**ЖАСАНДЫ ИНТЕЛЛЕКТ ҚЫЗМЕТТЕРІН ПАЙДАЛАНУДАҒЫ ҚАУІПСІЗДІК**

Жасанды интеллект (AI) құралдары CHATGPT 2022 жылдың қарашасында іске қосылғаннан бері пайдаланушылардың назарын кеңінен аударды. Қолдану біздің жұмысымызды немесе өмірімізді жақсартса да, құралдарды қолданатын және пайдаланушыларды зиянды бағдарламаны жүктеуге немесе құпия ақпаратпен бөлісуге алдау үшін қылмыскерлер болуы мүмкін. Сондықтан ықтимал қауіпсіздік мәселелерін елемеуге болмайды.

**AI құралдарын пайдалану кезіндегі қауіпсіздік кеңестері:**

**1.Сенімді AI құралын пайдаланыңыз:** AI құралдарына тек олардың ресми арнасы арқылы қол жеткізіңіз. Қолданбаларды тек ресми қолданбалар сайттарынан орнатыңыз.

**2.Бірегей және күрделі құпия сөзді пайдаланыңыз:** Егер сіз AI құралдарына кіру үшін есептік жазбаны қолдансаңыз, бірегей және күшті құпия сөзді қолданғаныңызға көз жеткізіңіз. Бірнеше есептік жазба үшін бірдей құпия сөзді пайдаланудан аулақ болыңыз және мүмкіндігінше екі факторлы аутентификацияны (2FA) қосыңыз.

**3. Қауіпсіз желіні пайдаланыңыз:** Жасанды интеллект құралдарына қол жеткізген кезде *Жалпыға Ортақ* Wi-Fi желілерін немесе басқа қорғалмаған желілерді пайдаланудан аулақ болыңыз. Қауіпсіз Wi-Fi немесе VPN желісін пайдаланыңыз.

**4. Жеке немесе құпия ақпаратты бөліспеңіз:** AI құралдарымен аты, мекен-жайы, телефоны, құпия сөздері, жеке сәйкестендіру нөмірі немесе әлеуметтік қауіпсіздік нөмірі немесе қаржылық Ақпарат және т.б. сияқты құпия немесе құпия ақпаратты бөліспеңіз.

**5.Браузер тарихын тазалаңыз:** AI құралдарын қолданғаннан кейін жазбаларыңыздың немесе сөйлесулеріңіздің барлық іздерін жою үшін шолғыш тарихын, кэшті және cookie файлдарын тазалаңыз.

**6.Жасанды интеллект құралдары шығаратын нәтижелерді тексеру:** жасанды интеллект құралдары шығаратын нәтижелерді тексеру ұсынылады, өйткені жасанды интеллект құралдары ұсынатын нәтижелер жеткілікті дәл болмауы мүмкін. Сондай-ақ, қосымшаны жасаушылар зиянды кодтың жоқтығына көз жеткізу және қауіпсіздіктің бұзылуын болдырмау үшін AI құралдары жасаған немесе тексерген бағдарламалық кодты екі рет тексеруі керек.

**7.Бағдарламалық жасақтаманы үнемі жаңартып отырыңыз:** киберқылмыскерлер пайдалана алатын осалдықтардың алдын алу үшін әрқашан операциялық жүйені, қолданбаларды және антивирустық бағдарламалық құралды жаңартыңыз, соңғы қауіпсіздік патчтарын орнатыңыз.

**8.Фишинг әрекеттерінен сақ болыңыз:** Киберқылмыскерлер басқалардың кейпіне ену және жеке ақпаратты ұрлау үшін жасанды интеллект құралдарын қолдануға тырысуы мүмкін. Әрқашан заңды AI құралдарымен өзара әрекеттесіп жатқаныңызды тексеріңіз және күдікті сілтемелерді басудан немесе тіркемелерді жүктеуден аулақ болыңыз.

**Жасанды интеллект арқылы таратылатын дезинформациямен күресу**

**Дезинформация** – бұл қасақана жасалған, алдауға бағытталған және зиян келтіру мақсатында таратылатын жалған немесе жаңылыстыратын ақпарат.1 Бұл жай ғана қате ақпараттан (мизинформация) өзгеше, өйткені оның негізінде зиян келтіру ниеті жатыр. Жасанды интеллект (ЖИ) контекстінде бұл ұғым озық алгоритмдер мен машиналық оқыту әдістерін пайдалана отырып, осындай жалған ақпаратты жасау және тарату процесін білдіреді. ЖИ дезинформацияны жасау мен таратуды автоматтандырып, оның ауқымын кеңейтеді және дәстүрлі әдістермен салыстырғанда оны анықтауды айтарлықтай қиындатады.

ЖИ тудырған дезинформацияның негізгі қаупі оның бұрын-соңды болмаған көлемінде, таралу жылдамдығында және шынайылық деңгейінде жатыр. Бұрын жалған ақпаратты жасау және тарату белгілі бір ресурстарды, уақытты және арнайы дағдыларды қажет ететін болса, қазіргі ЖИ құралдары, мысалы, ChatGPT сияқты үлкен тілдік модельдер (LLM), бағдарламалау дағдылары жоқ адамдарға да аз шығынмен мыңдаған, тіпті миллиондаған жалған хабарлама нұсқаларын лезде жасауға мүмкіндік береді. Сол сияқты, дипфейктерді (deepfakes) – фото, видео және аудио манипуляцияларды жасау бұрын техникалық білім мен қымбат жабдықты талап етсе, қазір тегін немесе арзан онлайн қызметтер арқылы қолжетімді бола бастады. Дезинформация жасау процесінің бұлайша "демократизациялануы" зиянкестердің санын арттырып, олардың мүмкіндіктерін кеңейтеді, бұл өз кезегінде ақпараттық кеңістіктің ластануын күшейтеді.

Бұл мәселе тек саяси алаңмен немесе қоғамдық пікірмен шектелмейді. ЖИ тудырған дезинформация академиялық ортаға да елеулі экономикалық және репутациялық қауіп төндіреді.Академиялық институттардың негізгі миссиясы – ақиқатты іздеу және білімді тарату болғандықтан, жалған ақпараттың таралуы олардың іргетасына нұқсан келтіреді. Институттар ЖИ дезинформациясын анықтайтын қымбат құралдарға және қызметкерлер мен студенттерді осы қауіпке қарсы тұруға үйретуге қомақты қаржы бөлуге мәжбүр. Бұл ресурстар оқыту, зерттеу және қоғамға қызмет ету сияқты негізгі академиялық қызметтерден алынады. Сонымен қатар, егер академиялық мекеменің аты жалған немесе бұрмаланған ақпаратпен байланыстырылса, бұл оның беделіне нұқсан келтіріп, қоғамдық сенімді жоғалтуға, студенттерді тартуда қиындықтарға, қаржыландырудың азаюына және стратегиялық серіктестіктердің бұзылуына әкелуі мүмкін. Бұл академиялық қауымдастықтың тиімді жұмыс істеуі және зерттеулердің қоғамға оң әсер етуі үшін қажетті іргелі сенімді бұзады.

**ЖИ-дің дезинформацияны жасау және таратудағы рөлі**

Жасанды интеллект дезинформацияны жасаудың және таратудың бірнеше негізгі механизмдерін күшейтеді:

● Дипфейктер (Deepfakes): Бұл ЖИ, әсіресе терең оқыту (deep learning) және нейрондық желілер арқылы жасалған, адамның бет-әлпетін, денесін немесе дауысын өте шынайы түрде өзгертетін немесе толығымен жасайтын медиа контент. Дипфейктер адамдардың ешқашан айтпаған сөздерін айтқызып, жасамаған іс-әрекеттерін жасағандай етіп көрсете алады. Бұл технология көбінесе Генеративті Жарыспалы Желілерге (Generative Adversarial Networks - GANs) негізделген, мұнда бір желі (генератор) жалған контент жасауға тырысса, екінші желі (дискриминатор) оны шынайыдан ажыратуға тырысады, осылайша генератордың сапасы артады.8 Дипфейктер саясаткерлерді (мысалы, АҚШ президенті Джо Байденнің дауысымен жасалған жалған телефон қоңырауы , Дональд Трамптың , Молдова президенті Майя Сандудың , Камала Харристің 10 жалған бейнелері), атақты адамдарды (мысалы, Тейлор Свифттің 6, Том Круздың жалған бейнелері), жаңалықтар жүргізушілерін және басқа да қоғам қайраткерлерін нысанаға алуы мүмкін. Дауысты клондау (Voice cloning) технологиясы да осы санатқа жатады және адамдарды алдау, қаржылық алаяқтық жасау немесе сенімді бұзу үшін қолданылуы мүмкін. Мысалы, басшысының дауысымен қоңырау шалған алаяқтарға ақша аударған компания басшысы немесе ұлының дауысымен көмек сұраған қоңырауға сеніп қалған ата-ана туралы оқиғалар бар.

● Автоматтандырылған контент генерациясы: Үлкен тілдік модельдер (LLM) сияқты ЖИ жүйелері адам жазғандай көрінетін мәтіндерді – жаңалықтар мақалаларын, әлеуметтік желідегі жазбаларды, электрондық хаттарды, тіпті академиялық жұмыстарды жаппай және жылдам жасай алады.Бұл мүмкіндік жалған жаңалық сайттарын құру, спам тарату, фишингтік шабуылдар жасау және саяси пропаганданы кеңінен тарату үшін белсенді қолданылуда. Мысалы, Венесуэланың мемлекеттік БАҚ-тары жоқ халықаралық арнаның ЖИ арқылы жасалған жүргізушілері бар видеоларды про-үкіметтік хабарламалар тарату үшін пайдаланған. ЖИ сондай-ақ автоматты түрде суреттерді 6, музыканы және басқа да медиа түрлерін жасай алады. Бұл технологияның қолжетімділігі (көбінесе тегін немесе арзан) кез келген адамға аз күш жұмсап, үлкен көлемде контент жасауға мүмкіндік береді.

● Мақсатты кампаниялар (Targeted Campaigns): ЖИ үлкен деректер жиындарын (пайдаланушылардың онлайн мінез-құлқы, демографиялық мәліметтер, қызығушылықтар) талдау арқылы дезинформациялық хабарламаларды белгілі бір топтарға немесе тіпті жеке тұлғаларға бағыттап, олардың сенімдеріне, қорқыныштарына немесе осалдықтарына әсер ететіндей етіп бейімдей алады. Бұл "микро-таргетинг" немесе "гипер-персонализация" деп аталады және әлеуметтік медиа платформаларының алгоритмдерімен және автоматтандырылған бот желілерімен үйлескенде өте қауіпті болуы мүмкін. ЖИ осылайша сайлау процестеріне араласу, қоғамдағы әлеуметтік жіктерді тереңдету , институттарға деген сенімді бұзу және тіпті зорлық-зомбылықты өршіту үшін қолданылуы мүмкін.

ЖИ-дің дезинформацияны таратудағы рөлі оны жасаумен ғана шектелмейді. Көптеген әлеуметтік медиа платформалары контентті сұрыптау және ұсыну үшін ЖИ-ге негізделген алгоритмдерді пайдаланады. Бұл алгоритмдер кейде пайдаланушылардың назарын аударатын, бірақ жалған немесе зиянды контенттің вирустық таралуын байқаусызда тездетуі мүмкін. Сонымен қатар, ЖИ арқылы басқарылатын боттар (автоматтандырылған аккаунттар) белгілі бір хабарламаларды немесе хэштегтерді жасанды түрде күшейтіп, олардың танымалдылығын арттырып, шынайы пікірталас ретінде көрсету үшін қолданылады. Осылайша, ЖИ дезинформацияның жасалуынан бастап, оның кең аудиторияға жетуіне дейінгі бүкіл тізбекті күшейтеді.

Бұл процестердің тағы бір қауіпті салдары – "өтірікшінің дивиденді" (Liar's Dividend) феномені. ЖИ арқылы жасалған контенттің шынайылығы соншалықты жоғары болғандықтан, адамдар кез келген күмәнді немесе өздеріне ұнамайтын ақпаратты, тіпті ол шындық болса да, "дипфейк" немесе "ЖИ жасаған" деп оңай жоққа шығара алады. Бұл жалпы ақпараттық экожүйеге, БАҚ-қа, сарапшыларға және ресми дереккөздерге деген сенімді одан әрі бұзады. Нәтижесінде, қоғам шындық пен жалғандықты ажыратуда қиналатын "пост-шындық" дәуіріне тереңдей түседі. Бұл дезинформациямен күресу стратегияларын әзірлеуде тек жалғандықты анықтау ғана емес, сонымен қатар шынайы ақпараттың сенімділігін қорғау қажеттілігін де тудырады.

**ЖИ арқылы жасалған дезинформацияны техникалық анықтау**

ЖИ арқылы жасалған дезинформацияның таралуына қарсы тұрудың маңызды бағыттарының бірі – осындай жалған контентті техникалық құралдармен анықтау. Бұл әдістер контенттің түріне (бейне, аудио, мәтін, бот белсенділігі) байланысты әртүрлі болады.

**Дипфейктерді анықтау әдістері**

**Дипфейктерді** **анықтау** – белсенді зерттеу саласы болып табылады және бірнеше негізгі тәсілдерді қамтиды:

● Артефакттарға негізделген әдістер: Бұл тәсіл ЖИ арқылы жасалған бейнелер мен аудиоларда жиі кездесетін, адам көзіне немесе құлағына байқала бермейтін, бірақ техникалық талдау арқылы анықталатын сәйкессіздіктерді немесе "артефакттарды" іздеуге негізделген. Оларға мыналар жатуы мүмкін: бет қимылдарының (әсіресе көздің жыпылықтауы, еріннің қимылы) табиғи еместігі, тері текстурасының тым тегіс немесе бұлыңғыр болуы, жарық пен көлеңкенің физикалық заңдылықтарға сәйкес келмеуі, фонның бұрмалануы немесе "дірілдеуі", көзілдіріктегі шағылысудың дұрыс болмауы, шаш немесе тіс сияқты ұсақ бөлшектердің анық еместігі, аудиодағы табиғи емес интонация, дауыс тембрінің өзгеруі немесе фондық шудың сәйкессіздігі. Бұл әдістер генерация технологияларының ерте кезеңдерінде тиімді болғанымен, технология жетілген сайын артефакттар азайып, оларды анықтау қиындай түседі.

● Машиналық оқыту және терең оқыту: Қазіргі заманғы дипфейк детекторларының көпшілігі машиналық оқытуға, әсіресе терең оқытуға негізделген. Бұл модельдер шынайы және жалған медианың үлкен дерекқорларында үйретіледі. Олар контенттің көрінбейтін статистикалық заңдылықтарын, пиксель деңгейіндегі немесе жиілік доменіндегі ерекшеліктерді талдау арқылы жалғандықтың нәзік белгілерін анықтауға үйренеді. Бұл салада Конволюциялық нейрондық желілер (CNNs), Transformer негізіндегі архитектуралар (мысалы, Vision Transformer - ViT), CLIP сияқты іргелі модельдерді бейімдеу және басқа да күрделі желілер қолданылады. Бұл модельдер артефакттарға негізделген әдістерге қарағанда жалпылау қабілеті жоғары болуы мүмкін.

● Мультимодальды анықтау: Кейбір дипфейктер тек визуалды немесе тек аудио деректерді талдау арқылы анықталмауы мүмкін. Мысалы, видеодағы бет кескіні шынайы көрінгенімен, оның ерін қимылдары аудиодағы дыбыспен сәйкес келмеуі мүмкін (lip-sync inconsistency). Мультимодальды тәсілдер видео, аудио және кейде мәтіндік метадеректерді бірге талдап, осындай сәйкессіздіктерді анықтауға тырысады. Бұл тәсіл күрделі манипуляцияларды анықтауда тиімдірек болуы мүмкін.

● Жаңа тәсілдер: Генерация технологияларының дамуымен (мысалы, диффузиялық модельдер немесе Neural Radiance Fields - NeRFs) детекторлар да жаңа әдістерді қажет етеді. Зерттеушілер білімді енгізу (Knowledge Injection), қабатты басу және контрастты жоғалту (layer-wise suppression and contrast losses), диффузиялық модельдердің кері процесін талдау, әрекет бірліктеріне негізделген талдау (Action Unit-guided analysis) сияқты жаңа тәсілдерді ұсынуда.

Дипфейктерді анықтаудағы қиындықтар: Бұл саладағы басты қиындық – детекторлардың жалпылану қабілетінің төмендігі. Яғни, белгілі бір генерация әдісімен жасалған фейктерді анықтауға үйретілген детектор жаңа, бұрын кездеспеген әдіспен жасалған фейктерді анықтауда тиімсіз болуы мүмкін. Сонымен қатар, шынайы өмірдегі деректердің ("in-the-wild") күрделілігі – бейне сапасының төмендігі, сығылу артефакттары, әртүрлі жарықтандыру жағдайлары – детекторлардың жұмысын қиындатады. Пост-өңдеу техникалары, мысалы, видеоның сапасын жақсарту (super-resolution), детекторлардың тиімділігін айтарлықтай төмендетуі мүмкін. Сондай-ақ, зиянкестер детекторларды алдау үшін арнайы қарсылас шабуылдар (adversarial attacks) жасай алады. Нәтижесінде, академиялық зерттеулерде жоғары дәлдік көрсететін көптеген детекторлар шынайы өмірде қолданғанда әлдеқайда нашар нәтиже көрсетуі мүмкін.

Бұл жағдай дипфейктерді анықтаудың үздіксіз "мысық пен тышқан" ойыны екенін көрсетеді. Генерация технологиялары жетілген сайын, ескі детекторлар тиімсіз болып қалады. Сондықтан, статикалық шешімдер жеткіліксіз; үнемі жаңартылып, жаңа қауіптерге бейімделе алатын динамикалық жүйелер қажет. Сонымен қатар, зертханалық жағдайдағы жоғары нәтижелерге алданып қалмау керек. Детекторлардың шынайы өмірдегі тиімділігінің шектеулі болуы мүмкін екендігі саясаткерлер мен платформалардың тек технологиялық шешімдерге ғана сенбей, медиа сауаттылық, реттеу және басқа да стратегияларды қатар дамытуы керектігін көрсетеді.

**Синтетикалық мәтінді анықтау құралдары**

ЖИ арқылы жасалған мәтіндерді (мысалы, LLM шығарған) анықтау да маңызды міндет болып табылады, бірақ оның өзіндік қиындықтары бар. Негізгі тәсілдер:

● Статистикалық және лингвистикалық талдау: Бұл әдістер адам жазған және машина жазған мәтіндер арасындағы статистикалық немесе стильдік айырмашылықтарды іздейді. Олар мәтіннің күрделілігін (мысалы, оқылу жеңілдігі индекстері), сөздік қорының байлығын немесе шектеулілігін, белгілі бір сөздердің немесе фразалардың қайталану жиілігін, сөйлем ұзындығының вариациясын, тыныс белгілерін қолдану ерекшеліктерін және басқа да лингвистикалық белгілерді талдайды. Кейбір зерттеулер LLM-дердің мәтін генерациялау кезінде белгілі бір ықтималдық үлестірілімдерін ұстанатынын және осы заңдылықтарды анықтауға болатынын көрсетеді.

● Модельге негізделген әдістер: Бұл тәсілдер мәтінді жіктеу үшін нейрондық желілерді пайдаланады. Көбінесе BERT, RoBERTa, DeBERTa сияқты алдын ала үйретілген тілдік модельдерге негізделген жіктегіштер қолданылады. Басқа бір қызықты тәсіл – RAIDAR деп аталатын әдіс, мұнда күдікті мәтін басқа бір LLM арқылы қайта жазылады (парафразаланады), содан кейін бастапқы және қайта жазылған мәтіндер арасындағы айырмашылықтар талданады. Сондай-ақ, LLM-нің өзінен берілген мәтінді генерациялау ықтималдығын сұрау арқылы (zero-shot detection) оның ЖИ арқылы жасалғанын бағалауға болады.

● Су таңбалау (Watermarking): Бұл – мәтінді генерациялау процесі кезінде оған байқалмайтын сандық із немесе "су таңбасын" енгізу әдісі. Бұл белгі кейіннен мәтіннің ЖИ арқылы жасалғанын және қай модельмен жасалғанын дәлелдеуге мүмкіндік береді. Су таңбасы сөздерді таңдау ықтималдығына, тыныс белгілерінің қолданылуына немесе басқа да лингвистикалық ерекшеліктерге енгізілуі мүмкін.

Синтетикалық мәтінді анықтаудағы қиындықтар: LLM технологияларының үздіксіз дамуы олардың жасайтын мәтіндерін адам жазғаннан ажыратуды барған сайын қиындатуда. Көптеген детекторлар белгілі бір модельге немесе деректер жиынтығына бейімделгендіктен, олардың жаңа модельдерге, басқа тілдерге немесе әртүрлі тақырыптық домендерге жалпылануы төмен болуы мүмкін. Зиянкестер мәтінді парафразалау немесе басқа да қарсылас шабуылдар арқылы детекторларды оңай алдай алады. Сонымен қатар, детекторлардың қате оң нәтиже (false positive) беру қаупі жоғары, яғни адам жазған мәтінді ЖИ жазған деп белгілеуі мүмкін.51 Бұл әсіресе академиялық ортада немесе журналистикада елеулі салдарға әкелуі мүмкін. Жалпы алғанда, қазіргі қолда бар мәтін анықтау құралдарының дәлдігі мен сенімділігі әлі де жеткіліксіз деңгейде.

Мәтінді анықтаудың дипфейктерді анықтаудан да күрделі болуының себебі, мәтінде визуалды немесе аудио артефакттар сияқты айқын "іздер" болмайды. Анықтау көбінесе статистикалық немесе стильдік ерекшеліктерге сүйенеді, ал заманауи LLM-дер бұл жағынан адам жазуына өте жақын, тіпті кейде асып түсетіндей деңгейге жетті. LLM-дер жақсарған сайын, олардың туындыларын анықтау да қиындай береді. Су таңбалау техникалық тұрғыдан перспективалы шешім болып көрінгенімен, оның кеңінен енгізілуі барлық негізгі LLM әзірлеушілерінің ынтымақтастығын және ортақ стандартты қабылдауын талап етеді, бұл бәсекелестік жағдайында қиынға соғуы мүмкін. Сонымен қатар, су таңбаларын парафразалау немесе басқа да өңдеу әдістері арқылы жоюға немесе айналып өтуге болады, бұл оның практикалық тиімділігін шектейді.

2.3. Автоматтандырылған бот белсенділігін анықтау

Әлеуметтік медиада дезинформацияны таратуда автоматтандырылған аккаунттар немесе боттар маңызды рөл атқарады. Оларды анықтау үшін келесі әдістер қолданылады:

● Ерекшеліктерге негізделген әдістер (Feature-based): Бұл ең кең таралған тәсілдердің бірі. Ол боттар мен адамдардың мінез-құлқындағы айырмашылықтарды көрсететін әртүрлі ерекшеліктерді талдауға негізделген. Бұл ерекшеліктер бірнеше санатқа бөлінеді:

○ Аккаунт метадеректері: Аккаунттың жасы, жазылушылар мен жазылымдар санының арақатынасы, профиль суретінің болуы/болмауы, профиль сипаттамасының толықтығы, верификация статусы.

○ Контент ерекшеліктері: Жазбалардың мазмұнын табиғи тілді өңдеу (NLP) арқылы талдау, қайталанатын немесе шаблонды хабарламалар, хэштегтерді немесе сілтемелерді шамадан тыс пайдалану, жазбалардың тональдылығы (sentiment analysis).

○ Уақыттық белсенділік: Жариялау жиілігінің өте жоғары немесе тым тұрақты болуы (периодтылық), белсенділік уақыты (мысалы, тәулік бойы белсенді болу), басқа аккаунттармен үйлестірілген әрекеттер (мысалы, бір уақытта бірдей контентті тарату).

○ Желілік құрылым: Аккаунттың әлеуметтік графтағы орны, оның жазылушылары мен жазылымдарының сипаты (олар да бот па?), басқа аккаунттармен өзара әрекеттесу үлгілері (лайк, репост, пікір алмасу).

● Графқа негізделген әдістер: Бұл әдістер жеке аккаунттарды емес, олардың арасындағы байланыстарды талдауға бағытталған. Боттар көбінесе топтасып жұмыс істейтіндіктен (ботнеттер), олардың арасында тығыз байланыстар немесе ерекше граф құрылымдары (мысалы, "жұлдыз" немесе тығыз кластерлер) байқалуы мүмкін. Бұл әдістер үйлестірілген зиянды әрекеттерді анықтауда тиімдірек болуы мүмкін.

● Кластерлеу және қалыптан тыс жағдайларды анықтау: Бұл тәсілдер пайдаланушылардың мінез-құлқын ұқсастықтары бойынша топтастырып, қалыпты мінез-құлық үлгілерінен айтарлықтай ауытқитын аккаунттарды немесе топтарды "қалыптан тыс" немесе "күдікті" деп белгілейді. Бұл әдістер алдын ала белгіленген деректерді қажет етпеуі мүмкін (unsupervised).

● Адам мен боттың айырмашылықтарын талдау: Кейбір зерттеулер боттар мен адамдардың коммуникация стиліндегі айырмашылықтарға назар аударады. Мысалы, боттар көбінесе автоматтандыруға оңай лингвистикалық белгілерді (көп хэштег, позитивті немесе негативті терминдерді жиі қолдану) пайдаланса, адамдар диалогты түсінуді және контекстке жауап беруді қажет ететін әрекеттерді (мысалы, пікірлерге жауап беру, талқылауға қатысу) жиірек жасайды. Сондай-ақ, олардың әлеуметтік желідегі өзара әрекеттесу құрылымдарында да айырмашылықтар болуы мүмкін.

Боттарды анықтаудағы қиындықтар: Боттар үнемі жетілдіріліп, адам мінез-құлқын жақсырақ имитациялайтын болып келеді, бұл оларды анықтауды қиындатады. Кейбір күрделі боттарды адамдардың өздері де ажырата алмайды. Боттарды анықтауға арналған дерекқорлардың сапасы, толықтығы және әртүрлілігі де маңызды мәселе болып табылады; бір дерекқорда үйретілген модель басқа жағдайда тиімсіз болуы мүмкін. "Бот" ұғымының өзі әртүрлі түсіндіріледі, және барлық автоматтандырылған аккаунттар зиянды емес. Botometer сияқты танымал құралдар болғанымен, олардың нақты жағдайлардағы дәлдігі мен сенімділігіне қатысты күмәндар бар. Ең басты қиындықтардың бірі – боттардың үнемі эволюциясы; детекторлар жаңа буын боттарына қарсы тұру үшін үнемі жаңартылып отыруы керек.

Боттарды анықтау саласындағы маңызды түйін – бұл тек техникалық мәселе емес, сонымен қатар концептуалды мәселе екендігінде. "Бот" дегеніміз не және оның қандай түрі зиянды?. Кейбір автоматтандырылған аккаунттар жаңалықтарды тарату, ауа райын болжау немесе клиенттерге қызмет көрсету сияқты пайдалы функцияларды орындайды. Сондықтан, тек аккаунттың автоматтандырылған екенін анықтау жеткіліксіз. Оның ниетін, мінез-құлқын және экожүйеге әсерін бағалау маңыздырақ болады. Бұл тек техникалық анықтаудан гөрі, күрделірек, контекстке негізделген талдауды қажет етеді.

Осыған байланысты, боттарды анықтаудағы ең перспективалы бағыттардың бірі – жекелеген аккаунттардың ерекшеліктеріне тым көп сүйенуден гөрі, үйлестірілген әрекеттерді (coordinated activity) және топтық мінез-құлықты анықтауға көшу болып табылады. Жекелеген күрделі боттар адамдардан айнымауы мүмкін, бірақ олардың жаппай, синхронды және үйлестірілген әрекеттері (мысалы, бірнеше аккаунттың бір уақытта бірдей контентті таратуы немесе бір-біріне жазылуы) көбінесе анықтауға болатын із қалдырады. Бұл тәсілдер (group-based detection) боттардың эволюциясына төзімдірек болуы мүмкін және болашақ зерттеулер мен әзірлемелер үшін негізгі бағыт болып саналады.