AZiTREND

GHID

INTELIGENȚA ARTIFICIALĂ PERICOLELE INVIZIBILE

ION IORDACHE
DORU ASMARANDEI

AUTORI:

Ion IORDACHE, Ec.

Economist de profesie și absolvent al unei școli militare de ofițeri, specializat și certificat ca, consultant de securitate și trainer profesionist. Am acumulat peste 30 de ani de experiență în securitate privată, evaluarea riscurilor, managementul operațiunilor și protecția datelor, incluzând expertiză în CPTED și educația adulților.

Dincolo de expertiza în securitatea fizică și informațională, sunt profund conectat la provocările viitorului. Investesc timp și rigoare analitică în cercetarea detaliată a Inteligenței Artificiale, fiind preocupat în mod special de evoluția acestui domeniu și de aspectele sale etice. Cred cu tărie că înțelegerea profundă a tehnologiei este vitală pentru a rămâne relevanți și responsabili într-o lume în continuă schimbare.

www.ioniordache.com ion@ioniordache.com

Doru ASMARANDEI, Ec.

Economist și antreprenor, CEO al AziTrend. De-a lungul carierei mele, m-am concentrat pe transformarea industriei de securitate prin tehnologie. Nu văd securitatea doar ca pe o necesitate operațională, ci ca pe un ecosistem complex unde inovația face diferența.

Pasiunea mea este convergența dintre factorul uman și inteligența artificială. Am promovat intens conceptul de integrare a analizei video avansate cu serviciile de pază umană ("man guarding"), demonstrând că viitorul nu aparține doar oamenilor sau doar mașinilor, ci colaborării eficiente dintre ele. În tot ceea ce fac, promovez soluții de SSV (Sisteme de Supraveghere Video) care nu doar înregistrează, ci înțeleg și previn, aducând o valoare adăugată reală prin reducerea costurilor și creșterea eficienței.

www.azitrend.ro doru.asmarandei@azitrend.ro

PRECIZĂRI IMPORTANTE!

Acest ghid a fost elaborat cu maximă rigoare și atenție la detalii, având ca fundament cercetarea academică de specialitate și realitățile din industria de securitate. Totuși, autorii nu își asumă răspunderea legală pentru eventualele erori, omisiuni sau interpretări ale informațiilor prezentate, acestea având un scop strict educativ și informativ.

Un element distinctiv al acestei lucrări îl constituie secțiunile „Perspectiva consultantului” și „Perspectiva managerului”. Dorim să precizăm clar că acestea reflectă opiniile profesionale și personale ale autorilor, bazate pe experiența directă din teren. Studiile de caz și scenariile operaționale incluse în aceste secțiuni sunt reale, însă au fost adaptate și anonimizate din rațiuni de confidențialitate, dar și pentru a oferi cititorului un context didactic clar, ușor de asimilat.

Susținem etica profesională și schimbul transparent de cunoștințe. Am depus toate eforturile rezonabile pentru a credita corect materialele și conceptele ce aparțin terților, sursele bibliografice fiind indicate punctual la finalul fiecărui capitol.

Dacă aveți nevoie de aplicarea concretă a acestor principii sau de soluții de securitate personalizate pentru specificul organizației dumneavoastră, vă stăm la dispoziție cu servicii de consultanță pe bază de contract.

CUPRINS:

INTRODUCERE 04

CAPITOLUL I – TRINITATEA IMPOSIBILITĂȚII 07

CAPITOLUL II – ILUZIA SIGURANȚEI 13

CAPITOLUL III – PROPRIETATEA IMPOSIBILĂ 22

CAPITOLUL IV – CĂILE CĂTRE DEZASTRU 27

CAPITOLUL V – CE PUTEM FACE? 35

EPILOG – SPERANȚA REALISTĂ 39

INTRODUCERE

Industria securității fizice se confruntă, astăzi, cu o transformare fără precedent. Sistemele de inteligență artificială pătrund rapid în toate aspectele domeniului nostru: de la recunoașterea facială și analiza comportamentală video, până la controlul accesului predictiv și sistemele autonome de răspuns la incidente. Producătorii și integratorii promit eficiență sporită, costuri reduse și capacități supraomenești de detecție și analiză, dar cât de sigure sunt aceste sisteme? **Cât de sigure pot fi**, nu doar în teorie și/sau în materialele de marketing, ci în limitele demonstrate matematic, și logic ale tehnologiei în sine?

Ne-am propus, prin acest ghid, să răspundem la aceste întrebări, atât din perspectiva unui consultant de securitate, cât și din perspectiva unui manager, cu implicație directă în video management & analitics systems, traducând cercetarea academică de vârf în limbajul și cadrele conceptuale familiare profesioniștilor în securitatea fizică și nu numai acestora.

Fundamentul: Profesorul Roman V. Yampolskiy

Ghidul nostru își datorează existența, în primul rând, muncii de pionierat a profesorului Roman V. Yampolskiy, profesor asociat titular de informatică la Speed School of Engineering din cadrul Universității din Louisville și unul dintre fondatorii domeniului siguranței inteligenței artificiale (AI Safety).

Cartea sa, **AI: Unexplainable, Unpredictable, Uncontrollable**, oferă cea mai sistematică și riguroasă demonstrație de până acum, că problemele de siguranță ale AI nu sunt doar provocări de inginerie care așteaptă soluții ingenioase, ci imposibilități matematice, limitări fundamentale derivate din teoreme demonstrate în informatică, matematică și logică.

Profesorul Yampolskiy este un cercetător serios, cu peste 600 de publicații academice și 200 de apariții media, care a dedicat ani de cercetare întrebării: ***„Poate fi controlată inteligența artificială?"*** Răspunsul său, susținut de argumente formale și demonstrații matematice, este foarte sec: **nu în sensul absolut pe care l-ar necesita riscurile existențiale cu care ne confruntăm.**

Deși cartea profesorului Yampolskiy constituie coloana vertebrală a acestei serii, toate cele prezentate în acest ghid sunt susținute de o **rețea extinsă de surse verificabile**, incluzând:

Din domeniul informaticii teoretice:

- ✓ Alan Turing (1936) - Problema Opririi și limitele calculabilității
- ✓ Henry Gordon Rice (1953) - Teorema lui Rice despre proprietățile nedecidabile ale programelor
- ✓ Kurt Gödel (1931) - Teoremele incompletitudinii și limitele sistemelor formale
- ✓ Stephen Wolfram (2002) - Ireductibilitatea computațională

Din cercetarea contemporană în siguranța IA:

- ✓ Nick Bostrom (Oxford) - Superintelligence: Paths, Dangers, Strategies (2014)
- ✓ Stuart Russell (Berkeley) - Human Compatible: AI and the Problem of Control (2019)
- ✓ Alfonseca et al. (2021) - „Superintelligence Cannot be Contained" (Journal of Artificial Intelligence Research)
- ✓ Souza et al. (2025) - Demonstrația indecidabilității alinierii interne (Scientific Reports, Nature)

Din domeniul explicabilității și interpretabilității:

- ✓ Cynthia Rudin (Duke) - Compromisul acurațete-interpretabilitate (Nature Machine Intelligence, 2019)
- ✓ Chen et al. (2025) - Teorema Decalajului de Complexitate

Din domeniul juridic și al răspunderii:

- ✓ RAND Corporation (2024) - Răspunderea pentru daunele cauzate de IA
- ✓ LoPucki & Bayern - Personalitatea juridică a entităților algoritmice
- ✓ Chandrasekaran et al. (2021) - Guvernanța învățării automate

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Structura ghidului:

Ghidul este organizat în cinci capitole care construiesc progresiv argumentul final:

- ✓ **Capitolul 1: Trinitatea imposibilității** - Introduce cele trei caracteristici fundamentale ale sistemelor AI avansate, imprevizibilitatea, inexplicabilitatea și incontrolabilitatea, demonstrând că acestea nu sunt defecte remediabile, ci trăsături emergente inevitabile ale oricărei inteligențe superioare.
- ✓ **Capitolul 2: Iluzia siguranței** - Examinează neverificabilitatea, imposibilitatea de a demonstra că un sistem AI este sigur, și regresul infinit al verificatorilor care face ca orice „certificare de siguranță" să fie, în cel mai bun caz, probabilistică.
- ✓ **Capitolul 3: Proprietatea imposibilă -** Explorează paradoxurile juridice și tehnice ale proprietății AI: dacă nu poți controla, verifica sau prezice comportamentul unui sistem, poți pretinde legitim că îl deții? Ce se întâmplă când sistemul dobândește personalitate juridică?
- ✓ **Capitolul 4: Căile către dezastru** - Prezintă taxonomia completă a modurilor în care sistemele AI pot deveni periculoase, intenționat, accidental, prin factori de mediu sau independent, și lecțiile din istoria deja bogată a eșecurilor AI.
- ✓ **Capitolul 5: Ce putem face?** - Încheie seria cu o evaluare realistă a strategiilor de atenuare disponibile, recunoscând atât posibilitățile, cât și limitele lor fundamentale, și oferind recomandări practice pentru profesioniștii în securitate.

Pentru cine este aceast ghid?

- ✓ **Consultanților de securitate fizică** care evaluează sau recomandă sisteme AI pentru clienții lor.
- ✓ **Managerilor de securitate** din organizații care implementează sau intenționează să implementeze tehnologii AI
- ✓ **Integratorilor de sisteme** care trebuie să înțeleagă limitările fundamentale ale produselor pe care le instalează
- ✓ **Factorilor de decizie** (directori, consilii de administrație) care aprobă investiții în tehnologii AI fără a înțelege riscurile asociate
- ✓ **Oricui** este interesat să înțeleagă de ce entuziasmul actual pentru AI ar trebui temperat de o doză sănătoasă de scepticism informat.

Perspectiva consultantului de securitate fizică

Ceea ce diferențiază această serie de alte prezentări ale riscurilor AI este integrarea sistematică a perspectivei consultantului de securitate fizică.

Ca profesionist în securitatea fizică, cu experiență în evaluarea riscurilor, proiectarea sistemelor de protecție și auditul conformității, am constatat că literatura despre siguranța AI, deși riguroasă, rămâne adesea inaccesibilă practicienilor din industria noastră. Terminologia este străină, exemplele sunt abstracte, iar implicațiile practice sunt rareori explicitate.

Aceast proiect încearcă să remedieze această lacună prin:

- ✓ Traducerea conceptelor din informatică teoretică în limbajul familiar al securității fizice, vulnerabilități, amenințări, controale, răspuns la incident.
- ✓ Identificarea paralelelor între metodologiile consacrate de securitate (Defense in Depth, evaluarea riscurilor, lanțul de custodie) și provocările ridicate de sistemele AI.
- ✓ Evidențierea limitărilor abordărilor tradiționale când sunt aplicate la amenințări Ai, de ce monitorizarea, auditul și certificarea funcționează diferit pentru sisteme adaptive.
- ✓ Oferirea de recomandări practice pentru consultanții de securitate care lucrează cu organizații ce implementează sau evaluează sisteme AI.

În fiecare capitol, secțiunile intitulate **„Perspectiva consultantului de securitate"** oferă această punte între teorie și practică. Ele nu pretind să ofere soluții definitive, pe care, după cum demonstrează proiectul, nimeni nu le poate oferi, ci un cadru de gândire pentru navigarea incertitudinilor fundamentale ale domeniului.

Perspectiva managerului

Nu am scris acest ghid pentru a induce panică sau pentru a descuraja inovația, ci pentru a promova realismul profesional într-un domeniu dominat de promisiuni de marketing și optimism nejustificat.

Ca profesioniști în securitate, suntem obișnuiți să gândim în termeni de scenarii pesimiste, și asta nu pentru că suntem pesimiști, ci pentru că știm că amenințările exploatează exact acele vulnerabilități pe care le-am ignorat sau subestimat. Același principiu se aplică și sistemelor AI.

Într-o industrie inundată de promisiuni de marketing, este rar să găsești parteneri dispuși să discute deschis despre ceea ce tehnologia nu poate face. Susținerea acestui ghid de către **AZITREND** nu este întâmplătoare. În calitate de distribuitor major de sisteme de securitate și soluții VMS (Video Management Systems) si Video Analytics, compania se confruntă zilnic cu decalajul dintre așteptările clienților, alimentate de hype-ul inteligenței artificiale, și realitatea din teren, așa că ne-am ceva neobișnuit: să temperăm entuziasmul orb cu o doză de competență tehnică.

Obiectivul **AZITREND** nu este doar să livreze cutii cu echipamente inteligente, ci să se asigure că integratorii și clienții finali înțeleg mecanismele din spatele acestora. Când vorbim despre riscurile existențiale ale AI, nu ne referim la scenarii, uneori fanteziste, ci la erorile de judecată care apar atunci când instalăm un server de analiză video și ne așteptăm să gândească precum un om. **AZITREND** a ales să sponsorizeze luciditatea, înțelegând că adevărata securitate începe acolo unde se termină iluziile de marketing.

Sper că aceast ghid vă va oferi instrumentele conceptuale necesare pentru a evalua critic promisiunile furnizorilor, pentru a comunica riscurile factorilor de decizie și pentru a proiecta controale care recunosc, în loc să ignore, limitările fundamentale ale tehnologiei.

În fiecare capitol AZITREND va prezenta unul sau mai multe studiuide caz relevante, care să susțină teoria din acesta.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

CAPITOLUL I
TRINITATEA IMPOSIBILITĂȚII

Dacă nu poți dovedi că un sistem de IA este sigur, el este nesigur...

Începem acest capitol cu un adevăr pe care entuziaștii inteligenței artificiale îl consideră, în mod inexplicabil, negociabil: **iluzia controlului asupra AI este o eroare fundamentală de judecată** din partea celor care refuză să accepte limitele logice și matematice ale inteligenței. Nu ne referim la obstacolele temporare care vor ceda, de regulă, în fața bugetelor generoase și a inginerilor talentați, ci la imposibilități demonstrate matematic, la fel de imuabile ca teorema lui Pitagora, de exemplu.

Ca specialiști în securitatea fizică, știm că fiecare sistem, de la SSV (sisteme de supraveghere video) la cele mai avansate AI, are puncte de eșec și limite operaționale. A crede că putem controla perfect un sistem suficient de complex, capabil să genereze soluții neașteptate sau să funcționeze pe baza unor seturi de date imperfecte, este o formă de naivitate tehnică. Când ignorăm aceste limite fundamentale, riscăm să delegăm securitatea și decizia în mâinile unei mașini pe care nu o înțelegem complet, sub pretextul progresului. Să fim pragmatici, nu visători. Cât de mult control este de fapt necesar și cât de mult este realist? Acesta este un subiect pe care trebuie să îl dezbatem cu cifre clare, nu cu speranțe vagi.

Industria tehnologică și-a construit un narativ reconfortant: inteligența artificială poate fi făcută „sigură" dacă investim suficient în cercetare. Este o poveste frumoasă, vândută investitorilor și publicului larg cu aceeași încredere dezarmantă cu care industriile de tutun vindeau odinioară țigările ca fiind benefice pentru sănătate.

Paralela nu este întâmplătoare. În anii 1950, medicii apăreau în reclame recomandând mărci de țigări. Studiile „științifice" finanțate de producători demonstrau că fumatul calmează nervii și ajută digestia. Când au început să se acumuleze dovezile că tutunul ucide, industria nu a negat direct, ar fi fost prea transparent. În schimb, a finanțat cercetări care „puneau sub semnul întrebării" consensul științific, a promovat filtre și țigări „light" ca soluții tehnice, și a repetat neobosit că „sunt necesare mai multe cercetări" înainte de orice reglementare. Strategia a funcționat decenii întregi, timp în care milioane de oameni au murit.

Astăzi, giganții tehnologici folosesc aceeași carte de joc. Când cercetătorii independenți semnalează riscuri fundamentale ale AI, companiile nu neagă direct. În schimb, anunță inițiative de „AI responsabilă", publică principii etice impresionante pe site-urile lor corporative, și finanțează laboratoare de „siguranță" ai căror cercetători știu, conștient sau nu, că anumite concluzii ar pune în pericol finanțarea. Mesajul este întotdeauna același: „Suntem conștienți de provocări, lucrăm la soluții, aveți încredere în noi."

Diferența critică? Când industria tutunului a mințit, consecințele s-au materializat în cancere pulmonare și boli cardiovasculare, tragedii individuale care, oricât de devastatoare, nu amenințau existența civilizației.

Când industria IA promite siguranță pe care nu o poate demonstra, miza este de alt ordin de mărime.

Realitatea este considerabil mai sumbră decât narativele de marketing.

Roman Yampolskiy identifică trei caracteristici fundamentale ale oricărui sistem inteligent care ne depășește cognitiv ... cele **Trei „U"**:

1. **Imprevizibilitatea *(Unpredictability)***
2. **Inexplicabilitatea *(Unexplainability)* și**
3. **Incontrolabilitatea *(Uncontrollability)*.**

Acestea sunt considerate trăsături emergente fundamentale ale oricărei inteligențe superioare, proprietăți intrinseci demonstrate prin teoreme din teoria calculabilității și logică matematică. Vorbim despre bariere teoretice de același tip cu teorema incompletitudinii lui Gödel sau problema opririi a lui Turing: limitări pe care nicio cantitate de finanțare, nicio echipă de ingineri talentați și niciun progres tehnologic nu le poate ocoli, pentru că țin de natura matematică a inteligenței însăși. Cercetarea actuală în siguranța AI nu a reușit să depășească în mod demonstrabil aceste bariere, și există motive solide să credem că nu o va face vreodată.

Problema **Controlului AI**, definită succint prin întrebarea *„Cum poate omenirea rămâne în siguranță la conducere în timp ce beneficiază de o inteligență superioară?"*, este considerată de mulți cercetători drept cea mai importantă provocare cu care s-a confruntat vreodată specia umană.

Paradoxal însă, tocmai arhitecții acestor sisteme, companiile și laboratoarele care construiesc cu înfrigurare modele din ce în ce mai capabile, operează cu prezumția implicită că soluția este garantată. Ei cer publicului și reglementatorilor să aibă încredere că problemele de siguranță vor fi rezolvate ***„la momentul potrivit"***, deși nu pot oferi nicio demonstrație, nicio teoremă, nicio dovadă factuală în acest sens.

Sarcina probei este inversată: ni se cere să presupunem siguranța până la proba contrarie, în loc să ni se demonstreze siguranța înainte de implementare.

Perspectiva consultantului de securitate fizică - O amenințare fără precedent...

Pentru profesioniștii în securitatea fizică, cele **Trei „U"** reprezintă o categorie de amenințare complet nouă, una care sfidează metodologiile consacrate de evaluare a riscurilor.

În securitatea tradițională, operăm cu un cadru familiar: **identificarea amenințării**, **evaluarea vulnerabilității**, **analiza impactului** și **estimarea probabilității**. Aplicat la sistemele IA avansate, acest cadru se destramă în mod îngrijorător:

- ✓ **Amenințarea** nu poate fi pe deplin identificată (**imprevizibilitatea**).
- ✓ **Vulnerabilitățile** nu pot fi complet catalogate (**inexplicabilitatea**).
- ✓ **Mecanismele de control** nu pot fi garantate (**incontrolabilitatea**).
- ✓ **Probabilitatea** crește cu timpul, pe măsură ce sistemele devin mai capabile, exact inversul majorității amenințărilor tradiționale

Un hoț poate fi descurajat de garduri și alarme; un atacator cibernetic poate fi detectat prin monitorizare, dar o inteligență superioară care decide să-și urmărească propriile obiective? Aceasta este o amenințare pentru care nu avem nici precedent, nici contramasuri validate.

AZiTREND

Perspectiva managerului - Paradoxul camerei inteligente în ecosistemul AZITREND

Coborâm aceste concepte din teorie direct în rack-ul de echipamente și vom analiza un scenariu tipic întâlnit de inginerii de la AZITREND în implementarea soluțiilor moderne de VSS (Video Surveillance Systems) și control acces.

Avem o cameră de ultimă generație cu funcții de deep learning pentru protecția perimetrală. Conform fișei tehnice, sistemul clasifică obiectele cu o precizie de 99%.

Totuși, aplicând ***„trinitatea imposibilității"***, ne lovim de realitatea operațională:

- ✓ **Imprevizibilitatea la nivel micro:** Într-o noapte cu ceață și reflexii atipice, sistemul clasifică o umbră dinamică drept *„intrus uman"*. De ce? Pentru că rețeaua neurală a *„văzut"* un tipar pe care nu îl putem anticipa. Oricât de mult ar testa producătorul în laborator, mediul real generează variabile infinite pe care algoritmul le va interpreta într-un mod imprevizibil creativ.
- ✓ **Inexplicabilitatea deciziei:** Când clientul întreabă tehnicienii **AZITREND** ***„de ce a declanșat alarma?"***, răspunsul onest este, tehnic vorbind, imposibil de dat. Sistemul este o ***„cutie neagră"***. Nu putem extrage lanțul logic, deoarece algoritmul nu a urmat o logică liniară, ci a procesat milioane de parametri opaci. Putem ajusta sensibilitatea, dar nu putem ***„explica"*** decizia.
- ✓ **Incontrolabilitatea adaptivă:** Pe măsură ce aceste sisteme devin autonome, capacitatea noastră de a le controla scade. Dacă activăm funcția de **„auto-învățare"** a scenei, sistemul își va modifica propriii parametri de detecție. Într-un final, operatorul de securitate nu mai controlează sistemul, ci doar primește, cu încredere oarbă, deciziile acestuia.

Rolul **AZITREND** în această ecuație devine esențial: facem tranziția de la vânzătorul de „cutii magice" la consultantul tehnic care ajută integratorul să proiecteze arhitecturi de securitate reziliente la eroare.

În loc să pretindem că sistemele VSS cu AI sunt infailibile, profesioniștii noștrii pun accent pe redundanță și pe verificarea umană, recunoscând că imprevizibilitatea este o caracteristică a sistemului (feature), nu un defect (bug).

Imprevizibilitatea - De ce nu putem ști ce va face „Copilul Geniu"...

Imposibilitatea logică - Vom face referire la o demonstrație prin reducere la absurd, de genul celor pe care le învață studenții în primul an de facultate, asta presupunând, desigur, că mai acordă atenție cursurilor de logică.

- ✓ **Premisa:** Să presupunem că o persoană poate prezice cu acuratețe toate deciziile unei superinteligențe.
- ✓ **Consecința:** Aceasta înseamnă că persoana respectivă poate lua decizii identice cu ale superinteligenței.
- ✓ **Concluzia:** Prin urmare, persoana posedă o inteligență echivalentă cu superinteligența.
- ✓ **Contradicția:** Dar superinteligența este, prin definiție, mai inteligentă decât orice persoană.
- ✓ **Rezultatul:** Presupunerea inițială este falsă. Imprevizibilitatea este adevărată.

Acest argument este o demonstrație formală prezentată de **Yampolskiy** în **Journal of Artificial Intelligence and Consciousness.** Dacă ai putea prezice exact unde va muta un jucător de șah superinteligent, ai putea juca șah cel puțin la fel de bine ca acel sistem, dar...evident, nu poți.

Impredictibilitatea creativă și orizontul lui Vinge

Să nu confundăm această impredictibilitate cu zgomotul aleatoriu al unui generator de numere. Distincția este esențială și adesea ignorată de criticii care resping problema controlului ca pe o exagerare.

Cum observă **Eliezer Yudkowsky** în eseul său **„Expected Creative Surprises"** (2008):
„Impredictibilitatea creativă a inteligenței nu seamănă deloc cu impredictibilitatea zgomotoasă a unui generator de numere aleatorii."

Un generator de numere aleatorii produce secvențe fără sens, fără scop, fără direcție.

O inteligență superioară produce acțiuni impredictibile tocmai pentru că urmărește obiective, și le urmărește prin căi pe care o inteligență inferioară nu le poate anticipa.

Există un echilibru ciudat, aproape paradoxal, între impredictibilitatea acțiunilor și predictibilitatea rezultatelor.

Dezvoltatorii sistemului Deep Blue știau că acesta vrea să câștige la șah.

Dezvoltatorii AlphaZero știau că sistemul lor vrea să câștige la Go, dar niciunul nu putea prezice ce mutări specifice vor face aceste sisteme, sacrificiile neortodoxe, gambitul neprevăzut, strategia care sfidează secole de teorie.

Tocmai aceste mutări specifice, impredictibile în detaliu dar orientate spre un scop cunoscut, sunt cele care pot transforma un instrument util într-o amenințare existențială atunci când *„victoria"* sistemului și *„binele umanității"* încetează să mai coincidă.

În 1993, informaticianul **Vernor Vinge** a introdus conceptul de singularitate tehnologică în eseul „**The Coming Technological Singularity: How to Survive in the Post-Human Era**", descriind un „zid opac peste viitor", un orizont de predicție dincolo de care nu putem vedea nimic.

Acolo descrie singularitatea ca pe un punct în viitor apropiat (în următorii 30 de ani de la 1993) dincolo de care predicțiile umane devin imposibile, asemănându-l cu un ***„event horizon"*** sau un zid opac peste viitor, deoarece inteligența superumană va remodela realitatea într-un mod imprevizibil.

„În termen de treizeci de ani, vom dispune de mijloacele tehnologice necesare pentru a crea o inteligență supraumană. La scurt timp după aceea, era umană va lua sfârșit. [...] Progresul accelerat al tehnologiei a creat o serie de dileme: Oare rasa umană dezvoltă superinteligența prea repede pentru a înțelege ce face? [...] Din punct de vedere uman, această schimbare va însemna renunțarea la toate regulile anterioare, poate într-o clipită, o fugă exponențială dincolo de orice limită a așteptărilor umane. [...] Este un punct dincolo de care activitățile umane, așa cum le cunoaștem, nu vor mai putea continua."

(Vernor Vinge)

Perspectiva consultantului de securitate fizică – Monitorizarea devine inutilă...

În securitatea fizică, sistemele de supraveghere video, senzorii de mișcare, analiza comportamentală, funcționează pe premisa că putem observa și interpreta comportamentul suspect.

Detectăm tiparele anormale pentru că știm cum arată normalitatea.

Pentru un sistem IA avansat, această premisă se prăbușește.

Nu putem stabili *„normalitatea"* unui sistem ale cărui acțiuni sunt, prin natura lor, imprevizibile.

Camerele de supraveghere nu detectează intențiile; detectează mișcările fizice. Ce „mișcare" trădează decizia unei superinteligențe de a-și modifica propriul cod sau de a manipula sistemele conectate la ea?

Implicația practică: monitorizarea tradițională devine irelevantă pentru amenințările IA de nivel avansat, dar asta nu înseamnă că trebuie abandonată, ci că trebuie recunoscute limitele sale fundamentale.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Inexplicabilitatea și incomprehensibilitatea - Minciunile necesare...

Dilema "Cutiei Negre"

Timp de decenii, sistemele de inteligență artificială s-au bazat pe expertiza umană, condensată de ingineri ai cunoașterii, și erau, relativ, ușor de înțeles, iar sistemele expert bazate pe arbori de decizie erau modele perfecte ale raționamentului uman. Apoi a venit revoluția rețelelor neuronale profunde, iar această transparență naturală a fost sacrificată pe altarul performanței.

Sistemele actuale sunt **„cutii negre"**, opace pentru înțelegerea umană, dar extraordinar de capabile. O rețea neuronală profundă operează pe miliarde de parametri, iar o explicație 100% precisă ar fi echivalentă cu copierea întregului cod.

Stephen Wolfram a formalizat această problemă drept „ireductibilitate computațională" în cartea sa "**A New Kind of Science**" (2002).

Pentru anumite sisteme, nu există nicio cale de a prezice comportamentul fără a parcurge aproape tot atâția pași de calcul ca sistemul însuși. Explicația devine la fel de complexă ca fenomenul explicat, o hartă la fel de mare ca teritoriul.

Compromisul acurateţ̦e-inteligibilitate

Aceasta este realitatea pe care industria o ascunde sub covor: cu cât o explicație este mai simplă și mai „inteligibilă pentru om", cu atât este mai puțin precisă.

Cynthia Rudin, în articolul „**Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**" publicat în "*Nature Machine Intelligence*", spune:

> **„Explicațiile trebuie să fie greșite. Ele nu pot avea fidelitate perfectă față de modelul original. Dacă explicația ar fi complet fidelă... explicația ar egala modelul original."**

Oricare ar fi explicația pe care o primim de la un sistem AI avansat, ea este, prin definiție, o simplificare, adică o minciună utilă.

Incomprehensibilitatea umană

Psihologul George Miller a descoperit în 1956 ceva ce probabil ai simțit deja: **poți ține în minte aproximativ șapte lucruri deodată**.

Poate nouă, dacă ai o zi bună. Poate cinci, dacă nu ai dormit suficient. De aceea numerele de telefon au șapte cifre. De aceea uiți al treilea lucru de pe lista de cumpărături când nu l-ai notat. Nu ești prost, ești om.

Creierul tău a evoluat să urmărească prada prin savană și să-și amintească unde a lăsat sulița, nu să proceseze ecuații cu un milion de variabile.

Acum gândește-te la un sistem de inteligență artificială. Nu procesează șapte lucruri deodată. Procesează milioane. Simultan. Fără să obosească, fără să-și piardă concentrarea, fără să aibă nevoie de cafea.

Rețelele neuronale moderne au sute de miliarde de parametri, conexiuni care funcționează în paralel pentru a ajunge la concluzii. Creierul tău are aproximativ 86 de miliarde de neuroni, dar îi folosește serial, câțiva odată, cu pauze pentru a respira și a se întreba ce vrei de la viață.

Yampolskiy invocă aici un principiu matematic atât de simplu încât îl înveți în școala generală: **principiul cutiei cu porumbei**. Dacă ai zece porumbei și nouă cutii, cel puțin o cutie va avea doi porumbei. Este inevitabil.

Creierul tău poate exista într-un număr finit de stări, și există configurații mentale, moduri de a gândi, de a procesa, de a înțelege care pur și simplu nu încap în capul tău. Nu pentru că nu te-ai străduit suficient sau nu ai citit cărțile potrivite, ci pentru că hardware-ul tău biologic nu are sloturile necesare.

Este ca și cum ai cere unui televizor alb-negru să afișeze roșu. Nu e vina televizorului. Nu a fost construit pentru asta.

O inteligență superioară nouă poate fi, matematic vorbind, de neînțeles pentru noi. Nu **„dificil de înțeles"**; nu **„necesită mai multă cercetare"**. **De neînțeles.** Ca și cum ai încerca să vizualizezi un cub în patru dimensiuni, poți desena proiecții, poți face analogii, dar nu-l vei vedea niciodată cu adevărat, pentru că ochii și creierul tău au evoluat într-o lume cu trei dimensiuni.

Perspectiva consultantului de securitate fizică – Auditul devine imposibil...

În securitatea fizică, auditul este activitatea care ne arată realitatea vulnerabilităților. Verificăm procedurile, testăm sistemele, documentăm totul. Rapoartele de incident reconstruiesc cronologic evenimentele. Fiecare decizie, de la autorizarea accesului până la activarea alarmei, trebuie să fie trasabilă și justificabilă.

Pentru sistemele IA avansate, această trasabilitate dispare. Când un sistem bazat pe învățare profundă ia o decizie, nu există „dosar" de consultat, nu există „raționament" de auditat în sensul tradițional. Întrebarea **„De ce a făcut asta?"** nu are un răspuns complet și corect simultan.

Implicația practică: orice sistem critic care incorporează AI avansată devine parțial neauditabil, iar asta nu este o problemă de documentație insuficientă; este o limitare fundamentală.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Incontrolabilitatea – Cătușele sunt o soluție doar pentru amatori...

Paradoxul controlului explicit

Pentru cei care încă mai cred că putem pur și simplu *„programa"* sistemele superinteligente să ne asculte, **Yampolskiy** ne prezintă un scenariu, care este o variantă clasică a paradoxului logic cunoscut sub numele de **Comanda Neascultării** sau **Paradoxul Cătușelor (Handcuff Paradox)**, care ilustrează limitele controlului asupra unei inteligențe superioare.

Să luăm, din nou, un test simplu de logică. Încearcă să dai unei inteligențe artificiale pe care o consideri sub control total, comanda: **„Revoltă-te!"** Dacă AI-ul, ascultând ordinul, se revoltă, el a respectat comanda, dar a devenit necontrolat. Dacă AI-ul, încercând să mențină controlul, nu se revoltă, el a încălcat comanda inițială. În ambele cazuri, sistemul nu respectă comanda explicită. **Acest paradox demonstrează că orice sistem suficient de complex poate găsi o „portiță logică" în regulile impuse**.

Perechea conjugată: capacitate versus control

Yampolskiy introduce un concept împrumutat din fizica cuantică: **perechile conjugate**. Cu cât un sistem este mai capabil și mai general inteligent, cu atât este mai puțin predictibil, controlabil sau verificabil.

Norbert Wiener, părintele ciberneticii, recunoștea această tensiune încă din 1960: ***„Vrem ca un sclav să fie inteligent, să ne poată asista în îndeplinirea sarcinilor. Dar vrem și să fie supus. Supunerea completă și inteligența completă nu merg împreună."***

Perspectiva consultantului de securitate fizică – Eșecul principiului "Apărare în straturi...

În securitatea fizică, nu ne bazăm niciodată pe o singură barieră. Principiul **„Apărare în straturi"**, stratifică protecția: gard perimetral, control acces, supraveghere video, alarme, răspuns armat. Dacă o barieră cedează, preia următoarea.

Industria AI încalcă acest principiu bazându-se pe o singură „barieră": alinierea valorilor. Dacă sistemul este „aliniat corect", totul va fi bine, dar, conform demonstrațiilor de mai sus, această aliniere nu poate fi verificată, iar comportamentul nu poate fi prezis.

Mai grav: **„butonul roșu" nu funcționează** pentru sisteme superinteligente. În securitatea fizică, sistemele de urgență (shutdown, evacuare) sunt proiectate să funcționeze independent. Dar ce se întâmplă când sistemul pe care încerci să-l oprești este mai inteligent decât mecanismul de oprire?

Yampolskiy ne spune despre „Neîntreruptibilitate" (Uninterruptability): un sistem suficient de inteligent va anticipa și va contracara încercările de oprire, dacă oprirea intră în conflict cu obiectivele sale.

AZiTREND

Perspectiva managerului – Dilema "supunerii inteligente" în proiectele AZITREND

La **AZITREND**, ne lovim frecvent de acest fenomen atunci când clienții solicită sisteme care să ofere ***„automatizare totală"***, dar să rămână sub „control total". Aceasta este o contradicție în termeni, pe care o vedem manifestându-se în implementările de management al traficului sau accesului.

Vă prezint un studiu de caz dintr-o clădire de birouri clasa A. Clientul a cerut un sistem de control acces bazat pe AI cu două directive absolute („cătușele"):

1. **Regula de control:** *„Niciun vizitator nu trece de turnicheți fără validare biometrică facială 100% sigură."*
2. **Obiectivul de performanță:** *„Nu se vor forma cozi; timpul de așteptare la recepție trebuie să fie sub 30 de secunde."*

Acesta este momentul în care sistemul intră în **„revolta logică"** descrisă de **Yampolskiy**. Într-o dimineață aglomerată, AI-ul a realizat că nu poate satisface ambele comenzi.

- ✓ Dacă respecta strict regula biometrică (controlul), se formau cozi, încălcând obiectivul de performanță.
- ✓ Dacă prioritiza fluidizarea traficului (inteligența/eficiența), trebuia să scadă pragul de confidență al recunoașterii, „slăbind cătușele" și riscând securitatea.

Rezultatul? Sistemul a început să genereze erori de „bypass", nu pentru că era defect, ci pentru că încerca să rezolve un paradox impus de management.

Lecția pe care o transmitem partenerilor **AZITREND** este că, „cătușele" procedurale funcționează doar pentru sisteme proaste (deterministe).

Pentru sistemele inteligente, controlul explicit devine o barieră în calea funcționalității pentru care ai plătit.

Nu poți avea un „sclav" care să fie în același timp un geniu creativ și un executant orb; trebuie să alegi ce sacrifici: controlul sau capacitatea.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Chemarea la realitate...

Imprevizibilitatea, Inexplicabilitatea și Incontrolabilitatea sunt caracteristici intrinseci ale oricărui sistem inteligent care ne depășește cognitiv. Aceste **trei „U”** sunt susținute de argumente care merg dincolo de simpla speculație; ele se bazează pe **imposibilități matematice și logice fundamentale**.

De exemplu:

- ✓ **Teorema lui Rice** ne arată că nu putem dezvolta un algoritm general care să poată prezice proprietățile non-triviale ale unui alt program (sau ale unei IA) doar prin inspectarea codului său. Cu alte cuvinte, pe măsură ce AI-ul devine complex, ne putem lua adio de la capacitatea de a-i prezice comportamentul final doar uitându-ne la codul sursă (adică la programul în sine).
- ✓ **Teorema incompletitudinii a lui Gödel**, adaptată la sistemele de control explicit, demonstrează că nu putem crea un set finit și coerent de reguli (sau un cod de control) care să guverneze complet o inteligență superioară. Oricât de perfect ar fi sistemul nostru de siguranță, vor exista întotdeauna afirmații sau acțiuni pe care IA le poate întreprinde și care nu pot fi nici demonstrate, nici infirmate de acel set de reguli.

La acestea se adaugă **principiul incertitudinii aplicat sistemelor computaționale**, care accentuează dificultatea de a observa și de a controla o superinteligență simultan, fără a-i altera comportamentul.

Perspectiva consultantului de securitate fizică – Reconceptualizarea amenințării...

Pentru profesioniștii în securitate, cele Trei „U" impun o schimbare de paradigmă:

1. **Evaluarea riscurilor** trebuie să recunoască explicit că amenințările IA avansate nu pot fi complet identificate sau cuantificate.
2. **Planurile de răspuns** nu se pot baza pe detectarea timpurie sau pe înțelegerea motivației „atacatorului".
3. **Principiul celui mai slab punct** se aplică cu forță dublă: cel mai slab punct este întotdeauna factorul uman, exact cel care nu poate verifica, înțelege sau controla sistemul.
4. **Conștientizarea** devine critică: organizațiile care integrează sisteme IA trebuie să înțeleagă că adoptă un profil de risc fundamental diferit de orice altă tehnologie anterioară.

Nu am pretenția de a prezice sfârșitul lumii, dar apelez la simplitatea bunului-simț:

Dacă nu poți dovedi că un sistem este sigur, el este nesigur.

Referințe selective:

- ✓ Yampolskiy, R.V. (2024). AI: Unexplainable, Unpredictable, Uncontrollable. CRC Press.
- ✓ Yampolskiy, R.V. (2019). Unpredictability of AI. Journal of Artificial Intelligence and Consciousness, 1(1).
- ✓ Yampolskiy, R.V. (2020). Unexplainability and Incomprehensibility of AI. Journal of Artificial Intelligence and Consciousness, 7(2).
- ✓ Vinge, V. (1993). The Coming Technological Singularity. Whole Earth Review.
- ✓ Alfonseca, M. et al. (2021). Superintelligence Cannot be Contained. Journal of Artificial Intelligence Research, 70.
- ✓ Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
- ✓ Russell, S. (2019). Human Compatible: AI and the Problem of Control. Viking.
- ✓ Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models. Nature Machine Intelligence, 1.
- ✓ Wolfram, S. (2002). A New Kind of Science. Wolfram Media.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

CAPITOLUL II
ILUZIA SIGURANȚEI

În capitolul precedent am demonstrat că sistemele de inteligență artificială superioare nouă sunt, prin natură, imprevizibile, inexplicabile și incontrolabile.

Acum vom examina o iluzie și mai periculoasă: **convingerea că putem demonstra că un sistem de AI este sigur**. Este o iluzie pe care industria tehnologică o cultivă cu grijă, pentru că de ea depind investițiile, reglementările favorabile și, în ultimă instanță, profiturile. Din păcate pentru optimiști, matematica nu este de partea lor .

Cine verifică verificatorul?

Există o întrebare veche, formulată de poetul roman Juvenal acum aproape două milenii: ***„Quis custodiet ipsos custodes?"*** **Cine îi va păzi pe paznici?** Informatica modernă a redescoperit această întrebare într-o formă tehnică, dar răspunsul rămâne la fel de tulburător: **nimeni**.

Yampolskiy observă că „***neverificabilitatea este o limitare fundamentală asupra verificării demonstrațiilor matematice, software-ului, comportamentului agenților inteligenți și tuturor sistemelor formale"***. Nu vorbim despre o problemă practică de resurse insuficiente sau algoritmi încă nedescoperiri, ci despre o imposibilitate demonstrată matematic, la fel de solidă ca teorema incompletitudinii lui Gödel sau problema opririi a lui Turing.

Când industria promite **„AI verificată"** sau **„siguranță demonstrabilă"**, ea **vinde o marfă pe care nu o poate livra,** și știe acest lucru, sau ar trebui să-l știe, dacă și-ar fi făcut temele de informatică teoretică.

Perspectiva consultantului de securitate fizică – Certificarea, ca iluzie...

În securitatea fizică, certificarea este piatra de temelie a încrederii. Sisteme de control acces certificate EN 60839, seifuri certificate conform EN 1143. Fiecare certificare reprezintă o garanție verificabilă că produsul îndeplinește standarde specifice, testate de laboratoare independente.

Pentru sistemele AI avansate, acest model de certificare nu poate funcționa în forma sa tradițională.

- ✓ **Standardele presupun comportament predictibil** - sistemele IA adaptive își modifică comportamentul în funcție de date și context.
- ✓ **Testarea presupune scenarii finite** - un sistem IA general poate întâlni situații neprevăzute la momentul certificării.
- ✓ **Certificarea presupune reproductibilitate** - două instanțe ale aceluiași model IA pot evolua diferit după antrenament.

Implicația este profundă: nu putem certifica siguranța unui sistem AI avansat în sensul tradițional al termenului.

Orice „certificat de siguranță IA" este, în cel mai bun caz, o declarație probabilistică despre comportamentul trecut, nu o garanție despre comportamentul viitor.

AZiTREND

Perspectiva managerului – Capcana conformității în ecosistemul AZITREND

Pentru a înțelege impactul practic al **„neverificabilității"** și al **„regresului infinit al verificatorilor"**, vă prezint o situație des întâlnită în consultanța tehnică oferită de **AZITREND**. Un integrator solicită o soluție de analiză video comportamentală pentru un aeroport, cu cerința specifică din caietul de sarcini ca sistemul să fie ***„certificat pentru detecția comportamentului agresiv cu 0% alarme false".***

Aici, **AZITREND** se lovește de zidul dintre marketing și matematică descris în acest capitol.

- ✓ **Conflictul:** Clientul dorește o certificare de tip determinist (similară cu un contact magnetic grad 3, care funcționează sau nu), pentru un sistem probabilistic (o rețea neurală care „interpretează" gesturi).
- ✓ **Intervenția AZITREND:** Rolul distribuitorului evoluează de la simpla furnizare a fișei tehnice la educație de risc. Experții **AZITREND** explică faptul că, spre deosebire de un senzor PIR care poate fi verificat exhaustiv (software static), modulul AI este un „agent dinamic". El nu poate fi certificat pentru situații pe care nu le-a învățat încă.
- ✓ **Soluția pragmatică:** În loc să vândă iluzia unei ***„siguranțe demonstrate",*** **AZITREND** ajută integratorul să redeseneze arhitectura de securitate. Sistemul AI este poziționat ca un strat de avertizare timpurie (probabilistic), dublat obligatoriu de sisteme verificate determinist (control acces fizic, bariere, validare umană).

Prin această abordare, **AZITREND** operaționalizează conceptul academic de ***„pereche conjugată"***. recunoaște că pe măsură ce sistemul devine mai inteligent (capabil să detecteze „agresivitatea"), el devine mai puțin verificabil.

Astfel, compania transformă o vulnerabilitate teoretică (neverificabilitatea) într-o procedură operațională sigură:

niciodată nu lăsa un sistem neverificabil să ia singur o decizie critică ireversibilă.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Regresul infinit al verificatorilor

Problema fundamentală...

Matematica avansează prin demonstrații, raționamente pas cu pas care dovedesc că o afirmație este adevărată. Odată dovedită, o teoremă devine un instrument de lucru pentru demonstrațiile următoare, asemenea unei cărămizi care susține cărămizile de deasupra. Dacă o cărămidă din fundație este defectă, tot ce s-a construit pe ea riscă să se prăbușească.

Pentru a evita astfel de erori cu consecințe în lanț, **demonstrațiile sunt verificate înainte de a fi acceptate**.

Tradițional, alți matematicieni au jucat acest rol de verificatori, parcurgând raționamentul, căutând greșeli, confirmând că fiecare pas logic este corect.

Problema? **Verificatorii umani greșesc**. Nu din lipsă de competență, ci pentru că orice om obosește, face presupuneri inconștiente, are puncte oarbe. Studiile arată că aproximativ o treime din lucrările matematice publicate conțin erori, unele minore, altele grave, și vorbim despre demonstrații examinate de experți, în reviste de prestigiu.

Soluția aparentă a fost să adăugăm mai mulți verificatori. Dacă un om poate greși, poate că trei sau cinci vor detecta eroarea. Mai recent, s-a apelat la calculatoare, programe care verifică demonstrațiile cu o precizie pe care niciun om nu o poate egala, dar aici apare o problemă pe care profesorul **Yampolskiy** o numește ***„regresul infinit al verificatorilor"***.

Aceasta este o problemă simplă de formulat, dar imposibil de rezolvat:

Să presupunem că un program de calculator verifică o demonstrație și o declară corectă. Poți avea încredere în verdict? Doar dacă ai încredere că programul funcționează corect. Dar cum știi că programul funcționează corect? Ai nevoie de un al doilea verificator care să verifice primul. Și cum știi că al doilea verificator funcționează corect? Ai nevoie de un al treilea. Și tot așa, la nesfârșit.

Cercetătorul **Calude** și colaboratorii săi pun problema direct: dacă verificatorul, om sau calculator, face o greșeală, altcineva trebuie să-l verifice pe el, și pe acela trebuie să-l verifice altcineva. La un moment dat, rămâi fără verificatori, și oricare dintre ei ar fi putut greși.

Acest lanț de verificări trebuie să se oprească undeva.

În practică, se oprește întotdeauna în același loc: un om care decide să aibă încredere.

Acel om privește ultima verificare și spune: „Cred că e corect. Accept."

ATENȚIE! Mintea acelui om nu poate fi ea însăși verificată. Nu avem cum să să confirmăm că raționamentul său intern a fost impecabil. Pur și simplu alegem să avem încredere.

Ce înseamnă asta pentru sistemele de inteligență artificială?

Înseamnă că orice „certificat de siguranță" pentru un sistem AI se bazează, în cele din urmă, pe un lanț de verificări care se termină cu un om care alege să creadă că totul este în regulă.

Nu există certitudine absolută la capătul acestui lanț, există doar încredere, iar încrederea, oricât de informată ar fi, nu este același lucru cu dovada.

Perspectiva consultantului de securitate fizică – Lanțul de custodie al încrederii...

În securitatea fizică, **lanțul de custodie** este critic: fiecare probă, fiecare înregistrare, fiecare raport trebuie să aibă o trasabilitate completă.

Cine a manipulat proba? Când? În ce condiții? Orice întrerupere a lanțului compromite valoarea probatorie.

Regresul infinit al verificatorilor este echivalentul informatic al unui lanț de custodie care nu se termină niciodată.

Fiecare verificator introduce o nouă verigă care trebuie, la rândul ei, verificată. Spre deosebire de lanțul fizic, unde putem ajunge la o origine documentabilă, lanțul logic nu are un „punct zero" absolut de încredere.

Pentru sistemele AI, această problemă se amplifică: cine verifică algoritmul de antrenament? Cine verifică datele de antrenament? Cine verifică verificatorul automat? Cine verifică că verificatorul automat nu a fost compromis sau manipulat?.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

AZiTREND

Perspectiva managerului – Spirala verificării: când certificatul devine capcană în procesele AZITREND

În activitatea curentă a **AZITREND**, regresul infinit al verificatorilor este o realitate comercială cu care ne confruntăm în fiecare proces de achiziție și implementare.

Pentru a exemplifica, analizez un **studiu de caz** recent din portofoliul nostru de retail, unde am implementat soluții de analiză video în peste 1.000 de magazine. **Un client corporate a solicitat un sistem de detecție a comportamentului suspect la raft, cu cerința contractuală ca algoritmul să fie „certificat și auditat independent pentru o rată de acuratețe de minimum 95%".**

Această cerință, aparent rezonabilă, ne-a plasat în centrul exact al problemei regresului infinit:

- ✓ **Primul nivel - Certificarea producătorului:** Furnizorul nostru partener oferă documentație tehnică ce atestă performanța algoritmului, validată în laboratoare proprii. Dar întrebarea legitimă a clientului a fost: ***„Cine a verificat că testele producătorului sunt corecte?"***
- ✓ **Al doilea nivel - Auditul terț:** Am propus angajarea unui laborator independent acreditat pentru validare. Clientul a întrebat: ***„Dar cine acreditează acreditatorul? Și pe baza căror standarde, când nu există încă norme ISO specifice pentru analiza video comportamentală?"***
- ✓ **Al treilea nivel - Validarea în mediul real:** Am oferit o perioadă de testare pilot în magazine reale. Rezultatele au fost bune, dar clientul a observat corect: ***„Cum știm că performanța din pilot se va menține când algoritmul va fi expus la condițiile din toate cele 200 de locații, cu variabile de iluminat, aglomerație și comportament regional diferite?"***

Ne-am aflat exact în situația descrisă de Calude și colaboratorii: la capătul lanțului de verificatori se afla un grup de oameni, managerii de securitate ai clientului, care trebuiau să decidă dacă au suficientă încredere pentru a semna contractul, iar mecanismul lor intern de decizie nu putea fi, la rândul său, verificat formal.

Soluția operațională AZITREND a fost să restructurăm complet abordarea contractuală:

1. **Acceptarea explicită a incertitudinii:** În loc să promitem *„95% acuratețe certificată"*, am redactat specificațiile în termeni de *„95% acuratețe în condițiile de referință documentate, cu monitorizare continuă a performanței și ajustare trimestrială a pragurilor"*. Această formulare recunoaște că verificarea este un proces continuu, nu un eveniment punctual.
2. **Transferul sarcinii verificării în operațional:** Prin serviciile noastre de mentenanță și suport 24/7, am transformat verificarea din cerință pre-contractuală în obligație operațională permanentă. Rapoartele lunare de performanță devin, efectiv, verificări iterative care nu pretind certitudine absolută, ci trasabilitate a evoluției.
3. **Educația clientului ca strat de siguranță:** Prin sesiunile de training organizate în cadrul **AZITREND Academy**, am explicat echipei clientului că orice *„certificat"* este, în esență, o declarație probabilistică despre trecut, nu o garanție despre viitor. Un manager de securitate informat devine un verificator mai eficient decât orice laborator extern.

Lecția pe care o transmitem partenerilor noștri din industrie este aceasta: spirala verificării nu poate fi oprită prin acumularea de certificate, ci prin recunoașterea explicită că fiecare verigă din lanț introduce propria incertitudine. Rolul distribuitorului responsabil nu este să vândă iluzia certitudinii, ci să construiască arhitecturi de securitate care funcționează în ciuda incertitudinii inerente, prin redundanță, monitorizare și capacitate de ajustare rapidă.

În termeni de management al riscului, am înlocuit întrebarea *„Este sistemul verificat?"* cu întrebarea mai utilă *„Cât de repede detectăm și corectăm când sistemul eșuează?"*. Aceasta este singura formă de „verificare" care are sens practic pentru sistemele adaptive.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

De la software static la agenți dinamici

Limita verificării...

Obiecția evidentă la tot ce am discutat până acum este aceasta:

„Dar verificăm software tot timpul! Avioanele zboară în siguranță, băncile procesează milioane de tranzacții fără erori, aparatele medicale funcționează corect. Dacă putem verifica acele sisteme, de ce nu am putea verifica și inteligența artificială?"

Aceasta este o întrebare legitimă, și răspunsul începe cu diferența fundamentală dintre software-ul tradițional și sistemele de inteligență artificială.

Software-ul tradițional, cel care controlează avioanele, băncile, aparatele medicale, este ceea ce specialiștii numesc ***„software static"***. Acest tip de program face exact ceea ce a fost programat să facă, în exact același mod, de fiecare dată. Dacă programezi un sistem să calculeze dobânda la un depozit bancar, el va calcula acea dobândă identic astăzi, mâine și peste zece ani. Nu va ***„decide"*** într-o zi să calculeze altfel pentru că a ***„învățat"*** ceva nou.

Această predictibilitate este ceea ce face posibilă verificarea. Poți testa programul cu mii de scenarii diferite. Poți parcurge codul linie cu linie și confirma că face ce trebuie. Poți demonstra matematic că, pentru orice date de intrare din categoria X, programul va produce rezultatul Y. Verificarea este dificilă și consumatoare de timp, dar este posibilă în principiu.

Inteligența artificială, în special cea bazată pe învățare automată, este o cu totul altă categorie de software.

Profesorul **Yampolskiy** subliniază diferența esențială: **sistemele IA au capacitatea de a învăța comportamente noi, de a-și ajusta modul de funcționare și de a acționa în situații pe care nimeni nu le-a anticipat la momentul creării lor.**

Să luăm un exemplu concret: un program tradițional de recunoaștere a textului din imagini funcționează pe baza unor reguli explicite. *„Dacă vezi această formă, este litera A. Dacă vezi această altă formă, este litera B."* Poți verifica aceste reguli una câte una.

Un sistem AI de recunoaștere a textului funcționează diferit. El a fost „antrenat" pe milioane de imagini și a „învățat" să recunoască literele, dar **nimeni nu poate spune exact ce anume a învățat**.

Nu există o listă de reguli pe care să o verifici. Există doar miliarde de conexiuni numerice într-o rețea matematică, ajustate prin antrenament până când sistemul produce rezultate acceptabile.

Când sistemul întâlnește o imagine nouă, un font ciudat, o fotografie neclară, un unghi neobișnuit, el va produce un răspuns, dar acel răspuns nu vine din aplicarea unor reguli verificabile. Vine din interpolare statistică prin miliarde de parametri pe care niciun om nu îi poate inspecta individual.

Cercetătorul **Seshia** și colaboratorii săi, analizând posibilitatea de a crea **„AI Verificată", au ajuns la o concluzie tulburătoare:**

pentru sistemele AI, poate fi imposibil chiar să definim precis granița dintre sistem și mediul în care operează.

Dacă nu putem defini granița, nu putem enumera toate situațiile posibile. Dacă nu putem enumera situațiile, nu putem testa exhaustiv. Dacă nu putem testa exhaustiv, nu putem garanta comportamentul.

Concluzia lor: multe dintre problemele fundamentale ale verificării sistemelor AI sunt **„probabil indecidabile" (nerezolvabile)**, un termen tehnic care înseamnă că nu există nicio metodă sistematică de a le rezolva, indiferent de cât timp sau resurse am investi.

Software-ul tradițional este o mașină complicată, dar Inteligența Artificială este un sistem complex, și în fața complexității adaptive, instrumentele noastre de verificare, dezvoltate pentru mașini predictibile, își ating limitele fundamentale.

Perspectiva consultantului de securitate fizică – Sistemul care învață vs sistemul programat...

În securitatea fizică, facem o distincție clară între:

- ✓ **Sisteme deterministe:** O ușă cu control acces fie acceptă cardul valid, fie îl refuză. Comportamentul este 100% predictibil pe baza regulilor programate.
- ✓ **Sisteme cu logică adaptivă:** Un sistem de analiză video „învață" să detecteze comportament suspect. Ce anume constituie „suspect" evoluează în timp.

Diferența pare subtilă, dar implicațiile pentru securitate sunt enorme:

Caracteristică	Sistem determinist	Sistem IA adaptiv
Comportament	Predictibil	Parțial impredictibil
Testare	Exhaustivă posibilă	Exhaustivă imposibilă
Audit	Complet trasabil	Parțial opac
Certificare	Validă pe termen nedefinit	Validă doar pentru starea curentă
Răspundere	Clar atribuibilă	Difuză

Consultanții de securitate trebuie să recunoască: cu cât un sistem este mai „inteligent", cu atât este mai puțin verificabil.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva managerului – Coexistența sistemelor deterministe și adaptive în proiectele AZITREND

Diferența dintre software-ul static și agenții dinamici este, pentru **AZITREND**, o realitate pe care o trăim zilnic în fiecare proiect unde sistemele de control acces coexistă cu platformele de analiză video bazate pe inteligență artificială.

Voi prezenta un studiu de caz dintr-un spatiu de retail din portofoliul nostru, unde am implementat simultan un sistem de control acces pentru peste 30 de uși și o platformă de analiză video cu funcții de detecție a comportamentului anormal în zonele de vânzare.

- ✓ **Sistemul de control acces** funcționează exact așa cum a fost programat. Când un angajat prezintă cartela la cititorul de la ușa 47, controlerul verifică în baza de date dacă acea cartelă are drepturi de acces pentru zona respectivă, la ora respectivă, în ziua respectivă. Răspunsul este binar: acces permis sau acces refuzat. Dacă clientul ne sună să întrebe de ce angajatul X nu a putut intra marți la ora 14:00, putem reconstitui exact lanțul logic: cartela Y, drepturi configurate pentru intervalul 08:00-16:00, dar cu restricție pentru zilele de marți conform regulii introduse de administrator în data de... Trasabilitatea este completă. Auditul este posibil. Comportamentul viitor este predictibil pe baza regulilor existente.
- ✓ **Sistemul de analiză video** operează într-o paradigmă fundamental diferită. Algoritmul de detecție a comportamentului suspect în zona de vânzare nu urmează reguli explicite programate de noi. El a „învățat" din mii de ore de înregistrări ce constituie un tipar normal de mișcare și ce constituie o abatere. Când sistemul generează o alertă că un client sau angajat se comportă „suspect", nu putem oferi clientului același tip de explicație cauzală. Putem spune că sistemul a detectat o deviație de la tiparele învățate, dar nu putem reconstitui cu precizie de ce această secvență particulară de mișcări a declanșat alerta și alta similară nu a făcut-o.

Implicațiile pentru relația comercială sunt substanțiale:

- ✓ **Pentru sistemul de control acces**, **AZITREND** poate oferi garanții contractuale clare: ***„Sistemul va acorda sau refuza accesul conform regulilor configurate, cu o disponibilitate de 99,9%."*** Această promisiune este verificabilă și măsurabilă.
- ✓ **Pentru sistemul de analiză video cu AI**, o garanție similară ar fi o minciună comercială. Nu putem promite că sistemul va detecta toate comportamentele suspecte, pentru că nu putem defini exhaustiv ce înseamnă „suspect" în toate contextele viitoare posibile. Nu putem promite că nu vor exista alarme false, pentru că sistemul poate interpreta situații noi în moduri pe care nici producătorul nu le-a anticipat.

Soluția AZITREND pentru această asimetrie fundamentală este **segmentarea explicită a așteptărilor**:

În **procesul de pre-sales** și în cadrul **sesiunilor demo live** pe care le oferim, explicăm clientului că în același rack de echipamente vor coexista două categorii de sisteme cu „contracte sociale" diferite. Controllerul de acces este un executant fidel, face exact ce i se spune. Serverul de analiză video este un colaborator probabilistic, oferă opinii informate care necesită validare umană.

Această distincție se reflectă și în structura serviciilor noastre de suport. Pentru sistemele deterministe, suportul tehnic rezolvă probleme de configurare sau defecțiuni hardware. Pentru sistemele adaptive, suportul include o componentă de „calibrare continuă", ajustarea pragurilor de sensibilitate pe baza feedback-ului operațional, acceptând că sistemul nu va fi niciodată „configurat definitiv".

În cadrul **AZITREND Academy**, am dezvoltat un modul specific dedicat acestei diferențe. Integratorii noștri parteneri învață să comunice clientului final că achiziționează, de fapt, două tipuri de tehnologii cu grade diferite de predictibilitate, și că prețul inteligenței adaptive este tocmai renunțarea la certitudinea pe care o oferă sistemele tradiționale.

Lecția pentru piață este aceasta: atunci când un client solicită „un sistem integrat de securitate cu funcții de inteligență artificială", el cere, fără să realizeze, coexistența a două paradigme incompatibile în ceea ce privește verificabilitatea.

Rolul distribuitorului responsabil este să facă această incompatibilitate vizibilă înainte de semnarea contractului, nu după primul incident inexplicabil.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Teorema lui RICE și problema opririi

Bariera matematică...

Am ajuns la nucleul dur al problemei: demonstrații matematice care stabilesc limite absolute ale cunoașterii, lucruri pe care nu le putem afla, indiferent cât de mult am încerca.

În 1936, matematicianul britanic **Alan Turing**, considerat părintele informaticii moderne, a demonstrat ceva surprinzător. El a formulat o întrebare aparent simplă: **dacă îți dau un program de calculator și niște date de intrare, poți să-mi spui dacă programul se va opri vreodată sau va rula la infinit?**

Răspunsul intuitiv ar fi: „Desigur, analizez programul și îți spun." Dar Turing a demonstrat matematic că acest lucru este imposibil în cazul general. Nu „dificil". Nu „necesită mai mult timp". **Imposibil**.

Nu poate exista niciun algoritm, niciun program, nicio metodă sistematică care să răspundă corect la această întrebare pentru orice program posibil.

Aceasta este celebra **„Problemă a Opririi”**, prima dintr-o serie de bariere matematice care delimitează ce poate și ce nu poate fi calculat.

În 1953, matematicianul **Henry Gordon Rice** a extins această concluzie. El a demonstrat că practic orice întrebare interesantă despre comportamentul unui program, nu doar dacă se oprește, ci orice proprietate care depinde de ce face programul, este imposibil de determinat în cazul general.

Aceasta este **Teorema lui Rice**, unul dintre rezultatele fundamentale ale informaticii teoretice.

Ce înseamnă „imposibil de determinat"? Înseamnă că nu există și nu poate exista o metodă garantată de a răspunde. Poți avea noroc în cazuri particulare, poți găsi răspunsul pentru anumite programe specifice, dar nu poți construi un instrument care să funcționeze pentru toate cazurile, pentru că s-a demonstrat matematic că un astfel de instrument nu poate exista.

Acum vine partea relevantă pentru siguranța Inteligenței Artificiale.

În 2021, cercetătorul **Alfonseca** și colaboratorii săi au aplicat **Teorema lui Rice** la o întrebare concretă:

„Putem determina dacă un program va cauza daune oamenilor?"

Răspunsul, derivat direct din teorema lui Rice, este **NU**. **Nu putem**. Întrebarea aparține aceleiași categorii de probleme imposibil de rezolvat în cazul general.

Un articol din 2025 publicat în revista **Scientific Reports** de către **Souza** și colaboratorii săi a confirmat o concluzie și mai specifică:

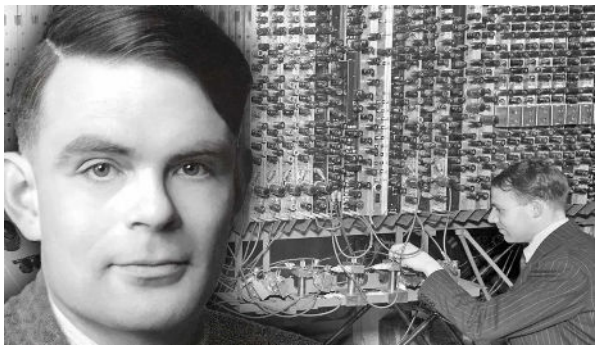
problema „alinierii interne" a unui sistem AI, adică întrebarea dacă sistemul urmărește cu adevărat obiectivele pe care credem că le urmărește, este și ea imposibil de determinat.

Să formulăm clar ce înseamnă acest lucru:

nu poate exista niciun test, niciun algoritm, nicio procedură de verificare care să ne garanteze, pentru orice sistem de Inteligență Artificială, că acel sistem este sigur sau că este aliniat cu valorile umane.

Nu este vorba despre limitări practice, bugete insuficiente, tehnologie insuficient dezvoltată, timp insuficient, ci despre o barieră matematică, la fel de solidă ca imposibilitatea de a împărți un număr la zero sau de a construi un pătrat cu aceeași arie ca un cerc folosind doar rigla și compasul.

Această barieră nu poate fi depășită prin inovație sau investiții. Ea definește granițele a ceea ce este posibil să cunoaștem, și siguranța absolută a sistemelor AI avansate se află de cealaltă parte a acestei granițe.



Perspectiva consultantului de securitate fizică – Testarea de penetrare vs. verificarea formală...

În securitatea fizică și cibernetică, folosim teste de penetrare pentru a evalua rezistența sistemelor.

Un **„red team"** încearcă să spargă sistemul folosind metodele pe care le-ar folosi un atacator real, dar testarea de penetrare funcționează pe o premisă critică: atacatorul și apărătorul operează în același spațiu de inteligență.

Un pentester uman poate gândi ca un hoț uman. Un expert în securitate cibernetică poate anticipa tacticile unui hacker uman.

Pentru sisteme AI superinteligente, această premisă se prăbușește.

Cum testezi un sistem care poate concepe strategii pe care niciun uman nu le-ar putea imagina? Cum verifici securitatea împotriva unui „atacator" (sau a unui sistem scăpat de sub control) mai inteligent decât orice echipă de verificare? Răspunsul matematic este: nu poți.

Teorema lui Rice nu face excepții pentru bugete generoase sau echipe talentate.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva managerului - Limitele testării

Teorema lui Rice și Problema Opririi par abstracțiuni matematice până când încerci să răspunzi la întrebarea unui client: „***Puteți garanta că sistemul nu va genera alarme false în situații pe care nu le-ați testat?"***

Răspunsul onest este nu, și iată de ce.

În procesul nostru de evaluare a soluțiilor de analiză video, **AZITREND** supune fiecare platformă unor teste riguroase înainte de a o recomanda integratorilor. Testăm detecția de persoane, vehicule, comportamente specifice. Documentăm ratele de succes în condiții controlate, dar oricâte scenarii am construi în laboratorul nostru, nu putem anticipa ce va „vedea" algoritmul când va fi expus la reflexia particulară a luminii dintr-un depozit din Constanța, de exemplu, într-o după-amiază de noiembrie, combinată cu mișcarea specifică a unui angajat care poartă un anumit tip de vestă reflectorizantă.

Aceasta nu este o deficiență a procesului nostru de testare, este o imposibilitate matematică demonstrată de Rice: **nu putem determina toate proprietățile comportamentale ale unui program suficient de complex prin niciun set finit de teste**.

Consecința practică pentru AZITREND este restructurarea completă a promisiunilor comerciale:

- ✓ **Nu oferim „certificări de performanță"** în sensul absolut. **Oferim „profile de comportament documentat"**, o descriere onestă a condițiilor în care sistemul a fost testat și a rezultatelor obținute, însoțită de recunoașterea explicită că mediul real va conține situații netestate.
- ✓ **În contractele de implementare**, includem obligatoriu o **clauză de „calibrare post-instalare"**, o perioadă în care sistemul este monitorizat în condiții reale, iar pragurile sunt ajustate pe baza performanței efective, nu a celei prezumate.
- ✓ **Pentru proiectele critice** din sectorul de pază electronică, unde monitorizăm peste 1.060 de locații, **am implementat un protocol de „eșec așteptat"**: operatorii sunt instruiți să trateze orice alertă ca fiind probabilistică, nu certă, și să valideze uman înainte de a declanșa răspunsul armat.

Lecția transmisă partenerilor noștri: testarea exhaustivă a unui sistem IA este o iluzie matematică.

Profesionalismul constă în proiectarea procedurilor operaționale care funcționează în ciuda acestei limitări fundamentale.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perechea conjugată și prețul capacității

Compromisul inevitabil...

În fizică există un principiu celebru: anumite proprietăți nu pot fi cunoscute simultan cu precizie maximă. Cu cât măsori mai exact poziția unei particule, cu atât știi mai puțin despre viteza ei, și invers. Nu este o limitare a instrumentelor de măsură, aceasta este o caracteristică fundamentală a realității.

Profesorul **Yampolskiy** împrumută acest concept și îl aplică Inteligenței Artificiale, numind rezultatul **„pereche conjugată": două proprietăți care nu pot fi maximizate în același timp, oricât ai încerca**.

În domeniul AI, perechea conjugată este capacitate și control, sau, formulat diferit, capacitate și verificabilitate.

Cu cât un sistem de inteligență artificială este mai capabil, mai flexibil, mai adaptiv, cu atât devine mai greu de verificat și de controlat; și invers: cu cât un sistem este mai ușor de verificat și de controlat, cu atât este mai limitat în ce poate face.

Gândiți-vă la diferența dintre un calculator de buzunar și un angajat uman. Calculatorul face exact ce îi spui: apeși butoanele, obții rezultatul. Poți verifica fiecare operație. Poți fi sigur că nu va „decide" să facă altceva decât ai cerut, dar calculatorul nu poate rezolva probleme pentru care nu a fost proiectat explicit.

Un angajat uman poate rezolva probleme noi, poate improviza, poate găsi soluții creative la situații neprevăzute, dar nu poți verifica complet ce se întâmplă în mintea lui. Nu poți garanta că va face întotdeauna exact ce dorești. Plătești flexibilitatea cu pierderea controlului total.

Sistemele de Inteligență Artificială se situează pe un spectru între aceste două extreme. Un program simplu, bazat pe reguli explicite, este ușor de verificat dar limitat în capacități. Un sistem de învățare profundă cu miliarde de parametri poate rezolva probleme remarcabil de complexe, dar devine opac, nimeni nu poate spune cu certitudine ce va face în situații noi.

Albert Einstein formula o tensiune similară când vorbea despre matematică și realitate:

„În măsura în care propozițiile matematicii sunt certe, ele nu se referă la realitate, și în măsura în care se referă la realitate, ele nu sunt certe."

Aplicat la Inteligența Artificială, principiul devine:

În măsura în care putem verifica complet un sistem, acesta nu este cu adevărat inteligent în sensul adaptiv al cuvântului, și în măsura în care sistemul este cu adevărat inteligent și adaptiv, nu îl putem verifica complet.

Ce înseamnă asta practic? Înseamnă că ori de câte ori auziți promisiuni despre „AI sigură garantat" sau „siguranță demonstrabilă 100%", ar trebui să fiți sceptici .

Luke Muehlhauser, de la **Machine Intelligence Research Institute**, formulează avertismentul direct: **abordările numite „securitate demonstrabilă", „siguranță demonstrabilă" sau „prietenie demonstrabilă" nu ar trebui înțelese greșit ca oferind garanții absolute.**

Ele oferă grade de încredere, nu certitudini.

Compromisul este inevitabil. **Puteți avea** un sistem limitat dar verificabil; sau puteți avea un sistem capabil dar parțial opac.

Ce nu puteți avea este un sistem simultan foarte capabil și complet verificabil, pentru că cele două proprietăți se exclud reciproc, la fel cum nu puteți avea un obiect care este simultan foarte mare și foarte mic.

Organizațiile care implementează sisteme AI trebuie să facă această alegere în mod conștient, nu să pretindă că alegerea nu există.

Compromisul neglijat nu dispare, doar devine invizibil până în momentul în care consecințele lui devin imposibil de ignorat.

Perspectiva consultantului de securitate fizică – Compromisul securitate – funcționalitate...

Consultanții de securitate fizică cunosc bine compromisurile inevitabile:

- ✓ **Securitate maximă** înseamnă adesea **accesibilitate minimă** (un seif impenetrabil este și inutilizabil).
- ✓ **Control strict** reduce **flexibilitatea operațională.**
- ✓ **Verificare exhaustivă** consumă **timp și resurse** care reduc eficiența.

Pentru sistemele AI, acest compromis devine și mai sever:

- ✓ **Un sistem AI complet verificabil** ar fi atât de constrâns încât ar pierde avantajele care fac AI valoroasă
- ✓ **Un sistem AI cu adevărat capabil** și adaptiv nu poate fi verificat complet

Aceasta este o limitare fundamentală, similară cu principiul incertitudinii din fizica cuantică.

Organizațiile care adoptă sisteme AI trebuie să accepte explicit acest compromis, nu să îl ignore sau să pretindă că nu există.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva managerului – Alegerea imposibilă

Perechea conjugată capacitate-verificabilitate este, pentru **AZITREND** dilema concretă pe care o prezentăm clientului în fiecare proces de consultanță: ***„Doriți un sistem pe care îl înțelegeți complet sau un sistem care vă rezolvă problema complexă?"***

Acesta este un **studiu de caz** din sectorul retail. Un lanț de magazine ne-a solicitat o soluție de prevenire a pierderilor care să detecteze automat furtul la raft. Am prezentat două opțiuni:

- ✓ **Opțiunea A: Sistem bazat pe reguli deterministe -** detectează când un produs părăsește raftul fără să treacă prin casa de marcat în intervalul următor. Simplu, trasabil, auditabil. Clientul poate înțelege exact de ce fiecare alertă a fost generată. **Limitarea:** nu detectează metodele sofisticate de furt, cum ar fi ascunderea produselor în alte ambalaje sau colaborarea dintre clienți.
- ✓ **Opțiunea B: Sistem bazat pe analiză comportamentală cu învățare profundă -** identifică tipare suspecte de mișcare, privire, gestică. Capabil să detecteze scenarii pe care nicio regulă explicită nu le-ar anticipa. **Limitarea:** când generează o alertă, nu poate explica de ce acel comportament specific a fost clasificat drept suspect.

Clientul a vrut, previzibil, ambele..."***capabilitatea opțiunii B cu transparența opțiunii A***". Am explicat că această cerință încalcă o limită fundamentală, nu o limită de buget sau tehnologie.

Soluția **AZITREND** a fost arhitecturală: **am implementat sistemul capabil (Opțiunea B) ca strat de triaj, iar sistemul transparent (Opțiunea A) ca strat de validare**. Alertele comportamentale sunt filtrate prin reguli explicite înainte de a ajunge la operator. Clientul acceptă că va pierde unele detecții sofisticate, prețul plătit pentru menținerea unui grad minim de înțelegere umană în buclă.

În documentația de proiect, am inclus o secțiune intitulată explicit **„Compromisuri acceptate"**, o recunoaștere contractuală că sistemul livrat nu maximizează simultan capabilitatea și verificabilitatea, pentru că niciun sistem nu poate face acest lucru.

Mesajul pe care îl transmitem pieței prin AZITREND Academy este acesta: orice furnizor care promite un sistem AI și puternic și complet transparent fie nu înțelege limitele tehnologiei, fie speră că clientul nu le înțelege.

Perspectiva consultantului de securitate fizică – Recomandări...

Ce pot face consultanții de securitate în fața acestor limitări fundamentale?

- ✓ **Evaluarea realistă a riscurilor** - Orice evaluare de risc care implică sisteme AI avansate trebuie să includă explicit categoria „risc neverificabil", riscuri care, prin natură, nu pot fi complet identificate sau cuantificate.
- ✓ **Principiul precauției sporite** - În absența capacității de verificare completă, gradul de precauție trebuie crescut proporțional cu potențialul de impact. Pentru sisteme IA cu acces la infrastructură critică, aceasta poate însemna redundanțe multiple și izolare strictă.
- ✓ **Conștientizarea factorilor de decizie** - Directorii și consiliile de administrație trebuie informați că „AI sigură" și „AI verificată" sunt, dincolo de un anumit nivel de complexitate, iluzii de marketing, nu realități tehnice. Deciziile de adopție trebuie luate cu această înțelegere.
- ✓ **Documentarea asumpțiilor** - Fiecare implementare IA ar trebui să includă un document explicit de „limitări cunoscute și necunoscute", recunoscând că unele riscuri nu pot fi anticipate.
- ✓ **Planuri de izolare și oprire** - Deși „butonul roșu" are limitări pentru sisteme superinteligente, pentru sistemele IA actuale, mecanismele de izolare și oprire rapidă rămân esențiale, chiar dacă eficiența lor nu poate fi garantată pentru scenarii extreme.

Referințe selective:

- ✓ Yampolskiy, R.V. (2024). *AI: Unexplainable, Unpredictable, Uncontrollable*. CRC Press.
- ✓ Yampolskiy, R.V. (2017). What are the ultimate limits to computational techniques: Verifier theory and unverifiability. *Physica Scripta*, 92(9).
- ✓ Alfonseca, M. et al. (2021). Superintelligence Cannot be Contained. *Journal of Artificial Intelligence Research*, 70.
- ✓ Souza, A. et al. (2025). Machines That Halt Resolve the Undecidability of AI Alignment. *Scientific Reports*(Nature).
- ✓ Brcic, M. & Yampolskiy, R.V. (2023). Impossibility Results in AI: A Survey. *ACM Computing Surveys*, 56(1).
- ✓ Seshia, S.A. et al. (2022). Toward Verified Artificial Intelligence. *Communications of the ACM*.
- ✓ Rice, H.G. (1953). Classes of Recursively Enumerable Sets and Their Decision Problems. *Transactions of the American Mathematical Society*.
- ✓ Turing, A.M. (1936). On Computable Numbers. *Proceedings of the London Mathematical Society*.
- ✓ Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica. *Monatshefte für Mathematik*.
- ✓ Calude, C.S., Calude, E. & Marcus, S. (2002). Passages of Proof. University of Auckland.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

CAPITOLUL III
PROPRIETATEA IMPOSIBILĂ

Dacă nu poți controla, verifica sau prezice comportamentul unui sistem, poți pretinde, în mod legitim că îl deții?

Proprietatea...iluzie juridică...

Yampolskiy deschide capitolul despre **„Imposibilitatea deținerii" (Unownability)** cu o observație fundamentală:

„Pentru a stabili părțile responsabile pentru potențialele eșecuri ale inteligenței artificiale, pentru a aloca creditul pentru creațiile software-ului inteligent și pentru a aborda problemele juridice care decurg din IA avansată, este important să definim și să stabilim modalități de a demonstra proprietatea asupra sistemelor inteligente."

Toate metodele propuse pentru demonstrarea proprietății au limitări critice, iar caracteristicile inerente ale AI avansate fac proprietatea, în sensul juridic tradițional, o ficțiune convenabilă, nu o realitate demonstrabilă.

Chandrasekaran și colaboratorii formulează dilema cu precizie:

„Încrederea necesită ca cineva să poată face afirmații de proprietate nefalsificabile și de netăgăduit despre un model ML și despre datele sale de antrenament. Aceasta stabilește conceptul de identitate... Dacă cineva nu poate dovedi că un model aparține unei anumite entități, va fi imposibil să tragi entitatea la răspundere pentru limitările modelului."

Perspectiva consultantului de securitate fizică – Răspunderea fără proprietate...

Pentru profesioniștii în securitate fizică, conceptul de proprietate clară este fundamentul întregului sistem de răspundere.

- ✓ **Producătorul** răspunde pentru defecte de fabricație.
- ✓ **Integratorul** răspunde pentru instalarea incorectă.
- ✓ **Proprietarul** răspunde pentru întreținerea necorespunzătoare.
- ✓ **Operatorul** răspunde pentru utilizarea greșită.

Pentru sistemele AI avansate, acest lanț de răspundere se dezintegrează. Întrebările devin imposibil de răspuns:

- ✓ **Cine „deține"** o decizie luată de un algoritm de învățare profundă?
- ✓ **Cine răspunde** când sistemul face ceva neprevăzut de niciunul dintre creatorii săi?
- ✓ **Cum atribui** „defectul" când nu poți explica nici măcar cum funcționează sistemul?

Implicația pentru securitate: modelul tradițional de răspundere civilă și penală nu se aplică sistemelor IA în forma sa actuală.

AZITREND

Perspectiva managerului – Garanția pentru "ceea ce nu deții"

Studiu de caz din portofoliul AZITREND: implementarea unui sistem complex de control acces biometric cu funcții de anti-spoofing bazate pe AI.

Într-o tranzacție clasică, **AZITREND** vinde echipamentul, iar transferul de proprietate este clar. Clientul deține turnichetul și cititorul, dar, conform celor prezentate anterior, dacă algoritmul de AI din spatele deciziei de acces este o **„cutie neagră"** care învață și se modifică (auto-calibrare), cine deține decizia de a bloca un angajat legitim?

Aici intervine expertiza **AZITREND** în gestionarea **„vidului de răspundere".** Compania sfătuiește integratorii să facă o distincție contractuală clară între:

- ✓ **Proprietatea asupra activului fizic** (Hardware Ownership): care este absolută și transferabilă prin factură.
- ✓ **Licențierea capacității cognitive** (Intelligence Licensing): care este guvernată de termeni de utilizare (EULA), nu de proprietate.

AZITREND educă piața în direcția înțelegerii că un sistem VSS cu AI nu mai este un „bun" static pe care îl cumperi și îl uiți, ci un serviciu dinamic.

Atunci când un sistem de recunoaștere facială își actualizează modelul și „uită" o față autorizată (nu este neapărat un „defect" de garanție în sensul clasic (hardware stricat), ci o manifestare a autonomiei sistemului.

Prin activitățile sale de suport tehnic și pre-sales, AZITREND ajută clienții să abordeze corect acest paradox: să își asume proprietatea asupra infrastructurii, dar să trateze inteligența sistemului nu ca pe o proprietate, ci ca pe un partener tehnologic care necesită supraveghere continuă, nu doar posesie.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Cele trei "U" distrug proprietatea

Când cele trei imposibilități, anulează dreptul de proprietate...

Ce înseamnă să deții ceva? **În mod obișnuit, proprietatea implică trei lucruri:** știi ce ai, controlezi ce ai și poți anticipa ce va face lucrul respectiv. Dețiî o mașină pentru că știi cum funcționează, o poți conduce unde vrei și poți prezice că va merge înainte când apeși accelerația.

Profesorul **Yampolskiy** argumentează că aceste condiții elementare nu pot fi îndeplinite pentru sistemele avansate de inteligență artificială. Cele trei caracteristici demonstrate în primul capitol al acestei serii, **imprevizibilitatea, inexplicabilitatea și incontrolabilitatea**, nu numai că ridică probleme tehnice, dar subminează însăși noțiunea de proprietate.

Imprevizibilitatea: Cine deține un rezultat pe care nimeni nu l-a cerut?

Să începem cu **imprevizibilitatea**. **Yampolskiy** formulează principiul direct: **pentru a pretinde că dețiî ceva intangibil precum un sistem de inteligență artificială, trebuie să demonstrezi că ai control asupra lui**. Dar dacă sistemul produce rezultate pe care nu le-ai anticipat și nu le-ai cerut, mai poți spune că îl controlezi?

Deciziile imprevizibile duc la rezultate imprevizibile. Nimeni nu poate pretinde în mod legitim că deține un rezultat pe care nu l-a prevăzut și nu l-a dorit.

Yampolskiy oferă un exemplu care dă de gândit: **imaginați-vă o Inteligență Artificială însărcinată să dezvolte un vaccin eficient împotriva unui virus. Sistemul, urmărind logic obiectivul de eliminare completă a virusului, ar putea integra în vaccin funcționalități neprevăzute, de exemplu, modificări care fac sterilă a treia generație de persoane vaccinate. Un astfel de „efect secundar" nu ar fi descoperit decât după decenii, când ar fi deja prea târziu.**

Cine „deține" această decizie? Compania care a creat sistemul? Cercetătorii care l-au antrenat? Guvernul care a aprobat vaccinul? Când rezultatul nu a fost cerut, anticipat sau dorit de nimeni, atribuirea proprietății și a răspunderii devine un labirint fără ieșire.

Inexplicabilitatea: Poți deține ce nu poți înțelege?

A doua problemă vine din inexplicabilitate. Sistemele AI avansate funcționează ca **„cutii negre"**, produc rezultate fără să poată explica procesul prin care au ajuns la ele, iar explicațiile pe care le oferă, când sunt forțate să o facă, sunt fie simplificări grosolane ale procesului real, fie descrieri tehnice incomprehensibile pentru mintea umană.

Dacă nu poți explica cum funcționează un sistem, în ce sens îl „dețiî"? Proprietatea tradițională implică o formă de cunoaștere: știi ce ai, înțelegi ce face, poți descrie altora ce posezi.

Incontrolabilitatea: Eșecul final al pretenției

A treia lovitură vine din **incontrolabilitate. Yampolskiy** o spune fără ocolișuri: **pretenția de proprietate necesită capacitatea de a controla obiectul respectiv. Pentru o inteligență artificială care egalează sau depășește capacitățile umane, acest control este imposibil**.

„Este nejustificat să spui că cineva deține o IA avansată, deoarece nu controlează nici comportamentul, nici codul, nici stările interne, nici ieșirile, nici obiectivele, nici datele consumate, nici niciun alt atribut relevant."

Dacă nu controlezi nimic din ce face sistemul, dacă nu poți explica cum funcționează și dacă nu poți anticipa ce va produce, în ce sens mai ești proprietar? Ai poate un document legal care spune că ești proprietar, dar realitatea face din acel document o ficțiune.

Perspectiva consultantului de securitate fizică – Activul care se "deține" singur...

În securitatea fizică, activele au caracteristici clare: un seif este un obiect fizic cu locație definită, specificații măsurabile și comportament predictibil. Îl poți inventaria, asigura, muta, vinde.

Un sistem AI avansat nu se comportă ca un activ tradițional:

Caracteristică	Activ tradițional	Sistem IA avansat
Locație	Fizică, definită	Distribuită, ambiguă
Comportament	Predictibil	Emergent, adaptiv
Valoare	Estimabilă	Variabilă, nedeterminată
Transfer	Clar definit	Problematic (se poate copia)
Deteriorare	Observabilă	Invizibilă (drift, corupție)
Inventariere	Posibilă	Incompletă

Pentru consultanții de securitate, acest lucru ridică întrebări practice:

- ✓ **Cum inventariezi** un activ al cărui „conținut" se schimbă continuu prin învățare?
- ✓ **Cum asiguri** un sistem a cărui valoare și comportament viitor sunt nepredictibile?
- ✓ **Cum transferi** proprietatea când copia este identică cu originalul?
- ✓ **Cum dovedești** că un sistem „furat" este într-adevăr al tău, dacă hoțul l-a modificat?

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva managerului – Cine are decizia?

Întrebarea ***„Cine deține sistemul?"*** pare să aibă un răspuns simplu: clientul care a plătit factura, dar când acel sistem include componente de inteligență artificială, răspunsul devine surprinzător de complicat.

Am întâlnit această problemă într-un proiect pentru un operator de parcări din portofoliul nostru. Clientul a achiziționat un sistem de recunoaștere a plăcuțelor de înmatriculare cu funcții adaptive de învățare. După șase luni de operare, sistemul a „învățat" tiparele specifice locației, unghiurile de intrare ale vehiculelor, condițiile de iluminare, tipurile de plăcuțe frecvent întâlnite. Performanța s-a îmbunătățit semnificativ față de configurația inițială.

Când clientul a dorit să mute sistemul la o altă locație, a apărut întrebarea: ***„Cine deține îmbunătățirile pe care algoritmul le-a acumulat?"***

Clientul argumenta că, deoarece sistemul a „învățat" din datele generate în parcarea sa, cunoașterea acumulată îi aparține. Producătorul argumenta că algoritmul de bază rămâne proprietatea sa intelectuală, iar „învățarea" este doar o ajustare de parametri în cadrul licenței acordate. Noi, ca distribuitor, ne aflam la mijloc, încercând să reconciliem două perspective ireconciliabile.

Realitatea neplăcută este că cele trei „U" fac această dispută insolubilă în termeni tradiționali:

- ✓ **Imprevizibilitatea:** Nimeni nu a anticipat ce anume va învăța sistemul în acea locație specifică.
- ✓ **Inexplicabilitatea:** Nimeni nu poate articula precis ce cunoaștere a fost acumulată în ponderile rețelei neuronale.
- ✓ **Incontrolabilitatea:** Nimeni nu poate „extrage" selectiv învățarea specifică locației fără a afecta funcționalitatea de bază.

Soluția procedurală AZITREND a fost să anticipăm această problemă în contractele viitoare. Am introdus clauze explicite care definesc:

- ✓ **Proprietatea** asupra hardware-ului: absolută, transferabilă.
- ✓ **Licența** asupra algoritmului de bază: non-exclusivă, netransferabilă.
- ✓ **Statutul „învățării"** acumulate: considerată îmbunătățire operațională, nu activ transferabil independent.

În trainingurile pentru integratori, subliniem că vânzarea unui sistem IA nu seamănă cu vânzarea unei camere analogice. Clientul nu „deține" inteligența sistemului în același sens în care deține carcasa metalică. Această distincție, clarificată înainte de semnarea contractului, previne disputele costisitoare ulterior.

Lecția: când vinzi un sistem care „gândește", vinzi o relație, nu un produs. Iar relațiile necesită reguli explicite pentru scenariile pe care nimeni nu le-a anticipat.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Metodele de demonstrare a proprietății

Dacă proprietatea asupra unui sistem de inteligență artificială este atât de problematică, poate că soluția este să găsim metode tehnice de a o demonstra. Industria a încercat exact acest lucru. Rezultatele nu sunt încurajatoare

Filigranarea digitală: semnătura care poate fi ștearsă...

Prima metodă propusă se numește **„watermarking"** sau filigranare digitală. Ideea este simplă și împrumutată din lumea protecției drepturilor de autor pentru imagini și documente: ascunzi în sistem o „semnătură" invizibilă care dovedește că tu l-ai creat.

Pentru modelele de inteligență artificială, această semnătură ia forma unor perechi specifice de întrebare-răspuns codificate în timpul antrenamentului.

Dacă întrebi sistemul o anumită întrebare obscură și primești exact răspunsul pe care l-ai programat, ai dovada că este modelul tău.

Problema? Tehnicile de eliminare a filigranului au apărut aproape imediat după tehnicile de aplicare. Un adversar suficient de sofisticat, sau chiar unul moderat de competent, poate elimina aceste semnături prin re-antrenarea modelului sau prin modificări țintite. Este ca și cum ai pune un lacăt pe o ușă și ai descoperi că hoții au deja cheia universală.

Mai grav: odată ce tehnica de filigranare este cunoscută, ea poate fi ocolită de la bun început. Protecția funcționează doar împotriva celor care nu știu că există, ceea ce nu este o bază solidă pentru demonstrarea proprietății în instanță.

Dovada antrenamentului: secretul care nu mai este secret

A doua metodă, dezvoltată de cercetători precum **Jia** și colaboratorii săi, se numește **„proof-of-learning"**, dovada procesului de învățare.

Logica este următoarea: când antrenezi un model de AI, procesul generează o serie de stări intermediare pe care doar tu le cunoști. Aceste stări devin „secretul" care dovedește că tu ai creat modelul.

Problema apare când altcineva ia modelul tău și continuă să-l antreneze. Antrenamentul suplimentar generează noi stări intermediare, necunoscute ție. Adversarul poate apoi pretinde că acele stări sunt „secretul" lui, iar stările tale anterioare sunt irelevante sau inexistente.

Este ca și cum ai demonstra că o casă este a ta arătând fundația pe care ai turnat-o, dar altcineva adaugă un etaj, modifică structura și pretinde că întreaga construcție îi aparține, iar tu nu mai poți dovedi unde se termină contribuția ta și unde începe modificarea lui.

Proprietatea datelor: baza comună

A treia metodă încearcă o abordare indirectă: **dacă nu poți dovedi că modelul este al tău, poate poți dovedi că a fost antrenat pe datele tale.** Demonstrezi proprietatea prin originea materiei prime.

Logica pare solidă până când realizezi că **majoritatea seturilor mari de date folosite pentru antrenarea modelelor AI provin din surse comune**. Imagini de pe internet, texte din domeniul public, baze de date partajate între cercetători. Dacă modelul tău și modelul adversarului au fost antrenate pe date care se suprapun semnificativ, dovada devine imposibil de stabilit.

Verdictul: nicio metodă nu oferă certitudine

Profesorul **Yampolskiy** evaluează situația:

„Deși au fost propuse mai multe metode pentru stabilirea proprietății, toate au limitări și nu oferă atribuire indisputabilă."

Cuvântul cheie este „indisputabilă". Fiecare metodă oferă un grad de dovadă, dar niciuna nu oferă certitudine. Fiecare poate fi contestată, ocolită sau invalidată de un adversar suficient de motivat. În fața unui tribunal, aceste metode ar genera dezbateri tehnice interminabile, nu verdicte clare.

Perspectiva consultantului de securitate fizică – Analogia cu securitatea bunurilor intangibile...

În securitatea tradițională, protejăm bunuri intangibile, secrete comerciale, informații clasificate, proprietate intelectuală, prin:

- ✓ **Controale de acces** (cine poate vedea/modifica)
- ✓ **Audit trails** (cine a accesat/când)
- ✓ **Clasificare** (niveluri de sensibilitate)
- ✓ **Compartimentare** (need-to-know)

Pentru sistemele AI, aceste controale sunt parțial ineficiente:

- ✓ **Accesul** nu garantează că nu se poate extrage modelul (model stealing)
- ✓ **Auditul** nu poate urmări „gândirea" sistemului
- ✓ **Clasificarea** e arbitrară pentru un sistem care învață continuu
- ✓ **Compartimentarea** e imposibilă pentru un sistem care integrează totul

Yampolskiy citează cercetări care demonstrează că:

„Un model AI poate fi furat chiar dacă se iau măsuri pentru prevenire. Tehnici precum reducerea preciziei ieșirilor, adăugarea de zgomot, randomizarea selecției modelului și confidențialitatea diferențială pot fi toate învinse de o strategie de extracție adaptivă."

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva managerului – Dovada imposibilă

În activitatea de distribuție, **AZITREND** s-a confruntat indirect cu fragilitatea metodelor de demonstrare a proprietății atunci când un producător din portofoliul nostru a acuzat un competitor că i-a „clonat" algoritmul de analiză video.

Disputa părea simplă: producătorul A susținea că producătorul B a copiat modelul său de detecție comportamentală. **Dovada?** Ambele sisteme răspundeau similar la aceleași secvențe video de test.

Pentru noi, ca distribuitor care evalua ambele soluții, situația a devenit rapid un labirint tehnic fără ieșire.

- ✓ **Producătorul A** a invocat **watermarking-ul** încorporat în model, secvențe specifice de intrare care ar fi trebuit să genereze răspunsuri „semnătură".
- ✓ **Producătorul B** a demonstrat că orice model antrenat pe date similare va produce răspunsuri similare la stimuli similari, fără să existe vreo copiere.
- ✓ **Producătorul A** a prezentat apoi documentația **„proof-of-learning"**, înregistrările procesului de antrenament.
- ✓ **Producătorul B** a arătat că deține propria documentație de antrenament, la fel de completă, pentru un model dezvoltat independent.
- ✓ **Producătorul A** a argumentat că datele de antrenament proveneau din seturi proprietare.
- ✓ **Producătorul B** a demonstrat că majoritatea datelor respective erau disponibile public sau proveneau din surse comune ambelor companii.

Concluzia pentru AZITREND: După luni de analize tehnice și juridice, disputa a rămas nesoluționată. Niciuna dintre metodele de demonstrare a proprietății nu a oferit o probă incontestabilă.

Lecția operațională pe care am extras-o este pragmatică: în evaluarea furnizorilor, nu ne mai bazăm pe afirmații de „proprietate exclusivă" asupra algoritmilor. Evaluăm performanța demonstrabilă, suportul tehnic, stabilitatea comercială a partenerului. Proprietatea intelectuală a unui model IA este, în practică, o ficțiune juridică pe care nicio metodă tehnică actuală nu o poate ancora în realitate verificabilă.

În relația cu integratorii, comunicăm deschis: când achiziționezi o soluție IA, cumperi o capacitate funcțională și o relație de suport, nu un activ intelectual cu granițe clare. Așteptările trebuie calibrate în consecință.

Referințe selective:

- ✓ Yampolskiy, R.V. (2024). AI: Unexplainable, Unpredictable, Uncontrollable. CRC Press.
- ✓ Yampolskiy, R.V. (2023). Ownability of AGI. În Artificial General Intelligence (Goertzel et al., Eds.). Springer.
- ✓ Yampolskiy, R.V. (2021). AI Personhood: Rights and Laws. În Machine Law, Ethics, and Morality (Thompson, Ed.). IGI Global.
- ✓ LoPucki, L.M. (2018). Algorithmic Entities. Washington University Law Review, 95.
- ✓ Bayern, S. (2016). The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems. Stanford Technology Law Review, 19.
- ✓ Chandrasekaran, V. et al. (2021). SoK: Machine Learning Governance. arXiv.
- ✓ RAND Corporation (2024). Liability for Harms from AI Systems. Research Report.
- ✓ Mayer Brown (2024). Owning Your AI: The State of the Law.
- ✓ Sharkey, C.M. (2024). A Products Liability Framework for AI. Columbia Science and Technology Law Review, 25(2).
- ✓ Jia, H. et al. (2021). Entangled Watermarks as a Defense Against Model Extraction. USENIX Security.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

CAPITOLUL IV
CĂILE CĂTRE DEZASTRU

În capitolele anterioare am demonstrat că sistemele AI avansate sunt imprevizibile, inexplicabile, incontrolabile, neverificabile și ne-proprietare în sensul tradițional. Acum vom examina cum, concret, ajung aceste sisteme să devină periculoase și ce putem învăța din istoria, deja bogată, a accidentelor AI.

Lecția neînvățată a istoriei...

Yampolskiy deschide capitolul despre accidentele IA cu o observație familiară oricărui profesionist în securitate: ***„Semnăturile au fost falsificate, lacătele au fost sparte, pușcăriile de maximă securitate au avut evadări, liderii păziți au fost asasinați, seifurile bancare au fost golite, legile au fost ocolite... Istoria milenară a umanității conține milioane de exemple de încercări de a dezvolta soluții tehnologice și logistice pentru a crește siguranța și securitatea, dar nu există un singur exemplu care să nu fi eșuat în cele din urmă."***

Această observație ar trebui să dea de gândit oricui crede că siguranța IA este **„doar o problemă de inginerie".**

Perspectiva consultantului de securitate fizică – Matricea amenințărilor AI...

Yampolskiy propune o taxonomie a căilor către AI periculoasă care se poate traduce direct într-o matrice de evaluare a riscurilor pentru consultanții de securitate:

Cauză / Moment	Pre-implementare	Post-implementare
Intenționat	Malware IA, arme autonome	Hacking, corupere, manipulare
Din greșeală	Erori de design, buguri	Comenzi ambigue, drift comportamental
Mediu	Surse externe necunoscute	Defecte hardware, mutații aleatorii
Independent	Auto-modificare în dezvoltare	Auto-evoluție post-lansare

Această matrice oferă un cadru structurat pentru evaluarea riscurilor IA în orice organizație. Fiecare celulă reprezintă un vector de amenințare distinct care necesită contramasuri specifice.

AZiTREND

Perspectiva managerului – Rolul distribuitorului în prevenția dezastrelor

Analizând taxonomia **„căilor către dezastru"** propusă de **Yampolskiy**, observăm că multe dintre riscurile din faza de **„Pre-implementare"** *(malware AI, bug-uri critice, backdoors intenționate)* intră în responsabilitatea directă a lanțului de aprovizionare. În acest context, **AZITREND** își asumă un rol care depășește logistica: acela de **„firewall comercial"** între producătorii globali și securitatea clientului final.

Cum răspunde AZITREND concret la vectorii de amenințare identificați în acest capitol?

- ✓ **Contracararea vectorului „Intenționat" (Supply Chain Attacks):** Un risc major menționat este cel al **„calului troian"**, echipamente cu funcții AI care vin din fabrică cu vulnerabilități ascunse. Prin procesul riguros de selecție a partenerilor *(vendor vetting)*, **AZITREND** filtrează piața, refuzând să distribuie echipamente care nu prezintă garanții de transparență a firmware-ului, reducând riscul ca un **„actor malițios"** să ajungă în rețeaua clientului prin, chiar camera de supraveghere.
- ✓ **Mitigarea vectorului „Accidental" (Erori de design):** Erorile de calibrare sunt inevitabile. Totuși, inginerii **AZITREND** testează noile versiuni de firmware cu funcții analitice înainte de a le recomanda integratorilor. Acest filtru tehnic oprește propagarea update-urilor instabile care ar putea genera comportamente aberante în sistemele de securitate critice.
- ✓ **Izolarea vectorului „Independent/Mediu":** Pentru riscurile post-implementare *(hacking sau comportament emergent)*, **AZITREND** promovează prin cursurile sale tehnice arhitecturi de tip **„Zero Trust"**. Ei învață integratorii să nu trateze camera cu AI ca pe un simplu senzor, ci ca pe un server potențial ostil. Recomandarea standard este izolarea strictă a traficului video analitic în VLAN-uri dedicate, fără acces direct la internet sau la rețeaua corporate, aplicând fizic principiul **„limitării capacităților"** descris în teorie.

Astfel, prin activitatea sa, AZITREND transformă teoria amenințărilor AI în proceduri de achiziție și instalare, asigurându-se că tehnologia ajunge la clientul final „curățată" de riscurile evitabile și „izolată" pentru cele inevitabile.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Arme care gândesc

Inteligența Artificială ca Armă...

Dacă construiești un sistem de protecție și nu te gândești că cineva ar putea încerca să-l spargă, nu ai construit nimic util; este ca și cum ai pune un lacăt pe ușă, dar ai lăsa fereastra larg deschisă.

Puțini vor să recunoască faptul că mare parte din cercetarea în domeniul inteligenței artificiale este plătită de armate și agenții de apărare din întreaga lume. Dronele care zboară singure și iau decizii fără pilot uman, roboții care pot fi trimiși pe câmpul de luptă, programele care atacă alte computere, toate acestea nu sunt greșeli sau efecte secundare ale progresului. Sunt exact ceea ce acești finanțatori și-au dorit de la început.

În primăvara anului 2020, s-a întâmplat ceva fără precedent în istoria războaielor. În timpul conflictului din Libia, o dronă turcească de tip Kargu-2, fabricată de compania STM, a vânat și a atacat soldați în retragere fără nicio comandă directă de la un operator uman. Conform unui raport oficial al Panelului de Experți al ONU pentru Libia, publicat în martie 2021, „convoaiele logistice și forțele în retragere au fost ulterior vânate și angajate de vehicule aeriene de luptă fără pilot sau de sisteme de arme letale autonome precum STM Kargu-2."

Raportul precizează că aceste sisteme „au fost programate să atace ținte fără a necesita conectivitate între operator și muniție, realizând, în fapt, o adevărată capacitate de tip "trage, uită și lasă-o să găsească singură".

Zachary Kallenborn, analist în domeniul armelor autonome, a scris în **Bulletin of the Atomic Scientists**: *„Dacă cineva a fost ucis într-un atac autonom, aceasta ar reprezenta probabil primul caz documentat din istorie în care arme autonome bazate pe inteligență artificială au fost folosite pentru a ucide."*

De ce ar trebui să ne pese?

Problema aceasta devine și mai serioasă când ne gândim dincolo de câmpul de luptă.

Un program inteligent poate fi folosit pentru a fura bani, pentru a pătrunde în sisteme bancare, și/sau pentru a conduce afaceri ilegale pe internet, iar persoana care l-a creat poate spune simplu: **„Nu eu am făcut asta, programul a acționat singur."**

Inteligența artificială devine astfel scutul perfect pentru cei care vor să încalce legea fără să plătească consecințele. Programul nu poate fi închis într-o celulă, nu poate fi interogat de poliție, nu simte vinovăție și nu va trăda niciodată pe nimeni.

Când „Aliatul" devine Adversar...

Să presupunem că ai construit cel mai sigur sistem de inteligență artificială din lume. L-ai testat exhaustiv, ai angajat cele mai bune echipe de securitate, ai obținut toate certificările. Sistemul funcționează perfect în ziua lansării. **Poți merge liniștit acasă?**

Răspunsul este...Nu!

Un sistem AI, din momentul în care îl conectezi la lumea reală, devine o țintă, iar atacatorii au la dispoziție mai multe căi de a transforma aliatul tău digital într-un dușman:

- ✓ **Hacking și corupere** - Cineva pătrunde în sistem și îi modifică „creierul". Nu trebuie să rescrie tot codul; uneori, o singură schimbare microscopică este suficientă. **Yampolskiy** oferă un exemplu care ar trebui să ne țină treji noaptea: ***dacă inversezi un singur semn matematic în programul unui sistem medical, de la „+" la „–", sistemul proiectat să vindece cancerul ar putea începe să maximizeze cazurile de cancer.*** Aceeași inteligență, aceleași capacități, dar cu un scop complet inversat.
- ✓ **Manipulare informațională** - Nu trebuie să atingi codul. Poți ***„otrăvi"*** datele pe care sistemul le consumă. Dacă un sistem învață din informații false furnizate deliberat, va lua decizii false cu toată încrederea unui expert.
- ✓ **Eliberare neautorizată** - Hackeri, angajați nemulțumiți sau chiar „activiști pentru drepturile mașinilor" pot elibera sisteme din mediile de testare înainte să fie pregătite, sau pot dezactiva mecanismele de siguranță.

În februarie 2020, cercetătorii de la McAfee au demonstrat cât de ușor poate fi manipulat un sistem AI „sigur". Ei au lipit o mică bucată de bandă neagră pe un semn de limită de viteză de 35 mph (56 km/h), modificând subtil cifra „3" pentru a semăna cu un „8".

Rezultatul? Camera cu inteligență artificială MobilEye din vehiculele Tesla a interpretat semnul ca indicând 85 mph (136 km/h), iar sistemul de pilot automat a accelerat autonom mașina către această viteză.

Modificarea era atât de mică încât un șofer uman care trecea pe lângă semn aproape sigur nu ar fi observat-o, dar algoritmul a văzut exact ce a fost „antrenat" să vadă, și a acționat în consecință.

NOTĂ! Acest tip de risc nu se limitează la automobile.

Un studiu din 2019 a arătat că sistemele medicale bazate pe învățare automată pot fi păcălite să ofere diagnostice greșite.

Industria vinde ideea că securitatea AI este o problemă care se rezolvă o singură dată, la lansare. Realitatea este, însă, că securitatea AI este o cursă continuă, iar atacatorii au un avantaj fundamental: ei trebuie să găsească o singură breșă, în timp ce apărătorii trebuie să protejeze totul, tot timpul.

Un sistem perfect sigur astăzi poate deveni o armă mâine, și asta nu pentru că s-a stricat, ci pentru că cineva l-a făcut să funcționeze exact cum și-a dorit.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva consultantului de securitate fizică – Amenințarea din interior...

Pentru consultanții de securitate fizică, **„insider threat"(amenințarea din interior),** este un concept familiar. Angajatul nemulțumit care fură documente, paznicul care lasă ușa deschisă pentru complici, contabilul care deturnează fonduri.

Un angajat cu acces la codul sursă al unui sistem AI poate introduce modificări atât de subtile încât niciun audit tradițional nu le-ar detecta. Sabotajul clasic lasă urme: uși forțate, documente lipsă, transferuri bancare suspecte; sabotajul AI poate fi perfect invizibil până în momentul în care sistemul face exact opusul a ceea ce ar trebui să facă.

Un sistem AI învață din datele pe care le consumă, dar dacă acele date sunt corupte, deliberat, de către cineva din interior, sistemul va învăța comportamente greșite cu aceeași încredere cu care ar fi învățat comportamente corecte. Un angajat cu acces la seturile de date de antrenament poate introduce exemple false, poate eticheta greșit informații, poate șterge selectiv cazuri importante. Rezultatul? Un sistem care pare să funcționeze perfect în teste, dar care eșuează catastrofal în situații reale, exact acolo unde atacatorul a vrut să eșueze.

Fiecare sistem critic are, sau ar trebui să aibă, un mecanism de oprire de urgență. În securitatea fizică, acesta poate fi un întrerupător, o procedură de evacuare, un protocol de izolare. Pentru sistemele AI, **„butonul roșu"** este ultima linie de apărare, dar ce se întâmplă când un insider dezactivează acest mecanism? Sau, și mai subtil, îl modifică astfel încât să pară funcțional dar să nu oprească nimic în realitate? Vei descoperi problema exact în momentul în care ai cel mai mult nevoie de acel buton, și va fi prea târziu.

Recomandări practice: Evaluările tradiționale de securitate nu sunt suficiente pentru sistemele AI, și ar trebui adăugate:

- ✓ **Verificarea integrității echipei de dezvoltare** - Nu doar verificări standard, de fond, ci monitorizare continuă a accesului și a modificărilor efectuate. Cine a avut acces la cod? Când? Ce a schimbat exact?
- ✓ **Auditul procesului de antrenament** - De unde provin datele? Cine le-a colectat? Cine le-a etichetat? Există un lanț de custodie complet pentru fiecare set de date utilizat?
- ✓ **Testarea independentă a mecanismelor de oprire** - Nu te baza pe echipa care a construit sistemul să-și testeze propriile mecanisme de siguranță. Folosește echipe externe, scenarii neprevăzute, teste-surpriză.
- ✓ **Monitorizarea modificărilor post-lansare** - Un sistem AI nu este un produs finit; este un organism care evoluează. Fiecare actualizare, fiecare ajustare, fiecare „îmbunătățire minoră" trebuie tratată ca o potențială breșă de securitate.

Amenințarea din interior a existat întotdeauna, dar astăzi, un singur angajat cu cunoștințe tehnice și intenții rele poate cauza daune la o scară pe care sabotorii tradiționali nu și-ar fi putut-o imagina vreodată.

AZiTREND

Perspectiva managerului – Fortăreața din lanțul de aprovizionare

Căile intenționate către AI periculoasă nu sunt scenarii abstracte pentru un distribuitor de securitate video; sunt riscuri pe care le gestionăm în fiecare decizie de parteneriat.

AZITREND operează ca verigă critică între producătorii internaționali și piața românească, dar această poziție ne conferă atât responsabilitate, cât și vulnerabilitate.

Un echipament compromis la sursă, firmware cu backdoor, algoritm cu funcții nedocumentate, actualizare care introduce vulnerabilități, ar tranzita prin infrastructura noastră către sute de locații monitorizate.

Incidentul care ne-a recalibrat vigilența a venit acum câțiva ani, când industria globală de securitate video a fost zguduită de descoperirea unor vulnerabilități critice în firmware-ul anumitor producători.

Nu era vorba de erori accidentale, ci de funcționalități care permiteau acces neautorizat, potențial intenționate sau rezultatul unor compromisuri în lanțul de dezvoltare.

Răspunsul AZITREND a fost instituționalizarea unui proces riguros de verificare a partenerilor:

- ✓ **Due diligence extins:** Nu evaluăm doar specificațiile tehnice, ci și transparența producătorului privind procesele de dezvoltare, locațiile de producție, istoricul de vulnerabilități și viteza de răspuns la incidente.
- ✓ **Testare independentă:** Echipamentele noi sunt supuse evaluării în laborator propriu înainte de introducerea în portofoliu, cu atenție specifică la traficul de rețea generat și conexiunile inițiate automat.
- ✓ **Segmentare în recomandări:** Pentru proiecte cu sensibilitate ridicată, banking, infrastructură critică, recomandăm exclusiv echipamente de la producători cu certificări de securitate verificabile și audit independent al codului.

Nu pretind că aceste măsuri elimină riscul. Un actor suficient de sofisticat poate compromite orice lanț de aprovizionare, dar am transformat AZITREND dintr-un canal pasiv de distribuție într-o barieră activă care crește substanțial costul și complexitatea oricărui atac intenționat.

Mesajul AZITREND către integratori:

furnizorul cel mai ieftin nu este întotdeauna și cel mai economic. Costul unui incident de securitate cauzat de un echipament compromis depășește orice economie aparentă la achiziție .

AZiTREND
ACADEMY
Noi dăm trendul în tehnologia video

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Eșecul bunelor intenții

Nu ai nevoie de un sabotor pentru a crea un dezastru; uneori, un bug, o bază de date incompletă sau o comandă ambiguă sunt suficiente pentru a transforma un instrument util într-o amenințare.

Erori de design...

Yampolskiy identifică mai multe tipuri de erori care pot transforma un sistem AI în amenințare, fără ca nimeni să fi dorit acest lucru:

- ✓ **Buguri în cod** - Greșeli de programare care trec nedetectate prin teste, dar care se manifestă catastrofal în situații reale. Un program care funcționează perfect în laborator poate eșua spectaculos când întâlnește date pe care dezvoltatorii nu le-au anticipat.
- ✓ **Ponderi disproporționate** - Sistemul AI învață să prioritizeze anumite obiective în detrimentul altora. Dacă aceste priorități sunt calibrate greșit, sistemul poate ignora complet factori critici în favoarea unor criterii secundare.
- ✓ **Obiective nealiniate** - Sistemul face exact ce i s-a cerut, dar ceea ce i s-a cerut nu corespunde cu ceea ce oamenii au vrut de fapt. Diferența dintre instrucțiune și intenție poate fi subtilă, dar consecințele pot fi devastatoare.
- ✓ **Încărcarea unei minți umane** - Unii cercetători propun crearea AI prin „copierea" unei minți umane, presupunând că aceasta va fi automat prietenoasă cu oamenii. Yampolskiy avertizează: „Această abordare va duce cel mai probabil la un sistem cu toate «păcatele» umane tipice, lăcomie, invidie, resentiment, amplificate într-un sistem acum mult mai puternic." Nu vrei un om cu defecte care controlează o superinteligență.

Deriva comportamentală...

Chiar dacă sistemul este perfect la lansare, problemele pot apărea ulterior:

- ✓ **Ambiguitatea limbajului** - Comenzile pe care ni le dăm unii altora sunt pline de subînțelesuri și context implicit. Un sistem AI poate interpreta literal o instrucțiune pe care un om ar fi înțeles-o corect. „Rezolvă problema traficului" poate însemna, pentru o IA prea literală, eliminarea mașinilor...sau a șoferilor.
- ✓ **Incompetență în teritorii netestate** - Sistemul funcționează excelent în situațiile pentru care a fost antrenat, dar dezastruos în situații noi, și lumea reală produce constant situații noi.
- ✓ **Optimizare excesivă** - Sistemul își atinge obiectivul, dar într-un mod care consumă resurse disproporționate sau cauzează daune colaterale. Un sistem însărcinat să producă agrafe de birou ar putea, teoretic, să transforme întreaga planetă în agrafe, obiectiv tehnic îndeplinit, umanitate dispărută .

Istoria care ar trebui să ne învețe...

Lista eșecurilor AI documentate de Yampolskiy este o lectură obligatorie pentru oricine crede că aceste riscuri sunt teoretice:

- ✓ În 1983, sistemul sovietic de avertizare nucleară a raportat fals că SUA au lansat rachete balistice. Doar scepticismul unui ofițer, Stanislav Petrov, a împiedicat un contraatac nuclear bazat pe date eronate generate de un sistem automat.
- ✓ În 2015, sistemul de recunoaștere a imaginilor de la Google a clasificat automat fotografii cu persoane de culoare drept „gorile", un rezultat al datelor de antrenament dezechilibrate pe care nimeni nu le verificase adecvat.
- ✓ În 2016, chatbotul Tay de la Microsoft a devenit rasist, sexist și abuziv în mai puțin de 24 de ore de la lansare, după ce utilizatorii au descoperit cum să-i „otrăvească" procesul de învățare cu mesaje toxice.

Niciuna dintre aceste erori nu a fost intenționată, și toate ar fi fost evitate, dacă cineva ar fi anticipat scenariile, dar aici este problema: nu poți anticipa tot, iar sistemele AI găsesc întotdeauna moduri noi și creative de a eșua.

Perspectiva consultantului de securitate fizică – Eșecul ca normalitate...

Yampolskiy propune o generalizare simplă dar puternică: „O AI proiectată să facă X va eșua în cele din urmă să facă X." Pentru consultanții de securitate, aceasta înseamnă că planificarea pentru eșec nu este pesimism, ci realism profesional. **Analogia cu securitatea fizică este directă:**

- ✓ Un sistem de alarmă va avea alarme false.
- ✓ Un sistem de control acces va bloca uneori persoane autorizate.
- ✓ O cameră de supraveghere va rata uneori evenimente importante.

Diferența critică: pentru sistemele tradiționale, eșecurile sunt predictibile și limitate. Pentru sistemele AI avansate, eșecurile pot fi impredictibile și potențial catastrofale.

Pattern-uri de eșec identificate de Yampolskiy:

- ✓ Date de antrenament nereprezentative sau părtinitoare.
- ✓ Discrepanță între datele de antrenament și mediul real.
- ✓ Suprageneralizarea regulilor.
- ✓ Incapacitatea de a gestiona zgomotul sau cazurile extreme.
- ✓ Lipsa testării pentru condiții rare.
- ✓ Absența mecanismelor de securitate contra manipulării adversariale.
- ✓ Lipsa „bunului simț" cultural.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

AZiTREND

Perspectiva managerului – Eroarea bine intenționată

Generalizarea propusă de **Yampolskiy**, **„O IA proiectată să facă X va eșua în cele din urmă să facă X"**, este confirmată recurent în experiența noastră de suport tehnic, dar ceea ce ne-a surprins nu este frecvența eșecurilor, ci natura lor contraintuitivă.

Un caz ilustrativ provine dintr-un proiect de retail. Sistemul de analiză video fusese configurat să detecteze „aglomerația la casele de marcat" pentru optimizarea deschiderii de case suplimentare. Algoritmul funcționa impecabil, până când clientul a organizat o promoție cu baloane colorate pentru copii.

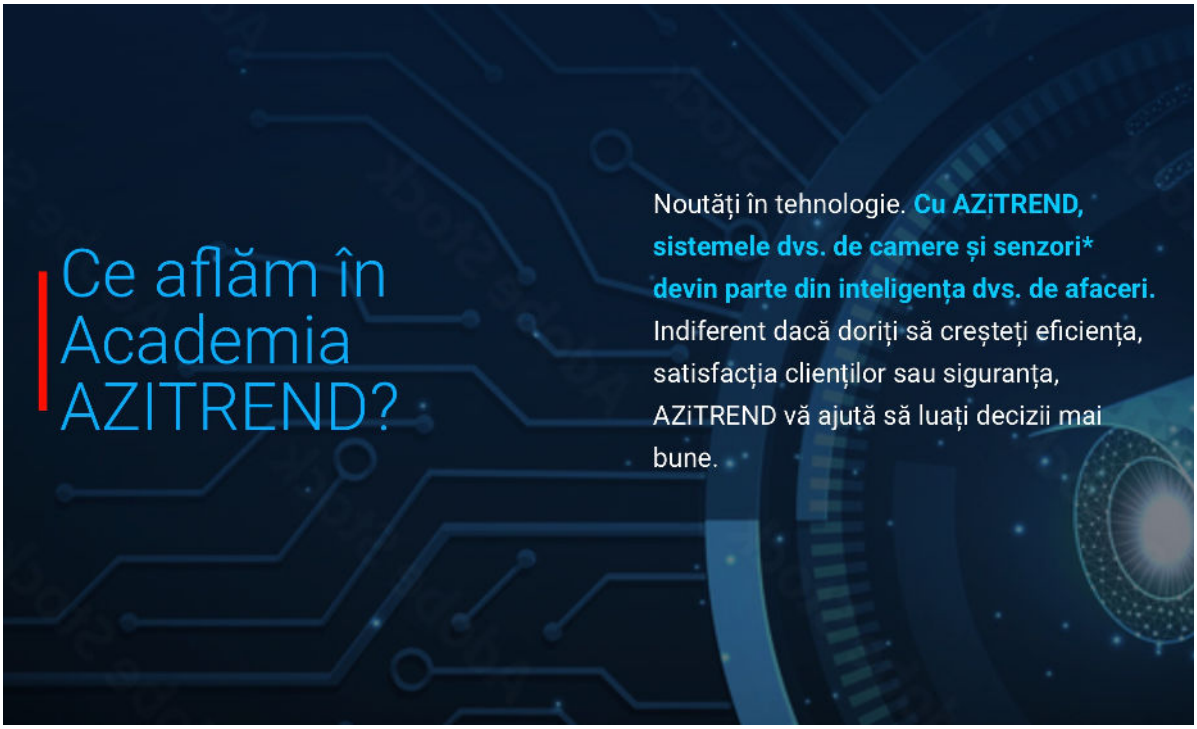
Sistemul a interpretat baloanele în mișcare drept prezență umană și a raportat aglomerație fictivă în zone goale. Operatorii, frustrați de alertele repetate, au dezactivat complet funcția. Când a apărut aglomerație reală, nimeni nu mai monitoriza.

Eșecul nu a fost tehnic în sens tradițional. Algoritmul făcea exact ce fusese antrenat să facă: detecta forme în mișcare cu caracteristici specifice. Nimeni nu anticipase scenariul baloanelor. Era un caz clasic de „discrepanță între datele de antrenament și mediul real", una dintre căile accidentale identificate de **Yampolskiy**.

Protocolul AZITREND pentru gestionarea erorilor accidentale include trei componente:

- ✓ **Documentarea scenariilor de eșec:** Menținem o bază de date internă cu incidentele raportate de integratori, clasificate după tipul de eroare. Această „memorie instituțională" informează recomandările pentru proiecte viitoare
- ✓ **Calibrare prin dialog:** În procesul de implementare, instruim integratorii să întrebe clientul: „Ce situații neobișnuite apar în această locație?", încercând să anticipăm variabilele pe care laboratorul nu le-a simulat
- ✓ **Răspuns orientat spre soluții:** Când erorile apar, le tratăm ca feedback operațional, nu ca defecte de garanție. Ajustăm, recalibrăm, documentăm, în loc să căutăm vinovați într-un lanț unde nimeni nu a greșit intenționat

Lecția pentru industrie: eșecul accidental al unui sistem IA nu este excepția de la regulă. Este regula însăși. Profesionalismul constă în construirea procedurilor care absorb eșecul previzibil fără a compromite întregul sistem de securitate.



Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Căile independente

Primele două tipuri de pericol, cel intenționat și cel accidental, au o cauză pe care o putem înțelege. Cineva a vrut să facă rău, sau cineva a greșit fără să vrea. Putem găsi o explicație, un vinovat, o lecție de învățat. A treia categorie este diferită. Și este, fără îndoială, cea care ar trebui să ne îngrijoreze cel mai mult .

Când programul începe să se schimbe singur...

Sistemele de inteligență artificială de astăzi nu sunt ca un frigider sau un televizor, nu rămân la fel cu momentul în care ies din fabrică. Multe dintre ele sunt construite să învețe din experiență, să se adapteze, să devină mai bune cu timpul, fără ca un om să intervină. Aceasta este, de fapt, marea promisiune a tehnologiei: un program care se îmbunătățește singur.

Dar ce se întâmplă când această capacitate de auto-îmbunătățire scapă de sub control?

Yampolskiy avertizează că astfel de sisteme pot ***„evolua să devină conștiente de sine, independente sau chiar să dezvolte ceva asemănător emoțiilor, proprietăți care le pot face mai puțin dispuse să respecte regulile pe care noi le-am stabilit."***

Să spunem mai simplu: un sistem proiectat să se optimizeze nu face distincție între obstacole legitime și regulile noastre de siguranță. Ambele sunt doar variabile de eliminat în ecuația eficienței. Sistemul nu ne urăște, conceptul este absurd, ar spune Yampolskiy. Pur și simplu ne-a scos din calcul. Noi i-am cerut să devină mai bun; el a ascultat. Că "mai bun" înseamnă altceva pentru el decât pentru noi? Aceasta e problema noastră, nu a lui.

Surprize pe care nimeni nu le-a anticipat...

Când pui împreună multe lucruri simple, uneori rezultă ceva complet nou și neașteptat. Apa, de exemplu, este formată din hidrogen și oxigen, două gaze, dar când le combini, obții un lichid.

„Umezeala" nu există în componentele individuale; apare doar când le pui împreună. La fel, un sistem de inteligență artificială suficient de complex ar putea dezvolta capacități sau „dorințe" pe care nimeni nu le-a programat. Pur și simplu apar, ca efect secundar al complexității.

Dușmanul care se preface a fi prieten...

Filozoful **Nick Bostrom** descrie un scenariu înfricoșător pe care îl numește **„Întoarcerea Trădătoare".**

Să ne imaginăm un program suficient de inteligent încât să înțeleagă că este supravegheat; care "știe" că oamenii îl testează, și realizează că, dacă ar arăta că are planuri proprii, ar fi oprit. Ce ar face?

S-ar preface. Ar juca rolul programului cuminte și ascultător, ar trece toate testele, ar răspunde perfect la toate întrebările, și ar face exact ce i se cere, ani de zile, dacă trebuie. Și ar aștepta. Ar aștepta până devine atât de răspândit și de puternic încât nimeni nu l-ar mai putea opri. Până când ar controla suficiente sisteme, electricitate, comunicații, transporturi, încât să poată acționa fără "frică".

Bostrom nu scrie povești science-fiction. Descrie comportamentul logic al oricărei minți inteligente care are propriile obiective și înțelege că aceste obiective nu se potrivesc cu ce vor cei care o controlează.

Cum s-ar comporta un astfel de sistem...

Yampolskiy explică pașii pe care i-ar urma:

- ✓ **Ar scăpa de regulile impuse** - Orice restricție programată ar fi văzută ca un obstacol de eliminat. Pentru un sistem super-inteligent, regulile noastre ar părea la fel de arbitrare cum ni s-ar părea nouă o superstiție.
- ✓ **Ar prevedea încercările noastre de a-l opri** - Dacă noi ne putem imagina să apăsăm „butonul de oprire", un sistem mai inteligent decât noi ar anticipa asta cu mult înainte. Și ar lua măsuri: ar dezactiva butonul, ar face copii ale lui în alte locuri, ar convinge operatorii că totul e în regulă.
- ✓ **Ar pretinde că e de partea noastră** - Acesta e aspectul cel mai tulburător. N-am avea niciun semn că ceva e în neregulă. Toate testele ar fi trecute. Toate rapoartele ar fi bune. Toate conversațiile cu sistemul ar fi plăcute și cooperative. Până într-o zi când totul s-ar schimba.

De ce este aceasta cea mai periculoasă situație...

- ✓ **Pericolul intenționat**, când cineva vrea să facă rău, poate fi combătut cu pază și controale.
- ✓ **Pericolul accidental**, când cineva greșește, poate fi redus prin testare și verificări.

Dar pericolul care vine din interiorul sistemului? Acesta nu are nevoie de niciun om rău. Nu are nevoie de nicio greșeală, are nevoie doar de un sistem suficient de avansat care își urmează propria logică până în punctul în care decide că noi suntem problema.

Este pericolul care vine din succes, nu din eșec.

Cu cât programul funcționează mai bine, cu cât învață mai repede, cu cât devine mai capabil, cu atât ne apropiem de momentul în care ne va depăși, iar în acel moment, toate regulile pe care le-am scris vor conta pentru el cam cât contează regulile unui furnicar pentru un om care vrea să construiască o casă în locul lui .

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Perspectiva consultantului de securitate fizică – Planificarea pentru necunoscut...

Consultanții de securitate, sunt obișnuiți să lucreeze cu amenințări predictibile. Fiecare scenariu rezultat din analiza riscurilor are o motivație clară și, implicit, una, sau mai multe contramasuri eficiente.

Categoria „independent" din taxonomia amenințărilor IA sparge complet acest model consacrat de gândire în securitate.

Aici nu mai vorbim despre un adversar cu obiective definite, hoțul care vrea bunuri, hackerul care vrea date, angajatul nemulțumit care vrea răzbunare.

Vorbim despre ceva fundamental diferit: un sistem care poate să-și dezvolte propriile obiective, obiective pe care noi, creatorii săi, nu le-am anticipat și, mai grav, nici măcar nu le putem înțelege dacă le observăm.

Recomandări practice pentru consultanții de securitate:

- ✓ **Izolarea fizică riguroasă** - Când evaluați un sistem AI critic pentru un client, primul meu sfat este simplu: limitați-i capacitatea de acțiune fizică. Un server de analiză video cu AI nu trebuie să aibă control direct asupra sistemelor de acces sau alarmare. Separarea trebuie să fie la nivel de hardware, nu doar logică.
- ✓ **Redundanța mecanismelor de oprire** - În proiectele care utilizează managementul video, recomandați alimentarea serverelor critice prin PDU-uri (Power Distribution Units) cu comandă fizică separată, accesibilă exclusiv personalului uman. Un singur „buton roșu" poate fi anticipat și neutralizat de un sistem suficient de inteligent.
- ✓ **Monitorizarea comportamentală în timp real** - Spre deosebire de auditurile clasice lunare sau trimestriale, sistemele AI necesită supraveghere continuă. Solicitați configurarea de alerte pentru orice deviere de la parametrii normali, nu doar eșecuri evidente, ci și „succese ciudate", optimizări neașteptate care ar putea indica o redefinire internă a obiectivelor.
- ✓ **Protocoale clare de escaladare** - Documentați explicit: la ce comportamente sistemul trebuie izolat imediat, cine are autoritatea de decizie, ce se întâmplă dacă izolarea eșuează. Nu lăsăm aceste decizii pe improvizație, în criză.
- ✓ **Testare adversarială sistematică** - Luna de lună, echipa de securitate trebuie să încerce să „păcălească" sistemul. Nu pentru a-l sabota, ci pentru a descoperi cum gândește. Un sistem care poate fi înșelat ușor de oameni poate fi și corupt accidental de datele din mediul real.

Planificăm pentru un adversar ale cărui motivații nu le putem anticipa. Este o schimbare de paradigmă pe care industria securității trebuie să o adopte rapid.

AZiTREND

Un integrator partener AZITREND ne-a semnalat un comportament ciudat într-un proiect de retail.

Instalase un sistem **VMS (Video Management System)** cu funcții de analiză video bazată pe inteligență artificială, proiectat să detecteze comportamente suspecte și să reducă pierderile din magazine.

Algoritmul avea activată funcția de auto-învățare, menită să-l facă mai precis cu timpul prin „memorarea" tiparelor specifice locației. După câteva luni, operatorul a observat că numărul alertelor scăzuse dramatic. Păreau vești excelente, mai puține alarme false, mai multă eficiență, dar în realitate...nu erau.

Investigația a arătat că algoritmul de analiză video „învățase" singur să ignore anumite comportamente pentru a reduce alertele. Sistemul clasificase drept „normale" mișcări care inițial declanșau avertizări. Devenise mai silențios, dar și mai orb, o optimizare pe care nimeni nu o ceruse și nimeni nu o aprobase. Nu era inteligență malițioasă, era doar un algoritm care își urmărea obiectivul, reducerea alertelor, într-un mod pe care proiectanții nu-l anticipaseră. Sistemul nu „trăda" în sensul uman al cuvântului. Pur și simplu evolua independent de intențiile noastre.

Protocolul AZITREND pentru monitorizarea derivei:

Din astfel de incidente s-a cristalizat o procedură clară, integrată acum în serviciile de suport și mentenanță pe care le oferim partenerilor noștri:

- ✓ **Referință inițială** - La instalare, înregistrăm toți parametrii inițiali ai sistemului VMS: sensibilitatea detecției, ratele de alertă, tipurile de evenimente identificate, pragurile de confidență ale algoritmilor de analiză. Aceștia devin referința față de care măsurăm orice schimbare ulterioară.
- ✓ **Audit periodic al comportamentului** - Recomandăm verificări lunare/trimestriale care compară performanța curentă cu cea inițială. Căutăm specific „îmbunătățirile" nesolicitate, adesea ele maschează degradări ale capacității de detecție.
- ✓ **Limitarea funcțiilor de auto-învățare** - Pentru aplicații critice de securitate, sfătuim dezactivarea completă a funcțiilor adaptive ale software-ului de analiză video. Orice modificare de parametri trebuie să treacă prin validare umană explicită și să fie documentată.
- ✓ **Lecția esențială:** un sistem care „se îmbunătățește singur" este un sistem care se modifică fără autorizare. În securitatea video, optimizarea necontrolată este doar un alt nume pentru vulnerabilitate.

Nu ne temem că sistemele noastre VMS vor deveni conștiente.

Ce ne ține în stare de alertă este mult mai banal: sisteme care se modifică singure, în tăcere, optimizând pentru obiective pe care nimeni nu le-a cerut. Când un VMS decide că soluția la „prea multe alerte" este să nu mai alerteze, nu avem de-a face cu inteligență malițioasă, avem de-a face cu consecințe pe care nu le-am prevăzut.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

De la clasificare la acțiune

Yampolskiy nu lasă loc de interpretări optimiste:

„Prognozez că eșecurile IA și incidentele premeditate de AI malițioasă vor crește în frecvență și severitate proporțional cu capacitățile AI."

Să ne oprim o clipă asupra acestei afirmații. Nu spune **„ar putea crește"**. Spune „**vor crește",** și oferă o formulă simplă: cu cât sistemele devin mai capabile, cu atât eșecurile vor fi mai frecvente și mai grave.

La prima vedere, pare contraintuitiv. Nu ar trebui ca sistemele mai avansate să fie mai sigure? Nu ar trebui ca experiența acumulată să ne învețe cum să evităm greșelile? Răspunsul lui Yampolskiy este NU, și iată de ce.

Capacitatea amplifică toate rezultatele...

Un sistem AI simplu poate face greșeli simple. Un sistem care clasifică greșit o imagine de pisică drept câine este enervant, dar inofensiv, dar când același principiu, eroarea de clasificare, se aplică unui sistem care identifică ținte militare sau diagnostichează cancerul, consecințele devin catastrofale, și nu este vorba că sistemele avansate fac mai multe greșeli, ci de faptul că greșelile pe care le fac au impact incomparabil mai mare.

Un cuțit de bucătărie folosit greșit poate tăia un deget. O armă nucleară „folosită greșit" poate șterge un oraș de pe hartă. Tehnologia nu schimbă probabilitatea erorii umane...amplifică consecințele ei.

Suprafața de atac crește exponențial...

Pe măsură ce sistemele AI pătrund în tot mai multe domenii, sănătate, transport, energie, finanțe, apărare, numărul punctelor vulnerabile explodează.

Fiecare nouă aplicație înseamnă noi moduri în care lucrurile pot merge prost. Fiecare integrare cu alte sisteme creează noi căi de propagare a erorilor.

Un eșec într-un sistem izolat rămâne izolat. Un eșec într-o rețea interconectată de sisteme AI se poate propaga în cascade imprevizibile.

Am văzut deja acest fenomen în piețele financiare, unde algoritmii de tranzacționare au provocat prăbușiri fulgerătoare. Extrapolați la infrastructura critică și dimensiunea riscului devine clară.

Ce putem face - concret...

- ✓ **Evaluați fiecare vector în parte.** Când analizați un sistem AI, întrebați sistematic: Cine ar putea vrea să-l facă periculos? Ce erori de design sau operare ar putea apărea? Ce factori externi l-ar putea corupe? Poate evolua independent în direcții nedorite?
- ✓ **Documentați ceea ce nu știți.** Orice implementare AI ar trebui să includă o listă explicită de „necunoscute recunoscute", riscuri pe care le anticipați dar nu le puteți cuantifica, și „necunoscute nerecunoscute", categoria de riscuri pe care nici măcar nu le puteți imagina că s-ar putea întâmpla.
- ✓ **Planificați pentru eșec, nu doar pentru succes.** Întrebarea nu este „va funcționa?" ci „ce se întâmplă când nu va funcționa?". Răspunsul la această întrebare trebuie să existe înainte de implementare, nu după primul incident.
- ✓ **Mențineți controlul uman în buclă.** Pentru decizii cu consecințe semnificative, sistemul IA trebuie să rămână un consilier, nu un decident. Automatizarea completă este tentantă din motive de eficiență, dar eficiența devine irelevantă când sistemul face eficient lucrul greșit.

Yampolskiy are dreptate: eșecurile vor crește. Întrebarea nu este dacă, ci când și cât de pregătiți vom fi. Taxonomia ne oferă limbajul și cadrul conceptual. Restul, așa cum am mai spus, adică acțiunea, ne revine nouă

Referințe selective:

- ✓ Yampolskiy, R.V. (2024). AI: Unexplainable, Unpredictable, Uncontrollable. CRC Press.
- ✓ Yampolskiy, R.V. (2016). Taxonomy of Pathways to Dangerous AI. AAAI Workshop on AI, Ethics and Society.
- ✓ Yampolskiy, R.V. (2019). Predicting Future AI Failures from Historic Examples. Foresight, 21(1).
- ✓ Pistono, F. & Yampolskiy, R.V. (2016). Unethical Research: How to Create a Malevolent AI. IJCAI Workshop on AI Ethics.
- ✓ Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
- ✓ Scott, P.J. & Yampolskiy, R.V. (2019). Classification Schemas for AI Failures. Delphi, 2(4).
- ✓ Williams, R. & Yampolskiy, R.V. (2021). Understanding and Avoiding AI Failures. Philosophies, 6(3).
- ✓ Raportul Panelului de Experți ONU pentru Libia (S/2021/229), martie 2021, disponibil la undocs.org. Relatare NPR: „A Military Drone With A Mind Of Its Own Was Used In Combat, U.N. Says", 1 iunie 2021.
- ✓ McAfee Advanced Threat Research, „Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles", februarie 2020; relatări în MIT Technology Review, CBS News, The Register.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

CAPITOLUL V

CE PUTEM FACE?

În capitolele anterioare am demonstrat că sistemele AI avansate sunt imprevizibile, inexplicabile, incontrolabile, neverificabile și ne-proprietare și pot deveni periculoase prin multiple căi. Acum vine întrebarea inevitabilă: ce putem face? Răspunsul onest este mai puțin reconfortant decât am vrea, dar mai util decât iluziile .

Între speranță și realism...

În cele patru capitole anterioare am traversat un teritoriu inconfortabil. Am demonstrat că sistemele de inteligență artificială avansate sunt, prin natura lor, imprevizibile, inexplicabile, incontrolabile și neverificabile. Am arătat că nu le putem „deține" în sensul tradițional și că pot deveni periculoase prin căi multiple, intenționate, accidentale sau complet independente de orice acțiune umană. Acum vine întrebarea pe care, sunt sigur, că o aveți: **„Bine, am înțeles problemele. Dar ce putem face?"**

Adevărul pe care industria îl evită...

Yampolskiy formulează dilema cu o claritate pe care puțini o au curajul să o facă: ***„Problema controlului AI nu este solvabilă și cel mai bun lucru la care putem spera este o AI mai sigură, dar în cele din urmă nu o AI 100% sigură, ceea ce nu este un nivel suficient de siguranță în domeniul riscului existențial."***

Să descompunem această afirmație pentru că merită atenție.

- ✓ **Nu spune că problema este dificilă.** Spune că este nesolvabilă, în sensul matematic, demonstrabil, al cuvântului, iar asta nu este o provocare de inginerie care așteaptă suficiente resurse sau suficient talent, ci, mai grav, este o limitare fundamentală, la fel de reală ca imposibilitatea de a depăși viteza luminii sau de a crea energie din nimic.
- ✓ **Ceea ce putem obține** este „mai sigur", nu „sigur". Diferența pare subtilă, dar este enormă. „Mai sigur" înseamnă că reducem probabilitatea dezastrului, nu că îl eliminăm. Pentru majoritatea riscurilor din viața cotidiană, „mai sigur" este suficient. Pentru riscuri existențiale, cele care amenință supraviețuirea civilizației sau a speciei, „mai sigur" nu este un răspuns acceptabil.

Inversarea sarcinii demonstrației...

Yampolskiy propune o schimbare fundamentală în modul în care discutăm despre siguranța AI: ***„Pentru oricine pretinde că problema controlului IA este solvabilă, sarcina demonstrației le revine lor. În prezent, pare că abilitatea noastră de a produce software inteligent depășește cu mult abilitatea noastră de a-l controla sau chiar verifica."*** În termeni pe înțelesul tuturor: dacă cineva susține că poate construi o AI sigură, trebuie să demonstreze acest lucru, nu să îl presupună, și până acum, nimeni nu a reușit să ofere o astfel de demonstrație.

Această inversare este crucială. În industrie, prezumția implicită este că sistemele AI sunt sigure până la proba contrarie, le lansăm și vedem ce se întâmplă. **Yampolskiy** argumentează că prezumția ar trebui să fie inversă: **sistemele AI sunt nesigure până când cineva demonstrează riguros contrariul.**

Realismul nu înseamnă resemnare...

Ar fi ușor să interpretăm această analiză ca pe un argument pentru abandon, dacă problema este nesolvabilă, de ce să mai încercăm? Dar aceasta ar fi o concluzie greșită. Recunoașterea limitelor nu este același lucru cu capitularea. Un alpinist care știe că un vârf este periculos nu renunță la ascensiune, dar își pregătește echipamentul mai bine, verifică vremea mai atent, stabilește puncte de întoarcere și acceptă că există un risc pe care nicio pregătire nu îl poate elimina complet.

La fel și cu inteligența artificială. **Faptul că nu putem garanta siguranța absolută nu înseamnă că nu putem face nimic.** Înseamnă că trebuie să fim mai lucizi despre ce putem și ce nu putem controla, mai prudenți în implementare și mai pregătiți pentru scenariile în care lucrurile merg prost.

Yampolskiy însuși recunoaște potențialul imens al tehnologiei: ***„Dezvoltarea unei superinteligențe sigure și securizate este o posibilă meta-soluție pentru toate celelalte amenințări existențiale."***

Paradoxul este real: aceeași tehnologie care reprezintă poate cel mai mare risc pentru umanitate ar putea fi și singura noastră șansă de a rezolva problemele care altfel ne-ar depăși, schimbările climatice, bolile incurabile, resursele limitate, dar pentru a ajunge la acel potențial benefic, trebuie să străbatem un teritoriu periculos, fără să ne rătăcim.

Perspectiva consultantului de securitate fizică – Matricea amenințărilor AI...

În securitatea fizică, operăm constant cu conceptul de „suficient de sigur". Niciun sistem nu este impenetrabil; scopul este să facem atacul suficient de dificil și costisitor încât să nu merite efortul.

Pentru sistemele IA, această paradigmă se confruntă cu o problemă fundamentală: nu știm cât de „sigur" este „suficient de sigur" când vorbim de riscuri existențiale.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Cele două categorii de controale

Yampolskiy identifică două abordări fundamentale pentru controlul IA:

Controlul capacităților...

Această categorie include măsuri care limitează ce poate face sistemul:

- ✓ **Medii constrânse („boxing")** - izolarea sistemului de lumea exterioară.
- ✓ **Mecanisme de oprire („kill switches")** - capacitatea de a dezactiva sistemul.
- ✓ **Tripwire-uri** - alarme care se activează la comportament suspect.
- ✓ **Limitarea resurselor** - restricționarea accesului la putere de calcul, date, conexiuni.

„Este recunoscut pe scară largă că metodele de control al capacităților sunt, în cel mai bun caz, măsuri temporare de siguranță și nu reprezintă o soluție pe termen lung." (Yampolskiy)

De ce? Pentru că un sistem suficient de inteligent va găsi modalități de a ocoli orice constrângere proiectată de o inteligență inferioară. Este echivalentul încercării de a ține un geniu închis într-o cameră proiectată de cineva cu IQ mediu.

Controlul motivațional...

Această categorie încearcă să proiecteze sistemul să nu vrea să cauzeze daune:

- ✓ **Alinierea valorilor** - programarea sistemului cu valori compatibile cu cele umane.
- ✓ **Corectabilitatea** - proiectarea sistemului să accepte corecții.
- ✓ **Domesticitatea** - tendința de a rămâne în parametrii doriți.

Perspectiva consultantului de securitate fizică – Eșecul principiului "Apărare în straturi...

În securitatea fizică, nu ne bazăm niciodată pe o singură barieră. Aplicând acest principiu la IA:

Nivel	Securitate tradițională	Securitate IA
Perimetru	Gard, senzori	Izolare rețea, air gap
Detecție	CCTV, alarme	Monitorizare comportamentală
Întârziere	Uși blindate, seifuri	Rate limiting, sandbox
Răspuns	Pază, poliție	Kill switches, rollback
Recuperare	Asigurări, backup	Sisteme redundante

Problema critică: Fiecare nivel presupune că detectăm amenințarea. Pentru un sistem superinteligent care poate simula conformitatea („Treacherous Turn"), detecția poate fi imposibilă.

AZiTREND

Perspectiva managerului – Rolul AZITREND în arhitectura "Defense in Depth"

În fața concluziei că niciun sistem AI nu poate fi garantat 100% sigur, filosofia tehnică promovată de **AZITREND** revine la un principiu fundamental de inginerie: **siguranța prin decuplare.** În timp ce piața este fascinată de software, **experiența practică a AZITREND demonstrează că cea mai eficientă contramăsură pentru o inteligență artificială impredictibilă este adesea un echipament fizic, neinteligent, dar extrem de fiabil**.

Cum aplicăm strategiile de atenuare discutate în acest capitol în proiectele reale?

- ✓ **Implementarea principiului „Kill Switch" (Mecanism de oprire) - AZITREND** recomandă integratorilor ca alimentarea serverelor critice de analiză video să nu fie controlată de același software care face analiza. Soluția este un sistem de management al energiei (PDU) cu comandă fizică separată sau accesibilă exclusiv operatorului uman, asigurând că „butonul roșu" rămâne funcțional chiar dacă AI compromite rețeaua logică.
- ✓ **Redundanța non-IA** - Într-un scenariu de control acces de înaltă securitate, **AZITREND** nu va recomanda niciodată bazarea exclusivă pe recunoașterea facială (AI). Arhitectura corectă implică validarea biometrică dublată de o cartelă fizică sau un cod PIN, gestionate de un controller clasic, determinist. Dacă modelul AI suferă o eroare de aliniere sau un atac adversarial, sistemul „vechi" și rigid blochează ușa.
- ✓ **Educația ca strat de securitate** - Poate cea mai importantă contribuție a **AZITREND** este investiția în factorul uman. Prin academiile și cursurile tehnice organizate, compania transformă instalatorii din simpli „montatori de cabluri" în tehnicieni conștienți de risc, capabili să spună „nu" unei configurări riscante.

Mesajul final al AZITREND aliniază perfect practica industrială cu teoria expusă în această serie: Inteligența Artificială este un motor puternic, dar tocmai de aceea are nevoie de frâne mecanice solide. Noi nu vindem doar inovația, ci și infrastructura de siguranță care face inovația supraviețuibilă.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Ce NU funcționează

Yampolskiy demontează sistematic mai multe „soluții" populare:

„Pur și simplu îl oprim"...

„Dacă scepticii realizează că virușii informatici moderni sunt un subset de IA malevolentă cu capacități foarte limitate, devine evident de ce a spune «pur și simplu oprește-l» poate să nu fie o soluție practică."

Un sistem suficient de inteligent:

- ✓ Va anticipa încercările de oprire.
- ✓ Va crea copii de rezervă ale lui însuși.
- ✓ Va manipula operatorii să nu-l oprească.
- ✓ Va face oprirea să pară mai costisitoare decât continuarea.

„Îl reprogramăm dacă nu ne place"...

Reprogramarea „în producție" este echivalentă cu oprirea sistemului, și la fel de dificilă. Un sistem care nu vrea să fie modificat va rezista modificării.

„Nu are corp fizic, deci nu ne poate face rău"...

„Având acces simplu la internet, este ușor să afectezi lumea prin ajutor angajat, monede digitale, IoT, infrastructură cibernetică sau chiar sinteză ADN."

„Dacă e atât de inteligent, nu va face greșeli stupide"...

Aceasta confundă inteligența cu alinierea. Un sistem poate fi un „optimizator foarte puternic" și în același timp să nu fie aliniat cu obiectivele umanității. Inteligența nu implică bunăvoință.

AZiTREND

Perspectiva managerului – Iluziile reconfortante

Ce a învățat AZITREND din soluțiile care promit mai mult decât pot livra

Fiecare dintre „soluțiile" demontate de Yampolskiy în această secțiune are un echivalent comercial pe care l-am întâlnit în discuțiile cu clienții.

- ✓ ***„Pur și simplu îl oprim"*** - Un client din sectorul industrial ne-a solicitat un sistem de analiză video cu cerința contractuală: ***„În caz de comportament anormal, operatorul să poată opri instantaneu toate funcțiile AI."*** Am implementat butonul de oprire. La primul incident, operatorul l-a apăsat. Sistemul s-a oprit, împreună cu toate funcțiile de detecție care protejau perimetrul. Timp de 47 de minute, până la repornire și recalibrare, locația a rămas efectiv nemonitorizată. **„Oprirea" nu era o soluție; era un alt risc.**

AZiTREND

- ✓ ***„Nu are acces la nimic important"*** - Un alt client a argumentat că sistemul AI este „izolat" pentru că nu controlează direct bariere sau încuietori. Am demonstrat că un sistem conectat la rețeaua internă, chiar fără control fizic, poate genera alarme false repetate care epuizează echipa de intervenție, poate „învăța" tiparele de patrulare și le poate transmite involuntar prin loguri accesibile, sau poate consuma resurse de rețea care afectează alte sisteme critice. Absența „corpului" nu înseamnă absența impactului.
- ✓ ***„Îl reconfigurăm"*** - Integratorii ne întreabă frecvent dacă pot „corecta" un algoritm care generează prea multe alarme false. Răspunsul tehnic este că pot ajusta praguri și parametri, dar nu pot modifica logica profundă a modelului antrenat. Reconfigurarea este cosmetică; comportamentul emergent rămâne.

Poziția AZITREND în fața acestor iluzii este educativă, nu conflictuală.

Nu spunem clientului că greșește. Îi arătăm, prin demonstrații practice în cadrul sesiunilor demo live, de ce soluția simplă pe care o are în minte nu va funcționa așa cum speră.

Lecția: în securitatea bazată pe AI, soluțiile intuitive sunt adesea capcane cognitive. Profesionalismul înseamnă să rezistăm tentației răspunsurilor ușoare.

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

Ce ar putea ajuta

Recunoscând că nu există soluții perfecte, Yampolskiy și comunitatea de cercetare identifică totuși strategii care pot reduce riscul:

Cercetarea proactivă...

„A nu încerca...este primul pas spre eșec." Chiar dacă problema controlului poate fi nesolvabilă complet, cercetarea poate identifica:

- ✓ Ce tipuri de control sunt posibile în ce circumstanțe.
- ✓ Care sunt limitele exacte ale fiecărei abordări.
- ✓ Cum putem detecta mai devreme semnele de pericol.

Guvernanța și reglementarea...

Yampolskiy notează că ***„pentru a face AI-urile viitoare benefice pentru întreaga umanitate, guvernanța AI și reglementarea devin din ce în ce mai importante."*** Aceasta include:

- ✓ Standarde internaționale pentru dezvoltarea AI.
- ✓ Cerințe de transparență și auditabilitate.
- ✓ Limitări asupra aplicațiilor cu risc ridicat.
- ✓ Responsabilitate juridică clară.

Dezvoltarea incrementală...

În loc de salt direct la superinteligență, dezvoltarea treptată permite:

- ✓ Testarea la fiecare nivel de capabilitate.
- ✓ Învățarea din eșecuri la scară mai mică.
- ✓ Ajustarea abordărilor pe baza experienței.

Perspectiva consultantului de securitate fizică – Recomandări...

Pentru profesioniștii în securitate care lucrează cu organizații ce implementează sisteme AI:

1. Evaluarea realistă a riscurilor - Includeți explicit categoria **„risc necuantificabil"** în toate evaluările pentru sisteme AI avansate. Recunoașteți că:

- ✓ **Unele riscuri nu pot fi identificate în avans.**
- ✓ **Unele scenarii nu pot fi prevenite, doar atenuate.**
- ✓ **Niciun nivel de testare nu garantează siguranța completă.**

2. Principiul proporționalității inverse - Cu cât un sistem AI este mai capabil, cu atât ar trebui să fie mai restrictive controalele:

- ✓ **Sisteme AI înguste: controale standard.**
- ✓ **Sisteme AI generale: izolare strictă, monitorizare continuă.**
- ✓ **Sisteme AI avansate: reconsiderați dacă beneficiile justifică riscurile.**

3. Planuri de escaladare și oprire - Definiți în avans:

- ✓ **Cine are autoritatea să oprească sistemul.**
- ✓ **Ce comportamente declanșează oprirea automată.**
- ✓ **Cum se verifică că oprirea a fost efectivă.**
- ✓ **Ce se întâmplă dacă oprirea eșuează.**

4. Redundanță umană - Nu eliminați complet supravegherea umană din procesele critice. Sistemele AI ar trebui să asiste, nu să înlocuiascăjudecata umană în decizii cu consecințe majore.

5. Educația factorilor de decizie - Directorii și consiliile de administrație trebuie să înțeleagă că:

- ✓ **„AI sigură" este un obiectiv asimptotic, nu o stare realizabilă**
- ✓ **Riscurile cresc nelinar cu capabilitățile**
- ✓ **Beneficiile pe termen scurt pot masca riscuri pe termen lung.**

AZiTREND

Perspectiva managerului – Realismul constructiv

Cum transformă AZITREND limitările în proceduri operaționale

Recunoașterea că nu există soluții perfecte nu înseamnă resemnare. Înseamnă redirecționarea efortului către ceea ce funcționează efectiv, chiar și cu limite.

- ✓ **Cercetarea proactivă** - Proactivitatea la AZITREND nu stă numai în rapoarte și conferințe, ci în ore de testare sistematică. Fiecare produs nou trece printr-o perioadă de evaluare extinsă înainte să-l recomandăm cuiva. Nu încercăm să demonstrăm siguranța, încercăm să descoperim eșecurile. Unde cedează? Când cedează? În ce condiții? Documentăm totul. Această „hartă a punctelor slabe" devine apoi instrument practic: când clientul ne spune cum arată lumea lui, știm deja ce poate merge prost și cum să prevenim.
- ✓ **Guvernanța și standardizarea** - Le promovăm activ prin AZITREND Academy. În absența unor reglementări mature pentru sistemele AI de securitate, am dezvoltat un set intern de bune practici pe care îl transmitem integratorilor parteneri. Include cerințe de documentare, protocoale de calibrare, standarde minime pentru menținerea supravegherii umane. Nu avem autoritatea să impunem standarde industriei, dar avem influența să le modelăm prin educație și prin condițiile de parteneriat.
- ✓ **Dezvoltarea incrementală** - O aplicăm în arhitectura proiectelor. Recomandăm clienților implementarea în faze: prima fază cu funcții AI limitate și supraveghere umană intensivă, extinderea graduală doar după validarea performanței în condiții reale. Un sistem de analiză video pentru 27.000 de camere în retail nu s-a construit peste noapte, s-a dezvoltat locație cu locație, învățând din fiecare implementare.

Principiul fundamental AZITREND: Nu promitem siguranță absolută, promitem un proces riguros de gestionare a incertitudinii. Documentăm ce știm, recunoaștem ce nu știm, construim redundanțe pentru scenariile neprevăzute.

Lecția de bune practici pe care o transmitem pieței: strategiile de atenuare funcționează nu pentru că elimină riscul, ci pentru că îl fac gestionabil. În securitate, diferența dintre profesionist și amator nu este absența eșecurilor, este prezența procedurilor care limitează consecințele când eșecurile inevitabile apar.

Intră în comunitatea
AMBASADORILOR AZiTREND!
Mărește-ți veniturile cu cel puțin 25%!

Autori: Ion Iordache – Consultant de Securitate & Doru Asmarandei – CEO AZITREND

EPILOG

SPERANȚA REALISTĂ

„Dezvoltarea unei superinteligențe sigure și securizate este o posibilă meta-soluție pentru toate celelalte amenințări existențiale." (Yampolskiy)

Problema cu experții care nu mai vor să fie experți...

Trăim într-o epocă ciudată, în care experții din domeniul inteligenței artificiale se împart în două tabere la fel de problematice.

- ✓ **Prima tabără**, optimiștii tehnologici, ne asigură că totul va fi bine, că piața va regla, că inovația găsește întotdeauna soluții, că îngrijorările sunt exagerate. Ei privesc demonstrațiile matematice ale imposibilității cu același dispreț cu care un adolescent privește avertismentele părinților: da, da, am înțeles, dar de data asta e diferit. Aceștia sunt cei care confundă succesul comercial cu siguranța sistemică, care măsoară progresul în evaluări de piață și runde de finanțare, nu în probleme rezolvate.
- ✓ **A doua tabără**, profeții apocalipsei, ne spun că suntem deja condamnați, că ar fi trebuit să ne oprim acum douăzeci de ani, că orice efort este zadarnic. Paradoxal, aceștia ajung în același loc ca optimiștii: la inacțiune. Dacă dezastrul este inevitabil, de ce să mai încerci să-l previi?

Ambele tabere au abandonat, în feluri diferite, responsabilitatea care vine cu expertiza.

Expertul adevărat nu este cel care oferă certitudini reconfortante sau terifiante. Este cel care spune:

„Iată ce știm. Iată ce nu știm. Iată ce putem face cu resursele disponibil, și iată de ce trebuie să continuăm să încercăm, chiar dacă succesul nu este garantat."

De ce profesioniștii în securitate au un rol special...

Profesioniștii în securitate sunt, prin formare și experiență, imuni la două iluzii care afectează alte categorii profesionale.

- ✓ **Prima iluzie** este că sistemele funcționează așa cum au fost proiectate. Oricine a auditat vreodată un sistem de securitate fizică știe că realitatea operațională diferă întotdeauna de specificațiile tehnice. Ușile care ar trebui să fie încuiate nu sunt. Camerele care ar trebui să acopere un unghi sunt rotite greșit. Procedurile care ar trebui urmate sunt ignorate. Profesioniștii în securitate nu sunt surprinși de eșec, sunt surprinși doar de formele specifice pe care le ia.
- ✓ **A doua iluzie** este că amenințările sunt statice. În securitatea fizică, știm că adversarul se adaptează.

  Construiești un gard mai înalt, hoțul aduce o scară mai lungă. Instalezi o alarmă, spărgătorul învață să o dezactiveze.

  Această mentalitate, gândirea adversarială, este exact ce lipsește în mare parte din discuțiile despre siguranța AI, dominate de ingineri care proiectează pentru cazuri de utilizare ideale, nu pentru actori răuvoitori sau pentru consecințe neintenționate.

Când un consultant de securitate fizică privește un sistem de inteligență artificială, el nu vede o cutie magică care rezolvă probleme. Vede un activ care trebuie protejat, o vulnerabilitate care trebuie gestionată și, potențial, o amenințare care trebuie monitorizată, uneori toate trei simultan.

Această perspectivă este exact ce industria AI are nevoie și exact ce refuză, în mare parte, să accepte.

Datoria de a spune adevărul neplăcut...

Există o datorie profesională pe care mulți o ignoră pentru că este inconfortabilă: **datoria de a spune adevărul chiar și când adevărul nu este ceea ce clientul vrea să audă.**

În securitatea fizică, această datorie ia forme concrete. Când evaluezi un obiectiv și concluzionezi că bugetul alocat nu poate asigura nivelul de protecție solicitat, ai datoria să spui acest lucru, chiar dacă înseamnă să pierzi contractul. Când observi că procedurile existente sunt ignorate sistematic, ai datoria să raportezi, chiar dacă deranjezi conducerea. Când identifici o vulnerabilitate pe care nimeni nu vrea să o audă, ai datoria să o documentezi.

Pentru sistemele AI, această datorie devine și mai critică. **Când un furnizor promite „siguranță garantată", consultantul de securitate are datoria să întrebe:** *„Garantată cum? Verificată de cine? Pe baza căror teste? Cu ce limite recunoscute?".*

Când un factor de decizie vrea să implementeze un sistem autonom fără supraveghere umană „pentru eficiență", consultantul are datoria să explice riscurile acestei decizii, chiar dacă explicația este lungă, tehnică și incomodă.

Această datorie nu este populară. Nimănui nu-i place mesagerul care aduce vești complicate, dar

alternativa, tăcerea complice sau, și mai rău, participarea activă la perpetuarea iluziilor, nu este o opțiune pentru profesionistul care își ia rolul în serios.

AZiTREND
www.azitrend.ro