

Bundan başqa, digər bir məxfilik təhdidi saxlama funksiyasının pozulması riskindən irəli gəlir. Əgər təcavüzkar saxlama funksiyasına xələl yetirərsə, verilənlər subyektlərinin məxfiliyi pozula bilər.

8.3.3 Verilənlərin ilkin emalı və modelləşdirilməsi

Verilənlərin emalı prosesində potensial məxfilik təhdidləri mövcuddur:

- maşın öyrənməsi və Sİ üsullarından istifadə edərək qeyri-həssas verilənlərdən həssas verilənlərin çıxarılması;
- fərdi verilənlər bir neçə mənbədən əldə edilir. Verilənlərin deidentifikasiya edilmiş olmasına baxmayaraq, Sİ başqa mənbələrdən əldə olunmuş verilənlərə əsasən çıxarılan nəticələrdən istifadə edir. Bu zaman Sİ verilənləri təkrar identifikasiya edə bilər.

8.3.4 Model sorğusu

Qeyri-münasib səbəblərə görə modeli sorğuya məruz qoymaqla model oğurluğu 8.2.4-cü bənddə təsvir edilmişdir. Bu cür təhlükəsizlik hücumları maşın öyrənməsi modelinin təmin etdiyi məxfi informasiyanı üzə çıxartmaq üçün nəzərdə tutulmuşdur. Bu cür hücum fərdlər haqqında həssas informasiyanı üzə çıxartmaq üçün istifadə oluna bilər. Beləcə, bu onu məxfilik hücumuna çevirir.

Bu cür hücumlar modelin bütün həyat dövrü ərzində, o cümlədən onun tərtib edilməsi, quraşdırılması və istismarı mərhələlərində baş verə bilər. Hücumlar həm modeli sorğuya çəkmək səlahiyyəti olan aktorlar, həm də ilk olaraq, modelə giriş əldə etmək üçün təhlükəsizliyi pozmalı olan digər şəxslər tərəfindən edilə bilər.

Daha bir təhdid fərdlərin qəbul etmədiyi modeldən qeyri-münasib istifadə ilə əlaqədardır. Məsələn, onların profilləşdirilməsi, çeşidlənməsi və ya klassifikasiyası ilə. Bu isə onların sosial həyatına (məsələn, sosial xidmətlərinə, kredit kartlarına) təsir göstərə bilər.

8.4 Qərəz

Qərəz başqaları ilə müqayisədə bəzi şeylər, insanlar və ya qruplara üstünlük vermək deməkdir. Qərəz, adətən, insanın koqnitiv qərəzi, sosial qərəz və statistik qərəz (məsələn, seçim qərəzi, nümunəgötürmə qərəzi, əhatə qərəzi) kimi mənbələrdən və ya sadəcə, texniki xətalardan irəli gəlir. Qərəz SI sisteminin hazırlanmasının müxtəlif mərhələlərində meydana çıxır. Bu, nişanlara, təlim verilənləri çoxluqlarına, çatışmayan xüsusiyyətlərə/nişanlara, verilənlərin emalı problemlərinə təsir edən verilənlər qərəzi formasında ola bilər. Yaxud modellərə və ya model birləşmələrinə təsir edən arxitektura problemləri şəklində olması da mümkündür. Qərəz özünü avtomatlaşdırma qərəzi, yəni Sİ sistemlərinin tövsiyələrindən həddən artıq asılı olma kimi də göstərə bilər. Verilənlərdə qərəzin olmasının effektləri modelə təsir edə bilər. Bu zaman dəqiqliyin azalmasından tutmuş təsnifat tapşırıqları üçün tamamilə səhv təsnifata qədər arzuolunmaz nəticəyə gətirib çıxarması ehtimalı yaranır. Bu qərəzlərin aradan qaldırılması həmişə mümkün olmur və bu, yanlış nəticələrin əldə olunmasına səbəb ola bilər. Təlim verilənləri çoxluğunun səbəb olduğu qərəz çox vaxt statistik metod və qaydaların yanlış tətbiqi və ya nəzərə alınmamasına əsaslanır.

Qərəzin qiymətləndirilməsi üçün konkret obyekt qrupları kontekstində sistemin performans göstəriciləri müəyyən edilməli və ölçülməlidir. Bu məqsəd üçün spesifik olan bir sıra göstəricilər mövcuddur. Bununla belə, istifadə ediləcək göstəriciləri seçərkən çox diqqətli olmaq vacibdir, çünki mürəkkəb nisbətlər gözlənilməz nəticələrə səbəb ola bilər. Elə nümunələr var ki, konkret (obyektlər) qrup üçün qərəzi kompensasiya etmək cəhdləri digər qrup kontekstində qərəzin artması ilə nəticələnmişdir.

Qərəzin səbəbləri, habelə Sİ sistemlərində müvafiq təzahürlər və əlaqədar təsir azaltma tədbirlərinə dair çoxsaylı nümunələr var. Bu mürəkkəb mövzunun ətraflı təsviri hazırda hazırlanma mərhələsində olan ətraflı texniki hesabatda təqdim olunacaq.

8.5 Proqnozlaşdırılmazlıq

Proqnozlaşdırma imkanı Sİ sistemlərinin məqbul olmasında mühüm rol oynayır. Proqnozlaşdırma imkanı anlayışı elə bir qabiliyyətdir ki, bu zaman insan müəyyən bir mühitdə Sİ sisteminin növbəti hərəkətləri haqqında nəticə çıxarmağı bacarır. Texnologiyaya güvən çox vaxt proqnozlaşdırma imkanına əsaslanır: sistemə o halda etibar edilir ki, onun konkret vəziyyətdə nə etdiyi barədə nəticə çıxarmaq mümkün olsun (hətta bunu niyə etdiyinin izahını vermək mümkün olmasa belə). Əksinə, sistem tanış ssenarilərdə proqnozlaşdırma imkanı olmadan işlədikdə ona güvən azalacaq.

Sİ sisteminin insanlarla qarşılıqlı əlaqədə olduğu və insanların təhlükəsizliyinin həmin qarşılıqlı əlaqədən asılı olduğu bir əməliyyat mühitində sistemin proqnozlaşdırma imkanı təkcə arzuolunan deyil, həm də zəruridir. Sİ-dən avtonom nəqliyyat vasitələrində istifadə aşkar bir istifadə ssenarisidir. Çünki Sİ-yə əsaslanan avtonom nəqliyyat vasitələrinin geniş surətdə yayılması, çox güman ki, bu cür nəqliyyat vasitələrinin proqnozlaşdırılması mümkün olan tərzdə davrana bilməsinə əsaslanır. Analoji olaraq, insanlarla birbaşa qarşılıqlı əlaqədə olan kollaborativ robotların məqbul olması üçün insan operatorlar robotların davranışını proqnozlaşdıra bilməlidirlər. Beləcə, operatorların təhlükəsizliyini təmin etməlidirlər [55].

İnsanların idarə etdiyi avtomobillər ssenarisində koqnitiv mexanizmlərimiz təcrübəmizə, təkrarlara və oxşar ssenarilərdə olmağımıza əsaslanır. Bu yolla da ətrafımızdakı insan və obyektlərin ehtimal olunan hərəkətləri haqqında cəld və demək olar ki, qeyri-iradi qərarlar verməyimizə imkan verir. Kənardakıların davranışdakı hətta kiçik dəyişikliklər belə, təcrübəmizlə ziddiyyət təşkil edəcək dərəcədə proqnozlaşdırma imkanının olmamasına gətirib çıxara bilər. Məsələn, insan sürücü avtonom avtomobilin gələcək hərəkətlərini qabaqcadan görə bilmir. Elə görə də özü idarə olunan avtomobillə avtonom olmayan

avtomobil toqquşa bilər və əksinə.

Maşın öyrənməsi alqoritmləri daha ənənəvi proqramlaşdırma üsulları ilə müqayisədə proqnozlaşdırma imkanı baxımından spesifik problemlər yaradır. Maşın öyrənmə alqoritmləri böyük həcmdə verilənləri təhlil edərək, yeni qanunauyğunluqlar və həllər aşkar edərək öyrənir. Təlim verilənlərinin tərkibi və qanunauyğunluqların reallaşdırıldığı əsas modellərdəki variasiyalar parametrləri təyin edir. Bu parametrlər Sİ sistemlərinin düzgün emal edə biləcəyi bir sıra giriş verilənləri üçün nəzərdə tutulur. Maşın öyrənməsi modelini çaşdıran ssenarilər bunlarla nəticələnə bilər: proqnozlaşdırılması mümkün olmayan davranış, nasazlıqlara qarşı dayanıqlı mexanizmlərin və ya avtomatik idarəetmə sisteminin dayandırılması üçün digər mexanizmlərin olmaması. Bundan əlavə, fasiləsiz və ya ömür boyu öyrənmə ssenarilərində maşın öyrənməsi sistemlərinin qərar qəbuletmə üçün əsaslandığı “məntiq”in zamanla dəyişməsi mümkündür. Beləcə, bu “məntiq” proqnozlaşdırma imkanının olmamasını alqoritmik baxımdan artıra bilər.

8.6 Qeyri-şəffaflıq

Sİ sistemləri bir sıra formalarda qeyri-şəffaflıq nümayiş etdirə bilər. Birincisi, Sİ modelinin özü texniki cəhətdən qeyri-şəffaf ola bilər. Çünki onun istifadə etdiyi qərar qəbuletmə prosedurları, mahiyyət etibarilə, insanlar tərəfindən asanlıqla interpretasiya oluna bilmir. İkincisi, verilənlər və verilənlərin mənbələri şəffaf olmadıqda bütün sistemin davranışı kənar müşahidəçi üçün qeyri-şəffaf hal alır. Üçüncüsü, Sİ sistemi həmişə təşkilati proseslər kontekstində həyata keçirilir, məsələn: verilənlərin toplanılması, idarə edilməsi, Sİ nəticələrinin praktik tətbiqi və sistemin işlənməsi. Bu proseslər açıqlanmadıqda hətta interpretasiya oluna bilən Sİ modeli istifadəçilər və digər kənar maraqlı tərəflər üçün qeyri-şəffaf bir sistemə çevrilir.

Təsir azaldıcı tədbirlərin müzakirəsi üçün bax: 9.2.

8.7 Sİ sistemlərinin spesifikasiyası ilə bağlı problemlər

Əksər məhsul uğursuzluqları spesifikasiya mərhələsində baş verir. Bu mərhələdə bütün məhsul, o cümlədən onun imkanları və mühiti müəyyən edildiyinə və onun çıxış verilənləri həyata keçirmə mərhələsi üçün giriş verilənləri olduğuna görə bu mərhələdə baş verən uğursuzluqlar böyük təsirə malikdir. Bu uğursuzluqların məhsulun həyat dövrünün sonrakı mərhələlərində düzəldilməsi çətin və ya qeyri-mümkündür.

Xüsusilə məhsul mühiti tam və ya kifayət qədər təfərrüatlı şəkildə təhlil edilmədikdə bu mərhələdə xətalar baş verir. Tam təhlilə məhsulun nəzərdə tutulan funksional imkanlarına təsir göstərə biləcək bütün ətraf mühit amilləri, habelə texniki və fiziki təhlükəsizlik təhdidlərinin nəzərə alınması, eləcə də məhsulla əlaqədar hüquqi, normativ və etik bazanın tədqiq edilməsi daxildir. Bundan başqa, quraşdırılma mühitində təyinatlı istifadə üçün performans və istifadəyə yararlılıq aspektlərinin diqqətə alınması vacibdir. Eləcə də müxtəlif istifadəçi qruplarında baş verən istənilən dəyişikliklər nəzərə alınmalıdır.

Maşın öyrənməsi metodlarına əsaslanan Sİ sistemlərindən irəli gələn potensial təhlükə və risklər mövcuddur. Onları təhlil etmək üçün təlim verilənlərində qərəzi və ya alqoritmin səhv təlimi səbəbindən yaranan sistem nasazlığını nəzərə almaq vacibdir. Bundan əlavə, risklərin qarşısını sonradan istifadə olunan metodların xüsusiyyətlərini nəzərə almaqla və tapşırığı buna uyğun müəyyən etməklə almaq olar.

Hüquqi sistemlər daxilində risk və hüquqi məsuliyyətin bölüşdürülməsi mürəkkəb bir tapşırıqdır. Ehtimal var ki, Sİ-dən istifadə atributların təyin edilməsi prosesini çətinləşdirir və ya riski/öhdəliyi necə qiymətləndirməyimizdə dəyişikliklərə səbəb olur.

Sİ sistemlərindən heterogen mühitlərdə mürəkkəb problemlərin həllində istifadə olunur. Elə bu səbəbdən tam və düzgün spesifikasiyanın yaradılması vacibdir. Məqsədlərin və izah oluna bilmə səviyyəsinin müəyyən edilməsi (daha ətraflı məlumat üçün bax: 9.3) istənilən Sİ sistemi spesifikasiyasının mühüm komponentidir.

Spesifikasiya müəyyən edildikdən sonra layihənin ayrı-ayrı iştirakçılarının diqqətinə çatdırılmalıdır. Beləliklə, spesifikasiya tərifl ərinin kifayət qədər əsaslı olmaması daha bir uğursuzluq mənbəyinə çevrilir. Bunun səbəbi onun çoxsaylı qeyri-rəsmi ötürmələrə görə spesifikasiyanın yanlış interpretasiya olunması və saxtalaşdırılması riskini artırmasıdır. Ümumiyyətlə, spesifikasiyada yazılan ideyalar həmin spesifikasiya əsasında sistemi yaratmalı olan şəxs tərəfindən interpretasiya olunur. İdeal spesifikasiya ilə yekun tərtibat arasında ilk uyğunsuzluq bundan yaranır. Maşın öyrənməsindən istifadə edərkən əlavə uyğunsuzluqlar yaranır. Bu da son alqoritmin verilənlərə əsaslanan təlim vasitəsilə əldə olunan konsepsiyalar əsasında yaradılmasından qaynaqlanır.

Bu qeyri-müəyyənliklərin hamısı belə bir zərurət yaradır ki, spesifikasiya verifikasiya və validasiya oluna bilən tələbləri (bunlar son nəticədə əldə olunacaq məhsulun test edilməsi üçün əsas götürüləcək) ehtiva etsin.

8.8 Sİ sistemlərinin tətbiqi ilə bağlı problemlər

8.8.1 Verilənlərin toplanması və hazırlanması

Verilənlərin toplanılması mərhələsində problemin həlli üçün lazım olan verilənlər mənbələri müəyyən edilir. Həll olunmalı problemin mürəkkəbliyindən asılı olaraq, səciyyəvi verilənlər çoxluğunu tapmaq çətin ola bilər. Verilənlər bir neçə mənbədən əldə olunduqda normallaşdırma və ya çəki əmsallarının təyin edilməsi problemləri meydana çıxa bilər. Qeyd etmək vacibdir ki, (model üçün giriş verilənləri olan) verilənlər mənbələrinin emalı zamanı meydana çıxan qüsurların modelin qiymətləndirilməsi prosesində aşkar edilməsi qeyri-mümkündür. Çünki həmin proses çox vaxt inteqrasiya olunmuş deyil, təcrid olunmuş mühitdə aparılır.

Bu mərhələdə modelləri öyrətmək üçün istifadə olunan verilənlər çoxluğu və ya çoxluqları yoxlanılır və sənədləşdirilir. Bu zaman onların modelin işləməsi üçün əsas götürülən verilənləri nə dərəcədə yaxşı əks etdirməsinə baxılır.

Tələb olunan verilənlər tapıldıqdan sonra çox vaxt verilənlərin hazırlanması tapşırıqları adlanan bir sıra tapşırıqlar yerinə yetirilir. Bu tapşırıqlar verilənlərin təmizlənməsi və modelin istifadə edəcəyi münasib formata uyğunlaşdırılmasını ehtiva edir. Bu mərhələdə vacib aspekt verilənlərin keyfiyyətini (məsələn, verilənlərin çatışmaması, kopiyalanması, uyğunsuz və ya yanlış formatda olmasını) yoxlamaqdır.

8.8.2 Modelləşdirmə

8.8.2.1 Ümumi müddəalar

Modelləşdirmə mərhələsində müvafiq modelləri generasiya etmək üçün seçilmiş alqoritmlər öyrədilir. Verilənlər üzərində öyrədildiyi zaman model maşın öyrənməsi alqoritminin inikasına çevrilir. Əslində, bir neçə model qurmaq, daha sonra isə təhlil edilən konkret biznes probleminə münasibətdə ən yaxşı performans nümayiş etdirəni seçmək tövsiyə olunur. Dəqiq bir model generasiya etmək üçün ümumi üsul mövcud verilənləri üç qrupa bölməkdən ibarətdir: təlim verilənləri, validasiya verilənləri, test verilənləri. Təlim verilənləri çoxluğu, adından da göründüyü kimi, verilənlərin elə bir hissəsidir ki, onunla modeli öyrətmək və onun

məqsədə uyğunluğu təmin edilir. Validasiya verilənləri çoxluğundan təlim keçmiş modelin proqnozlaşdırma qabiliyyətini növbəti mərhələdə yoxlamaq və modeli tənzimləmək üçün istifadə olunur. Nəhayət, test verilənləri çoxluğundan istifadə test mərhələsində təlim keçmiş, məqsədəuyğun və tənzimlənmiş modelin yekun qiymətləndirilməsini təmin etmək üçün nəzərdə tutulur.

Müasir texnologiyalar nəzərə alınmaqla, maşın öyrənməsi ilə Sİ sisteminin qurulması 8.8.2.2–8.8.2.5 bəndlərində təsvir edilən mərhələlərdən ibarətdir. Bu mərhələlər çox vaxt iterativ proses şəklindədir.

8.8.2.2 Xüsusiyyətlərin işlənməsi

Maşın öyrənməsində xüsusiyyətlər modelin proqnozlar vermək üçün istifadə etdiyi dəyişən giriş verilənləridir. Xüsusiyyətlərin qurulması elə bir prosesdir ki, burada emal olunmamış verilənlər proqnozlaşdırılan modellər üçün əsas problemi ən yaxşı şəkildə əks etdirən xüsusiyyətlərə çevrilir. Bu, öz növbəsində, görünməyən verilənlərə əsaslanan modelin dəqiqliyinin artması ilə nəticələnir [56]. Funksiyalar, adətən, proqramlaşdırmadan müəyyən qədər istifadə olunaraq verilənlərin transformasiyası addımlarının (miqyasın dəyişdirilməsi, diskretizasiya, normallaşdırma, verilənlərin inikası, aqreqasiyası və nisbətləri kimi) ardıcıllığı vasitəsilə yaradılır. Nəzərə almaq lazımdır ki, xüsusiyyətlər verilənlərin transformasiyasının çoxsaylı mərhələlərinin nəticəsi ola bilər. O zaman proses diqqətlə sənədləşdirilməsə, emal olunmamış ilkin verilənlər ilə əlaqəni bərpa etməkdə çətinlik yarana bilər.

Xüsusiyyətlərin qurulması modelin performansına böyük təsir göstərir. Bu təsir müsbət və ya mənfi ola bilər. Məsələn, modelin proqnozlaşdırılmasında üstün dərəcədə iştirak edən bir xüsusiyyət modelin dayanıqlığına təsir göstərə bilər. Bunun səbəbi son proqnozun bütün xüsusiyyətlərlə mütənasib şəkildə əlaqəli olmaqdan daha çox, yalnız həmin dəyişənin qiymətindən həddsiz asılı olmasıdır. Bu, qeyri-dəqiq nəticələrə səbəb ola bilər.

8.8.2.3 Model təlimi

Təlim verilənləri çoxluğu proqnozlaşdırılan dəyişən (hədəf dəyişən) ilə əlaqəli bəzi informasiyanı ehtiva etdikdə hədəf sızması (verilənlərin sızması da adlanır) baş verir. Bu isə istehsal mühitində mövcud olmur. Bu o zaman baş verə bilər ki, məsələn, təlim verilənləri proqnozlaşdırma zamanı mövcud olmayan informasiyanı ehtiva edir (çünki müvafiq dəyişən/xüsusiyyət yalnız hədəf qiymət proqnozlaşdırıldıqdan sonra yenilənir). Bu o zaman da baş verə bilər ki, hədəf dəyişən xüsusiyyət kimi daxil edilə bilməyən proksi dəyişən vasitəsilə giriş verilənlərindən nəticə çıxarsın. Hədəf sızması olan modellər, adətən, qiymətləndirmə mərhələsində çox dəqiq olur, lakin istehsal mərhələsində zəif performans nümayiş etdirir.

Modeli qurmaq üçün seçilmiş alqoritm təlim verilənlərindən istifadə edilməklə öyrədilməlidir. Bu mərhələ ilə əlaqədar çətinlik elə bir model qurmaqdır ki, həll olunan problemə və ya uyğun modelə münasibətdə təlim verilənlərinin yaxşı təqdim olunmasını təmin etsin. Təlim prosesində həddən artıq təlim keçmiş və ya kifayət qədər təlim keçməmiş model yaratmaq riski var. Həddən artıq təlim keçmiş model həddsiz çox təfərrüat öyrənmiş və əsas verilənlər modelinə (o cümlədən onun küyünə və ya verilənlər çoxluğunun daxili xətasına) uyğun gələn modeldir. Bu isə onun yeni verilənlər gəldikdə proqnozlar verməkdə zəif performans nümayiş etdirməsi ilə nəticələnir. Çox vaxt bu problem model üçün həddən artıq çox xüsusiyyət (məsələn, giriş verilənləri) seçildikdə baş verir. Digər tərəfdən, model verilənlərin əsas qanunauyğunluqlarını əks etdirmədikdə və buna görə də yaxşı proqnozlar vermək üçün həddsiz ümumi xarakter daşıdıqda kifayət qədər uyğunluğun olmaması baş verir. Bu, modelin kifayət qədər müvafiq xüsusiyyətləri olmadıqda daha tez-tez təkrarlanır. Buna görə

də münasib miqdarda proqnoz verilənlərlə münasib xüsusiyyətləri seçmək vacibdir. O zaman model həddən artıq spesifik (hədsiz çox sayda xüsusiyyətlə) və ya həddən artıq qeyri-müəyyən (kifayət qədər xüsusiyyətləri olmayan) olmur. Modelin uyğunlaşdırılması təlim mərhələsində yaranan bir problemdir. Hərçənd proqnozların keyfiyyəti validasiya və statistik verilənlər əsasında test mərhələlərində ölçülür.

Modelin seçilməsi və tərtib edilməsi üçün digər yanaşmalara daxildir: yeni tapşırığı öyrənmək üçün bir tapşırıq haqqında biliklərdən istifadə etmək məqsədi daşıyan transfer öyrənmə və yeni modelləri paylanmış, həmçinin birgə şəkildə öyrənmək məqsədi daşıyan federativ öyrənmə.

8.8.2.4 Modelin sazlanması və hiperparametrlərin optimallaşdırılması

Təlim mərhələsində modellər onların hiperparametrlərinin tənzimlənməsi ilə kalibrlənir/sazlanır. Hiperparametrlərə nümunə olaraq, qərar ağacı alqoritmində ağac dərinliyini, təsadüfi meşə alqoritmində ağacların sayını, *k*-orta alqoritmində *k* klasterlərin sayını, neyron şəbəkəsində layların sayını və s. qeyd etmək olar. Yanlış hiperparametrlərin seçilməsi proqnoz modellərinin uğursuzluğuna səbəb ola bilər.

Modelin keyfiyyəti təkcə onun strukturu, təlim alqoritmi və verilənlərindən asılı deyil. Modelin hiperparametrlərinin seçimi də həlledici amildir. Bəzi tətbiqlərdə hiperparametrlərin optimallaşdırılması müasir səviyyəyə daha çox çatmışdır, nəinki öyrənmə alqoritmləri. Hiperparametrlərin optimallaşdırılması, adətən, öyrənmə prosesinin xarici dövrəsini təşkil edir.

8.8.2.5 Modellərin validasiyası və qiymətləndirilməsi

Sazlandıqdan sonra modellər validasiya verilənləri çoxluqları ilə müqayisə edilməklə qiymətləndirilir. Bu onların performansını təlim üçün istifadə edilən verilənlər çoxluğundan fərqli olan verilənlər üzərində yoxlamaq məqsədini daşıyır. Sadə model validasiyası üsulunda yalnız bir validasiya verilənləri çoxluğundan istifadə edilir. Lakin daha etibarlı modellər qurmaq üçün K-qat çarpaz validasiya üsulundan istifadə edilə bilər. Bu üsul verilənləri *k* altçoxluqlara bölməkdən ibarətdir. Onların hər birindən validasiya çoxluğu kimi istifadə olunur, qalan *k*-1 altçoxluqları birləşdirilərək təlim çoxluqları formalaşdırılır. *K* validasiya testlərinin nəticələri ən yaxşı performanslı və ən etibarlı modeli (təlim verilənlərində mövcud olan küyə həssaslıq baxımından) müəyyən etmək məqsədilə müqayisə olunur. Modelin seçilməsi onun digər modellərlə müqayisədə performansına əsaslanır. Modelin performansını qiymətləndirmək üçün istifadə edilən statistik göstəricilərə nümunə olaraq, ROC AUC (əyri altındakı sahə), uyğunsuzluqlar matrisi (proqnozlaşdırılan qiymətləri test verilənləri çoxluğundan əldə olunan faktiki qiymətlərlə müqayisə edir) və ya F-1 göstəricisini (uyğunsuzluqlar matrisi əsasında hesablanır, dəqiqlik və tamlıq arasında ideal sərhədi əks etdirir) qeyd etmək olar.

Ayrı bir təkrar test mərhələsində modelləşdirmə mərhələsindən sonra seçilmiş model yekun məntiqi ardıcıllıq üçün yeni verilənlər (test verilənləri çoxluğu) ilə yenidən test edilir. Modelin sazlanması üçün bəzi yekun parametrlər (məsələn, bu və ya digər kateqoriyaya aid olma ehtimalını, beləliklə də, yanlış müsbətlərlə yanlış mənfilər arasında qarşılıqlı nisbəti müəyyən edən klassifikasiya problemlərində hədd) biznes istifadəçiləri ilə birgə müəyyən edilir. Bunun səbəbi onların konkret biznes tətbiqindən asılı olmasıdır.

İstehsal mühitində quraşdırılma, adətən, təkrar test mərhələsindən sonra baş verir.

8.8.3 Model yeniləmələri

Model istehsal mühitində quraşdırıldıqdan sonra onun yeni əldə edilmiş verilənlər əsasında yenilənməsi tələb oluna bilər. Modelin nə vaxt yenidən təlim keçməli/yenilənməli olduğunu müəyyən etmək lazımdır. Bunu operativ şəkildə etmək məqsədilə modelin performansını/dəqiqliyini davamlı olaraq nəzarətdə saxlamaq vacibdir. Modellərin yenilənməsi, adətən, modeli daha etibarlı etmək və (və ya) onu müxtəlif tapşırıqlar üçün ümumiləşdirmək və ya yeni verilənlər çoxluqlarına münasibətdə dəqiqliyini artırmaq məqsədini daşıyır.

Modeli yeniləməyin qeyri-mürəkkəb metodlarından biri modelə yenidən təlim keçmək üçün sadəcə həm ilkin, həm də yeni verilənlərdən istifadə etməkdir. Bu yanaşma tətbiq edildikdə problemlər yarana bilər. Məsələn, bütün verilənlərin mərkəzi yerdə toplanması, eləcə də artmaqda olan daha böyük həcmli verilənlər əsasında yeni modelə təkrar təlim keçmək üçün xeyli hesablamaların aparılması. Bu problemlərin həlli üçün yanaşmalardan biri tədricən öyrənmədir. Bu halda, mövcud modellər yeni verilənlər əsasında genişləndirilir. Ümumiyyətlə, modellərin yeni verilənlər əsasında yenilənməsi və genişləndirilməsi üçün verilənlər və hesablama baxımından səmərəli alqoritmlər sahəsində aktiv tədqiqatlar aparılmaqdadır.

Modeli yeniləyərkən, əsas risk kimi, performansa təsir nəzərə alınmalıdır. Yenilənmiş model yoxlanılmalı və təkrar test edilməlidir. Bu, ilkin tapşırığa əsaslanan daha əvvəlki performansa nisbətən pisləşmənin olmadığını və yeni tapşırıqlara əsaslanan performansın konkret biznes tətbiqinə uyğunluğunu təmin etmək məqsədilə olunur. Bundan başqa, istehsal mühitində quraşdırılan modellərin müxtəlif versiyalarının tam izlənilə və audit oluna bilməsini təmin etmək vacibdir.

8.8.4 Proqram təminatı səhvləri

Süni intellekt metodları proqram təminatında alqoritmlərin həyata keçirilməsinə əsaslanır. Bu səbəbdən tərtibat prosesinin istənilən proqram təminatının hazırlanması prosesi ilə eyni mənfi cəhətləri var. Bu zaman proqram təminatı qüsurları baş verə bilər, məsələn: yaddaşa səhv giriş və yaddaşın səhv işlənməsi, səhv giriş və çıxış verilənləri, səhv verilənlər və idarəetmə axınları. Sİ alqoritmləri adətən, çoxnüvəli sistemlərdə tətbiq olunur. Çünki onlar üçün çox vaxt xeyli hesablama resursları tələb olunur. Bu hallarda, yarış şəraiti, qarşılıqlı bloklamalar və ölçmə effektləri (“Heisenbug xətaları”) kimi paralellik (vaxt baxımından üst-üstə düşmə) xətaları da nəzərə alınmalıdır.

8.9 Sİ sistemlərinin istifadəsi ilə bağlı problemlər

8.9.1 İnsan-kompüter qarşılıqlı əlaqəsi (HCI) amilləri

İnsan amillərinə əsaslanan bir sıra çətinliklər mövcuddur. [57] istinadına əsasən, bu amilləri dörd əsas kateqoriyada qruplaşdırmaq olar:

1) istifadə (bu halda, avtomatlaşdırma insanlara öz məqsədlərinə çatmaq imkanını verir);

2) qeyri-münasib istifadə (bu halda, avtomatlaşdırmadan həddən artıq asılılıq gözlənilməz mənfi nəticəyə səbəb olur). Məsələn, bir şəxsin avtomatlaşdırmaya həddən çox güvənərək yola diqqət yetirməməsi qeyri-münasib istifadə ola bilər;

3) istifadəsizlik (bu halda, avtomatlaşdırmaya kifayət qədər güvənməmək mənfi nəticəyə səbəb ola bilər). Məsələn, bir şəxsin düzgün işləyən avtomatlaşdırma

sisteminin avtomatik idarəetmə sistemini dayandırması və qəzaya səbəb olması istifadəsizlik halı ola bilər;

4) sui-istifadə (bu halda, avtomatlaşdırılmış sistem son istifadəçinin maraqları nəzərə alınmadan quraşdırılır). Məsələn, bir şəxsin avtomatlaşdırılmış sistemin avtomatik idarəetmə sistemini asanlıqla dayandırmasına mane olan konstruksiya sui-istifadə hesab olunacaq.

8.9.2 Gerçək insan davranışı nümayiş etdirən Sİ sistemlərinin yanlış tətbiqi

Sİ sistemləri insanlara xas xüsusiyyət və davranışları (əl yazısı [58], səs [59] və şifahi və ya mətnə əsaslanan söhbətlər [60],[61] kimi) təcəssüm və ya təqlid edəcək şəkildə tərtib oluna bilər. Pis niyyətli şəxslər tərəfindən düzgün tətbiq edilmədikdə bu texnologiyalardan insanları aldatmaq üçün istifadə oluna bilər. Elə hallar var ki, çat-botlar və ya e-poçt botları [62] tanışlıq saytında üzv olan həqiqi insanla ünsiyyət təəssüratı yaratmaq üçün insan davranışını təqlid edir.

8.10 Sistemin texniki təminatının nasazlıqları

Sİ sistemləri üçün aparat vasitələri nasazlıqlara qarşı etibarlı şəkildə dözümlü olmalıdır. Aparat vasitələrində baş verən nasazlıqların istənilən alqoritmin düzgün icrasına mane olan səbəbə çevrilməsi mümkündür. Bu cür nasazlıqlar həm alqoritmə nəzarəti, həm də verilənlər axınını poza bilər. Sİ ilə əlaqədar olduqda aparat vasitələrinin nasazlıqları həm nəticə çıxarmanın, həm də təlimin düzgün alqoritmik icrasına mane olur.

Aparat vasitələrinin nasazlıqlarına dair sübutlar fərqlənir. Bu cür nasazlıqlar verilənlərin bütövlüyünün pozulması, verilənlərin itkisi və ya verilənlərin axını ilə əlaqədar müvəqqəti problemlər kimi xətaları əhatə edə bilər. Bu cür xətalar tək problem və ya müxtəlif növ nasazlıqların birləşməsinə görə yarana bilər. Onların Sİ kontekstində əlavə olaraq araşdırılması lazımdır.

Xüsusilə aparat vasitələrinin nasazlıqları, mahiyyət etibarilə, davamlı (sistemin komponenti və ya modulunda davamlı problem), keçici (keçib gedən müvəqqəti nasazlıqlar) və ya fasiləli (fasilələrlə davam edən problemlər) ola bilər. Aparat vasitələrinin nasazlıqları zərərsiz və ya zərərli ola bilər. Bunlar təsadüfi və ya sistematik səbəblərdən irəli gəlir.

Qurğunun işinin dayanmasına səbəb olan nasazlıqlar qüsurlu komponentlərə görə baş verən zərərsiz, klassik aparat vasitəsi nasazlıqları ola bilər. Daha gizli olan nasazlıqlar qurğunun məqbul görünən, lakin yanlış nəticə çıxarmasına və ya komponentin "zərərli şəkildə hərəkət etməsinə" səbəb olan nasazlıqlardır. Bu nasazlıqlar təhlükəsiz xətalardır. Adətən, qablaşdırmanın çürüməsi nəticəsində yaddaş özəkləri və ya məntiq komponentlərinin vəziyyətində arzuolunmaz müvəqqəti dəyişikliklər yaranır. Bunun səbəbi alfa hissəcikləri, neytronlar və elektromaqnit küy və elektromaqnit şüalar kimi xarici elektromaqnit müdaxiləsi sayılan mənbələrdən yüksək enerjili şüalanmadır. Həmçinin naqil yolları və ya komponentlər arasında daxili çarpaz əngəllər və ya saat nasazlığı kimi pis niyyətlə tətbiq olunan pozuntular da buna səbəb ola bilər.

Nasaz drayverlər xətalı proqram təminatına əsaslanan hesablamalarda aparat vasitələri nasazlığının təsirini daha da artıra bilər.

Bu cür xətaların mənbələri və təsiri həm Sİ tətbiqlərindən, həm də onun konkret kateqoriyaya aid sistemlərdə quraşdırılmasından asılıdır. Sİ tətbiqləri, əsasən, nəticə çıxarmaq üçün istifadə olunan son qurğularından tutmuş həm nəticə çıxarmaq, həm də təlim üçün istifadə olunan bulud kateqoriyalı hesablama və saxlama resurslarına qədər müxtəlif sistemlərdə

quraşdırılır. Sİ tətbiqinin həyat dövrü ərzində təlim keçmiş model Sİ tətbiqlərini konkret sistem platformasına, məsələn, resurslarla məhdudlaşdırılmış son qurğuya uyğunlaşdıran bir sıra transformasiyalardan keçir.

Müxtəlif nasazlıq modellərində mümkün xəta mənbələri və nəticə etibarilə, nasazlıqlara qarşı dözümlü effektiv strategiyalar nəzərə alınmalıdır. Məsələn, real zaman rejimli paylanmış sistemlər Sİ tətbiqlərini işə salmaq üçün, adətən, hazır aparat vasitəsi hissələri (məsələn, ümumi təyinatlı mərkəzi prosessor (CPU), qrafik prosessor (GPU) və proqramlaşdırılan inteqral sxemlər (FPGA)), proqram təminatı (məsələn, ümumi təyinatlı əməliyyat sistemləri) və protokollar (məsələn, TPC/IP stek protokolları) ehtiva edir. Xəta mənbələrinə aiddir: verilənlərin bütövlüyünün pozulması, mesajların qəsdən olmayaraq təkrarlanması, yanlış mesaj ardıcıllığı, verilənlərin itirilməsi, qəbuledilməz gecikmələr, mesajların daxil edilməsi. Aparat vasitələrində asinxron axın planlamasını dəstəkləyən sistemlərdə, məsələn, qrafik prosessorlarda (GPU) axın planlayıcısına təsir edən xətalar ciddi fəsadlara səbəb ola bilər. Bu cür xətalar, məsələn, sistemin iş yükünün son tarixlərini qaçırması ilə əlaqədar baş verə bilər. Bundan başqa, yaddaşın diaqnostikası sistem arxitekturasının konkret aspektlərinə görə çətin ola bilər.

Sistemin işləməsinə təsir edə biləcək digər xəta mənbələri Sİ tətbiqlərinin həyat dövrü ilə əlaqədardır. Məsələn, modelin transformasiyası zamanı, yəni təlim keçmiş model son qurğuya, deyək ki, resursla məhdudlaşan daxili sistemə uyğunlaşdırıldıqda, məsələn, verilənlərin ixtisaslaşması və ya modelin budanmasına ehtiyac olduqda əlavə xəta mənbələri meydana çıxır.

9 TƏSİRLƏRİN ARADAN QALDIRILMASI TƏDBİRLƏRİ

9.1 Ümumi müddəalar

Təsir azaldıcı tədbirlər 8-ci bölmədə təsvir edilən məlum Sİ zəifliklərini aradan qaldıra bilən mümkün nəzarət vasitələri və tövsiyələrdir. Qeyd etmək lazımdır ki, müəyyən nəzarət vasitələri və ya tövsiyə bir neçə zəifliyi aradan qaldırmağa kömək edə bilər.

Etibarlı olmayan şəkildə işləyən sistem güvənli sayılmayacaq. Bəzi hallarda, sistem düzgün işləyə bilər, lakin yeni daxil edilmiş, düzgün olmayan giriş verilənlərinə görə təhrif olunmuş çıxış verilənləri hasil edə bilər. Bu kontekstdə nəzarət məqamları təyin olunmalıdır. Məqsəd güvənin qorunub saxlanıldığını və ya saxlanılmadığını müəyyən etməkdir. Bu cür nəzarət məqamları Sİ sisteminin həyat dövrü ərzində müntəzəm olaraq və ya Sİ sistemi qərar qəbulu üçün istifadə edildiyi vaxtlarda təyin oluna bilər.

9.2 Şəffaflıq

Şəffaflıq Sİ sisteminin funksiya, komponent və prosedurlarının görünməsini təmin edir. İdeal hallarda, şəffaf AI sistemi təkrarlanan davranış nümayiş etdirməlidir. Şəffaflıq verilənlər, xüsusiyyətlər, modellər, alqoritmlər, təlim metodları və keyfiyyət təminatı proseslərinin kənar yoxlama üçün əlçatan olmasını təmin etməlidir. Şəffaflıq maraqlı tərəflərə Sİ sisteminin tərtib olunması və işləməsini qiymətləndirməyə imkan verir. Bu onların Sİ emalında görmək istədikləri dəyərlər əsasında baş verir. Həmin dəyərlər ədalətlilik və ya məxfilik məqsədlərinə əsaslana və ya konkret bir maraqlı tərəfin etik dünyagörüşündən, məsələn, məziyyət etikasından və ya digər qlobal dəyərlər sistemindən irəli gələ bilər.

Şəffaflığın Sİ proseslərinin bütün səviyyələrinə inteqrasiyası kömək edir ki, 8.6-cı bənddə təsvir edilən qeyri-şəffaflıq məsələlərinin səbəb olduğu problemlər azalsın. Şəffaf Sİ sistemi maraqlı tərəflərə hansı verilənlərin, xüsusən də fərdi məlumatların harada və nə üçün toplandığı barədə məlumat verir. Bir şərtlə ki, bu cür metaverilənlər verilənlər toplanarkən

qeydə alınmış olsun. Həmçinin qərar qəbuletmə avtomatlaşdırıldıqda şəffaf Sİ sistemi maraqlı tərəfləri məlumatlandıra və qərar qəbuletmə prosesini izah edə bilər. Fərdi məlumatlar işlənərkən maraqlı tərəflər üçün hüquqi təsiri olan qərarların qəbul edilməsi ilə nəticələnə bilər. Bu zaman məxfilik qaydalarına görə, şəffaf Sİ sisteminin qərar qəbuletmə prosesinə insan müdaxiləsinə dair sorğuları qəbul etməsi, beləliklə də, proseslə əlaqədar maraqlı tərəflərin fikirlərini nəzərə alması tələb oluna bilər. Müxtəlif Sİ sistemlərinin tərtib edilməsi üçün müəyyən olunan şəffaflığın bir neçə səviyyə və xüsusiyyəti var (məsələn, açıq verilənlər sahəsində).

Sİ sistemləri üçün reytinq simvolları, nişanları və ya işarələrinin istifadə edilməsi konkret maraqlı tərəf qrupları üçün şəffaflığı artırmağa kömək edə bilər. Məsələn, Dünya İqtisadi Forumu və UNICEF təşkilatının birgə təşəbbüsü olan "Generation AI" [63] uşaq oyuncaqlarında istifadə edilən Sİ sistemi üçün reytinq simvolları təklif edir. Əsas cəhət bu simvolların valideynlər/qəyyumlar üçün əlçatan olmasıdır. Bu cür şəffaflığa nail olmaq üçün sistemin izah edilə bilən olması vacibdir.

Sİ sistemlərinin şəffaflığı verilənlər, xüsusiyyətlər, alqoritmlər, təlim metodları və keyfiyyət təminatı proseslərinin maraqlı tərəflər tərəfindən kənar yoxlama üçün əlçatan olmasını təmin etməlidir. Bundan əlavə, yoxlamalar planlaşdırılarkən maraqlı tərəflərin baza bilik səviyyəsi nəzərə alınmalıdır. Bu, mütləq şəkildə olmasa da, aşağıda qeyd olunanların izahını ehtiva edə bilər:

— yoxlanılan Sİ mexanizmi, ümumiyyətlə necə işləyir, məsələn, qərar ağacı induksiyası necə işləyir;

— hansı model sinfi və parametrləşdirmədən istifadə olunur;

— modeldə hansı xüsusi dəyişən və ya funksiyalardan istifadə edilir;

— potensial dəyişənlər və ya xüsusiyyətlər çoxluğu necə seçilmişdir.

Maşın öyrənməsi sahəsində təlim keçmiş ekspert üçün model seçimi və dəyişənlərin seçilməsi prosedurunu əks etdirən qısa xülasə şəklində izah kifayət edər. Peşəkar olmayanlar üçün isə Sİ və verilənlərə əsaslanan modeldən nəticəçıxarma üzrə giriş kursu, eləcə də qərar ağacı modellərinin necə işlədiyinin izahı vacib olardı. Həmçinin xüsusiyyət və dəyişənlərin parametrləşdirilməsi və seçilməsinin təsirinin izahı gərəklidir.

Şəffaflıqla əlaqədar əsas məsələ onun necə işlədiyi və istifadə olunan alqoritmlərin məqsədəuyğun olub-olmamasıdır. Şəffaflıq verilən, funksiya, alqoritm və təlim metodlarının kənar yoxlama üçün əlçatan olmasını təmin edir. Şəffaflıq tədbirləri məxfilik və biznes maraqlarını tamamlamağa yönəldilməlidir [64]. Həmçinin Sİ-nin ümumi etimadı doğrultma xüsusiyyətini yaxşılaşdırmağa da. Qeyri-şəffaf modellər olduqda belə bir model üçün müəyyən dərəcədə şəffaflıq və ya izahlılıq təmin edən texniki metodlar mövcud ola bilər [65].

Sİ sisteminin şəffaflığı və izah edilə bilməsi ilə maraqlı tərəfin həmin sistemə olan güvən dərəcəsi arasında ciddi uyğunluq olmaya bilər. Bununla belə, şəffaflıq və izahlılıq maraqlı tərəflərə mühüm sübut və məlumat təmin edir. Bu onlara Sİ sisteminə duyduqları güvən barədə rəy formalaşdırmağa kömək edir.

9.3 İzahlılıq

9.3.1 Ümumi müddəalar

Sİ sisteminin şəffaflığına zəmanət vermək üçün təkcə izahlılıq kifayət etməsə də, bu, şəffaf Sİ sisteminin vacib komponentidir. Sİ sisteminin tərtib, tətbiq və istifadəsi ilə əlaqədar proseslərin, o cümlədən verilənlərin toplanılması üsulları, öz-özünü audit prosesləri, dəyər

öhdəlikləri və maraqlı tərəflərin cəlb edilməsinin izahları da rol oynayır. Sİ sistemlərinin izah oluna bilməsi korporativ sosial məsuliyyət çərçivəsində korporativ şəffaflığın bir növü hesab edilə bilər.

Sİ sistemlərinə verilən izahları bunlara görə təsnif etmək olar: izahın məqsədləri, o cümlədən kontekst, maraqlı tərəflərin ehtiyacları və arzuolunan anlayış növləri, eləcə də izah üsulu.

9.3.2 İzahın məqsədləri

İzah həmişə anlayışı çatdırmağa yönəlmiş bir cəhddir. İzahın effektivliyini onun formasını verildiyi kontekstə, o cümlədən hədəf auditoriyaya və çatdırmaq üçün nəzərdə tutulan anlayış səviyyəsinə uyğunlaşdırmaqla artırmaq olar [66].

Maraqlı tərəflərin aşağıdakılardan hansını nəzərdə tutmasından asılı olaraq, izah etmək cəhdi üçün bir neçə fərqli, lakin eyni dərəcədə etibarlı izahat üsulları təklif edilə bilər:

- nəticənin necə əldə edildiyinin səbəb-nəticə anlayışı;
- nəticənin əsaslandığı biliyin epistemik anlayışı;
- nəticənin etibarlı olması üçün təklif edilən əsasların əsaslandırıcı anlayışı.

Həm Sİ sisteminin özü, həm də sistemin təmin etdiyi nəticə izah predmeti ola bilər.

İzah edilə bilən Sİ sistemləri sistemlərin necə işlədiyini induktiv şəkildə müşahidə etməlidir. Bununla yanaşı, nəticələrinin həqiqiliyi, dəqiqliyi və ağlabatanlığını təmin edən proseslər barədə anlayış formalaşdırmağa yönəlməlidir. Müvafiq təlimat və standartlara riayət olunması maraqlı tərəflərə izahları anlamaqda kömək edə bilər.

Sİ sistemləri ilə əlaqədar izahlar onun nəticələrinin etibarlılığı, münasibliyi və legitimliyi, habelə həmin əsaslarla qəbul edilmiş qərar və tədbirlər üçün əsaslandırma təmin edə bilər. Bu cür izahlar Sİ sistemini daha tədqiq və mübahisə oluna bilən hala gətirmək məqsədini daşıyır. Xüsusən də nəticə kimi qəbul edilən qərar və tədbirlərdən təsirlənən maraqlı tərəflər üçün.

İzahlar absolyut deyil. Belə ki, onlar hədəf modelə və izahı qəbul edənə nisbətən müəyyən edilir. İzahlar müqayisəli kimi başa düşülür. İnformasiya insana elə bir şəkildə təqdim edilir ki, həmin insanın sistemə dair mental modelinin sistemin özünə uyğunluğunu artırsın [66],[71].

9.3.3 İlkin (ex-ante) və postfaktum (ex-post) izahat

İlkin (ex-ante) izahat sözügedən sistem istifadə edilməzdən əvvəl onun ümumi xassə və xüsusiyyətlərini izah edir. Sİ sistemi sözügedən sistemdən istifadə etməzdən əvvəl onun xassə və xüsusiyyətləri haqqında müvafiq məlumatı həm tərtibatçılara, həm də maraqlı tərəflərə təqdim edəcək şəkildə ilkin (ex-ante) formada izah edilir.

Qərar qəbuletmədə rol oynayan xassə və xüsusiyyətlərin postfaktum (ex-post) izahı. Simvollar Sİ-nin izahlılığını, beləliklə də, etimadı doğrultma xüsusiyyətini artırır.

İlkin (ex-ante) və postfaktum (ex-post) izahlar müxtəlif funksiyaları yerinə yetirir. İlkin (ex-ante) izahat sistemin yaxşı tərtib edildiyinə və öz məqsədinə xidmət edəcəyinə dair inam yaratmaq məqsədi daşıyır. Onun məqsədi istifadəçilər arasında inam yaratmaq və ilk növbədə, Sİ sistemindən istifadəni stimullaşdırmaqdır. Digər tərəfdən, postfaktum (ex-post) izahat konkret alqoritmik nəticələr və onların əldə olunduğu şəraiti izah etməyə imkan verir.

Yəni ilkin (ex-ante) izahat Sİ sisteminə inam yaratmaq üçün vacib olsa da, postfaktum (ex-post) izahata çıxış olmadan sistemin şəffaflığına nail olmaq mümkün deyil.

İdeal olaraq, Sİ sistemi ilkin (ex-ante) və postfaktum (ex-post) izahatların ardıcıllığını təmin etməlidir. İlkin (ex-ante) izahatda qeyd olunan xassə və xüsusiyyətlər postfaktum (ex-post) izahat vasitəsilə aşkar edilən konkret sistem alqoritmlərinin icrası ilə sübuta yetirilir.

9.3.4-cü yarımbənddə əsas diqqət postfaktum (ex-post) izahlılığa yönəldilir.

9.3.4 İzahlılığa yanaşmalar

Generasiya olunan izahatların mərhələsi, əhatə dairəsi və təfərrüatlarına görə, izahlılığa yanaşmaları kateqoriyalara ayırmaq olar. İzahatlar Sİ modelinin tərtib edilməsinin müxtəlif mərhələlərində generasiya oluna bilər:

1) modelləşdirmə öncəsi;

2) modelləşdirmə;

3) modelləşdirmə sonrası.

Modelləşdirmə öncəsi mərhələsi modeli qurmazdan əvvəl verilənləri anlamaq üçün nəzərdə tutulur. Sİ modellərinin sonrakı tərtibatının əsaslandırılmasında konkret verilənlər çoxluğunu anlamaq üçün bir qrup metodlar nəzərdə tutulur (məsələn, aspektlər, proyektorun quraşdırılması, verilənlər çoxluğunun standartlaşdırılması, verilənlər çoxluğunun riyazi cəhətdən anlaşılması) [72]-[76].

Modelləşdirmə mərhələsi öz qərarlarını izah edə və ya mahiyyət etibarilə, interpretasiya oluna bilən Sİ modellərinin tərtib olunmasına xidmət edir [77]-[79].

Modelləşdirmə sonrası mərhələsi interpretasiya oluna bilməyən Sİ modelinin qərarları barədə izahatlar generasiya etməyə xidmət edir [80]-[88].

Sİ modelinin izah edilməsi prosesinin aşağıdakı kimi təsviri mümkündür:

- lokal olaraq, konkret giriş/çıxış verilənləri cütü üçün model qərar qəbuletmə prosesini izah etməklə;
- qlobal olaraq, ümumi konsepsiyaya və ya nümunələr sinfinə münasibətdə modelin daxili məntiqini izah etməklə.

Bundan əlavə, izahatlar müxtəlif səviyyəli təfərrüatla generasiya oluna bilər. Dərin neyron şəbəkəsində bir izahat səviyyəsində proqnozlaşdırılan çıxış verilənlərinin hər bir layının rolunu müzakirə etmək olar. Konkret layda hər bir neyronun rolunu tədqiq etməklə daha təfərrüatlı anlayış əldə etmək olar [89]-[95].

9.3.5 Postfaktum (ex-post) izahat üsulları

9.3.5.1 Ümumi müddəalar

İzahat üsullarını səbəb-nəticə, epistemik və əsaslandırıcı kateqoriyalara aid etmək olar.

Bu üç izahat üsulu bir-birindən fərqlənə bilər, çünki təşkilat epistemik və ya əsaslandırıcı izahat vermədən səbəb-nəticə tipli izahat verə bilər. Məsələn, kredit qərarı alqoritmi nəticəsində gender xüsusiyyətinin yüksək əhəmiyyəti alqoritmin qərarı necə qəbul etməsi barədə qismən səbəb-nəticə izahatı təmin edir, lakin fərdin kredit qabiliyyəti üçün genderin

hansı funksional rolu oynadığı və ya bu əsasda kreditlə bağlı qərarın hansı standartlarla əsaslandırıla bilməsi barədə suallara cavab vermir.

Buna görə Sİ sisteminin tam izahı aşağıdakı xüsusiyyətlərin hamısını əhatə edə bilər:

— alqoritmin necə qərar qəbul etdiyini izləyən səbəb və nəticə zənciri;

— modelləşdirilən fenomendə ölçülən xüsusiyyətlərin funksional rolu;

— alqoritmik nəticənin əsaslandırıldığı etik və digər prinsip və standartlar.

Bu cür izahedici xüsusiyyətlərin seçilməsi izahatın kim üçün nəzərdə tutulduğundan, izahatın məqsədlərinin nədən ibarət olmasından və tətbiqdə hansı səviyyədə inamın gözlənildiyindən asılıdır.

9.3.5.2 Səbəb-nəticə izahatı: fəaliyyət mexanizmi

SI sisteminin öz nəticələrinə necə gəldiyini anlamaq məqsədini daşıyan izahat səbəb-nəticə əlaqələri zəncirindən ibarətdir. Bu zaman konkret bir nəticə əldə etmək üçün giriş xüsusiyyətlərinin emal edildiyi mexanizmlər izah olunur.

Maşın öyrənməsi prosesində səbəb-nəticə zəncirlərinin izlənilməsinin nəticəsi seçilmiş abstraksiya səviyyəsindən asılıdır. Yəni keyfiyyət xassələri (məsələn, obyektin forması), hesablama metaforları (məsələn, vektor qiyməti) və fiziki faktların (məsələn, prosessor registrində yükün vəziyyəti) hamısı Aİ prosesinin səbəb-nəticə tarixində rol oynaya bilər. Nəticənin necə əldə olunduğuna dair səbəb-nəticə izahı bu səviyyələrin istənilən birində davam edə bilər [96].

Hansı səviyyədə abstraksiyanın faydalı olması izahedici məqsəddən asılıdır. İnterpretasiya oluna bilən Sİ sistemi sayəsində sistemin istifadə etdiyi konkret qərar amillərinin abstraksiyasının ən yüksək səviyyəsi və onların yekun nəticəyə təsiri belə, insanlar üçün məna kəsb edən keyfiyyət xüsusiyyətləri ilə uyğunlaşdırıla bilər. Bundan başqa, səbəb-nəticə izahı faktın əksinə olan müdaxilələri dəstəkləyə bilər [97]. Yəni giriş xüsusiyyətləri modifikasiya olunarsa, əldə edilmiş nəticənin necə dəyişəcəyini başa düşməyi təmin edə bilər.

9.3.5.3 Epistemik izahat: fəaliyyət göstərdiyini haradan bilirik

Epistemik əsaslandırma məqsədi üçün effektiv izahat modelləşdirilən fenomendə funksional və ya məntiqi əlaqələri izləyir. Bu bir alqoritmin çıxardığı nəticənin niyə doğru olduğunun izahını verməyə xidmət edir. Yəni bu təsvir sistemin özünə aid deyil, sistemin həsr olunduğu aləmin xüsusiyyətlərinə aiddir.

9.3.5.4 Əsaslandırıcı izahat: hansı əsaslarla fəaliyyət göstərir

Sİ sisteminin avtomatik qərar verməsi üçün izahat nəticənin etibarlılığını əsaslandıra bilər. Buna görə ictimai kontekstdə səbəb-nəticə və funksional izahatdan kənara çıxmaq lazımdır. Bu, əldə olunan qərarın əsaslandığı prinsip, fakt və standartları çatdırmaq məqsədini daşıyır. Mövcud vəziyyəti nəzərə alınaraq, bu cür izahat qəbul edilmiş qərarın nəyə görə ədalətli, etibarlı və əsaslandırılmış olduğunu bildirir.

Bu cür əsaslandırıcı izahat Sİ sistemlərinin alqoritmlər, istifadə olunan verilənlər və qərar qəbuletmə xüsusiyyətləri kimi xassələrinə aid ola bilər. Lakin sistemin tətbiqi ilə əlaqədar institusional və sosial faktlara istinad etmədən natamam hesab olunur. Qeyd olunanlara istifadə ssenarisinə aid qaydalar, standartlar və təşkilati proseslər daxildir.

Effektiv əsaslandırıcı izahat nəticəni dəstəkləyən arqument funksiyasını yerinə yetirir. Bu izahat sistematik şəkildə əldə edilməlidir. Beləliklə, uğurlu izahat ətraflı şəkildə araşdırıla və ona qarşı etiraz edilə bilər. Analoji olaraq, sistemin nəticəsi yenidən qiymətləndirilə bilər. Bu zaman yaranmış vəziyyəti ləğv etməyə və ya düzəltməyə yönəlmiş mümkün əks-arqumentlər nəzərə alınmalıdır.

9.3.6 İzahlılıq səviyyələri

Sİ sisteminin müvafiq izahlılıq səviyyəsini sistemin tətbiq olunduğu istifadə ssenarisi kontekstində seçmək olar. Beləliklə, müxtəlif izah səviyyələri üçün aydın müəyyən edilmiş tələblər tətbiq üçün Sİ növünü seçməyə kömək edə bilər. Sözügedən tələblər izahedici gücünə əsaslanmalıdır. Məsələn, özünün işləməsinə dair mənalı səbəb-nəticə izahatı təmin etməyən, interpretasiya oluna bilməyən sistemlər yüksək səviyyədə izahlılıq tələb edən məhsul və ya xidmətlərdə istifadə üçün münasib deyil. Tətbiq edilə bilən izahlılıq səviyyəsi konkret haldan asılı olaraq qiymətləndirilir.

Sİ sistemi üçün müvafiq izahlılıq səviyyəsini seçərkən tətbiq mülahizələri aşağıdakıları ehtiva edə bilər:

— Sİ sistemi, giriş verilənləri kimi, ayrı-ayrı şəxslərin həssas fərdi məlumatlarından istifadə edir;

— Sİ sisteminin nəticələri insanların rifahına əhəmiyyətli dərəcədə təsir edəcək şəkildə istifadə olunur;

— Sİ əsaslı qərar qəbuletmədə uğursuzluqların əhəmiyyətli nəticələri olur;

— tətbiq istifadəçinin və ya üçüncü şəxslərin avtonomluğunun məhdudlaşdırılması ilə nəticələnə bilər;

— sistem kənar müşahidəçilərə və quraşdırıldığı daha geniş cəmiyyətə qeyri-trivial təsir göstərir, məsələn, müəyyən iş elanlarının yalnız kişilərə göstərilməsi gender bərabərsizliyinin artmasına səbəb olur.

İzahat səviyyələri müxtəlif maraqlı tərəf qruplarının konkret ehtiyaclarından, Sİ sistemi nəticələrinin əsaslandığı verilənlərin aspektlərindən, Sİ nəticələrinə əsaslanaraq qərarların qəbul edilməsində insan müdaxiləsinə ehtiyacdan asılı olaraq dəyişə bilər. Həmçinin maraqlı tərəflərin bu cür qərarlara dair şəxsi fikirlərini ifadə etmək və bu cür qərarlara etiraz etmək ehtiyacı da böyük rol oynayır.

9.3.7 İzahların qiymətləndirilməsi

İzahların keyfiyyətinin ölçülməsini də nəzərə almaq vacibdir. Bunun üçün aşağıdakı aspektlər nəzərə alınmalıdır:

— davamlılıq (bu halda, yaxınlıqdakı nöqtələrin proqnozları üçün müvafiq izahat, demək olar ki, ekvivalent olacaq);

— məntiqi ardıcıllıq (bu halda, modeli müəyyən xüsusiyyətin proqnozlaşdırılan nəticəyə töhfəsi artacaq şəkildə dəyişdikdə həmin funksiyanın izahlılıq metodu ilə qiymətləndirilən əhəmiyyət göstəricisi azalmayacaq);

— selektivlik (bu halda, əhəmiyyətə əsaslanan izahatlar üçün töhfə göstəricisinin generasiya olunan proqnoza ən çox təsir edən xüsusiyyətlər arasında bölüşdürülməsi arzuolunandır). Yəni ən yüksək uyğunluq göstəricisi olan

xüsusiyyətin (və ya xüsusiyyətlər çoxluğunun) silinməsi modelin nəticələrində kəskin dəyişikliyə səbəb olacaq. Bu, generasiya olunan izahatda düzgün xüsusiyyətlərin uyğun xüsusiyyətlər kimi müəyyənləşdirilməsini təmin edir.

İzahatın dəqiqliyi və anlaşıla bilməsi arasındakı qarşılıqlı nisbəti də nəzərə almaq vacibdir [98]-[105].

9.4 İdarəolunanlıq

9.4.1 Ümumi müddəalar

Operatora Sİ sistemindən idarəetmənin ötürülməsi üçün etibarlı mexanizmlər təmin etməklə idarəetmə imkanına nail olmaq mümkündür. İdarəetmə imkanına nail olmaq üçün, ilk növbədə, həll edilməli olan məsələlər bunlardan ibarətdir: prosesdə çoxsaylı maraqlı tərəflər, məsələn, xidmət təminatçısı və ya məhsul təchizatçıları, tərkib komponenti olan Sİ təminatçısı, istifadəçi və ya tənzimləmə səlahiyyəti olan qurum iştirak etdikdə kimə hansı Sİ sistemləri üzərində hansı idarəetmə imkanının təklif olunması.

Aşağıda biz Sİ sisteminin həyat dövrünə idarəetmə nöqtələrini inteqrasiya etmək ehtiyacını təsvir edirik. Bu inteqrasiya etibarlı qərarların qəbul edilməsi istiqamətində atılan bir addım hesab oluna bilər.

9.4.2 “Dövrdə insan” nəzarət nöqtələri

Eynilə idarəetmə konturunda operator ilə idarəetmə məqamları ilə bağlı olduğu kimi, Sİ sistemlərinin həyat dövründə insanların roluna münasibətdə iki rol xüsusilə vacibdir:

— Si sistemlərinin nəticələri insanlar tərəfindən qərar qəbuletmə prosesini gücləndirmək üçün istifadə olunduqda onları nəzərə almaq üçün yekun qərar qəbuletmə prosesində ixtiyar və avtonomluğa malik qərar qəbul edənlər;

— sistemin güvən səviyyəsini yenidən qiymətləndirmək, habelə onun işini yaxşılaşdırmaq üçün rəy bildirmək imkanı verilən konkret sahə üzrə ekspertlər. Bu kontekstdə konkret sahə üzrə ekspertlər tərəfindən yoxlanılan/kontekstləşdirilən nəticələr Sİ sistemləri üçün vacibdir. Bunun səbəbi konkret sahə üzrə ekspertlərin yanlış korrelyasiyaları aşkar edə və ya Sİ sistemində mövcud olmayan verilənlərlə sistemin nəyə görə müəyyən şəkildə işlədiyini izah edə bilməsidir.

9.5 Qərəzi azaltma strategiyaları

Qərəzlə mübarizə aparmaq üçün bir sıra strategiyalar mövcuddur:

- qərəzlə əlaqədar hüquqi və digər tələblərin nəzərə alınması sistem tələbləri müəyyən edilərkən, o cümlədən müvafiq hədd göstəriciləri təyin edilərkən aydın şəkildə ifadə oluna bilər;

- mənşə və tamlıq barədə verilənlər mənbələrinin təhlili sayəsində risklər müəyyən edilə və verilənlərin toplanılması və ya şərh edilməsi üçün istifadə olunan proseslər nəzərdən keçirilə bilər;

- modelə təlim keçmə prosesləri tərkibində texniki üsullardan qərəzi müəyyən etmək və azaltmaq üçün istifadə edilə bilər;

- qərəzi aşkar etmək üçün xüsusi test və qiymətləndirmə üsullarından istifadə edilə

bilər;

- real istifadə kontekstində qərəzlə əlaqədar problemləri aşkar etmək üçün sınaqlar və ya müntəzəm əməliyyat təhlillərindən istifadə edilə bilər.

Bu yanaşmaların hər birinin üstünlük və çatışmazlıqları var. Qərəzlə əlaqədar risklərin araşdırılması və təsirləri azaltma üsullarının sənədləşdirilməsi Sİ-nin etimadı doğrultma xüsusiyyətini artırmağa kömək edir.

9.6 Məxfilik

Fərdi məlumatların deidentifikasiyası üçün sintaktik (məsələn, k-anonimlik) və ya semantik metodlardan (məsələn, diferensial məxfilik) istifadə olunur [52]. Verilənlərin deidentifikasiya edilmiş olmasına baxmayaraq, fərdi məlumatlar bir neçə mənbədən əldə edildikdə Sİ başqa mənbələrdən əldə olunmuş verilənlərə əsasən çıxarılan nəticələrdən istifadə edir. Həmin vaxt Sİ verilənləri təkrar identifikasiya edə bilər. Məsələn, tədqiqatlar [53],[54] göstərir ki, k-anonimlik yetərli olmaya bilər.

İlkin deidentifikasiya yanaşmasından asılı olmayaraq, təkrar identifikasiyanın qalıq riskini verilənləri qəbul edən tərəflər arasında idarə etmək mümkündür. Bu, tərəflər arasındakı verilənlərdən istifadə haqqında razılaşmalar vasitəsilə baş tutur.

9.7 Etibarlılıq, dözümlülük və dayanıqlıq

Nəşr [30] etimadı doğrultmaya nail olmağın ən vacib komponentlərindən biri olaraq, sistemin “qabiliyyət”inə işarə edir. Qabiliyyət sistemin müəyyən bir tapşırığı yerinə yetirmək xüsusiyyəti kimi təsvir edilə bilər. Qabiliyyətin həm də bir neçə atribut, o cümlədən etibarlılıq, dözümlülük və dayanıqlıq baxımından qiymətləndirilməsi mümkündür.

Etibarlılıq sistemin və ya həmin sistem daxilindəki obyektin tələb olunan funksiyalarını müəyyən şərtlər altında müəyyən vaxt ərzində yerinə yetirmək qabiliyyətidir [106]. Başqa sözlə, etibarlı Sİ sistemi eyni giriş verilənlərini nəzərə alaraq ardıcıl şəkildə eyni nəticələr verir.

Digər növ proqram təminatı sistemlərində olduğu kimi, Sİ sistemlərində də aparat vasitələrinin nasazlıqları alqoritmin düzgün icrasına təsir göstərə bilər. Sistemdə fasilə, nasazlıq və uğursuzluqlar baş verdikdə sistemin potensial olaraq zəifləmiş imkanlarla işləməyə davam etmək qabiliyyəti nasazlıqlara qarşı dözümlülük adlanır. Bütün sistemin giriş verilənlərinə cavab olaraq, sistem və ya avadanlığın düzgün işləməsindən asılı olan aspekti, adətən, funksional təhlükəsizlik adlandırılır [107].

Dözümlülük bir insident baş verdikdən sonra sistemin tez bir zamanda öz iş vəziyyətini bərpa etmək qabiliyyətidir. Dözümlülük etibarlılıqla bağlıdır, lakin gözlənilən xidmət səviyyələri və gözləntilər fərqlidir. Dözümlülüklə gözləntilər, maraqlı tərəflərin müəyyən etdiyi kimi, daha aşağı ola, eləcə də bərpanı təklif edə bilər (bax: İstinad [42], 11.5-ci yarımbənd).

Çox vaxt sistemin istənilən şəraitdə, o cümlədən xarici müdaxilə və ya sərt ətraf mühit şəraitində performans səviyyəsini saxlamaq qabiliyyətini təsvir etmək üçün Sİ sistemləri üçün dayanıqlıqdan istifadə olunur. Dayanıqlığa dözümlülük, etibarlılıq və sistemin tərtibatçıları tərəfindən nəzərdə tutulduğu kimi, düzgün işləməsi ilə əlaqədar ola bilən digər atributlar daxildir. Aydındır ki, sistemin lazımi qaydada işləməsi konkret bir mühitdə/kontekstdə onun maraqlı tərəflərinin təhlükəsizliyi ilə birbaşa əlaqəlidir və ya onu təmin edir. Məsələn, maşın öyrənməsinə əsaslanan dayanıqlı Sİ sistemi həddən artıq öyrətmə kimi naməlum giriş verilənlərini ümumiləşdirə biləcək. Bunun üçün model və ya modellərə böyükhəcmli təlim verilənləri çoxluqlarından, o cümlədən küylü təlim

verilənlərindən istifadə edərək təlim keçmək vacibdir.

9.8 Sistemin texniki təminatın nasazlıqlarının aradan qaldırılması

Dayanıqlı və nasazlıqlara dözümlü sistemlər arxitekturaya və aparat vasitələrinin detallı layihələndirilməsinə, eləcə də bütün tərtibat prosesinə aid olan müxtəlif metodlar vasitəsilə yaradılır. Beləliklə, məhsulun həyat dövrünün hər bir mərhələsi, xüsusilə də layihələndirmə və spesifikasiya mərhələsi nəzarət altında olur.

Bu metodlardan biri ehtiyatdan istifadədir. Butrada məqsəd nasazlıqların gizlədilməsi və ya onlardan başqa şəkildə yan keçilməsi, bununla da, funksionallığı istənilən səviyyədə saxlamaqdır. Aparat vasitələrinin nasazlıqlarını bunlar vasitəsilə aradan qaldırmaq olar: aparat vasitələri (məsələn, davam olaraq və ya xırda struktur elementləri səviyyəsində *n* qədər artırma), informasiya (məsələn, yoxlama bitləri) və ya vaxt (məsələn, müxtəlif, adətən, təsadüfi vaxtlarda təkrarlanan hesablamalar) ehtiyatı. Proqram təminatının nasazlıqlarına qarşı qorunmaq üçün isə proqram təminatı ehtiyatından (məsələn, proqram təminatının müxtəlifliyi və ya hərəkətli hədəfin azaldılmasının digər formaları) istifadə olunur.

Birinci növ nasazlıqları aradan qaldırmaq məqsədilə nasaz komponentin təsirini aşkarlamaq və ya aradan qaldırmaq üçün konstruksiyaya əlavə aparat vasitələri daxil edilir. Aparat vasitələrinin ehtiyatının yaradılması statik və ya dinamik ola bilər. Beləliklə, bu, sadə təkrarlamadan tutmuş, aktiv qurğularda nasazlıq baş verdikdə ehtiyat qurğulara keçid olunan mürəkkəb konstruksiyalara qədər dəyişə bilər.

Ətraf mühit şəraitinin təsiri və ya konkret sensor texnologiyalarının zəifliyi kimi ümumi səbəblərdən yaranan nasazlıqlar da olur. Onların zərərli nəticələrinin qarşısını almaq üçün əlavə tədbirlər (məsələn, müxtəliflikdən istifadə) görülməlidir. Bundan başqa, icra mühitində xətaları aşkar etmək, əks tədbirlər görmək və ya sistemi təhlükəsiz vəziyyətə gətirmək üçün müxtəlif diaqnostik tədbirlər köməyə gələ bilər.

Nasazlıqlara qarşı dözümlü aparat vasitələrinin tətbiqi üçün metod və proseslərin ətraflı təsviri, eləcə də sertifikatlaşdırıla bilən funksional təhlükəsizlik səviyyələrinin təsviri IEC 61508 [107] standartında təqdim olunur.

9.9 Funksional mühafizəlilik

Sistemin funksional mühafizəliliyini təmin etmək üçün təhlükəsizliklə əlaqədar aspektləri yerinə yetirən xüsusi funksionallıq tətbiq oluna bilər. Bu cür funksionallıq sistemin idarəetmə funksiyalarının tərkib hissəsi və ya sözügedən sistemlərlə qarşılıqlı əlaqədə olan xüsusi sistem ola bilər. Sİ sistemləri üçün təhlükəsizliklə əlaqədar funksiyalar, məsələn, Sİ tərəfindən qəbul edilən qərarların məqbul diapazonda olmasını təmin etmək üçün onlara nəzarət edə bilər. Yaxud da problemli davranış aşkar edilərsə, sistemi müəyyən vəziyyətə gətirə bilər.

IEC 61508 [107] təhlükəsizlik funksiyalarının yerinə yetirilməsində istifadə edilən elektrik və (və ya) elektron və (və ya) proqramlaşdırıla bilən elektron elementlərdən ibarət sistemlər üçün ümumi yanaşma müəyyən edir. Bu onun həyat dövrü ərzində bütün təhlükəsizlik fəaliyyətlərini əhatə edir. Bu yanaşma təhlükəsizliklə əlaqədar olanbu cür sistemlərə aid məhsul və tətbiqlər üçün beynəlxalq standartların əsasını təşkil edir. ISO 26262 [108], IEC 62279 [109] və IEC 61511 [110] avtomobil, dəmir yolu və emal sənayeləri üçün xüsusi adaptasiyaların nümunələridir. IEC 61508 [107] standartı, tətbiqindən asılı olmayaraq, bütün elektrik və (və ya) elektron və (və ya) proqramlaşdırıla bilən elektron (E/E/PE) təhlükəsizlik sistemlərinə şamil edilir. Həmin standart, əsas etibarilə, nasazlıq baş verdikdə insanların və (və ya) ətraf mühitin təhlükəsizliyinə təsir göstərə bilən sistemlərə aiddir. Bununla belə, qəbul

edilir ki, nasazlığın nəticələri əhəmiyyətli iqtisadi fəsadlara da səbəb ola bilər. Belə hallarda, bu standartdan avadanlıq və ya məhsulların qorunması üçün istifadə olunan istənilən E/E/PE sisteminin müəyyən edilməsində istifadə olunması mümkündür.

9.10 Testetmə və qiymətləndirmə

9.10.1 Ümumi müddəalar

Sİ sistemlərinin test edilməsi və qiymətləndirilməsi üçün müxtəlif yanaşmalar mövcuddur. Onların tətbiq olunma imkanı və effektivliyi hər bir konkret hala görə fərqlənə bilsə də, məqbul səviyyədə etimadı doğrultmaya nail olmaq üçün, adətən, bir neçə yanaşmanın birləşdirilməsi tələb oluna bilər.

9.10.2 Proqram təminatının yoxlanılma və verifikasiya metodları

9.10.2.1 Ümumi müddəalar

Etimadı doğrultmaya nail olmaq üçün ənənəvi (Sİ olmayan) proqram təminatı sistemləri aşağıdakı iki oxa əsaslanır:

- mühüm əhəmiyyətli funksiyaların ehtiyatının yaradılması və ya monitorinqini təmin edən arxitektura;
- kodun validasiya və verifikasiyası (bu, icra edilə bilən kodun ehtiyaclara cavab verdiyinin və tam test edilmiş olduğunun mühəndislik yanaşmasından istifadə etməklə nümayiş etdirilməsindən ibarətdir). Nümayiş etdirmə hazırda sistemin davranışının məlum və deterministik olması faktına əsaslanır.

[111] saylı İstinada əsasən, validasiya “konkret təyinatlı istifadə və ya tətbiq üçün tələblərin yerinə yetirildiyinə dair obyektiv sübutların təqdim edilməsi yolu ilə təsdiqlənməsidir. Qeyd 1: düzgün sistem qurulmuşdur”. Verifikasiya “müəyyən edilmiş tələblərin yerinə yetirildiyinə dair obyektiv sübutların təqdim edilməsi yolu ilə təsdiqlənməsidir. Qeyd 1: sistem düzgün qurulmuşdur” [24].

Proqram təminatı sistemlərinin validasiya, verifikasiya və yoxlanılması metodları üçün də [111] saylı İstinadda müəyyən edilmiş qaydada formal metodlar tətbiq edilir. Proqram təminatı testlərinin əsas məqsədləri [111] saylı İstinadda qeyd edilmişdir:

“Test edilən elementin keyfiyyəti və test edilən elementin nə dərəcədə test edildiyinə əsaslanan hər hansı qalıq risk haqqında məlumat vermək; test edilən element istismara verilməzdən əvvəl onun qüsurlarını aşkar etmək; və maraqlı tərəflər üçün zəif keyfiyyətli məhsul ilə əlaqədar riskləri azaltmaq”.

Öz konstruksiyasına görə, Sİ sistemləri ənənəvi proqram təminatı sistemlərindən daha az deterministik xarakterə malikdir. Si sistemləri nadir hallarda tam izah oluna bilir. Sİ sisteminin proqram təminatına həm Sİ, həm də qeyri-Sİ komponentləri daxildir.

Sİ sisteminin bütün komponentləri onların düzgün işləməsi üçün proqram təminatı və aparat vasitələrinə dair qəbul edilmiş metodlara (modul və funksional testlər daxil olmaqla) uyğunlaşdırılmalıdır. Lakin onun Sİ komponentləri, aşağıda təsvir olunduğu kimi, həmin metodların modifikasiya olunmuş versiyasından istifadə edəcək.

Sİ sistemlərinə aid olduqda funksional testlər, münasib olduğu təqdirdə, qeyri-müəyyənliyi idarə edə bilməlidir. Qeyri-deterministik proqram təminatı komponentlərinə dair tələblərin

mövcud standart və metodlardan istifadə edərək müəyyənləşdirilməsi və test edilməsi çətindir. Bu, “orakul problemi” kimi məlumdur. Onu ayrılıqda bir testin öz müvəffəqiyyət meyarlarına uyğun olub-olmadığının müəyyənləşdirilməsində çətinlik kimi təsvir etmək olar. Bu problemin Sİ sistemlərində geniş yayılmış olması onu söyləməyə imkan verir ki, yeni verifikasiya və validasiya üsullarını təşviq etmək üçün yeni standartlaşdırma səyləri başlana bilər.

9.10.2.2 Formal metodlar

Proqram təminatının validasiya və verifikasiyası məqsədilə süni neyron şəbəkələrini test etmək və qiymətləndirmək üçün formal metodlardan yararlanmaq olar. Bunun üçün bir neçə göstəricidən istifadə etmək olar, məsələn:

- qeyri-müəyyənlik (bu, şəbəkənin reaksiyasında dəyişikliklərlə əlaqələndirilir və ümumiləşdirilməsinin qeyri-sabit davranışa gətirib çıxarmadığını yoxlamaq məqsədini daşıyır);
- maksimum sabit fəza (Sİ sisteminin yerinə yetirilən klassifikasiyanın təlim çoxluğunda sabit olacağını sübut etmək qabiliyyətinə uyğun gəlir).

9.10.2.3 Empirik testetmə

Proqram təminatının validasiya və verifikasiyası məqsədilə qeyri-deterministik həllərin empirik test edilməsi üçün müxtəlif üsullar mövcuddur. Buraya daxildir:

- metamorfik test: sistemin giriş və çıxış verilənləri arasında qarşılıqlı əlaqələr quran və nəticələrin test və müqayisə edilməsinin bir neçə iterasiyasının yerinə yetirilməsinə əsaslanan üsuldur. Bundan, adətən, orakul problemi [112] olan sistemlərdə istifadə olunur;
- ekspert komissiyaları: ekspert mülahizələrini əvəzləmək üçün Sİ sistemlərinin qurulduğu hallarda test nəticələrini nəzərdən keçirmək üçün komissiyalar yaradılır. Bu yanaşma əlavə problemlər yaradır (məsələn, ekspertlər razılığa gəlmədikdə) [113];
- bençmarkinq: diqqətlə tərtib edilmiş, ictimaiyyət üçün açıq olan və (və ya) müxtəlif sistemlərin rəqabətli test edilməsi üçün istifadə edilən verilənlər çoxluqları üzərində sistemin performansını ölçən üsuldur [114]. Bençmarkinq obrazların tanınmasında və Sİ metodlarının analoji tətbiq sahələrində müəyyən bir metoda münasibətdə inam yaratmaq üçün ümumi təcrübəyə çevrilmişdir [115].

9.10.2.4 İntellekt müqayisəsi

Avtomatlaşdırılmış qiymətləndirmə metodu olmadıqda Sİ sisteminin və insanın tətbiqi intellektual qabiliyyətlərinin müqayisəsi Sİ sisteminin funksionallığını təsdiqləməklə Sİ sisteminin keyfiyyətindən əminliyi təmin edə bilər. Bu yanaşma meyarların konkret hədd qiymətləri ilə müəyyən göstəricilərin müqayisəsinə əsaslanır. Daha sonra müxtəlif metodologiyaların (məsələn, uzlaşma əmsalı, "Pearson" testi) müxtəlif mühitlərdə (məsələn, “qumqabı” və ya fiziki məhdudlaşdırma) tətbiqi mümkündür.

9.10.2.5 Simulyasiya olunan mühitdə testetmə

Bəzi hallarda, Sİ sistemi tərəfindən yerinə yetirilməli olan tapşırıq ətraf mühitə fiziki təsirin göstərilməsi ilə xarakterizə olunur (məsələn, robota quraşdırılmış Sİ üçün). Bu zaman

performansın qiymətləndirilməsi və risklərlə əlaqədar tələblərə uyğunluq təhlili real və ya reprezentativ mühitdə aparılmalıdır. Bu cür intellektual mexatron sistemlərin məqbulluğunu təmin etmək üçün tələb olunan quraşdırılmış Sİ-nin iş perimetrini müəyyən etmək məqsədilə testlər idarə olunan mühitlərdə aparıla bilər. Süni iqlim kameralarında fiziki testlər, vibrasiya, zərbə və sabit təcilə dayanıqlıq testləri də keçirilə bilər. Bunlar sistemlərin performansını ekstremal şəraitdə qiymətləndirmək və işləməsi üçün hüdud şərtlərini dəqiq müəyyənləşdirmək məqsədini daşıyır. Demək olar ki, sonsuz sayda mümkün konfiqurasiyaya malik açıq və dəyişən mühitlərdə Sİ sistemlərini qiymətləndirmək olar. Bu zaman simulyasiya vasitəsilə validasiyaya imkan verən virtual test mühitlərinin tərtib edilməsi də faydalı ola bilər.

9.10.2.6 İstismar şəraitində aparılan testlər

Test mühitləri ilə faktiki iş şəraiti arasındakı fərqlərə əsasən, istismar şəraitində aparılan testlər çox vaxt quraşdırılmış sistemin performans, səmərəlilik və ya davamlılığını test etməklə onun keyfiyyətini yaxşılaşdırmaq üçün çox effektiv bir üsuldur.

Bəzi diqqətəlayiq nümunə və sahələr aşağıdakılardan ibarətdir:

- üz tanıma sınaqları [116];
- kənd təsərrüfatı tətbiqləri üçün qərar qəbuetməyə dəstək sistemlərinin test edilməsi [117];
- sürücüsüz avtomobillərin test edilməsi praktikası [118],[119];
- nitq və səsin tanınması sistemlərinin test edilməsi [120],[121];
- tibbi robotexnika [122];
- çatbotların (vokal köməkçilər və s.) koqnitiv iş yükünün ölçülməsi [123];
- intellektual tədris sistemlərinin test edilməsi [124].

Sİ sistemlərinin istismar sınaqları metodologiya, istifadəçilərin sayı və ya istifadə ssenariləri, məsul təşkilatın/şəxslərin statusu və nəticələrin sənədləşdirilməsindən asılı olaraq çox fərqlənir. İstismar sınaqlarından Sİ sistemlərinin keyfiyyətinin yaxşılaşdırılmasına yönəlmiş tədbir kimi istifadə oluna bilər. Bu o cür sistemlərin tətbiqi ilə əlaqədar risklərdən asılıdır. Bir sıra tətbiqlərdə A/B testlərindən istifadə edilir. Bu, sistemin performansını müqayisə etmək məqsədilə müxtəlif istifadəçilərə sistemlərin müxtəlif versiyalarını çatdırmaq üsuludur.

Tərtib edilən Sİ proqram təminatının səmərəliliyinə ciddi diqqət yetirməklə yanaşı, onun insanlar üçün məqbul olması da nəzərə alınmalıdır. İstismar sınaqları buna nail olmağa kömək edə bilər. Bundan əlavə, funksional test aparılarkən Sİ sisteminin uğursuzluğu lazımsız ola və ya onu aradan qaldırmaq qeyri-mümkün ola bilər. Dəyişən nəticələr nümayiş etdirən Sİ sistemləri nəzərdə tutulan məqsədlər üçün faydalı hesab oluna bilər. Hərçənd sistemin Sİ sisteminin planlaşdırılmış və arzuolunan nəticəsini əldə etmək qabiliyyətini ənənəvi proqram təminatı testləri ilə əlaqədar yanaşmalarından istifadə etməklə hər zaman ölçmək olmur.

Bir sıra Sİ sistemləri ilə ənənəvi sistemlər arasındakı digər əsas fərq bundan ibarətdir ki, ənənəvi sistemlər müəyyən spesifikasiyalara uyğunluğu ciddi şəkildə təmin edilməklə tərtib və istehsal olunmalı və keyfiyyətinə nəzarət edilməlidir. Ənənəvi proqram təminatı elə tərtib olunur ki, onun davranışını təkrarlamaq mümkün olsun. Sİ sistemləri isə əksinə, ümumiləşdirməyə meyillidir. Bu, empirik testlərin keçirilməsində problemlərə gətirib çıxarı. Odur ki, istismar sınaqları keyfiyyəti qiymətləndirmək üçün daha effektiv ola bilər.

Məhsulun nəticələrinin qeyri-müəyyənliyi və quraşdırılması ilə əlaqədar risklərlə mübarizə üsulları tibb sahəsində bir sıra normativ aktların mövzusudur. Tibbi təyinatlı Sİ sistemlərinin ISO 14155 [125] standartına uyğun olması tələb edilə bilər. Onların “klinik sınaqlar”a bənzər bir prosedur olan “klinik tədqiqatlar”dan keçməsi lazım ola bilər [126],[127]. Bu, nüvə sistemləri və uçuşları idarəetmə sistemləri kimi, digər sahələr üçün də keçərlidir.

9.10.2.7 İnsan zəkası ilə müqayisə

Sİ sistemini yoxlamaq üsullarından biri onu insan zəkasının imkanları ilə müqayisə etməkdir. Bu, Sİ sistemi verilənlərin emalı və qərar qəbuletmə ilə əlaqədar insan fəaliyyətini avtomatlaşdırmaq üçün nəzərdə tutulduğu halları ehtiva edir. Bu cür yanaşma müxtəlif maraqlı tərəflərə – Sİ sistemlərinin əsas istifadəçilərinə, kənar tərəflərə və Sİ tətbiq olunduğu sahədə tənzimləyicilərə (o cümlədən tənzimləyici səlahiyyətli orqanlara) verilənlərin emalı və qərar qəbuletmə ilə əlaqədar bəzi tətbiq tapşırıqlarını yerinə yetirmək üçün Sİ sistemlərinə güvənməyə imkan verə bilər. Halbuki əvvəllər bu, əsas etibarilə, insanlar tərəfindən yerinə yetirilirdi.

İnsan fəaliyyəti ilə müqayisənin faydalı olacağı hallara nümunə kimi, avtonəqliyyat vasitələrinin idarə edilməsi və ya səhiyyə kimi ənənəvi olaraq lisenziya tələb edən fəaliyyət sahələrini qeyd etmək olar. Özü idarə olunan nəqliyyat vasitələrinin şəhər küçələrində hərəkət etməsinə və ya avtonom sistemə hər hansı bir xidmət göstərməsinə icazə verilməsi yalnız o zaman baş verə bilər ki, həmin hərəkətləri yerinə yetirən Sİ sisteminin performansının insanınkından zəif olmadığına dair sübut olsun. Bu cür yanaşma aşağıdakılara nail olmağa imkan verir:

- Sİ sistemlərinin istifadəçiləri və maraqlı tərəfləri ümid edə bilərlər ki, informasiya emalı tapşırığını yerinə yetirərkən Sİ sisteminin keyfiyyəti eyni problemi həll edən insan-operatorun təmin etdiyi keyfiyyət qədər yaxşı olacaq;
- üçüncü tərəflər ümid edə bilərlər ki, Sİ sisteminin işi insanlara və ya maddi nemətlərə zərər verməyəcək.

Belə nəticəyə gəlmək olar ki, Sİ sisteminin statistik göstəriciləri, ən azı, müəyyən edilmiş həddi qiyməti qədər yaxşı olarsa, verilənlərin emalı və prosesin təhlükəsizliyi ilə əlaqədar bəzi tətbiq tapşırıqlarının yerinə yetirilməsində Sİ sisteminin performansı, ən azı, insan imkanları qədər yaxşıdır.

Belə həddi qiymətləri və Sİ sistemlərinin proses təhlükəsizliyinə inam əldə etmək üçün vacibdir ki, Sİ sisteminin və ya təbii insan zəkasının istifadə edildiyi müvafiq verilənlərin emalı tapşırığının xarakterini əks etdirən reprezentativ verilənlər nümunələrindən istifadə olunsun.

9.10.3 Dayanıqlığa dair mülahizələr

Dayanıqlığın tərifi “sistemin istənilən şəraitdə öz performans səviyyəsini saxlamaq qabiliyyəti” kimi verilir. Dayanıqlığın daha ümumi mənada nə olduğunu anlamaq üçün qeyd etmək lazımdır ki, Sİ sistemlərindən, adətən, biliklərdən nəticə çıxarılarkən (simvolik yanaşma) və ya verilənləri ümumiləşdirərkən (subsimvolik yanaşma) istifadə olunur.

Əsas prinsip ondan ibarətdir ki, Sİ sistemi əvvəlcədən məlum olmayan verilənlərlə əhəmiyyətli dərəcədə dəyişə bilən kontekstlərdə işləməyi bacarmalıdır. Gözlənilən odur ki, Sİ sistemi çoxvariasiyalı iş şəraitində işləməlidir və onun dayanıqlığı öz konstruksiyasına müvafiq olaraq işləməyə davam etmək qabiliyyətinə uyğundur. Sİ sisteminin növündən asılı olaraq, sistemin dayanıqlığını qiymətləndirmək üçün müxtəlif göstəricilər əsas götürülməlidir.

Sİ sistemindən interpolyasiyanın yerinə yetirilməsi üçün istifadə edildikdə onun dayanıqlığı "istənilən etibarlı giriş verilənə məqbul cavab amplitudası göstəricilərinə malik olmaq qabiliyyəti" kimi nəzərdə tutulur. Yəni gözlənilir ki, Sİ sistemi interpolyasiya yerinə yetirərkən qeyri-sabit davranış nümayiş etdirməyəcək.

Sİ sistemindən klassifikasiyanın aparılması üçün istifadə edildikdə onun dayanıqlığı "həm məlum giriş verilənləri, həm də müəyyən diapazonda olan giriş verilənləri üzrə ardıcıl klassifikasiya təyin etmək qabiliyyəti" kimi qəbul edilir. Yəni Sİ sistemindən həm məlum, həm də naməlum giriş verilənlərini düzgün qaydada klassifikasiya edəcəyi gözlənilir. Bir şərtlə ki, naməlum olanlar məlum giriş verilənlərindən çox fərqli olmasın.

Sİ sistemindən problem həll etmə tapşırığının yerinə yetirilməsi məqsədilə istifadə edildikdə onun dayanıqlığı "ilkin problemdə məqbul dəyişiklik həyata keçirildikdən sonra da effektiv həll təmin etmək qabiliyyəti" kimi qəbul olunur. Yəni Sİ sistemindən müxtəlif problemlər üçün məqbul həll yolları təmin edəcəyi gözlənilir. Bir şərtlə ki, onlar ilkin problemdən çox fərqli olmasın.

Sİ sistemindən bal hesablanmasında istifadə edildikdə onun dayanıqlığı "həm məlum giriş verilənləri, həm də məqbul diapazon daxilində olan giriş verilənləri üzrə sıralama üçün ardıcıl əminlik ölçüləri təyin etmək qabiliyyəti" kimi qəbul edilir. Yəni naməlum giriş və çıxış verilənləri halında, Sİ sistemindən məlum giriş verilənləri və naməlum giriş verilənlərinə təyin olunan baldan köklü şəkildə fərqlənməyən bal təyin etməsi gözlənilir. Bir şərtlə ki, naməlum giriş verilənləri məlum giriş verilənlərindən çox fərqli olmasın.

9.10.4 Məxfiliklə bağlı mülahizələr

Sİ sistemində məxfilik təhdidlərini aradan qaldırarkən məxfilik göstəriciləri məxfilik səviyyələrini və sistemin təmin etdiyi mühafizə dərəcəsini qiymətləndirməyə kömək edir. Məxfilik göstəricilərinin müəyyənləşdirilməsi və tətbiq edilməsi bu problemin həllinə yönəlmişdir. Sİ sistemlərində müxtəlif sahələrdə həssas verilənləri qorumaq üçün müxtəlif maşın öyrənməsi və ya məxfiliyin təkmilləşdirilməsi üsulları mövcuddur. Məxfilik göstəricilərinin müəyyən edilməsi məlumatların məxfilik səviyyəsinin kəmiyyətcə qiymətləndirilməsidir. Bu, öz növbəsində, konkret Sİ modeli daxilində məxfilik modelinin təkmilləşdirilməsi ilə nəticələnir [128],[129]. Texniki olaraq, məxfilik göstəricisi verilənlərin müxtəlif xassələrini nəzərə alır və sistemdə məxfilik səviyyəsini əks etdirən qiymət hasil edir. Məxfilik göstəricilərinin üstünlüyü məxfiliyi qoruyan müxtəlif üsulları müqayisə etmək, konkret sahədə müxtəlif metodları qiymətləndirmək və məxfiliyə mənfi təsirləri minimuma endirmək qabiliyyətidir. Məxfilik göstəriciləri həssas verilənlərin təcavüzkar tərəfindən təhdidə məruz qaldığı zaman faydalıdır. Məxfilik göstəriciləri verilənlərin mənbəyi, qiymətləndirilən məxfilik aspektləri və təcavüzkarın yaratdığı təhdiddən asılı olaraq fərqlənir.

9.10.5 Sistemin proqnozlaşdırılmasına dair mülahizələr

Yuxarıda təsvir edilən bəzi test və validasiya yanaşmaları Sİ sisteminin proqnozlaşdırıla bilməsini qiymətləndirmək üçün vacibdir. Proqnozlaşdırıla bilməni anketə əsaslanan eksperimentlər çərçivəsində subyektiv açıq rəylər vasitəsilə ölçmək olar. Həmin eksperimentlərdə iştirakçılardan məqsədləri müəyyənləşdirmək və məsələn, robotun gələcək hərəkətlərini proqnozlaşdırmaq xahiş olunur [130]. Digər göstəricilər də istifadə oluna bilər. Məsələn, istifadəçinin Sİ sisteminin niyyətini müəyyən etməsi və müvafiq şəkildə reaksiya verməsi üçün reaksiya müddəti. Bu halda, fərz edilir ki, daha qısa reaksiya müddəti proqnozlaşdırıla bilmə dərəcəsinin daha yüksək olduğunu göstərir. Diqqətli baxış davranışı da robotun proqnozlaşdırıla bilməsinin dolayı göstəricisini təmin edir. Fərz edilir ki, iştirakçı

robotu nə qədər tez-tez və uzun müddət müşahidə edirsə, robot o qədər az proqnozlaşdırıla bilər [131]. Sİ sisteminin çox sayda ətraf mühit şəraiti kombinasiyası üzərində test edilməsi onun davranışını daha dolğun xarakterizə etməyə imkan verir. İstifadəçilər bu xarakteristikaya əsaslanaraq Sİ sistemindən nə gözləyəcəklərini bilirlər, bu da proqnozlaşdırıla bilməni asanlaşdırır.

9.11 İstifadə və tətbiq imkanı

9.11.1 Uyğunluğun təmin edilməsi

Uyğunluğun təmin edilməsinə olan ehtiyac qismən müxtəlif sənaye sahələrində artıq tətbiq olunan müxtəlif standart və qaydalarla bağlıdır. Sİ sistemləri mövcud qaydaları və uyğunluq standartlarını nəzərə almalı və istifadə ssenarisindən ayrı olaraq qiymətləndirilməməlidir.

9.11.2 Gözləntilərin idarə edilməsi

Sistem qeyri-real gözləntilərə uyğun performans nümayiş etdirə bilmir. Elə bu səbəbdən güvənin zəifləməsinin qarşısını almaq üçün gözləntiləri idarə etmək lazımdır. Bunun üçün Sİ sisteminin real imkanları, o cümlədən etibarlı nəticə gözlənilən giriş verilənləri diapazonu və nəticələrin dəqiqliyi (xüsusilə statistik nəticə çıxarmaya əsaslanan sistemlər üçün) barədə aydın təsəvvür olmalıdır.

9.11.3 Məhsulların nişanlanmsı

Məhsul və sistemlərin etiketlənməsi (o cümlədən ən son yenilənmiş informasiyaya aid linklər) həm istifadəçilərin, həm də Sİ sistemlərinin təchizatçılarının təhlükəsizliyi üçün mühüm əhəmiyyət kəsb edə bilər:

— son istifadəçinin Sİ agenti ilə qarşılıqlı əlaqədə olması və Sİ sisteminin niyyətini/məqsədini bildirməsi;

— modelin risk və məhdudiyyətləri;

— ehtiyac olduqda təkrar təlim tezliyi;

— son performans qiymətləndirilməsinin aparıldığı tarix;

— təlim verilənlərinin mənbəyi və tarixi [132].

9.11.4 Koqnitiv elmi tədqiqatlar

Bu bölmədəki müzakirələrə əsasən, sistemlərin münasib etimad doğrultma səviyyəsinə və düzgün istifadəsinə nail olmaq üçün konkret tətbiqlər üzrə tövsiyə və mülahizələr vacibdir. Bununla belə, sistemin keyfiyyəti ilə onun qeyri-münasib istifadəsi və ya istifadəsizliyi ehtimalı arasında qarşılıqlı əlaqənin qəbul edilmiş bir neçə nümunəsi var. Məsələn, yüksək etibarlı sistemlərin həddən artıq etimad doğurduğu və nəticə etibarilə, qeyri-münasib istifadəyə səbəb olduğu (məsələn, aviasiyada)məlumdur. Etibarsız sistemlərin (məsələn, EHR) isə inamsızlığa və nəticə etibarilə, istifadəsizliyə səbəb olduğu məlumdur[133]. Bu qanunauyğunluq dayanıqlıq, dözümlülük və dəqiqlik kimi digər göstəricilər üçün də keçərlidir.

Daha yaxşı sistemlər güvən hissi oyadır, bu da avtomatlaşdırılmış sistemin idarə edə bilməyəcəyi vəziyyətlərdə sistem uğursuzluqlarına səbəb ola bilər.

Nəticədə, insan amillərini nəzərə alaraq, insan yönümlü Sİ sistemlərində bu göstəricilərin optimallaşdırılmasının nüanslarını nəzərə almaq daha yaxşıdır. Bu yanaşma faydalı nəticələr və potensial olaraq, elmi cəhətdən əsaslandırılan standartlar təmin etmək üçün insan-kompüter qarşılıqlı əlaqəsi (HCI) və koqnitiv elm tədqiqatlarına əsaslanacaq.

10 YEKUN NƏTİCƏLƏR

Müştərilər, istifadəçilər və bütövlükdə cəmiyyət arasında Sİ tətbiqlərinin etibarlılığı, effektivliyi, ədalətliliyi və hətta təyinatına güvənin olmaması Sİ sistemlərinin potensial faydalarının dərk edilməsinə mane ola bilər. Biznes, hökumət, sosial və etik xarakterli narahatlıq doğuran problemlər sistematik şəkildə həll edilmədikdə Sİ sistemlərinə güvəni sarsıda bilər. Bu cür narahatlıqlar maşın öyrənməsinə əsaslanan Sİ sistemlərinin qərəz, proqnozlaşdırılmanın qeyri-mümkünlüyü və qeyri-şəffaflıq kimi zəifliklərindən irəli gələ bilər. Verilənlərin məxfiliyi və verilənlərin idarə edilməsi ilə əlaqədar digər məsələlər, məsələn, verilənlərin mənşəyi və keyfiyyəti Sİ sistemlərinin qurulması və istifadəsi zamanı ciddi problemlərə çevrilə bilər. Bunun səbəbi maşın öyrənməsinə əsaslanan bir sıra tətbiqlərin böyük verilənlərlə işləməsidir. Sİ sistemlərinin etimadı doğrultma xüsusiyyətini yaxşılaşdırmaq üçün maşın öyrənməsi və verilənlər ilə əlaqədar zəifliklər əsas prinsiplərdə, proseslərdə və hər bir konkret halda dəqiqliklə nəzərə alınmalı və aradan qaldırılmalıdır.

Konkret bir halda, Sİ-dən istifadənin məqsədəuyğun olub-olmayacağına dair qərar vermək üçün Sİ zəifliklərinin maraqlı tərəflərə potensial təsiri araşdırılmalıdır. Sİ tərtib və ya istifadə edən təşkilat özü, tərəfdaşları, nəzərdə tutulan istifadəçilər və cəmiyyətə mümkün təsirləri müəyyən etmək və riskləri müvafiq qaydada azaltmaq üçün risklərə əsaslanan yanaşma tətbiq edə bilər. Maraqlı tərəflərin hamısının potensial risklərin və onların azaldılması üçün həyata keçirilən tədbirlərin mahiyyətini başa düşməsi vacibdir.

Sİ sistemlərinə inam yaratmaq və qorumağa kömək edə biləcək kəmiyyətcə keyfiyyət ölçmələri və təkrarlanan proses modelləri hələ yaradılmamışdır. Sİ sistemlərinin dayanıqlığını müəyyən edilmiş səviyyəyə çatdırmaq üçün mövcud və yeni metodologiyaları tətbiq etmək çox vacibdir.

Yekun olaraq, bu sənəddən məlum olur ki, Sİ sistemləri və texnologiyalarının etimadı doğrultma xüsusiyyəti həm maraqlı tərəflərin Sİ və onun verilənlərdən şəffaf və əlçatan şəkildə istifadəsi ilə əlaqədar narahatlıqlarının həllindən, həm də bütün həyat dövrü ərzində texniki cəhətdən dayanıqlı, idarə və verifikasiya edilə bilən Sİ sistemlərinin qurulmasından asılıdır.

QOŞMA A
(məlumat üçün)

İctimai məsələlərlə bağlı işlər

İctimai məsələlərə münasibətdə Si sistemlərinin etimadı doğrultma xüsusiyyəti ilə bağlı çoxsaylı sahələrə aid böyük həcmdə işlər görülmüşdür. O cümlədən texnoloji etika, tədqiqat və innovasiya etikası (çox vaxt məsuliyyətli tədqiqat və innovasiya adlandırılır), məsul alqoritmlər, habelə verilənlərin idarə edilməsi çərçivəsində verilənlər üzrə etika və verilənlərin qorunmasını da buraya aid etmək olar. Bu digər işlərin məqsədi və əhatə dairəsi əksər hallarda tərtibatçılar, işçilər, müştərilər, istifadəçilər və bütövlükdə cəmiyyətə onların Sİ-dən istifadə edən bu və ya digər konkret xidmət və ya məhsula göstərəcəkləri güvən səviyyəsini səmərəli və dəqiq müəyyənləşdirməyə imkan verən standartların təmin edilməsindən daha yüksək səviyyədədir.

Mövcud işlərə dair misallara aşağıdakılar daxildir.

— Elektrotexnika və elektronika mühəndisləri institutu (IEEE) avtonom və intellektual sistemlərin etik cəhətdən əsaslandırılmış layihələndirilməsi ilə əlaqədar problemlərin hərtərəfli təhlilini aparmışdır [134]. Bu, peşəkar etikanın inkişaf etdirilməsinə yönəlmişdir. Eyni zamanda hesabatlılıq, şəffaflıq, verifikasiya oluna bilmə, proqnozlaşdırıla bilmə, sosial normalara uyğunluq, dəyərlər və layihələndirmə metodologiyaları sahələrində müvafiq töhfə verir.

— Bütün elmi sahələri əhatə edən, lakin dövlət maliyyəsi ilə dəstəklənənlərə ilk növbədə diqqət yetirən məsuliyyətli tədqiqat və innovasiya (RRI) sahəsində geniş təcrübə mövcuddur və uyğunlaşdırma üzrə işlər artır [135]. Onlar Sİ sistemlərinə güvəni formalaşdırıb artırmaq məqsədilə lazım olan üsullar üçün əsaslandırılmış yanaşmalar təklif edir. Həmçinin bu, elmi nəşr, təlim və tədqiqatçılar arasında gender balansı, açıq tədqiqat verilənləri və heyvan testləri ilə bağlı məsələləri əhatə edir (sonuncular hazırkı sənədin əhatə dairəsi xaricindədir). Özəl mənbələrdən maliyyələşdirilən inkişaf layihələrində RRI yanaşmasının tətbiqi ilə əlaqədar təcrübə də prioritet olacaq [136],[137].

— ISO/IEC AWI 38507 [138] standartı Sİ tipli qərarları idarə etmək üçün istifadə kontekstinin təmin olunmasına kömək etmək məqsədi daşıyır.

ƏDƏBİYYAT

[1] ISO/IEC 27000:2018, *Information technology — Security techniques — Information security management systems — Overview and vocabulary*

[2] ISO/IEC 11557:1992, *Information technology — 3,81 mm wide magnetic tape cartridge for information interchange — Helical scan recording — DDS-DC format using 60 m and 90 m length tapes*

[3] ISO/IEC 2382:2015, *Information technology — Vocabulary*

[4] ISO/IEC/IEEE 15939:2017, *Systems and software engineering — Measurement process*

[5] ISO/IEC 21827:2008, *Information technology — Security techniques — Systems Security Engineering — Capability Maturity Model® (SSE-CMM®)*

[6] IEC 61800-7-1:2015, *Adjustable speed electrical power drive systems — Part 7-1: Generic interface and use of profiles for power drive systems — Interface definition*

[7] ISO 5127:2017, *Information and documentation — Foundation and vocabulary*

[8] ISO 9000:2015, *Quality management systems — Fundamentals and vocabulary*

[9] ISO/IEC 10746-2:2009, *Information technology — Open distributed processing — Reference model: Foundations — Part 2*

[10] ISO/IEC Guide 51:2014, *Safety aspects — Guidelines for their inclusion in standards*

[11] ISO 13702:2015, *Petroleum and natural gas industries — Control and mitigation of fires and explosions on offshore production installations — Requirements and guidelines*

[12] ISO 10018:2020, *Quality management — Guidance for people engagement*

[13] ISO 18115-1:2013, *Surface chemical analysis — Vocabulary — Part 1: General terms and terms used in spectroscopy*

[14] ISO 31000, *Risk management — Guidelines*

[15] ISO 18646-2:2019, *Robotics — Performance criteria and related test methods for service robots — Part 2: Navigation*

[16] ISO 8373:2012, *Robots and robotic devices — Vocabulary*

[17] ISO/IEC 38500:2015, *Information technology — Governance of IT for the organization*

[18] ISO/IEC/IEEE 15288:2015, *Systems and software engineering — System life cycle processes*

[19] ISO/IEC 25012:2008, *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*

[20] ISO/IEC 25010:2011, *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*

[21] ISO/IEC/TR 22417:2017, *Information technology — Internet of things (IoT) use cases*

[22] ISO/IEC/IEEE 24765:2017, *Systems and software engineering — Vocabulary*

[23] ISO 22316:2017, *Security and resilience — Organizational resilience — Principles and attributes*

[24] ISO/IEC/TR 29110-1:2016, *Systems and software engineering — Lifecycle profiles for Very Small Entities (VSEs) — Part 1: Overview*

[25] ISO 10303-11:2004, *Industrial automation systems and integration — Product data representationand exchange — Part 11: Description methods: The EXPRESS language reference manual*

[26] United Nations Environment Programme (UNEP) *Rio Declaration on Environmentand Development*, 1992 (http://www.jus.uio.no/lm/environmental.development.rio.declaration.1992/portrait.a4.pdf)

[27] ITU-T Y.3052, *Overview of Trust Provisioning in ICT Infrastructures and Services,* 2017 (https://www.itu.int/rec/T-REC-Y.3052/en)

[28] IEEE 1012:2016, *IEEE Standard for System, Software and Hardware Verification and Validation*

[29] Mohammadi N. et al. *Trustworthiness Attributes and Metrics for Engineering Trusted Internet-Based Software Systems*, 2014. Communications in Computer and Information Science. 453. 19-35. 10.1007/978-3-319-11561-0_2 (https://www.researchgate.net/publication/267390448_Trustworthiness_Attributes_and_Metrics_for_Engineering_Trusted_Internet-Based_Software_Systems)

[30] Colquitt J., Scott B., Le Pine J. A, Trust, Trustworthiness and Trust Propensity: A Meta-Analytic Test of Their Unique Relationships With Risk Taking and Job Performance, *The Journalof applied psychology*, 2007. **92**. 909-27.10.1037/0021-9010.92.4.909 (https://www.researchgate.net/publication/6200985_Trust_Trustworthiness_and_Trust_Propensity_A_Meta-Analytic_Test_of_Their_Unique_Relationships_With_Risk_Taking_and_Job_Performance)

[31] Marr B. *What Is Deep Learning AI? A Simple Guide With 8 Practical Examples*, 2018 (https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples)

[32] O'Keefe K., Daragh O Brien, *Ethical Data and Information Management: Concepts, Tools andMethods*, Kogan Page pp. 46-47, 214-218, 262-263, 2018

[33] Freeman R.E., McVea J. *A Stakeholder Approach to Strategic Management*, 2001, DardenBusiness School Working Paper No. 01-02

[34] The European Commission High Level Expert Group on Artificial Intelligence *Draft EthicsGuidelines for Trustworthy AI*, Working Document for stakeholder' consultation, Brussels,December 2018

[35] IEEE EAD v2, *Ethically Aligned Design — Version II Request for Input*, The IEEE Global Initiative onEthics of Autonomous and Intelligent Systems, 2018

[36] Borning A., Muller M. *Next steps for Value Sensitive Design,* 2012. Proceedings of CHI, 1125-1134. New York, NY, ACM Press, 2012

[37] ISO/IEC 38505-1:2017, *Information technology — Governance of IT — Governance of data —Part 1: Application of ISO/IEC 38500 to the governance of data*

[38] Winikoff M. *How to make robots that we can trust*, 2018 (http://theconversation.com/how-to-make-robots-that-we-can-trust-79525)

[40] Doshi-Velez F. et al. *Accountability of AI Under the Law: The Role of Explanation*, SSRN ElectronicJournal, 2017

[41] European Group on Ethics in Science and New Technologies *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, 2018 (doi: 10 .2777/ 531856)

[42] ISO/IEC 30141:2018, *Internet of Things (IoT) — Reference Architecture*

[43] Commission de Surveillance du Secteur Financier *Artificial Intelligence: opportunities, risks andrecommendations for the financial sector*, 2018

[44] Pei K. et al. *Towards Practical Verification of Machine Learning: The Case of Computer VisionSystems*, 2017

[45] Szegedy C. et al. *Intriguing properties of neural networks*, 2014

[46] Goodfellow I. et al. *Explaining and Harnessing Adversarial Examples*, 2015

[47] Brendel W., Rauber J., Bethge M. *Decision-Based Adversarial Attacks: Reliable Attacks AgainstBlack-Box Machine Learning Models*, 2018

[48] Akhtar N., Mian A., Threat of Adversarial Attacks on Deep Learning in Computer Vision*: ASurvey*, 2018

[49] Yuan X. et al. *Adversarial Examples: Attacks and Defences for Deep Learning*, July 2018

[50] Raghunathan A., Steinhardt J., Liang P. *Certified Defenses against Adversarial Examples*, 2018

[51] ISO/IEC 29100:2011, *Information technology — Security techniques — Privacy framework*

[52] ISO/IEC 20889:2018, *Privacy enhancing data de-identification terminology and classification of techniques*

[53] Truta T.M., Vinay B. *Privacy Protection: p-Sensitive k-Anonymity Property*, 22nd InternationalConference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006, pp. 94-94

[54] Rocher L., Hendrickx M., de Montjoye Y., Estimating the success of re-identifications inincomplete datasets using generative models, *Nature Communications* **10**, Article number:3069, 2019

[55] Bergman M., Van Zandbeek M. *Close Encounters of the Fifth Kind? Affective Impact of Speedand Distance of a Collaborative Industrial Robot on Humans, Human Friendly Robotics*, p. 127-137.Springer, Cham, 2018

[56] Brownlee J. Ph.D., *Discover Feature Engineering, How to Engineer Features and How to Get Goodat It,* Machine Learning Mastery, 2014 (https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/)

[57] Parasuraman R., Riley V.A., Humans and automation: Use, misuse, disuse, abuse, *HumanFactors*, vol. **39**, pp. 230–253

[58] Haines T. S.F., Mac Aodha O., Brostow G. J. *My Text in Your Handwriting*, University CollegeLondon, Transactions on Graphics, 2016 (http://visual.cs.ucl.ac.uk/pubs/handwriting/)

[59] van DEN Oord A. *WaveNet: A Generative Model for Raw Audio,* 2016 (https://deepmind.com/blog/wavenet-generative-model-raw-audio/)

[60] Wagner K. *Someone built chatbots that talk like the characters from HBO's 'Silicon Valley'*, 2016 (https://www.recode.net/2016/4/24/11586346/silicon-valley-hbo-chatbots-for-season-3-premier)

[61] Dormon B. *The AI that (almost) lets you speak to the dead*, 2016 (https://arstechnica.com/information-technology/2016/07/luka-ai-chatbot-speaking-to-the-dead-mind-uploading/).

[62] Newitz A. *Ashley Madison code shows more women and more bots*, 2015 (https://gizmodo.com/ashley-madison-code-shows-more-women-and-more-bots-1727613924)

[63] World Economic Forum & UNICEF Innovation Centre *Generation AI*, (https://www.weforum.org/projects/generation-ai, https://www.unicef.org/innovation/stories/generation-ai)

[64] Pearl J. *Theoretical Impediments to Machine Learning With Seven Sparks from the CausalRevolution*, ArXiv: 1801 .04016 [Cs, Stat], January 2018 (http://arxiv.org/abs/1801.04016)

[65] Guidotti R. et al., A Survey of Methods for Explaining Black Box Models, *ACM Computing Surveys* **51**:1–42, 2019

[66] Achinstein P., The Nature of Explanation, 1985, Oxford University Press, 2017 (https://books.google.fi/books?id=TkULPY-xMNsC)

[67] Adadi A., Berrada M. 2018, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), *IEEE Access*; **6**:52138-60

[68] Doshi-Velez F., Kim B. 2017, *Towards A Rigorous Science of Interpretable Machine Learning*, arXiv 1702.08608

[69] Lipton Z. C., The mythos of model interpretability, *Communications of the ACM*, vol. **61**, no. 10, pp. 36–43, 2018

[70] Dhurandhar A. et al. 2017, *TIP: Typifying the Interpretability of Procedures*, arXiv preprint arXiv: 1706. 02952

[71] Miller T. *Explanation in artificial intelligence: insights from the social sciences*, 2017, arXiv preprint arXiv: 1706. 07269

[72] https://pair-code.github.io/facets/

[73] https://projector.tensorflow.org/

[74] Holland S. et al. 2018, *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*, arXiv preprint arXiv: 1805. 03677

[75] Gebru T. et al. 2018, *Datasheets for Datasets*, arXiv preprint arXiv: 1803. 09010

[76] Seck I. et al. 2018, *Baselines and a datasheet for the Cerema AWP dataset*, arXiv preprint arXiv: 1806 .04016

[77] Angelino E. et al. 2018, Learning certifiably optimal rule lists for categorical data, *Journal of Machine Learning Research*, **18**(234), pp.1-78

[78] Vaughan J. et al. 2018, *Explainable Neural Networks based on Additive Index Models*, arXiv preprint arXiv: 1806. 01933

[79] Kim B., Khanna R., Koyejo O. 2016, *Examples are not enough, learn to criticize! criticism for interpretability*, Proc 30th Int Conf Neural Inf Process Syst: 2288–96

[80] Ribeiro M.T., Singh S., Guestrin C. 2016, August, *Why should I trust you?: Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144), ACM

[81] Kim B. et al. 2018, July, *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)*, International Conference on Machine Learning (pp. 2673-2682)

[82] Koh P.W., Liang P. 2017, *Understanding black-box predictions via influence functions*, arXiv preprint arXiv: 1703. 04730

[83] Maclaurin D., Duvenaud D., Adams R. 2015, June, *Gradient-based hyperparameter optimization through reversible learning*, International Conference on Machine Learning (pp. 2113-2122)

[84] Friedman J.H. 2001, Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232

[85] Lah C., Mordvintsev A., Schubert L. 2017, *Feature visualization*, Distill, 2(11), p.e7

[86] Olah C. et al. 2018, The building blocks of interpretability, *Distill*, **3**(3), p.e10

[87] Erhan D. et al. 2009, Visualizing higher-layer features of a deep network, *University of Montreal*, **1341**(3), p.1

[88] Lundberg S.M., Lee S.I. 2017, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems* (pp. 4765-4774)

[89] Hohman F.M. et al. 2018, Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers, *IEEE Transactions on Visualization and Computer Graphics*

[90] Yosinski J. et al. 2015, *Understanding neural networks through deep visualization*, arXiv preprint arXiv: 1506. 06579

[91] Strobelt H. et al. 2018, Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks, *IEEE transactions on visualization and computer graphics*, **24**(1), pp.667-676

[92] Strobelt H. et al. 2018, *Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models*, arXiv preprint arXiv: 1804. 09299

[93] Andrews R., Diederich J., Tickle A.B. 1995, Survey and critique of techniques for extracting rules from trained artificial neural networks, *Knowledge-based systems*, **8**(6), pp.373-389

[94] Ailesilassie T. 2016, *Rule extraction algorithm for deep neural networks: A review*, arXiv preprint arXiv: 1610. 05267

[95] Ribeiro M.T., Singh S., Guestrin C. 2018, *Anchors: High-precision model-agnostic explanations*, AAAI Conference on Artificial Intelligence

[96] List C. *Levels: descriptive, explanatory and ontological*, 2017 (http://philsci-archive.pitt.edu/13311/)

[97] Woodward J. *In Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in the Philosophy of Science. Oxford, New York: Oxford University Press, 2005

[98] Montavon G., Samek W., Müller K.R. 2017, Methods for interpreting and understanding deep neural networks, *Digital Signal Processing*

[99] Lundberg S.M., Lee S.I. 2017, *A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems,* pp. 4765-4774

[100] Gilpin L.H. et al. 2018, *Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning*' arXiv preprint arXiv: 1806. 00069

[101] Narayanan M. et al. 2018, *How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation*, arXiv preprint arXiv: 1802 .00682

[102] Hooker S. et al. 2018, *Evaluating Feature Importance Estimates*, arXiv preprint arXiv: 1806 .10758

[103] Samek W. et al. 2017, Evaluating the visualization of what a deep neural network has learned, *IEEE transactions on neural networks and learning systems*, **28**(11), pp. 2660-2673

[104] Arras L. et al. 2017, What is relevant in a text document?: An interpretable machine learning approach, PloS one, **12**(8), p.e0181142

[105] Bach S. et al. 2015, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one, **10**(7), p.e0130140

[106] ISO/IEC 27040:2015, *Information technology — Security techniques — Storage security*

[107] IEC 61508:2010, *Commented version, Functional safety of electrical/electronic /programmable electronic safety-related systems*

[108] ISO 26262 (all parts), *Road vehicles — Functional safety*

[109] IEC 62279:2015, *Railway applications — Communication, signalling and processing systems -Software for railway control and protection systems*

[110] IEC 61511:2018, *Functional safety — Safety instrumented systems for the process industry sector*

[111] ISO/IEC/IEEE 29119-1:2013, *Software and systems engineering — Software testing — Part 1: Concepts and definitions*

[112] Zhou ZQ., Sun L. 2019, Metamorphic testing of driverless cars, *Communications of the ACM* **62**:61–67

[113] Sojda R.S. 2007, Empirical evaluation of decision support systems: Needs, definitions, potential methods and an example pertaining to waterfowl management, *Environmental Modelling & Software* **22**:269–277

[114] Ngan M., Grother P. 2015, *Face Recognition Vendor Test (FRVT) - Performance of Automated Gender Classification Algorithms*, National Institute of Standards and Technology

[115] Kohonen, Barna, Chrisley, 1988, *Statistical pattern recognition with neural networks: benchmarking studies*, IEEE 1988 International Conference on Neural Networks. pp 61–68

[116] BSI *An investigation into the performance of facial recognition systems relative to their planned use in photo identification documents – BioP I.* Bundesamt für Sicherheit in der Informationstechnik (BSI), Bundeskriminalamt (BKA), secunet Security Networks AG. 2004

(https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Studies/BioP/BioPfinalreport_pdf)

[117] Burke J.J., Dunne B., Field testing of six decision support systems for scheduling fungicide applications to control Mycosphaerella graminicola on winter wheat crops in Ireland, *The Journal of Agricultural Science*, V **146** N 4, 2008

[118] UK-Government *The Pathway to Driverless Cars: A Code of Practice for testing*. Great Minster House, 33 Horseferry Road, London SW1 P 4DR: Department for Transport, 2015 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/446316/pathway-driverless-cars.pdf)

[119] Dressler F. et al. Inter-vehicle communication: Quo vadis, *IEEE Communications Magazine*, vol. **52**, no. 6, pp. 170-177, June 2014 (doi: 10.1109/MCOM.2014.6829960)

[120] Lamel L. et al. *Field trials of a telephone service for rail travel information*. In Proceedings Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, 111–116. IEEE, 1996

[121] Isobe T. et al. *Voice-activated home banking system and its field trial*. In Proceedings Fourth International Conference on Spoken Language, ICSLP 96, 1688–1691. IEEE 1996.

[122] Jayawardena C. et al. Socially Assistive Robot HealthBot: Design, Implementation and Field Trials, in *IEEE Systems Journal*, vol. **10**, no. 3, pp. 1056-1067, Sept. 2016 (doi: 10.1109/JSYST.2014.2337882)

[123] Strayer D. L. e t al. *The smartphone and the driver's cognitive workload: A comparison of Apple, Google and Microsoft's intelligent personal assistants*. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 71(2), 93, 2017

[124] Causo A. et al. *Design of robots used as education companion and tutor.* In *Robotics and Mechatronics* (pp. 75-84). Springer, Cham, 2016

[125] DIN EN ISO 14155:2012-01, *Clinical investigation of medical devices for human subjects — Good clinical practice (ISO 14155:2011 + Cor. 1:2011); German version EN ISO 14155:2011 + AC: 2011, p.1–68*, (https://www.beuth.de/de/norm/din-en-iso-14155/134954877)

[126] Läkemedelsverket *The Medical Products Agency's Working Groupon Medical InformationSystems*. Medical Products Agency (Läkemedelsverket) Sweden, 2009 (https://lakemedelsverket.se/upload/foretag/medicinteknik/en/Medical-Information-Systems-Report_2009-06-18.pdf)

[127] Florek H.-J. et al. *Results from a First-in-Human Trial of a Novel Vascular Sealant. Frontiers inSurgery*, V 2, 2015

[128] Wagner I., Eckhoff D. *Technical privacy metrics: a systematic survey*, ACM Computing Surveys (CSUR) 51.3 (2018): 57, 2018

[129] Wu F.-J. et al. *The privacy exposure problem in mobile location-based services*, GlobalCommunications Conference (GLOBECOM), IEEE. IEEE, 2016

[130] Lichtenthäler C., Kirsch A. *Goal-predictability vs. trajectory-predictability, which legibilityfactor counts*. In proceedings of the 2014 ACM/IEEE international conference on human-robotinteraction, 2014. p. 228-229

[131] Lichtenthäler C., Lorenz T., Kirsch A. *Towards a legibility metric: How to measure the perceived value of a robot*. In International Conference on Social Robotics, ICSR 2011. 2011

[132] Davenport J.H. *The debate about "algorithms"*, Mathematics Today 162, August 2017, pp. 162-165. (https://ima.org.uk/6910/the-debate-about-algorithms/)

[133] Wohleber R. et al. *The Impact of Automation Reliability and Operator Fatigue on Performanceand Reliance*, 2016

[134] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence andAutonomous Systems *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With ArtificialIntelligence And Autonomous Systems*, Version 1. IEEE, 2016 (http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html)

[135] Reijers W. et al. *Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations*, Science and Engineering, Ethics Journal, pp 1-45, Springer, 2017

[136] Gurzawska A., Mäkinen M., Brey P., Implementation of Responsible Research and Innovation (RRI) Practices in Industry: Providing the Right Incentives, *Sustainability* 2017, **9**, 1759 (doi: 10.3390/ su9101759)

[137] Lubberink R. et al., Lessons for Responsible Innovation in the Business Context: A SystematicLiterature Review of Responsible, Social and Sustainable Innovation Practices, *Sustainability*,2017, **9**, 721 (doi: 10 .3390/ su9050721)

[138] ISO/IEC AWI 38507, *Information technology — Governance of IT — Governance implications of theuse of artificial intelligence by organizations*