IDR

Romanian Diplomatic Institute

Policy Paper Nr. 35/2023

Departe de utopii și distopii. Impactul AI asupra securității cibernetice

Claudiu Codreanu

Ministerul Afacerilor Externe

DEPARTE DE UTOPII ȘI DISTOPII. IMPACTUL AI ASUPRA SECURITĂȚII CIBERNETICE

Claudiu Codreanu
Cercetător, Direcția pentru Furnizare Expertiză
Institutul Diplomatic Român

Editori: Dragoș C. Mateescu, Mihai Constantinescu

REZUMAT

- **Ultimele evoluții din domeniul inteligenței artificiale (AI) au provocat atât entuziasm, cât și îngrijorări.** Totuși, acestea nu constituie încă bazele unor schimbări fundamentale pentru societate.
- **Actori internaționali majori depun eforturi pentru a reglementa dezvoltarea și utilizarea inteligenței artificiale**. Parlamentul European a adoptat Legea privind inteligența artificială, China a adoptat deja o legislație pentru reglementarea algoritmilor și a AI, iar Statele Unite și Regatul Unit întreprind propriile eforturi.
- **Inteligența artificială nu are, încă, potențialul de a cauza schimbări fundamentale sau apocaliptice.** Cu toate acestea, anumite grupuri de persoane și mediul înconjurător deja încep să fie afectate.
- **Ultimele evoluții din sectorul AI au un impact și asupra securității cibernetice și a operațiunilor cibernetice**. În acest aspect, pot fi identificate trei dimensiuni: protejarea sistemelor AI de atacuri cibernetice, utilizarea AI pentru eficientizarea atacurilor cibernetice, și utilizarea AI pentru consolidarea securității cibernetice.
- **Politicile de securitate cibernetică adoptate de state trebuie să țină cont de ultimele evoluții și să fie adaptate pentru a lua în calcul protejarea sistemelor AI de atacuri.** Factori cheie în permiterea unei dezvoltări sustenabile a AI și utilizării cu riscuri limitate a acestuia sunt transparența instituțiilor, respectarea statului de drept și asigurarea libertății societății civile și a presei. Totodată, un rol important îl vor avea diplomația digitală și diplomația cibernetică.

INTRODUCERE

În ultimul an, evoluțiile tehnologice din domeniul inteligenței artificiale (*Artificial Intelligence* – AI) au captat atenția jurnaliștilor, mediului academic, politicienilor, și a publicului larg. Motivul principial este că evoluția modelelor largi de limbaj (*Large Language Models* – LLM) și AI generatoare de conținut (*generative AI*) a depășit așteptările și viteza guvernelor și organizațiilor de a adopta reglementări solide. Totuși, **ideea, momentan eronată, conform căreia AI va transforma întreaga funcționare a societății și fiecare aspect al vieții umane, similar cu electricitatea, inhibă măsurile prin care să fie impuse limite asupra utilizării AI** (Michel 2023, 13).

Totodată, **în 2024 vor fi organizate zeci de alegeri importante în toată lumea, cele mai relevante fiind cele din Statele Unite și cele pentru Parlamentul European**. În Europa, vor avea loc alegeri în Austria, România și Regatul Unit, iar în spațiul eurasiatic și în Asia vor avea loc alegeri în Georgia, Turcia, Coreea de Sud, India, Taiwan. În Africa vor avea loc alegeri în Africa de Sud, Algeria, Egipt, Ghana, Tunisia, iar în America Latină în Mexic și Uruguay. Statele enumerate aici nu sunt toate în care vor avea loc alegeri, ci doar o listă cu cele mai reprezentative. Ultimele alegeri din SUA și alte țări importante în arena internațională au fost vizate de diferite campanii cibernetice și informaționale, existând riscul ca în 2024 să iasă în evidență rolul inteligenței artificiale.

Acest studiu se axează pe rolul și influența ultimelor evoluții din AI asupra securității și operațiunilor cibernetice. Cercetarea este împărțită în două părți. Prima va discuta evoluțiile recente din domeniul AI, iar ce-a dea doua va evidenția/analiza impactul acestora asupra securității cibernetice. În final, va fi propusă o serie de recomandări pentru a putea gestiona și valorifica instrumentele AI pentru securitatea cibernetică.

Cercetarea pleacă de la întrebarea: Cum afectează noile evoluții din domeniul AI zona securității cibernetice? Pentru a aborda întrebarea de cercetare, studiul **testează validitatea următoarei aserțiuni**: Evoluțiile din domeniul AI vor produce avantaje serioase pentru operațiunile cibernetice ofensive, dar și pentru apărătorii cibernetici. În esență, **noile instrumente, malware-uri și software-uri care fructifică ultimele evoluții AI ajută atât atacul, cât și apărarea**. Pe de o parte, hackerii guvernamentali și non-statali vor fi ajutați de creșterea amplorii atacurilor (mult mai ușor de răspândit către cât mai multe dispozitive, spre ex. rețelele de boți), precum și de scăderea nivelului de dificultate al unor pași (prin automatizarea acestora, spre ex. scanări de vulnerabilități). În plus, instrumentele AI vor

contribui la eficientizarea eforturilor de *phishing* și dezinformare (ex. scanări în baze mari de date, generarea de mesaje automate etc.). Pe de altă parte, vor fi ajutate și eforturile de apărare cibernetică a guvernelor, organizațiilor (civice, economice etc.) sau companiilor de securitate cibernetică și ale indivizilor. În acest caz, cele mai importante contribuții ale AI vor fi automatizarea scanării de vulnerabilități, instrumente care să monitorizeze în timp real activitatea pe rețele și să învețe singure să caute tipare, precum și îmbunătățirea abilităților de a ataca cibernetic un actor malițios.

CONTEXT ȘI CONCEPTE

Termenul de „inteligență artificială" (*artificial intelligence*) a fost introdus în 1955 de către cercetătorul american John McCarthy (Bjola 2022, 79; Bonfanti & Kohler 2020, 1). **Inteligența artificială se referă la acțiunea prin care calculatoarele procesează cantități mari de date, utilizând algoritmi sofisticați care simulează comportamente sau raționamente umane** (Bjola 2022, 79). AI reprezintă o tehnologie generală ce are ca scop îmbunătățirea vitezei, preciziei și magnitudinii unor procese automate de luare a deciziilor. Inteligența artificială va contribui la îmbunătățirea performanțelor umane în activități precum prezicerea, optimizarea, recunoașterea și luarea de decizii, inclusiv în contexte strategice sau militare (Maas 2019, 285-286). Astfel, **AI este descrisă ca o tehnologie care permite realizarea altor activități sau îmbunătățirea altor tehnologii și sisteme, având în vedere că poate fi aplicată pe diferite domenii, inclusiv pentru obiective ce țin de securitatea cibernetică** (Bonfanti & Kohler 2020, 1).

Totuși, tehnologiile prezentate ca fiind „inteligență artificială" au doar o capacitate limitată de a reproduce inteligența umană, imitând doar unele aspecte ale acesteia (Michel 2023, 6). **Până acum, nicio mașină (ex. sistem, robot, software, ș.a.m.d.) nu a trecut testul Turing.** Testul a fost propus de cercetătorul britanic Alan Turing în 1950, o mașină trebuind să îndeplinească două condiții: să reacționeze adecvat variațiilor dialogului uman, și să prezinte intenții și o personalitate cât mai aproape de cele umane (Bjola 2022, 79). Astfel, au existat apeluri din mediul academic și tehnologic de a înlocui termenul-umbrelă de „inteligență artificială" cu o terminologie exactă și individualizată care descrie capabilitățile tehnice ale sistemului respectiv (Michel 2023, 15).

În schimb, **mecanismele de învățare automată (ML – *machine learning*) sunt formate din date, algoritmi de învățare și putere computațională pentru antrenarea algoritmilor**. O bună parte din ultimele evoluții din AI provin din **învățarea aprofundată**

(*deep learning*), utilizând rețele neuronale aprofundate. Rețelele sunt compuse din numeroase straturi de neuroni artificiali, fiecare transformând datele pe care le primește. ML se bazează pe cantități substanțiale de date, dar chiar și așa, mecanismele nu pot obține o precizie de 100%, adeseori făcând predicții greșite. Mai mult, **modelele largi de limbaj (LLM)** încă produc greșeli imprevizibile și texte false în ciuda faptului că sunt antrenate pe cantități vaste de texte (Bonfanti & Kohler 2020, 1; Bezombes, Brunessauax & Cadzow 2023, 10; Michel 2023, 18).

Totodată, un ***deepfake* reprezintă un material video, fotografic sau audio generat de sisteme algoritmice complexe (mecanisme de învățare automată și de învățare aprofundată), cu o monitorizare umană limitată sau chiar inexistentă**. *Deepfake*-urile permit amplificarea amenințărilor din zona criminalității cibernetice, mai ales în ceea ce privește furtul de identitate și ocolirea sistemelor de autentificare biometrică. Ținând cont de faptul că generează ceva nou și nu doar imită realitatea, *deepfake*-urile au potențialul de a induce în eroare oamenii, făcându-i să creadă că materialul respectiv reprezintă realitatea (Bray, Johnson & Kleinberg 2023).

În ultimul an, **popularitatea utilizării sistemelor AI generatoare de texte, precum ChatGPT sau Google Bard, și a celor generatoare de imagini, precum Stable Diffusion, Midjourney sau Dall-E, a pus în centrul atenției publice și politice chestiunea dezvoltării rapide a tehnologiilor AI**. Printre primii cercetători și reprezentanți ai industriei care au ieșit public să vorbească despre îngrijorările cu privire la AI a fost **Geoffrey Hinton**, unul dintre pionierii sistemelor de inteligență artificială. Hinton a renunțat la postul de la Google și a ieșit public să avertizeze despre pericolele ultimelor evoluții din AI (Knight 2023a). În aceeași perioadă, **Eliezer Yudkowsky**, unul dintre cercetătorii cei mai importanți din domeniul AI, a subliniat într-un articol din 2023 pentru revista *TIME* că, în condițiile actuale, dacă va fi dezvoltat un sistem AI mult prea puternic, se așteaptă ca „fiecare membru al speciei umane și toată viața biologică de pe Pământ să moară la scurt timp după”. Totodată, cercetători în securitate biologică și nucleară au atras atenția că informațiile furnizate de sistemele AI generatoare de conținut ar putea ajuta grupări teroriste la fabricarea de arme biologice, chiar dacă momentan riscurile sunt reduse (Service 2023).

În acest context, mai mulți reprezentanți ai industriei tech și AI au semnat în martie 2023 o **scrisoare deschisă care face un apel pentru suspendarea timp de 6 luni a activităților laboratoarelor de AI care antrenează sisteme mai puternice ca GPT-4**, dar și pentru reglementarea industriei și monitorizarea sistemelor AI puternice (Yudkowsky 2023; Milmo & Stacey 2023). În aceeași perioadă, premierul britanic Rishi Sunak a anunțat că

Regatul Unit va organiza un **summit global asupra siguranței inteligenței artificiale**, acesta fiind adresat „statelor cu opinii similare” (Milmo & Stacey 2023). Totodată, grupul de state G7 a căzut de acord să formeze „**procesul Hiroshima AI**”, un forum menit să dezbată actualele probleme legate de ultimele evoluții din domeniul AI (Milmo & Stacey 2023).

În iulie 2023, ONU a organizat la Geneva **conferința *AI for good***, unde au fost aduși să vorbească despre AI și nouă dintre cei mai bine dezvoltați roboți umanoizi, nefiind clar dacă răspunsurile lor au fost programate dinainte (Ferguson 2023). Mai apoi, Secretarul general al ONU, António Guterres, avertiza în timpul primei sesiuni a Consiliului de Securitate asupra AI, din iulie 2023, că utilizarea sistemelor AI ar putea cauza un număr „oribil” de decese și distrugeri, făcând un apel pentru formarea unui nou panel interguvernamental (Milmo & Farah 2023).

Tot în aceeași perioadă, Google, Microsoft, Anthropic și OpenAI au înființat un **forum pentru a discuta reglementarea dezvoltării instrumentelor AI**. Obiectivele acestuia sunt discutarea riscurilor de securitate cu mediul academic și cu cel politic, dezvoltarea unor standarde pentru evaluare și lansarea de modele AI avansate, precum și promovarea utilizării AI pentru combaterea schimbărilor climatice sau în domeniul medical. Membrii forumului, împreună cu Amazon, Meta și alți actori relevanți, au căzut de acord cu Casa Albă pentru adoptarea unor garanții de siguranță în dezvoltarea și promovarea sistemelor AI (ex. introducerea unor inscripționări care să marcheze conținutul creat de AI). Totuși, există și critici la adresa inițiativei, industria de tehnologie având un istoric bogat de a eșua să respecte angajamentele pentru auto-reglementare. (Milmo 2023a; Benson 2023)

EFECTELE ULTIMELOR EVOLUȚII ALE AI SUNT DEJA VIZIBILE

Impactul actual asupra vieții umane și a planetei

Până la distrugerea speciei umane, **utilizarea sistemelor AI are deja efecte dezastruoase pentru anumite persoane și grupuri de persoane, mai ales cele provenind din comunități marginalizate sau discriminate**. Spre exemplu, în noiembrie 2022, un tânăr afro-american a fost arestat în statul american Georgia, fiind închis pentru 6 zile după ce un soft de recunoaștere facială l-a identificat în mod eronat ca autorul unor jafuri din Louisiana (Malik 2023). Incidentul a evidențiat faptul că software-urile pentru recunoaștere facială nu sunt calibrate și antrenate pe baza a suficiente date pentru a identifica cât mai corect posibil persoanele de culoare, dar și faptul că algoritmii utilizați reproduc prejudecățile umane. Astfel, utilizarea AI în activități pentru care nu a fost proiectat tehnic fără o supraveghere umană

responsabilă sau în contexte în care nu are suficiente reglementări poate genera riscuri ridicate pentru membrii grupurilor vulnerabile (Michel 2023, 11).

O altă problemă care a fost relativ trecută cu vederea din cauza avertizărilor apocaliptice este și **impactul sistemelor AI asupra mediului**. Așa cum producerea criptomonedelor consumă o cantitate enormă de electricitate, și serverele *cloud* pe care se bazează AI și toate sistemele care le întrețin vor consuma o cantitate din ce în ce mai mare de energie electrică, la fel și fabricarea cipurilor ș.a.m.d. Spre exemplu, numai producerea criptomonedei Bitcoin consumă mai multă energie electrică decât Norvegia și Ucraina la un loc (Stokel-Walker 2023).

În plus, există și riscul unei **curse a dezvoltării armelor pe bază de AI** (Maas 2019, 286). În această zonă se încadrează și chestiunea **sistemelor de arme complet autonome**, acestea funcționând fără comenzi externe în întreg procesul de găsire, urmărire, selectare și atacare a țintelor (Rosert & Sauer 2021). Nu există încă o reglementare internațională clară cu privire la aceste sisteme.

Utilizarea sistemelor AI în campaniile de dezinformare

Utilizarea de imagini și video-uri modificate sau false în campaniile electorale nu reprezintă o noutate. De exemplu, guvernatorul Floridei și actual candidat la alegerile primare ale Partidului Republican, Ron DeSantis, a distribuit *online* imagini generate cu ajutorul AI care îl arată pe Donald Trump îmbrățișându-l pe Anthony Fauci, fostul consilier-șef al Casei Albe pe chestiuni medicale. Fauci este disprețuit de conservatorii americani pentru politicile de limitare a răspândirii virusului SARS-CoV-2 și de promovare a vaccinului împotriva COVID-19. La fel, au apărut mai multe *deepfake*-uri cu Președintele Joe Biden în care anunță mobilizarea americanilor pentru a lupta în Ucraina. (Milmo 2023b)

Până acum, astfel de video-uri create artificial nu au reprezentat puncte de cotitură și nu au profitat la maximum de potențialul dat de AI generator. Totuși, lucrurile ar putea să se schimbe în 2024, exact atunci când va conta mai mult, în mijlocul campaniilor electorale din diferite state. **Peste 70 de state vor desfășura alegeri naționale și/sau regionale în 2024, acestea vizând peste 2 miliarde de persoane, majoritatea în Sudul Global** (Madung 2023). Un risc major este provocat de AI generator, care poate produce texte, imagini, video-uri și audio-uri credibile pentru campanii de dezinformare în timpul alegerilor (Milmo & Stacey 2023). Spre exemplu, anchete ale unor jurnaliști din Africa au arătat că algoritmul Twitter (astăzi platforma X) a fost manipulat frecvent și fără dificultăți pentru a răspândi propagandă și dezinformări în Kenya și Nigeria în timpul ultimelor campanii electorale (Madung 2023).

AI va face mai ușor atât procesul de producere al dezinformărilor, cât și cel de diseminare în masă a acestora. Un studiu publicat de jurnalul *Science Advances* în 2023 arăta că *tweet*-urile produse de ChatGPT-3 (predecesor al actualului GPT-4) pot să informeze și să dezinformeze mai bine decât cele produse de oameni (Spitale, Biller-Andorno & Germani 2023). Astfel, sistemele AI au potențialul de a crește eficacitatea campaniilor de dezinformare, îmbunătățind abilitatea actorilor externi de a interfera în afacerile interne ale altor state (Kenny 2023, 224).

Pe măsură ce tehnologiile pe bază de AI avansează și devin disponibile la scară largă, utilizarea acestora în campanii de dezinformare va deveni din ce în ce mai răspândită (Kenny 2023, 224). În plus, **dezinformarea va deveni mult mai bine țintită pe anumite grupuri de oameni sau chiar pe indivizi folosind datele colectate din spațiul online** (Benson 2023). Sistemele AI permit actorilor să creeze un număr semnificativ de conturi false pe *social media*, utilizând mecanisme de învățare automată pentru a pune bazele unor profiluri false, de la date, postări și, inclusiv, fotografii. Conturile pot fi controlate tot prin intermediul rețelei de boți, în masă, dând impresia că un număr ridicat și credibil de cetățeni împărtășesc aceeași viziune (Kenny 2023, 235).

Noile instrumente pe bază de AI pot permite unui număr mult mai mare de persoane să creeze imagini, mesaje, video-uri mult mai credibile față de acum 2-3 ani, facilitând crearea de materiale false sau de propagandă. Totuși, încă nu este clar rolul pe care îl vor avea noile platforme pentru desfășurarea campaniilor electorale din 2024. Cel mai probabil, alegerile vor fi afectate, dar acest lucru nu indică și faptul că, în mod cert, vor provoca schimbări fundamentale în desfășurarea acestora sau în preferințele majorității alegătorilor, mai ales față de campaniile de dezinformare care au țintit alegerile de după 2015. Ce a devenit clar este că producerea de texte și imagini false a devenit aproape gratuită, facilă, accesibilă și nu necesită abilități tehnologice dezvoltate. Mai mult, materialele produse par credibile, și devine din ce în ce mai dificil ca oamenii și tehnologiile să distingă între ce a fost produs artificial și ce este real. În plus, ceea ce crește este magnitudinea operațiunilor de diseminare a dezinformărilor, putând fi create rețele de boți care produc automat conținut artificial și credibil, la o scală mult mai mare față de ceea ce se putea face acum câțiva ani.

Eforturi pentru reglementarea națională și internațională a AI

Majoritatea discuțiilor privind ultimele evoluții din AI s-a axat pe potențialul îndepărtat ca tehnologia să aducă finalul umanității, și nu pe chestiunile care deja se întâmplă: creșterea supravegherii în masă și a discriminării pe diferite criterii sau diseminarea de campanii de

dezinformare (Bhuiyan & Robins-Early 2023). Cu toate acestea, mai mulți actori internaționali au început să încerce reglementarea sistemelor pe bază de AI.

În iunie 2023, **Parlamentul European a adoptat Legea privind AI**. Aceasta prevede limite și interdicții pentru colectarea de date biometrice și interdicții privind instrumentele AI ce prezintă riscuri ridicate, precum recunoașterea facială și sisteme de *predictive policing*, adică analiză predictivă pentru poliție (Bhuiyan & Robins-Early 2023).

Legea a fost adoptată de Parlamentul European în iunie 2023 și ar putea intra în vigoare până la finalul anului curent, UE sperând că ar putea genera și un „efect Bruxelles", devenind standardul internațional. Legea clasifică sistemele AI în cinci categorii de risc pentru utilizatori. În primul rând, **sistemele care prezintă riscuri inacceptabile vor fi interzise**. **Acestea includ sisteme care manipulează oamenii, sisteme utilizate de guverne pentru a clasifica populația după caracteristici personale sau statut socioeconomic, dar și sisteme de analiză predictivă pentru poliție, sisteme de identificare biometrică, precum recunoașterea facială în timp real**. În al doilea rând, sistemele cu risc ridicat, cele care pot afecta negativ siguranța și drepturile fundamentale, vor fi atent monitorizate. Acestea includ operarea infrastructurilor critice, sisteme utilizate în educație, sau cele utilizate pentru controlul frontierelor și migrației. Sistemele cu risc limitat trebuie să respecte un set de cerințe minimale de transparență, iar utilizatorii trebuie avertizați că interacționează cu un AI (ex. sisteme AI generatoare, precum ChatGPT sau Midjourney). În ultimul rând, sistemelor cu risc minimal sau fără risc nu le sunt impuse alte obligații. Acestea sunt și cele mai uzuale, cum sunt cele utilizate în filtrele de spam pe mail sau în jocurile video. În plus, cei care integrează într-un domeniu sau activitate sisteme pe bază de AI vor trebuie să asigure o supraveghere umană a platformei. (Milmo 2023c)

În mod similar, **Statele Unite, Regatul Unit și China au introdus sau au în plan să introducă propriile legislații privind AI** (Milmo 2023c). În Statele Unite, eforturile sunt încă într-o fază incipientă. Lipsa unor reglementări de la Washington a generat un proces controversat prin care companiile cele mai mari din sectorul AI propun singure reglementări, aspecte alese atent pentru a nu afecta actualele activități ale companiilor (Bhuiyan & Robins-Early 2023).

Între timp, **China a inițiat una dintre primele reglementări din lume privind algoritmii și inteligența artificială**. Legislația chineză prevede formarea unui registru al algoritmilor, o bază centrală pentru autorități unde pot fi găsite informații cu privire la algoritmi, sursele pentru datele de instruire și potențialele riscuri de securitate. Totodată, reglementările prevăd ca furnizorii de algoritmi să obțină consimțământul persoanelor dacă

imaginile și vocile lor sunt manipulate pentru generarea de videoclipuri (ex. *deepfake*-uri) (O'Shaughnessy & Sheehan 2023).

În aprilie 2023, China a adoptat o lege care prevede reglementarea sistemelor AI generatoare de conținut, incluzând interdicții pentru a evita discriminarea unor grupuri, precum și impunerea de limitări sau responsabilitate legală pentru dezvoltatori. Totuși, legislația are ca obiectiv și consolidarea regimului autoritar al Chinei. Actele normative prevăd condiții pentru păstrarea intimității persoanelor și evitarea creării de profiluri pe bază activității utilizatorilor, cerințe pentru transparență și pentru responsabilitate. Totodată, **conținutul generat de AI trebuie să „reflecte valorile socialiste de bază", fiind interzis conținutul care „subminează autoritatea statului"**. În plus, reglementările se aplică doar în sectorul privat (O'Shaughnessy 2023).

IMPACTUL SISTEMELOR AI ASUPRA SECURITĂȚII CIBERNETICE

Într-un raport al agenției UE pentru securitate cibernetică, ENISA (Bezombes, Brunessauax & Cadzow 2023, 10), sunt identificate trei dimensiuni ale relației dintre AI și securitatea cibernetică. Prima dimensiune se referă la **securitatea cibernetică a AI** – siguranța și vulnerabilitățile modelelor AI și ale algoritmilor. A doua dimensiune constă în **utilizarea AI pentru securitatea cibernetică** – unelte și mijloace AI pentru a consolida securitatea cibernetică. Cea de-a treia dimensiune face referire la **utilizarea malițioasă a AI** – utilizarea malițioasă sau contradictorie a AI pentru a genera tipuri de atacuri mult mai sofisticate.

Atacuri cibernetice mediate de noile tehnologii AI

Conform unui studiu din 2023 al companiei de securitate cibernetică *Imperva*, 47,4% din tot traficul web înregistrat în 2022 a fost reprezentat de trafic automat (boți). **Faptul că majoritatea traficului web provine de la boți a generat noi îngrijorări, pe baza cărora s-a format și teoria conspirației a Internetului mort (*dead Internet theory*)**. Nimeni nu mai poate fi complet sigur că ceea ce vede online este real sau că oamenii cu care interacționează în online sunt reali. În ciuda progreselor realizate de companiile de social media privind utilizarea boților pentru comentarii, reacții, reclame etc., există și temerea ajungerii în punctul în care sistemele *online* bazate pe AI vor identifica comportamentul boților ca autentic și cel al oamenilor ca neautentic (Tiffany 2021).

Sistemele autonome vor putea produce avantaje majore în operațiunile cibernetice, mai ales în executarea rapidă și pe scară largă a unor sarcini repetitive și mai puțin complexe (Perez

2023, 187). Astfel, **AI are capacitatea de a crește cantitatea de amenințări cibernetice, dar și de a schimba caracteristicile acestora, introducând amenințări noi și necunoscute** (Bonfanti & Kohler 2020, 2). Capabilitățile cibernetice autonome pot fi folosite în mai multe etape ale unui atac cibernetic, inclusiv în fazele de recunoaștere, infiltrare, și în cea de comandă și control (Perez 2023, 188). Atacurile cibernetice nu se desfășoară în totalitate sub o supraveghere umană directă în toate fazele operațiunii. *Malware*-ul Stuxnet, utilizat de SUA și Israel pentru a afecta programul nuclear al Iranului, a acționat fără control uman direct, mai ales atunci când a trebuit să infecteze o rețea închisă pentru a ținti facilitățile nucleare iraniene (Perez 2023, 204). În plus, scăzând costurile și cunoștințele tehnologice, instrumente cibernetice utilizate pentru atacuri vor deveni disponibile pentru mult mai mulți actori față de situația ultimilor ani (Bonfanti & Kohler 2020, 2).

AI ar putea să crească nivelul de sofisticare în construirea de *malware* și în funcționarea acestora. Un *malware* ajutat de AI ar putea să răspundă în timp real și autonom la schimbările de comportament ale țintei și să evite măsurile de apărare cibernetică. Un *malware* autonom este capabil să învețe din mediul unde este propagat pentru a evita detectarea, să caute și să infecteze noi ținte, dar și să găsească date importante și să permită atacuri cibernetice noi. În 2018, cercetători ai IBM au dezvoltat un astfel de *malware*, deci tehnologia nu este una foarte recentă (Bonfanti & Kohler 2020, 3).

În paralel, așa cum era de așteptat, **infractorii cibernetici și-au creat propriile clone după ChatGPT**, susținând că platformele pot ajuta la abilitatea hackerilor de a scrie coduri pentru *malware* sau email-uri de *phishing*. Ar fi vorba despre două modele largi de limbaj (LLM) care imită funcțiile ChatGPT și ale platformei Bard a Google, generând texte în urma solicitărilor utilizatorilor. Totuși, sistemele ar putea fi chiar ele fraude, și să nu funcționeze de fapt. Oricum, rolul lor este, momentan, unul limitat. Încă de la începutul verii din 2023, FBI și Europol au avertizat că infractorii cibernetici caută modalități de a utiliza AI generator în activitățile lor. **Modelele largi de limbaj ar putea face ca tentativele de fraudă, furt de identitate și inginerie socială să fie derulate mai rapid, dar ar putea contribui și la îmbunătățirea textelor de *phishing* scrise în engleză**. (Burgess 2023)

Atacuri împotriva sistemelor AI

Pe lângă utilizarea AI pentru eficientizare operațiunilor cibernetice, chiar sistemele AI vor deveni ținte importante ale atacurilor cibernetice, atâta vreme cât vor servi unor funcții societale, guvernamentale, economice și de securitate importante (Whyte 2023, 313). Astfel,

devine esențială asigurarea securității sistemelor cibernetice care se bazează pe tehnologii AI (Bonfanti & Kohler 2020, 2).

ChatGPT și celelalte sisteme similare au fost modificate în repetate rânduri pentru a preveni exploatarea lor pentru generarea de informații personale, discursuri instigatoare la ură sau instrucțiuni pentru improvizarea explozibililor etc. Totuși, un grup de cercetători de la Universitatea Carnegie Mellon a demonstrat într-un studiu din iulie 2023 că anumite seturi de texte introduse în *chat* pot ocoli toate măsurile defensive. Astfel de atacuri se numesc **atacuri contradictorii** (*adversarial attacks*), acestea implicând introducerea de texte care să „convingă" treptat *chatbot*-ul să ocolească limitele impuse de dezvoltatori**. Atacurile contradictorii pot consta în „otrăvirea datelor" (*data poisoning*), injecții în datele de instruire de la baza AI, care determină algoritmul de învățare să facă greșeli, sau exemple contradictorii, materiale create pentru a introduce în eroare și pentru a le clasifica eronat** (Knight 2023b; Bonfanti & Kohler 2020, 2).

Încă nu există un consens asupra modalităților de interacțiune dintre autonomie și cibernetic, și nici cum ar putea fi aplicat dreptul internațional în acest caz. Spre exemplu, chiar dacă **armele autonome** sunt considerate a fi fizice, sistemele cibernetice sunt imateriale, în ciuda faptului că sunt produse de echipamente fizice. În mod similar, o altă provocare va fi (și poate deja este) reprezentată de **vehiculele autonome**. Alături de AI și de *Internet of Things* și alte sisteme care întrepătrund lumea fizică cu cea cibernetică, vehiculele autonome vor genera noi riscuri de securitate, acestea fiind vulnerabile la atacuri cibernetice care ar putea să le compromită buna funcționare. Modificarea unor semne rutiere sau utilizarea de **imagini contradictorii** (anumite modificări care derutează sistemele pe bază de mecanisme de învățare automată) ar putea genera riscuri majore de securitate cibernetică și chiar de siguranță publică (Perez 2023, 186; Blumenthal & Csernatoni 2022, 2-4).

Impactul noilor evoluții asupra securității cibernetice

Instrumentele AI ar putea fi folosite și pentru a îmbunătăți *software*-urile actuale de detectare și identificare de *malware*, dar doar ca o unealtă complementară, și nu să înlocuiască mecanismele tradiționale (Bonfanti & Kohler 2020, 3). **Utilizarea AI pentru eficientizarea proceselor de analiză și intruziune în operațiunile cibernetice va oferi oportunități majore atacatorilor**, la fel ca în cazul instrumentelor de analiză a unor cantități substanțiale de date (Whyte 2023, 309). Totuși, **și apărarea cibernetică va avea de câștigat prin utilizarea uneltelor pe bază de AI**. Ultimele evoluții din sectorul AI produc nu doar avantaje pentru atacatorii cibernetici, ci și pentru apărători, reprezentând o sabie cu două tăișuri pentru

securitatea cibernetică. În acest sens, validitatea aserțiunii de la care a pornit acest studiu, conform căreia evoluțiile din domeniul AI produc avantaje serioase pentru operațiunile cibernetice ofensive, dar și pentru apărătorii cibernetici, este confirmată în urma acestei cercetări. Inteligența artificială reprezintă *doar un instrument*, iar impactul acestui instrument depinde de modul în care este creat sistemul sau cum este folosit. În mare parte, platformele AI ar putea avantaja cel mai mult criminalitatea cibernetică. În ceea ce privește acțiunile actorilor statali, inteligența artificială nu va avea neapărat un rol fundamental, dar va eficientiza și crește viteza unor etape ale operațiunilor cibernetice. Apărarea cibernetică ar putea fi la rândul ei consolidată de instrumentele AI, prin asistarea programelor actuale și a eforturilor umane de scanare după intruziuni și vulnerabilități, dar nu le va putea înlocui.

AI ar putea îmbunătăți **capabilitățile de detectare a amenințărilor cibernetice, de analiză**, dar și de prevenire, mai ales pentru *spam*-uri, *phishing* sau detectarea de *malware* (Bonfanti & Kohler 2020, 3). Astfel, va crește viteza detectării unor activități cibernetice malițioase, dar și capacitatea de a scana un spațiu cât mai larg. Totodată, AI ar putea fi folosit pentru **scanarea și testarea automată de vulnerabilități într-un sistem sau o rețea** (Bonfanti & Kohler 2020, 3). În același timp, instrumentele AI ar putea avea un rol important în învățarea și detectarea unor tipare și anomalii într-un sistem sau o rețea, dar precizia rămâne chestionabilă.

AI ar putea fi folosit și pentru **monitorizarea spațiilor *online* și a platformelor de *social media*,** identificând tipare și semne ale unor campanii malițioase și activități ale rețelelor de boți (Bonfanti & Kohler 2020, 3). Sisteme AI ar putea fi utilizate pentru detectarea de conturi false create în masă, de mesaje de dezinformare lansate în campanii corelate, învățarea și detectarea de tipare având un rol important pentru aceste eforturi. Totuși, utilizarea sistemelor AI are o limită în acest aspect, fiind necesară o monitorizare umană atentă pentru a nu marca eronat texte drept dezinformări.

Inteligența artificială permite desfășurarea unor operațiuni cibernetice mult mai complexe și de o magnitudine mult mai mare față de perioada precedentă. La fel de importantă va fi și chestiunea atacurilor cibernetice împotriva sistemelor AI și a organizațiilor și platformelor care utilizează AI. Totuși, **impactul inteligenței artificiale asupra proceselor de securitate cibernetică nu va fi unul care să schimbe fundamental dinamica**. Anumite forme de AI sunt deja utilizate în majoritatea sistemelor de operațiuni cibernetice ofensive, dar și în cele defensive, inclusiv în sistemele antivirus și *firewall*. Totuși, în acest caz este vorba de sisteme mult mai complexe, inteligente, capabile să ia decizii singure fără o intervenție sau monitorizare umană.

CONCLUZII ȘI RECOMANDĂRI

Nu poate exista așteptarea ca instituții sau state unde inegalitățile structurale sunt înrădăcinate de multă vreme să își poată reconfigura întregul sistem de valori și să implementeze în mod responsabil sisteme pe bază de AI. În lipsa unor măsuri mai largi, care să protejeze drepturile întregii societăți în care operează respectivele sisteme AI, sunt șanse foarte slabe să poată fi implementat în mod etic un sistem pe bază de inteligență artificială. State care nu oferă drepturi egale tuturor cetățenilor (ex. restricții asupra libertății de exprimare, limitarea drepturilor femeilor, a minorității LGBTQ+ sau a minorităților etnice) au șanse slabe să poată implementa programe AI publice care să fie etice, imparțiale sau corecte (Michel 2023, 33).

Factori cheie pentru permiterea unei dezvoltări sustenabile a inteligenței artificiale și utilizării cu riscuri limitate a acesteia sunt transparența instituțiilor, respectarea statului de drept și asigurarea libertății societății civile și a mediei. Toate acestea reprezintă aspecte fundamentale pentru a se putea implementa în mod responsabil sisteme pe bază de AI (Michel 2023, 25). Astfel, legislația privind AI trebuie să conțină măsuri pentru asigurarea transparenței, pentru a permite societății civile și mediei să monitorizeze modalitățile în care sunt dezvoltate și utilizate sistemele respective (O'Shaughnessy 2023).

Asigurarea monitorizării umane a inteligenței artificiale, precizia acesteia, capacitatea de a putea fi explicată, precum și garantarea transparenței și a siguranței, sprijină și securitatea cibernetică a AI în sine, dar și a sistemelor ce au la bază AI. Aceste elemente reprezintă criterii de bază ale oricăror politici și strategii în acest sector (Bezombes, Brunessauax & Cadzow 2023, 11), reprezentând aspecte esențiale pentru ca sistemele să nu fie abuzate, să nu funcționeze eronat, dar și să nu reproducă prejudecăți și erori umane. Astfel, **supravegherea umană** a sistemelor AI în anumite medii este necesară atât pentru observarea bunei lor funcționări, dar și pentru a identifica comportamente neobișnuite din cauza unor atacuri cibernetice (Bezombes, Brunessauax & Cadzow 2023, 10). Monitorizarea sistemelor pe bază de AI trebuie asigurată și în cazul celor utilizate pentru securitatea cibernetică, pentru a evita atacuri care ocolesc capabilitățile de detectare ale AI sau pentru a evita identificări eronate de comportamente și *software* ca fiind malițioase.

Bazele de date și infrastructurile aferente ar trebui să fie construite și gestionate ținând cont de principii stricte privind intimitatea. Valoarea seturilor de date pentru tehnologiile AI depinde de aplicarea acestora și de capabilitățile de a fructifica datele.

Acumularea și diseminarea unui număr din ce în ce mai mare de date creează vulnerabilități care nu vor putea fi gestionate rapid de legislațiile pentru protejarea intimității persoanelor. Totuși, aceste reglementări pot fi oricând schimbate atunci când ajunge la putere un regim autoritar, care ar putea abuza datele care înainte erau protejate prin lege. Totodată, există și îngrijorarea ca actorii privați să abuzeze seturile largi de date (Michel 2023, 17-19).

În esență, **trebuie adoptate modele centrate pe om în privința utilizării, implementării și reglementării AI**. Politicile privind inteligența artificială ar trebui să pună accentul pe asigurarea intimității persoanelor, pe protejarea datelor indivizilor, dar și pe protejarea drepturilor grupurilor minoritare, ținând cont și de impactul pe care tehnologiile le au asupra mediului. Totuși, politicile etice privind inteligența artificială depind de sistemele și valorile sociopolitice ale statelor și organizațiilor care le propun (Michel 2023, 34).

Mai mult, **politicile de securitate cibernetică** ar trebui să fie reevaluate și modificate pentru a putea cuprinde influența ultimelor evoluții din sectorul inteligenței artificiale. În esență, trebuie consolidate politicile care au ca obiectiv asigurarea securității cibernetice a sistemelor AI. Totodată, trebuie adoptate cât mai rapid politici adaptate pentru a face față noilor episoade de dezinformare din 2024, înainte de începutul campaniilor electorale. Așadar, **actualele strategii și politici vor trebui să facă față noilor evoluții, unde cel mai important aspect este creșterea amplitudinii atacurilor cibernetice, mai ales cele ale actorilor non-statali și din sfera criminalității cibernetice**. Ultimele evoluții din domeniul AI ar trebui integrate în discuțiile internaționale din cadrul ONU pentru reglementarea internațională a activităților actorilor statali în spațiul cibernetic.

Diplomația digitală și cibernetică pot contribui la gestionarea la nivel internațional a evoluțiilor curente din sectorul AI și al securității cibernetice. În primul rând, discuțiile ONU privind reglementarea spațiului cibernetic ar trebui să fie accelerate, existând riscul ca sistemele AI să amplifice atacurile cibernetice la nivel global. În al doilea rând, discuțiile ONU privind sistemele de AI ar trebui să țină cont de aspectele de securitate cibernetică, dar prioritățile ar trebui să fie sistemele care au un impact major asupra vieții umane – sistemele de arme autonome și cele de control social, spre exemplu. În al treilea rând, diplomația digitală devine un mijloc esențial pentru combaterea campaniilor de dezinformare online, iar sistemele AI ar putea juca un rol important pentru detectarea campaniilor. **Totuși, discuțiile și politicile privind inteligența artificială nu ar trebui să se axeze doar pe riscuri și predicții apocaliptice, tehnologia având potențialul de a fi utilizată și pentru consolidarea securității cibernetice sau pentru accelerarea progresului în medicină și știință, spre exemplu.**

Rămâne de văzut cum vor influența ultimele evoluții din sectorul AI și exportul de autoritarism digital și de tehnologii ce permit represiunea digitală. Aceste evoluții vor fi cuprinse într-un studiu care va apărea în **următorul volum al revistei IDR, *România occidentală*, în luna decembrie 2023**, într-un articol semnat de Sînziana Dumitrescu și Claudiu Codreanu, care va urmări exporturile Chinei de autoritarism digital către Africa.

BIBLIOGRAFIE

Benson, T. 2023. This disinformation is just for you. *Wired*. https://www.wired.com/story/generative-ai-custom-disinformation/.

Bezombes, P., Brunessauax, S., & Cadzow, S. 2023. *Cybersecurity of AI and standardisation*. ENISA. https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation.

Bhuiyan, J., & Robins-Early, N. 2023. The EU is leading the way on ai laws. The US is still playing catch-up. *The Guardian* https://www.theguardian.com/technology/2023/jun/13/artificial-intelligence-us-regulation.

Bjola, C. 2022. AI for development: Implications for theory and practice. *Oxford Development Studies*, 50(1), 78-90. https://doi.org/10.1080/13600818.2021.1960960.

Blumenthal, M. S., & Csernatoni, R. 2022. *Computers on wheels: Automated vehicles and cybersecurity risks in Europe*. EU Cyber Direct. https://eucyberdirect.eu/research/computers-on-wheels-automated-vehicles-and-cybersecurity-risks-in-europe.

Bonfanti, M. E., & Kohler, K. 2020. Artificial Intelligence for cybersecurity. *CSS Analyses in Security Policy*, 265. https://doi.org/10.3929/ethz-b-000417116.

Bray, S. D., Johnson, S. D., & Kleinberg, B. 2023. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1). https://doi.org/10.1093/cybsec/tyad011.

Burgess, M. 2023. Criminals have created their own ChatGPT clones. *Wired*. https://www.wired.com/story/chatgpt-scams-fraudgpt-wormgpt-crime/.

Ferguson, D. 2023. Robots say they have no plans to steal jobs or rebel against human. *The Guardian*. https://www.theguardian.com/technology/2023/jul/08/robots-say-no-plans-steal-jobs-rebel-against-humans.

Imperva. 2023. 2023 Bad bot report. https://www.imperva.com/resources/resource-library/reports/2023-imperva-bad-bot-report/

Kenny, J. 2023. Advanced artificial intelligence techniques and the principle of non-intervention in the context of electoral interference. În F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Gery (ed.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 223-257). Routledge. https://doi.org/10.4324/9781003284093.

Knight W. 2023b. A new attack impacts major AI chatbots—and no one knows how to stop it. *Wired*. https://www.wired.com/story/ai-adversarial-attacks/.

Knight, W. 2023a. What really made Geoffrey Hinton into an AI doomer. *Wired*. https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/.

Maas, M. M. 2019. How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, 40(3), 285-311. https://doi.org/10.1080/13523260.2019.1576464.

Madung, O. 2023. AI hysteria is a distraction: algorithms already sow disinformation in Africa. *The Guardian*. https://www.theguardian.com/global-development/2023/aug/09/ai-chatgpt-doomerism-threat-already-here-big-tech-algorithms-sow-disinformation.

Malik, K. 2023. Fantasy fears about AI are obscuring how we already abuse machine intelligence. *The Guardian*. https://www.theguardian.com/commentisfree/2023/jun/11/big-tech-warns-of-threat-from-ai-but-the-real-danger-is-the-people-behind-it.

Michel, A. H. 2023. *Recalibrating assumptions on AI*. Chatham House. https://www.chathamhouse.org/2023/04/recalibrating-assumptions-ai.

Milmo, D. 2023a. Google, Microsoft, OpenAI and startup form body to regulate AI development. *The Guardian*. https://www.theguardian.com/technology/2023/jul/26/google-microsoft-openai-anthropic-ai-frontier-model-forum.

Milmo, D. 2023b. Doctored Sunak picture is just latest in string of political deepfakes. *The Guardian*. https://www.theguardian.com/technology/2023/aug/03/doctored-sunak-picture-is-just-latest-in-string-of-political-deepfakes.

Milmo, D. 2023c. TechScape: Can the EU bring law and order to AI?. *The Guardian*. https://www.theguardian.com/technology/2023/jun/27/techscape-european-union-ai.

Milmo, D., & Farah, H. 2023. Malicious use of AI could cause 'unimaginable' damage, says UN boss. *The Guardian*. https://www.theguardian.com/technology/2023/jul/18/malicious-use-of-ai-could-cause-huge-damage-says-un-boss.

Milmo, D., & Stacey, K. 2023. Rishi Sunak's AI summit: what is its aim, and is it really necessary?. *The Guardian*. https://www.theguardian.com/technology/2023/jun/09/rishi-sunak-ai-summit-what-is-its-aim-and-is-it-really-necessary.

O'Shaughnessy, M. & Sheenan, M. 2023. *Lessons from the world's two experiments in AI governance*. Carnegie. https://carnegieendowment.org/2023/02/14/lessons-from-world-s-two-experiments-in-ai-governance-pub-89035.

O'Shaughnessy, M. 2023. *What a Chinese regulation proposal reveals about ai and democratic values*. Carnegie. https://carnegieendowment.org/2023/05/16/what-chinese-regulation-proposal-reveals-about-ai-and-democratic-values-pub-89766.

Perez, L. 2023. Is Stuxnet the next Skynet? Autonomous cyber capabilities as lethal autonomous weapons systems. În F. Cristiano, D. Broeders, F. Delerue, F. Douzet, & A. Gery (ed.), *Artificial Intelligence and International Conflict in Cyberspace* (pp. 186-222). Routledge. https://doi.org/10.4324/9781003284093.

Rosert, E., & Sauer, F. 2021. How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy*, 42(1), 4-29. https://doi.org/10.1080/13523260.2020.1771508.

Service, R. 2023. „Could chatbots help devise the next pandemic virus?". *Science*, 380(6651), 1211. https://www.science.org/doi/epdf/10.1126/science.adj3377.

Spitale, G., Biller-Andorno, N. & Germani, F. 2023. „AI model GPT-3 (dis)informs us better than humans". *Science Advances*, 9(26), 1-9. https://www.science.org/doi/epdf/10.1126/sciadv.adh1850.

Stokel-Walker, C. 2023. TechScape: Turns out there's another problem with AI – its environmental toll. *The Guardian*. https://www.theguardian.com/technology/2023/aug/01/techscape-environment-cost-ai-artificial-intelligence.

Tiffany, K. 2021. Maybe you missed it, but the Internet 'died' five years ago. *The Atlantic*. https://www.theatlantic.com/technology/archive/2021/08/dead-internet-theory-wrong-but-feels-true/619937/.

Whyte, C. 2023. Learning to trust Skynet: Interfacing with artificial intelligence in cyberspace. *Contemporary Security Policy*, 44(2), 308-344. https://doi.org/10.1080/13523260.2023.2180882.

Yudkowsky, E. 2023. Pausing AI developments isn't enough. We need to shut it all down. *TIME*. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.