AI がサイバー攻防のバランスとサイバー紛争の性質に与える影響

アンドルー・J・ローン [1]

ジョージタウン大学 セキュリティ・新興技術センター

*drew.lohn@georgetown.edu

要約

他の紛争領域や、AI によるリスクが高いと予想される他の分野とは異なり、サイバー領域は本質的にデジタルであり、AI のトレーニングとサイバーアプリケーションの間には緊密なフィードバックループが存在する。サイバーは AI による最大かつ最も早い影響を受ける分野である可能性があるため、AI の進歩に伴いサイバー領域がどのように変化するか理解することが重要である。我々のアプローチでは、文献をレビューし、サイバー紛争における攻撃的優位性について提案されている 9 つの議論と、防御的優位性について提案されている 9 つの議論を収集した。 さらに、ヒーリー、ジャービス、ナンドラジョグが個別に収集した、サイバー紛争と競争の特性に関する 48 の主張も追加で取り上げた。次に、これらの主張や提案が、AI の進歩の程度によってどのように変化するかを考察した。AI が攻撃力や防御力を広く強化するかどうかについて、サイバー領域は多面的すぎて単一の答えは出せないことがわかった。AI は、ある側面を改善し、別の側面を阻害し、またある側面は変化させないだろう。 我々は、AI がサイバー攻防のバランスとサイバー紛争および競争の特性に与える影響について、44 の見解を収集し、提示する。

キーワード：AI、サイバー、攻防のバランス、紛争、危機

目次

要約 1

目次 2

1. 序論 5

2. 攻防のバランス 7

3. 攻撃者と防御者の種類 8

3.1 サイバー脅威アクター 8
3.2 防御側または攻撃対象 9

4. AI の改善レベル 11

4.1 現状 11
4.2 信頼性が高く自律的 11
4.3 専門家レベル 12
4.4 ハードリミット 12
4.5 限界の突破 13

5. 進歩へのアクセス 14

5.1 能力へのアクセス制御 14
5.2 能力へのアクセスに対する限定的な制御 14
5.3 拡散したモデルとエージェント 14

6. 評価時に考慮すべき AI 能力 15

6.1 長所 15
6.2 弱み 16

7. サイバー攻撃と防御の非対称性 18

7.1 防御側の非対称性 18

7.1.1 防御側がデジタル領域を決定する 18
7.1.2 防御はレジリエンスを持つ 19
7.1.3 防御側は自身のネットワークを監視できる 20
7.1.4 防御側はローカルリソースを調達できる 21
7.1.5 攻撃者は全ての防御を突破しなければならない 21
7.1.6 防御側は攻撃側よりも数が多い 22
7.1.7 攻撃には時間がかかる 22
7.1.8 攻撃はエスカレーションのリスクを伴う 23
7.1.9 攻撃者は正当性を欠く 24

7.2 攻撃の不均衡 24

7.2.1 攻撃側は一度の成功で十分だ 24
7.2.2 攻撃者は攻撃対象を選べる 25
7.2.3 攻撃者は攻撃のタイミングを選択する 26
7.2.4 攻撃者は目標を選択する 26
7.2.5 標的は容易に見つけられる 27
7.2.6 防御には信頼性が求められる 27
7.2.7 防御にはより多くの官僚主義が伴う 28

7.2.8 防御には体系的な脆弱性がある 28
7.2.9 防御には体系的な脆弱性がある 29

8. サイバーの特徴 30

8.1 戦術的にはネットワーク速度で動く 30
8.2 作戦・戦略レベルでは遅い 30
8.3 サイバー空間はスケールフリーネットワークである 31
8.4 サイバー空間は人為的かつ適応可能である 31
8.5 サイバー空間は計り知れないほど複雑である 32
8.6 サイバー空間には固有の脆弱性が存在する 32
8.7 能力による影響が小さく可逆的な効果 33
8.8 攻撃は諜報活動に含まれるより小さなケースである 33
8.9 技術変化の急速なペース 34
8.10 普遍的な相互接続性と依存性 34
8.11 許可不要な革新と接続 34
8.12 曖昧な境界 35
8.13 物理世界との結びつき 35
8.14 情報（及び能力）の複製容易性 35
8.15 民間セクターによる支配 35
8.16 全体的な攻撃者の優位性 36
8.17 迅速または正確な帰属の困難さ 36
8.18 直接観察が困難である 36
8.19 能力は一時的で予測困難な影響をもたらす 37
8.20 敵対勢力は絶え間なく接触し、作戦上の休止はほとんどない 37
8.21 国家の権力源に対する大陸間規模の即時的近接性 37
8.22 優位性は能力の保有ではなく使用によって得られる 38
8.23 敵対勢力は武力紛争レベルを下回る能力を日常的に使用する 38
8.24 参入障壁が低い 38
8.25 優位性は一時的なものである 39
8.26 能力は他の領域より大幅に安価である 39
8.27 能力は迅速に再生可能である 39
8.28 攻撃は壊滅的な結果を招く可能性がある 39
8.29 戦術的成功は機動力と主導権に依存する 40
8.30 戦略的成功は大胆さと主導権により強く結びつく可能性がある 40
8.31 抑止が困難である 40
8.32 防御の成功は攻撃者を挫かない 41
8.33 攻撃の警告は困難である 41
8.34 意図の伝達は問題を抱えている 41
8.35 先制攻撃兵器となる可能性が高い 42
8.36 意外性がより重要である 42
8.37 国家が代理勢力を利用しやすい 42
8.38 攻撃と防御は類似しており、相互に影響し合う 42
8.39 概念の混乱と明確な定義の欠如 43
8.40 不十分かつ競合する認可 43
8.41 指揮統制の困難さ 44
8.42 機密指定が過剰である 44
8.43 戦術的交戦が分析の基本単位である 44

8.44 紛争はサイバー空間内で水平的・垂直的にエスカレートするが、まだその外には及んでいない.....45
8.45 サイバー紛争はエスカレーション、誤算、不安定化を招く可能性がある.....45
8.46 インターネットには共通モード障害がある.....46
8.47 多数のデバイスが優位性を与える 46
8.48 サイバー空間は持続的に活動する者に有利である.....46

9. AI がサイバーに与える影響 48

9.1 デジタル生態系の変化.....48
9.2 デジタル環境の強化.....49
9.3 デジタル戦闘における戦術的側面 51
9.4 インセンティブと機会 53
9.5 紛争と危機への戦略的影響 54

10.結論 56

著者 57

謝辞 57

資金提供 57

著作権 57

参考文献 58

1. 序論

AI が影響を与えうるあらゆる領域の中で、サイバーセキュリティとサイバー紛争は最も重要なものかもしれない（Silver, 2022; Younis, 2023）。デジタル領域であるサイバーは、AI の進歩を取り入れるのに自然に適している。また、サイバーのデジタル的性質は、通常の運用、計画的な実験、自律的な自己対戦を通じて有用なデータが容易に生成される点で、AI の発展にも有益である。核戦争や生物兵器開発といった他の領域のリスクほど直接的な災害をもたらすことは通常ないが、サイバー攻撃は原則として、ほぼ全ての他の領域で災害を引き起こす可能性がある（Radoini & Siddiqui, 2021; S. Smith, 2021）。

AI が進化を続け、ほぼ全ての領域に組み込まれるにつれ、それらの領域がどう変化するかの手がかりはより価値を持つようになる。変化が良い方向か悪い方向かはともかく、変化が起こり得ることを知ることで、社会は適応と準備のための時間を得られる。技術の未来に関する予測は極めて不確実だが、可能性のある未来を考察することで、政策立案者、規制当局、そして社会全体が、リスクの少ない望ましい未来へと導く決断を下せるようになる。

本稿では、AI がサイバーセキュリティとサイバー紛争の性質に与える影響を包括的に考察し、AI が攻撃側か防御側に利益をもたらすかを問う。サイバー領域が歴史的に攻撃的バイアスか防御的バイアスかを巡る議論は数十年にわたり続いている。学者の見解は攻撃側に傾いているように見えるが、合意は形成されていない（Slayton, 2017a; Smythe, 2020）。 本報告書は、この歴史的議論を解決しようとはせず、また将来の AI の進歩によってサイバーが攻撃的か防御的かにバイアスを受けるかを解決しようともしない。我々は、AI が現在の状態よりもサイバーを攻撃的または防御的にバイアスを受ける可能性について、議論を提起することのみを試みる。 サイバーが現在、あるいは将来、攻撃的であるかどうかに関わらず、将来、攻撃的になるか防御的になるかを把握し、防御にさらにバランスを傾ける方法を見つけることは価値がある（Alperovitch & Healey, 2020; Janow, Rattray & Venables, 2017）。

方法論としては、サイバー攻撃と防御に関する文献をレビューし、この分野における著名な著者数名に協力を依頼した。このレビューの目的は、サイバー攻撃と防御のどちらにバイアスをかけるべきかについて、これまで提唱されてきた議論を抽出することである。議論を収集した後、AI が各議論に与える影響を考察し、AI がサイバー攻撃と防御のどちらにバイアスをかける方向にサイバーを導く可能性があるか、またその方法を評価した。

我々の文献レビューが意図せずサイバー攻防のバランスに関する側面を省略した可能性は十分にある。網羅的とは言えないが、最も顕著な議論は捉えていると考える。 また、このプロジェクト開始時点では存在しなかった、サイバーの攻防バランスという問題に直接的ではないものの、我々が発見した他のどの論文よりもはるかに広範な論文も発見することができた（Jason Healey、n.d.）。 ジェイソン・ヒーリー、ディヴィヤム・ナンドラジョグ、そして攻防のバランス理論の創始者とされる故ロバート・ジャービス（Jervis, 1978）によるこの論文は、サイバー紛争と競争の特徴を規定する要素をまとめたものである。この論文も、サイバー紛争の特徴を完全に網羅しているわけではないが、48 の要素をすべて検討する価値のある試みである。

この序論に続き、報告書は、サイバーにおける攻撃的および防御的バイアス、そして攻撃的または防御的バイアスを持つことの影響について記述している。報告書の第 3 節では、さまざまな脅威アクターと、さまざまな標的または防御者について記述している。第 4 節 では、AI の進歩の可能な程度について記述している。現状維持、信頼性が高く独立、専門家レベル、ハードリミット、リミットブレイクである。第 5 節では、AI の進歩へのアクセスを制御する可能性について記述している。 第六節では、AI がサイバー攻防の均衡やサイバー紛争の性質に及ぼす可能性についてブレインストーミングを行う際に用いた、様々な強みと弱みの概要を示す。第七節と第八節では、それぞれ AI の予想される影響、サイバー攻防の均衡、サイバー紛争の特性に関するアセスメントを記載する。第九節では、AI がサイバー領域に与える影響について、第七節と第八節の論理展開を統合する。第十節で結論を述べる。

第 7 章と第 8 章で評価した各要素は、それぞれ単独で詳細な報告書を要する。本稿での評価は予備的かつ不確実なものと見なすべきである。これらは、AI がサイバー攻防バランスとサイバー紛争の性質に及ぼす影響を大局的に把握しようとする試みである。本稿で特定した多くの個別要素について、今後より詳細な調査を実施する予定であり、研究コミュニティの他者にも独自のアセスメントを行うよう呼びかける。 個々の要素に関する結論の一部は、さらなる調査や時間の経過によりアセスメントの不確実性や無知が解消されることで変更される可能性がある。本報告書が、意

思決定者や防衛担当者を準備させるのに十分な、AI がサイバー領域に及ぼす影響の包括的な見解を提供すると同時に、サイバーおよび AI 研究者が調査すべき重要な問題の体系を提供することを望む。

2. 攻防の均衡

攻防の均衡は、サイバー領域という文脈よりも国際関係論においてより一般的な概念である。 これは元々、紛争発生の可能性に対するアセスメントおよび影響を与える手段として提唱された（Jervis, 1978）。より狭義のサイバー領域でも頻繁に再検討され、最近では AI が一般的な攻防バランスとサイバー固有の攻防バランスの双方に及ぼす影響が議論されている（Corsi, Kilian & Mallah, 2024; Tang, Saade & Kelly, 2024）。 技術や手法が攻撃よりも防御に大きな利益をもたらす場合、双方の攻撃意欲は低下し、逆もまた然りである（Garfinkel & Dafoe, 2021）。理論上、この均衡は安定性または不安定性の尺度を提供する。

一般的な軍事状況では、歴史的に防御が優先されてきた。経験則として、侵略軍は防衛軍よりも 3 倍以上の規模である必要がある（Mearsheimer, 1989）。サイバー領域では、その反対、つまり攻撃側が有利であるというのが一般的な見解だ（DoD, 2011; ヒーリー, 2021a; ロカテッリ, 2013）。これはコンセンサスではなく、地上侵攻において攻撃側が防御側の 3 倍の兵力を必要としないのと同じように、必ずしも常に当てはまるわけではない。

攻守のバランスは、安定と不安定、あるいは攻撃の傾向といった結果以外にも、さまざまな結果を指すことがある。例えば、金銭的な支出という観点から捉えることもできる。防衛支出が攻撃支出を上回っていることは、攻撃が有利であることを意味すると言われるが、攻撃は費用がかかり、その見積もりも大きく異なる（Greenfield et al., 2023）。[1]

攻防のバランスは損害の程度を指すこともある。金銭的意味では損害は甚大だった。軍事的意味では影響はより控えめだった。 多くの論者は、サイバー攻撃の影響は歴史的に小さく可逆的だと主張する。ただしこれは、貴重な情報を持続的に収集するサイバーの価値を過小評価している可能性がある。ロシアのウクライナ侵攻におけるサイバーの役割が限定的であり、彼らが従来型戦域を優先した事実は、サイバーが攻撃に提供する価値が従来考えられていたほど大きくないという主張に一定の信憑性を与えている（寺島、2022 年）。

あるいは、攻撃と防御のバランスは、脆弱性や強制力という観点から捉えることもできる。攻撃者は、攻撃のリスクを軽減しながら、影響力を高めることができる。例えば、サイバー空間への参入障壁が低いため、小規模で技術力の低い国でも、より先進的な国の多くのデジタル目標を脅かすことができる。 サイバー攻撃は攻撃者の影響力を高める一方で、「圧力解放弁」としての役割も果たすため、こうした攻撃の可能性や影響は減少する可能性がある（Healey & Jervis, 2020）。さらに、攻防のバランスは、困難さ、不便さ、効率性にも関連している。30 個ものパスワードを覚えておかなければならないことや、エアギャップのある施設で作業しなければならないという不満は、攻防のバランスが、望ましいよりも攻撃側に偏っている結果である。

関心のある結果の可能性をすべて検討すると、単一の攻防バランスを定義することは、おそらく実現不可能であり、望ましくもないだろう。攻防バランスを、そのさまざまな合理的な解釈や意味合いをすべて満たす形で定義することが可能であったとしても、そのバランスが攻勢に有利か、それとも防御に有利かを評価することは、おそらくあまり意味がないだろう。 攻防のバランスがどちらに傾いていようと、目標は技術と 政策を発展させ、バランスをさらに防御側に押しやることだ。従来の戦争領域である空・陸・海にサイバー領域が加わったことで、「戦略的に有用な問いは、サイバー能力の出現が国家間の攻防バランスをより広範にシフトさせたかどうか、そしてその方法である」（Huntley & Shives, 2024）。

サイバー領域に AI を加える場合、戦略的に有用な問いは、AI が攻防バランスを変化させるかどうか、そしてその方法である。この変化は国家間だけの問題ではない。国家は重要な攻撃者、標的、防衛主体だが、それだけが全てではない。次の節では、AI がサイバー領域に与える影響の文脈で考慮すべき脅威アクター、標的、防衛主体について述べる。

---

[1] 本報告書の著者の一人は以前、攻撃作戦のコスト見積もりが 6 桁も異なることを以下の文献の図 4.1 で示している：V.A. Greenfield, J.W. Welburn, K. Schwindt, D. Ish, A.J. Lohn, and G.S. Hartnett, “サイバーセキュリティとサプライチェーンリスクマネジメントは単に加算的ではない,” RAND Report A532-1 (2023)。

3. 攻撃者と防御者の類型

攻撃と防御のバランスがライバルの攻撃性を決定するという概念は、サイバーの文脈以外の国際関係学の分野で生まれたものである（Jervis, 1978）。その結果、文献は国家間の紛争に焦点を当てているが、他にも多くの攻撃者や防御者が存在する。一部の研究は、さまざまな攻撃者や標的を含むように拡大している（Healey, 2021b）。その中には、国際的な国境によって防御されている場合とそうでない場合があり、学校、病院、石油パイプラインなどをランサムウェアで攻撃する犯罪者も含まれる。さらに、サイバー初心者は、入手しやすいハッキングツールを使用して、世界中のサーバーを混乱させる可能性がある。また、国家のハッカーは、自国企業に経済的利益をもたらすために、個々の企業を標的にする場合もある。

AI がこれらのさまざまな攻撃者や防御者に与える影響はそれぞれ異なるため、その影響を区別するために、それぞれをいくつかのカテゴリーに分類する。この分類は明確なものではない。1 つの組織が 2 つ以上のカテゴリーに部分的に該当する場合もある。

3.1 サイバー脅威アクター

カナダ政府は脅威アクターのカテゴリーを次のように要約している：国家、サイバー犯罪者、ハクティビスト、テロリスト集団、スリル追求者、内部脅威（Cyber Centre, 2022）。IBM も同じ分類を使用しており（IBM, 2023）、本稿でもこれに従う。

国家

国家は世界で最も高度かつ秘密性の高い能力を有する。攻撃の分散や作戦全体の指揮統制を行う否認可能なネットワークなど、大規模なチームと重要なインフラを保有している。攻撃的サイバー活動の正当な使用を主張できる唯一の脅威アクターである。サイバー攻撃の目的は様々であり、外交的圧力への感受性も異なる。

サイバー犯罪者

最も破壊的で問題のある攻撃の多くは、通常は金銭的利益を目的として犯罪者によって行われている。代表的な例はランサムウェアやその他の恐喝手段だ。こうした犯罪者は、外国の被害者からの報復から防御される国に拠点を置くことが多く、ホスト国から指示や影響を受けることもあり、犯罪者と国家主体の境界が曖昧になる場合もある（CISA, 2022）。犯罪組織は比較的規模が大きく、高度な技術を持つ場合もあり、時には副業として活動する国家ハッカーで構成されることもある。

サイバーテロリスト

サイバーテロリストは、混乱や破壊を目的とする。ならず者国家はこのカテゴリーと国家カテゴリーを曖昧にすることがある。サイバーテロリストは比較的資金力がある可能性があるが、歴史的に見て洗練度や技術力は比較的低い。

ハクティビスト

ハクティビストは政治的・社会的イデオロギーを動機とする。その攻撃は国家やサイバー犯罪者が行うものと比べて破壊的ではないことが多い。典型的な攻撃はウェブサイトの改ざんや証拠となるデータの漏洩だ。ハクティビストは資金力が乏しく、技術的能力や洗練度も低い傾向がある。

スリル追求者

スリル追求者は好奇心や娯楽のためにハッキングを行う。技術的洗練度は低いが、公開されているハッキングツールにアクセスできるスクリプトキディの大半がこのカテゴリーに属する。また、無防備な標的を調査中に偶発的な損害を与える可能性のあるグレーハットハッカーもスリル追求者に含まれる。

内部者による脅威

内部関係者は、アクセス権限の高さゆえに、技術力が低くても特に大きな損害を与え得る。データ窃取やシステム・コードベースへの妨害行為が可能だ。内部関係者は不満を抱えた個人である場合もあれば、国家やイデオロギーに駆られたハクティビスト、テロリストに勧誘される場合もある。善意であっても、内部関係者が重大な過ちを犯す可能性もある。

3.2 防御側と攻撃対象

防御者と標的は必ずしも同一の事業体ではない。例えば防御業務の一部は外部委託される場合がある。標的となった組織は、ある企業製のアンチウイルスと別の企業提供の DDoS 対策を利用しているかもしれない。さらに防御はデジタルエコシステム全体で展開される。ソフトウェア開発者はセキュリティのベストプラクティスに従い、防御テストを実施し、防御パッチを発行する。災害復旧組織でさえ、サイバー防御の重要な側面であるレジリエンスの改善を提供している。

国家政府

政府を構成する省庁や機関は、一般的なサイバー攻撃の標的となる。国防総省の一部や核兵器備蓄を開発・管理する国立研究所が該当する。政府職員の機密情報を保持する人事管理局や、全市民の記録を管理する内国歳入庁も標的となり得る。さらに、連邦緊急事態管理庁やメディケアのような国内サービスを提供する組織も標的となる可能性がある。

政府は情報共有や標準設定、ネットワーク監視、侵入テスト、ツール開発など、防衛支援のための様々なサービスを提供する。国家政府はまた、敵対ネットワークへの侵入能力と権限を持つことで防衛において独自の役割を果たす。敵の能力や意図に関する警告を発する。さらに、発生中の攻撃を妨害したり、予測される攻撃を未然に防いだりするデジタル作戦を実行する。

重要インフラ

米国では重要インフラは、連邦政府ではなく州政府、地方政府、部族政府、準州政府によって主に管理されている。重要インフラは化学、水道、原子力、運輸など 16 の分野に分類される（CISA, n.d.）。これらのシステムは複雑で多様かつ数も多いため、全てを十分に防御することは困難だ。重要インフラはサイバー攻撃の標的としても影響力が大きいため、これらのシステムを特に効果的に防御することが重要となる。

大規模組織

大規模組織は通常、防御のための十分な資源を有しているが、高度で持続的な攻撃の標的となることが多い。防衛契約業者、大手銀行、大規模テクノロジー企業などがこれに該当する。各セクターは軍事的優位性、経済的混乱、知的財産窃盗など異なる理由で標的とされるが、防御面では共通点がある。いずれも多額の資金を投じて熟練したデジタル防御チームを維持し、24 時間体制でセキュリティ脅威や侵害の可能性を監視・対応する能力を有している。

こうした組織の中には、他の標的の防衛に貢献するものもある。サイバー戦が行われるソフトウェアやデジタルインフラを開発・管理しているのだ。攻撃の防止や阻止の多くは、顧客の監視、脆弱性の修復、情報共有を行うこれらの企業に委ねられている。

中小規模組織

小規模組織は、大規模な防御チームや最も優秀な防御要員を確保する余裕がない。それでもこれらの組織は重大なリスクを抱えており、例えば重要インフラ部門の大半を占めている。小規模組織はサプライチェーンの重要な拠点となり得るため、その機能不全は大規模なセクターを混乱させたり、攻撃者がより注目度の高い標的へ到達するための脆弱な侵入経路を提供したりする。

個人

国家も世界も、それぞれが標的となり得る個人で構成されている。個人は暗号通貨ウォレットや銀行口座から資金を盗まれる可能性がある。偽情報や窃盗の標的となることもある。選挙干渉キャンペーンやエクイファックスのハッキングのように集団的に標的とされることもあれば、ナイジェリア王子詐欺のように個別に狙われることもある。

個人はある程度、自身のデジタル防御に責任を負う。ウイルス対策ソフトのインストールや二要素認証の利用は、個人が取る一般的な防御策だ。人々は暗号化メッセージアプリを選択し、不審なメールやテキストメッセージをクリックしないようにする。個人が高度なスキャンを実行することは稀だが、バックグラウンドで気付かれずに動作する自律型防御ツールを採用することはあり得る。

4. AI の進化段階

AI は進歩が続くにつれ、様々な程度で発展するか、あるいは様々な発展段階を経る可能性がある。 技術的進歩とその進歩に対する社会学的反応を確信を持って予測するには、変数と不確実性が多すぎる。それらの変数には、独創性、自律性、信頼性、速度、拡張性の程度などがある。これらの変数は相互に関連している。例えば、速度や規模は、人間が判断の評価を行う時間がないという点で、ある程度の自律性を意味しうる。信頼性も、監視が少なくて済む場合、自律性をもたらし、それがさらに拡張性を可能にする。

本節では、AI 発展の四段階を説明する：現状維持、信頼性・自律性あり、専門家レベル、ハードリミット。これら四段階は、AI が技術的・社会学的側面で進歩する際に生じる複数の変数や不確実性を類型化することを目的とする。技術的・社会的側面の全てが同時に進歩する必要はない。 例えば、将来は専門家レベルの独創性を必要とする効果的な脆弱性発見技術が存在する一方で、現状維持から信頼性・自律性レベルへ進化できなかった信頼性の低いコード生成技術が共存する可能性がある。

4.1 現状維持

コンピュータは速度や記憶容量など様々な点で長年人間を凌駕してきたが、創造性においては人間に劣っていた。AI は既にほぼ全てのタスクにおいて人間の効率向上を可能にしており、技術的進歩がなくてもこの傾向は続くだろう。AI の利用拡大やサイバーセキュリティツール・タスクへの統合が進むにつれ、こうした効率化はさらに進む。 しかし現状、AI は完全な自律性を与えるには信頼性が低すぎる。また低スキル層を大幅に引き上げることもできない。AI はエラーを起こしやすく、改ざんや操作に脆弱である（A. J. Lohn, 2020b）。この信頼性の低さが、その独立性と拡張性を制限している。

現状レベルでは、AI はサイバーセキュリティ実務者の業務加速に寄与し得る。専門家が脆弱性を発見する支援を開始し、防御側がパッチを作成する手助けや攻撃側がエクスプロイトを仕掛ける支援を行っている。AI はプログラマーが 1 時間あたりより多くのコードを生成するのを助け、Security Copilot のようなツールは防御側が日常的に殺到する侵入の可能性を示すアラートを管理するのを支援する（Hulme, 2024）。

しかし進歩は鈍化し、こうした限定的で欠陥のある補助ツールを超える進展は見られない可能性がある。AI の進歩は様々な計算資源や資金調達上の課題に直面している（A. J. Lohn, 2023）。またデータ限界に近づきつつあり、インターネット上の全人類生成データに迫っているとも言える（Villalobos et al., 2024）。

特にサイバーのようなデジタル領域では、合成データセットを作成することでデータ制限を克服できる可能性がある。そのため、AI アルゴリズムやトレーニングのさらなる進歩がなくても、現状を超える進展が可能かもしれない（A. Lohn, Knack, Burke & Jackson, 2023）。しかし、そのデータがどれほど簡単に作成できるか、どれほど有用か、そしてサイバーセキュリティのどの側面が恩恵を受けられるかは不明だ。 例えば、攻撃者の戦術が適応する速度に AI の訓練が追いつくのは難しい。なぜなら、新たに開発された戦術や防御側によってまだ識別されていない戦術に関するデータは常に不足しているからだ（Moore, Kenneally, Collett & Thapa, 2019）。 これは、これまでサイバーセキュリティに AI を活用しようとした試みにおける課題であった。推論スケーリング、思考連鎖推論、合成データ生成といった技術が期待通りの成果を上げなければ、サイバーセキュリティ向け AI の側面は現状で停滞する可能性がある。

4.2 信頼性と独立性

AI は現在、高い誤り率のために制限されている。AI は通常、段階的な指示と、重要な出力のすべてに対する人間のレビューを必要とする。もし AI が信頼性が高く自律的であれば、人間の指導や監督が及ぶ限界をはるかに超えて拡大できるだろう。また、人間が監督できる速度よりも速く動作させることも可能になる。

このレベルの能力は、人間の専門家から得られる創造性や独創性を必要とする稀な側面よりも、人間の専門家から得られるデータが豊富なサイバーセキュリティの一般的な側面において達成しやすいかもしれない。サイバーセキュリティ

の一般的な側面には、ベストプラクティス文書や標準規格など、従うべき具体的な指針や指示が存在するケースも多い。一般的で十分に検証された指示に従うことは、独創性を示すことよりも達成可能な目標かもしれない。

AI モデルやシステムは人間より信頼性が高いかもしれないが、それでも重要システム向けに通常認証される構成要素と比べると、信頼性は桁違いに低い（A. J. Lohn, 2020a）。信頼性の高い AI コンポーネントを利用するためにも、新たな認証プロセスや標準が必要となる可能性がある。また、AI は意図的な操作や乗っ取りに対して著しく脆弱であることも知られている（A. J. Lohn, 2020b）。こうした欠点があるにもかかわらず、脆弱性のある AI でさえ、人間がしばしば最も脆弱な要素と見なされるサイバー環境においては、セキュリティ上の利点をもたらす可能性がある（Schneier, 2018）。

4.3 専門家

人間の専門家は、ほとんど指導や指示を必要としない。高レベルの指示と意図さえあれば、行動方針を決定し、ほとんど誤りなくその行動を実行できる。未指定の行動方針を構築する能力は、AI により多くの主体性を与えるだろう。人間の専門家はまた、創造性を活用して、稀で参照可能なデータが少ない個別のタスクを遂行できる。AI 専門家は、脆弱性の発見やステルス攻撃戦術の使用といった個別のタスクを、最高の人間と同等に効果的に遂行できるようになるだろう。

専門性は稀有で際立った特性であるため、専門性を拡大することは基礎的能力の拡大とは異なる効果をもたらす可能性がある。これにより防御側、国家、脅威アクター間の競争環境が大幅に平準化されるだろう。 小規模組織でも創造的で想像力豊かかつ多才な防御者を活用できるようになる一方、より広範な脅威アクターが高度な持続的脅威レベルの専門知識にアクセス可能になる。とはいえ、全ての脅威が高度に洗練されるわけではない。単純な攻撃が効果を発揮する事例は多く、防御効果を向上させなければこの状況は変わらない。

AI システム自体も、改ざんや操作に対して脆弱なまま残る可能性がある。歴史的に見て、AI モデルは能力が高まるほど操作されやすくなってきた（Ilyas et al., 2019）。人間の専門家も騙される可能性があり、AI システムと同様に、操作の試みが対象または類似対象で数十億回テストされた後に展開される場合、操作が容易になるかもしれない。AI モデルの能力が高まるほど、攻撃が潜伏できる入力空間が増え、攻撃者が利用できる内部計算も通常増加する。

4.4 ハードリミット

AI は最高の専門家さえも凌駕する能力を獲得する可能性がある。このシナリオは、訓練データが人間の生成に依存する必要がなく、 人間が全体を把握するには多様かつ複雑すぎる領域でより現実的だ。こうした領域では、AI は人間がこれまで見たことのない量のデータに触れ、個人が同時に理解できる範囲を超える数のサブ分野間の関連性を構築できる点で優位性を持つ。

サイバーセキュリティはまさにそのような領域だ。リバースエンジニアリング、ネットワークアーキテクチャ、コンパイラ設計、異常検知、ウェブインターフェースは、いずれも関連しつつも独立したサブ分野である。 人間の専門家がサイバー分野の全サブ分野を網羅することは不可能だ。また、主にデジタル領域であるため、人間が直接関与する必要のない膨大なデータを自律的に生成・収集することも可能である。チェスや囲碁を自ら学ぶように、AI エージェントは自身とのサイバー戦闘をシミュレートし、人間が教えられる以上のことを習得できる（A. Lohn et al., 2023）。

しかし、AI が万能になるわけではない。極めて高度な知能であっても、実現可能なことには厳しい限界が存在する。数学的に不可能と証明されるものや、実現不可能なリソースを必要とするものは存在する。例えば、ランダムなパスワードを総当り攻撃で解読するには平均的な推測回数が必要であり、利用可能な計算能力を超える速度での解読は不可能だ。あるいはもっと単純に、攻撃者は提供された帯域幅を超える速度で情報を流出させられず、提供されたメモリ容量を超える情報を保存できない。

より複雑なシステムでは、形式的に安全性が証明可能なソフトウェアが存在する。それらの証明は基盤となるハードウェアの信頼性といった仮定に依存するが、超知能 AI でさえソフトウェア自体を直接破ることは不可能だ。これにより防

御側はソフトウェアそのものではなく、仮定を保護するというより小さな作業に集中できる。一方、現在使用されている暗号化の大半は証明可能な安全性を有しておらず、専門家は単に解読が困難だと信じているに過ぎない。 証明可能な安全な手法に切り替えない限り、十分な知能を持つ AI は現行の暗号化を破る可能性がある。防御の観点からは、そのような知能が証明可能な安全な実装の設計に役立つかもしれない。

4.5 限界の突破

AI が我々が想定する限界を超越するほどの知能を獲得する可能性を提唱する理論もある。 本報告書では、このレベルの進展は対象外とする。AI が停止問題やライスの定理（Kozen, 1977）のような決定不能な問題を論理的に扱う可能性もある。あるいは、情報の保存や隠蔽に関するエントロピーの限界や信号対雑音比に関するシャノンの法則（Shannon, 1949）を破るかもしれない。カオスシステムの挙動を予測する可能性すらある。

AI がこれらの限界を破る可能性は低そうだ。むしろ、十分に知能的な AI は、限界そのものを破るよりも、限界に関する仮定やそれに依存するシステム自体を無効化する予期せぬ方法を創出する可能性が高い。

いずれにせよ、定義上限界を持たない知性について論じるのは難しい。

5. 進歩へのアクセス

AI の進歩の度合いだけでなく、その進歩がどのように実現されるかも重要だ。誰がアクセスできるか、コストはどれほどか、メモリ消費量はどうか、動作速度はどれほどか、そして最高・最速・最安のものとそれ以外の差がどれほどあるかが問題となる。

5.1 能力への制御されたアクセス

技術の開発は困難を伴うことがあり、市場シェアの維持や知的財産の防御、輸出規制などの理由で守られることもある。AI 技術は確かに普及するが、最も高性能な AI モデルやシステムは、アクセスを制限する一社あるいは少数の組織内に閉じ込められる可能性がある。「高性能」の定義には考慮すべき要素が多数存在する。通常は精度で測られるが、用途によっては創造性、速度、効率性などの特性がより重要となる場合もある。競争がスピード勝負なら、「賢い」モデルは愚かで速いモデルに敗れるかもしれない。 攻撃や防御の総量で勝負が決まる場合、安価で単純なモデルが優勢となることもある。また、あらゆる面で優れたモデルやエージェントであっても、タスクが単純すぎて劣ったモデルで十分なら、何の利点も生まない。

この制御されたアクセスシナリオは複数の形で生じうる。モデルやエージェントの訓練・運用に多額の費用がかかり、資金力のある少数の組織にしか実現不可能である場合だ。あるいは、新たなアルゴリズムやデータ生成手順といった秘密の概念的進歩や特許技術が差別化要因となる場合もある。新しいコンピュータチップ設計や製造技術が、模倣困難な優位性をもたらす可能性もある。

あるいは、サイバー応用における AI への制御されたアクセスは、AI 全般における寡占を必要としない場合もある。組織が、AI モデルやエージェントがサイバーセキュリティ応用で効果を発揮するために必要なツールやデータ群を開発する可能性がある。その場合、政府が機密データとツール群を保有することで、AI 全般をリードする必要なく、サイバー分野、特に攻撃的サイバー分野における AI で優位性を維持し続けることが想像できる。

5.2 能力へのアクセス制限

たとえ少数の組織しか有能な AI サービスを提供できなくても、その製品やサービスの使用方法を制限したり影響を与えたりする能力は限定的かもしれない。例えば、プロバイダはクライアントの受け入れ・拒否を選択し、どの要求を受け入れるか拒否するかを選択し、フォレンジック調査のためにユーザー活動を記録することができる（Ji, Goldstein & Lohn, 2023）。

しかし、膨大な数のユーザーとリクエストが存在する場合、こうした措置は事実上不可能となる可能性がある。

また、ユーザーが倫理観の低いプロバイダに移行できる場合、プロバイダの制約はさらに弱まる。代替プロバイダには敵対国のホストや、より寛容な方針・思想を持つプロバイダが含まれる可能性がある。プロバイダがサービスの制約や管理方法について意見が分かれる寡占状態でも、一定の合意形成の余地は残る。例えば敵対国同士でも、テロリストユーザーや活動、特定の犯罪行為を制限することには合意できるかもしれない。

5.3 拡散したモデルとエージェント

別の可能性として、サイバー分野における AI 能力が広く拡散し、不正なユーザーや AI 利用を制限・監視することが困難になるケースが考えられる。これは、競争力のあるモデルやエージェントを作成・複製するのに必要な費用や専門知識が比較的低い場合に発生しうる。また、開発者の一部が自らのモデルやエージェントを公に公開することを選択した場合にも起こりうる。 モデルを公開せずに、プロバイダはクローズドソースの製品を提供し、ユーザーのデバイスやサーバー上で動作させることも可能だ。検査や改変を不可能にする方法すら考えられる（Brundage et al., 2020）。その場合、製品やサービスへのアクセスを制限する手段はあるかもしれないが、その使用方法を制限したり監視したりする機会はほとんどないだろう。

6. 評価時に考慮すべき AI の能力

AI はコスト削減、速度向上、誤り率低減など、様々な面で性能向上を約束する（Pratt, 2024）。専門家はバイアス、セキュリティ脆弱性、誤り率増加などの潜在的な欠点も予測している（Bin Rashid & Kausik, 2024）。本節ではこうした提案されている長所と短所の一部を列挙する。本節の長所と短所は互いに重複したり、矛盾することもある。 我々は、いずれが真実であるかについて一切の主張を行わない。これらのリストは、後述する構造化されたブレインストーミングにおいて、AI がサイバーセキュリティをどのように変える可能性があるかについての導入として用いられた。

6.1 長所

速度

コンピュータは既に高速に動作している。攻撃的・防御的サイバーセキュリティの多くの側面は既に高度に自動化されており、AI によって加速されることはない。さらに、特に攻撃的活動など一部の活動は、不審な行動として検知されるのを避けるなどの理由で、意図的に遅く実行される。しかし、人間のアナリストやオペレーターがボトルネックとなっている領域では、AI によって既に高速なこの領域がさらに加速される可能性がある。

可用性

可用性とは主に、AI が疲労や警戒心の低下なく 24 時間 365 日稼働できる能力を指す。この利点は、交代勤務で 24 時間体制を維持できる人間のチームと比べた場合よりも、個々の人間と比較した場合により明確である。

専門知識へのアクセス

AI は、そうでなければ入手困難な専門知識や経験を提供する。それは複数プログラミング言語の知識、500 ページに及ぶサイバーセキュリティ標準の逐語的な内容の把握、あるいはファイアウォール設定のベストプラクティスなどである。これは特に、一流の人材を確保できない組織に適用可能だ。

拡張性

AI を含むデジタルシステムは通常、任意の回数だけ容易に複製できる。AI システムの数が増えるだけでなく、オペレーターの数や効率も向上し、人間の実効規模を拡大する。またコスト削減により、防御側・攻撃側双方が実施可能な対策の数を増やせる。システム数に加え、個々のシステムがより大量のデータを分析するなど、タスクや運用規模を拡大することも可能だ。

複雑性とサイロの打破

AI は人間が分析・理解できる範囲を超える複雑なデータやシナリオを解釈できる。また、より広範な情報源からのデータを考慮できるため、重要な情報が別々の担当者に分散している場合、人間が困難とする関連付けを可能にする。

精度の向上

AI の精度は、ますます多くの分野で人間を上回っている（Salem, Azzam, Emam & Abo 2024）。一方で、AI は虚構を生成しやすい（A. L. Smith, Greaves & Panch, 2023）。AI が人間より信頼性が高い場合でも、重要システム向けに通常認証される構成要素と比べると、信頼性は桁違いに低い場合がある（A. J. Lohn, 2020a）。 誤りを犯す可能性のある人間に対応するように設計されたシステムは、誤りを犯す可能性のある AI コンポーネントと共存可能にするために、再設計と再評価が必要となるかもしれない。

日常的な作業

コードの検査やリファクタリング、反復的な信号やアラートの監視といった一部のタスクは、人間が容易に遂行できるが、それらは面白みがない。そうしたタスクを遂行できる人間は、代わりに他のタスクを選ぶ。AI システムは、価値はあるが面白みのないタスクに取り組むことに何の抵抗もない。

6.2 弱点

スキルの低下

AI が人間が行うはずだった業務を遂行するにつれ、人間は本来身につけるべき技能を失う可能性がある（H.-P. Lee et al., 2025）。身近な例として、人間が地図の使い方や星による航法能力を失うことが挙げられる（Brumfiel, 2016）。人間の技能喪失に加え、システムが人間の操作を排除する形で設計される場合もある。 自動運転車にはハンドルがなく、店舗では地図や六分儀が販売されなくなり、AI 搭載の産業用制御システムには人間用インターフェースが組み込まれない可能性がある。こうした変化は、AI 搭載システムの攻撃に対する脆弱性を高め、レジリエンスや復旧を妨げる恐れがある。

バイアス

AI システムは、その訓練データに存在するバイアス を反映する。AI のバイアスは通常、人権や公平性の文脈で議論されるが、サイバーセキュリティにおいては、データや結果における体系的な傾向を指す場合もある。例えば、偏ったサイバーシステムは、特定の種類の脆弱性を他のものより優先するかもしれない。別の例として、可用性とセキュリティやプライバシーの間には正当な選好の違いがある。AI システムは、使いやすさをプライバシーより優先する、あるいはその逆の行動をとる可能性がある。

たとえ AI システムが人間よりもバイアスがかかりにくくても、単一のバイアスが高規模で体系的に伝播する場合、AI のバイアスは問題となる。なぜなら、多くの人間が関与すれば、より多様なバイアスとレジリエンスが期待できるからだ。

セキュリティ

AI モデルやシステムは、様々な攻撃に対して脆弱性があることが知られている。攻撃的・防御的 AI エージェントの双方が、破壊工作や乗っ取り、さらにはポイズニング攻撃のような改ざんに脆弱である可能性が高い（A. J. Lohn, 2020b）。

透明性

AI システムの内側の仕組みは人間には解釈不能だ（von Eschenbach, 2021）。人間の内側の仕組みも人間には解釈不能だが、共有された経験によって人間同士は理解し合える。

AI モデルやシステム自体の透明性に加え、透明性とは AI または人間によって生成される成果物自体を指す場合もある。例えば人間は人間が読める言語でコードを書くが、AI システムは機械にとって効率的だが人間には解釈困難な方法でコードを書いたりシステムを設計したりする可能性がある。同時に AI は複雑な概念やシステムの要約・説明に優れており、不透明なシステムを人間にとって理解しやすくする助けとなり得る。

人間の嗜好への整合性

AI システムは人間の経験、感情、嗜好、倫理観を理解するのに苦労する可能性がある。人間の指示や意図を解釈する能力が不十分な場合もある。また、自らの行動や選択に対する人間の反応を予測する能力も不十分な場合がある。しかし人間も他者の意図や価値観、予想される反応を解釈するのが難しい場合があるため、これらの領域において AI が人間の能力に匹敵あるいは凌駕する可能性は一部で存在する。

虚構の生成

AI システムは頻繁に自信を持って虚偽の発言を行う（Aljamaan et al., 2024; Magesh et al., 2024）。人間が犯さないような誤りを、検知が困難な形で生じさせることもある。

7. サイバー攻撃と防御の非対称性

サイバー領域における攻撃的優位性または防御的優位性について、多くの議論がなされてきた。我々はこれらの議論を擁護したり反駁したりせず、攻撃と防御のどちらが全体的に優位かを判断するためにそれらを評価しようともしない。我々は単に文献を調査し、これらの議論を収集するとともに、AI の進歩がそれらの議論を強化または弱体化させる可能性のある方法を考察した。表 7 に示したように、防御を支持する 9 つの議論と攻撃を支持する 9 つの議論を収集した。本節の残りの部分では、AI の進歩がこれらの議論にどのように影響し、サイバー領域のバランスを防御側または攻撃側へより傾ける可能性があるかを説明する。

防御側の優位性	攻撃側の優位性
• 防御側がデジタル戦場を決定する	• 攻撃側はたった一度の成功でよい
• 防御はレジリエンスを持つ	• 攻撃側は攻撃のタイミングを選べる
• 防御側はネットワークを監視できる	• 攻撃者は誰を攻撃するかを選ぶ
• 防御側はローカルリソースを調達できる	• 攻撃者は目標を選択する
• 攻撃者は全ての防御を突破しなければならない	• 標的は簡単に見つけられる
• 防御側は攻撃側より数が多い	• 防御には信頼性が求められる
• 攻撃には時間がかかる	• 防御にはより多くの官僚主義が存在する
• 攻撃はエスカレーションのリスクがある	• 防御には体系的な脆弱性がある
• 攻撃者は正当性を欠く	• 防御には体系的な脆弱性がある

7.1 防御的不均衡

7.1.1 防御側がデジタル地形を決定する

陸上戦では、防御側は塹壕を掘り、地雷を敷設し、城壁を築き、丘の上に砲台を設置できる。これにより、防御側には「防御者 1 人に対して攻撃者 3 人」という経験則上の優位性が生まれる（Mearsheimer, 1989）。デジタル領域は本質的に人工的で再構成可能であるため、サイバー紛争においては防御側が環境を定義する能力がさらに顕著になる可能性がある。

ネットワーク接続、帯域幅、エアギャップ、ファイアウォール、ブラックリスト、パスワード規則、アクセス権限、その他無数の設定は全て防御側が設定する。防御側は実装するシステムやサービスの要件や機能性を決定でき、欠陥や脆弱性のあるソフトウェアやシステムを使用しない選択もできる。防御側はさらに、インターネットのアドレス体系、暗号化コミュニケーション、さらにはケーブルやスイッチといった物理インフラといった、デジタル環境のよりグローバルな側面も設計する。

スピード

防御側は、システムを稼働させて攻撃者に晒す前に、時間をかけて防御設定を整えられる。ただし、ビジネスのスケジュールに追われる場合や、危機・紛争時に迅速な対応を迫られる場合もある。こうしたリアルタイムに近い調整により、足場を確保した攻撃者が、まさにそれを利用しようとした瞬間に到達不能になる可能性がある。あるいは、攻撃者がマルウェアやインプラントをインストールしようとしたまさにその瞬間に、新たな検索用語によって検知されることもある。

こうした対応の一部は AI なしでも可能だが、人間が AI の速度に追いつくには事前計画がより必要となる。人間は警戒レベルに応じた設定群を準備できる。その変更の一部は自動化可能であり、人間の指示下でも AI 指示下と同等の速度で実行される。 信頼性が高く独立した AI システムは事前計画よりも適応性が高いが、変更内容によってはリスクが高く委任すべきでない場合もある。被害者のコンピュータ上のプロセスを強制終了したり、デバイスを消去したり、サービスを一時停止したりする防御行動は全て危険を伴う。さらに防御エージェントが攻撃者に騙されたり操作されたりすれば、その再構成は害を及ぼす可能性すらある。

専門知識へのアクセス

防御が十分に構築された標的を攻撃するのは容易ではないが、防御を十分に構築した標的を作るのも容易ではない。大規模な防御専門家チームを擁する大企業は防御を適切に構築しており、攻撃が難しい。 重要インフラを含む大半の組織は、それほど堅牢ではない。現状の AI レベルでも、中程度のスキルを持つ防御者がシステムを強化するのに役立つだろう。専門家レベルの AI が広く利用可能になれば、あらゆる組織が大手企業並みの防御体制を構築できる。こうした進歩は防御の競争条件を平準化し、個人・中小組織・重要インフラに大きな利益をもたらす。攻撃対象となる領域が小さいため、小規模組織は大企業よりも防御が難しい標的になる可能性すらある。

堅牢な標的に対しては、攻撃者は侵入手法をより創意工夫する必要がある。例えば、より多くのゼロデイ脆弱性を悪用するといった方法だ。 攻撃者は既にゼロデイへの依存度を高める傾向にある。これは防御準備の改善が背景にある可能性がある（CISA, 2024）。既知のツールや技術に依存する攻撃者が堅牢な標的を攻略するのは困難だが、優れた攻撃者は十分な時間があればどんな標的も突破できるというのが通説だ。これは専門家レベルの AI 攻撃者が、専門家レベルの AI 防御者に対しても頻繁に成功する可能性を示唆している。

とはいえ、エキスパートレベル以上の攻撃者優位は必然ではない。攻撃者が発見できる新規のエクスプロイトや手法が多数存在することを前提としている。AI 規模での防御がそれらの大半を発見・修正すれば、エキスパートや超人的な AI 攻撃者でさえ手も足も出なくなる可能性がある。 防御側はまた、脆弱性の少ない、場合によっては証明可能な安全性を備えたシステム設計を行うこともできる。AI 防御側が証明可能な安全性を備えたシステムの割合を増やせれば、超人的 AI 攻撃者であっても攻撃経路は減少する。設計の改善により、防御側は防御タスクを段階的に縮小し、より少ない数の脆弱な経路に集中できるようになるかもしれない。

7.1.2 防御はレジリエンスを持つ

防御側は、どれだけのリスクを受け入れるかを決める。例として、安全なバックアップの維持、手動操作モード、選挙における紙の投票用紙などが、その限界を設定する手段となる。 核兵器をインターネットに直接接続したり、応答しない航空機を遠隔操縦したりするのは便利だが、サイバー攻撃の可能性すら許容できないほどリスクが高いものもある。代わりに防御側は、サイバー攻撃が引き起こし得る最大被害の限度を設定する（Gray Dr., 2013）。

スケーラビリティ

このレジリエンスは、システムが何を実行できるかではなく、システムが実行を阻止されるものから生まれる。AI は許容できないリスクを識別し、回避または管理するのに役立つ。 NIST サイバーセキュリティ枠組みの第 2 カテゴリーは「識別」であり、組織のサイバーセキュリティリスクを理解することを目的とする（NIST, 2024）。組織内のデジタル資産が膨大な数に上る可能性や、それらの悪用から生じうるシナリオの多さゆえに、これは困難な作業となり得る。

AI は膨大なデータを取り込み、多くのシナリオを評価する能力でこの問題を緩和できる可能性がある。仮説的なリスクのブレインストーミングを支援するなど、AI はさらなる大幅な改良なしにある程度これらの取り組みに貢献できるだろう。AI のさらなる進歩は、その精度と創造性を高めることで有用性を増す可能性がある。超人的知能は、特にリスクの完全なセットが容易に把握できない複雑なシステムにおいて、人間が決して考えつかないリスクを特定し排除できるかもしれない。

スキルの喪失

標的となるシステムや組織における人間の存在は、サイバー攻撃に対する一種のレジリエンスである。人間はしばしば脆弱な部分ではあるが、ハッカーによって直接制御されることはない。人間をデジタル事業体で置き換えることは、特に AI システムが他のデジタルシステムよりも乗っ取りや操作に対して脆弱なままである場合、それらのシステムをより脆弱にする恐れがある。

脆弱な構成要素で構築されていることに加え、AI の統合は復旧やレジリエンスの取り組みに悪影響を及ぼす可能性がある。例えば、ハンドルのない自動運転車は、攻撃を受けた場合に手動モードに戻すのが困難だ。AI の統合は、あらゆる分野で人間と機械のインターフェースが広く失われることにつながるかもしれない。たとえインターフェースが残ったとしても、人間はそれらを使用するスキルを失う可能性がある。

7.1.3 防御側はネットワークを監視できる

防御側はネットワークに監視装置を設置し、ログやスキャン情報をプロバイダできる。ネットワークに流入する全データ、エンドポイント上で実行中の全プロセス、全ユーザーの行動を監視可能だ。通常、監視対象となるエンドポイントやプロセスは膨大すぎて全てを監視できないが、防御側はほぼあらゆる対象を選択的に監視できる。また、スキャンや監視が性能を過度に低下させない限り、防御側は自らの行動の隠密性を気にする必要はない。

十分な時間と労力をかけさえすれば、攻撃者も調査対象のネットワークについて多くのことを学べる。NSA の攻撃サイバーチームの責任者だったロブ・ジョイスはこう述べている。「我々は時間をかけて…そのネットワークを設計した者や防御している者よりも深く理解するのだ… お前たちはそのネットワークで使うつもりだった技術を知っている。我々は実際にそのネットワークで使われている技術を知っている」と述べている（Zetter, 2016）。これは特に、人々が個人用と業務用デバイスの境界を曖昧にしている状況、とりわけ在宅勤務時には当てはまるかもしれない。

可用性

スキャン自体はすでに大規模に実施可能だが、スキャンを実行し解釈するスキルには限界がある。大企業にはネットワークを 24 時間監視するセキュリティ専門家がいるが、中小企業には AI なしではその余裕がない。 が自律的かつ確実にアラートを調査・対応したり、脅威環境に応じてスキャンやログを適応させたりできる AI は、防御にとって大きな助けとなるだろう。

複雑なデータとサイロ化の解消

人間の防御担当者を圧倒するほどのログ、データ、アラートは管理が困難だ。通常、管理可能にするためデータは切り詰められるため、有用な可能性のあるデータが人間の防御担当者を圧倒しないよう廃棄される。AI システムはより多くのデータを閲覧でき、人間が気づかない関連性を発見する可能性がある。AI は大量のデータを処理する能力では長年人間を凌駕してきたが、そのデータを理解する能力では必ずしもそうではない。

攻撃者と防御者の双方に影響する観測可能性には物理的な限界が存在する。信号検出理論の分野では、検出可能と検出不可能な境界を定量化する数学的成果が数多く存在する。これらの理論の多くはサイバーセキュリティ分野で生まれ、防御者が観測可能な範囲やステルス攻撃者の攻撃性の限界を定めるために広く応用されてきた。（Ding, Han, Xiang, Ge & Zhang, 2017）

セキュリティ

AI モデルは欺瞞的なパターンに脆弱であり、存在しないものを見たり、存在するものを見逃したりさせられる可能性がある（A. J. Lohn, 2020b）。この効果は画像処理モデルで特に悪名高いが、サイバーセキュリティを含む全ての AI カテゴリーで当てはまる。攻撃者がこの脆弱性を利用して行動を隠蔽できる場合、AI は防御対象のネットワークやシステムにおける可観測性の向上に苦労するかもしれない。 さらなる進歩により、AI システムは欺きにくくなるかもしれない。しかし歴史的に見て、入力や意思決定プロセスの複雑さを増すことで AI の能力を高めることは、同時に欺く方法の数を増やすことにもつながってきた。AI が高度になればなるほど、そのセキュリティ確保は困難になってきたのである。

7.1.4 防御側はローカルリソースを調達できる

防御側は、予算の範囲内で必要なだけのメモリ、演算能力、帯域幅を確保できる。疑わしい行動の分析や攻撃者への対応策立案に専念する高性能コンピュータへの高スループット接続を確立できる。 一方、攻撃者は被害者のローカルリソースを利用するか、遠隔のコマンド＆コントロールサーバーとデータをやり取りするかのいずれかだ。いずれの場合も、防御側に警戒されるリスクがある。防御側は攻撃者を排除したり、必要なら対象システムの電源を切るなど様々な方法でアクセスを遮断できる。

したがって攻撃者は、非常に少ないメモリと演算能力しか使用しない AI モデルやエージェントに制限されるか、遠隔モデルとの間で送受信する情報量を制限せざるを得ない。例えば防御側は 1 ギガバイトのソフトウェアを容易にスキャンして脆弱性を探せるが、攻撃者はその 1 ギガバイトをスキャンすることも、検知リスクを冒さずにスキャン用に送信することもできない。こうした攻撃側の制約により、防御側はより高性能な AI システムを選択し、情報へのアクセスが良好な適切なハードウェア上で効率的に稼働させるという優位性を得られる。

この不均衡は、防御側が制御するデジタル環境領域での交戦にのみ適用される。オープンソースソフトウェアの脆弱性発見など一部の攻撃活動は、被害マシンとの情報送受信を一切行わずに完全に遠隔で実行可能だ。MITRE の ATT&CK 枠組みの各段階には、 がこれらの制約に直面する要素と、直面しない要素が存在する（MITRE, 2025）。

可用性

防御側がリソースを調達できる能力は、少なくとも一部のケースにおいて、防御側が直面する攻撃者よりも高性能なモデルやエージェントを選択できることを意味する。最もリソースを消費するモデルやエージェントは、攻撃者が特定のタスクに利用できない場合がある。

しかし、攻撃者と防御者がこれほど異なる任務を遂行している以上、それだけでは攻撃者が不利とは言い切れない。例えば、実行中のプロセスに永続性を確立する作業と、永続性を提供するプロセスを識別して終了させる作業は、特に類似した仕事ではない。デジタル戦闘において AI の進歩が攻撃側と防御側のどちらに有利に働くかを判断するには、攻撃者のリソース制約にもかかわらず AI が攻撃を支援する能力を、おそらくタスクごとに評価する必要があるだろう。

7.1.5 攻撃者はあらゆる防御を突破しなければならない

防御側は単一の防御手段に依存しない。例えば攻撃者は、ファイアウォールを起動させずに被害マシンと通信し、アンチウイルスを起動させずにソフトウェアをインストールし、異常検知を起動させずに情報を転送またはコマンドを実行する必要がある。これらのいずれかの段階で失敗すれば攻撃は失敗する。

単一の脆弱性で攻撃が完結するケースもある。これはユーザーが中核システムやデータに直接アクセスできる公開サービスでより一般的だ。Heartbleed 脆弱性が好例で、攻撃者は他のユーザーの個人情報を取得できた（Durumeric et al., 2014）。この脆弱性が広まった背景には、ユーザーが自身の情報に直接アクセスできるウェブサイトが、同じサーバー上に他人の情報も保管しているケースが多かったことがある。システムはユーザーが自身の情報を取得するために設計されていたため、防御策はほとんど存在しなかったのだ。

スピード

防御の深度化が攻撃者を検知・排除できるほど十分に遅らせることを目的とするなら、攻撃と防御の速度が重要だ（A. Lohn, 2019）。 AI が防御側の侵入者検知・排除能力を、攻撃側の防御突破能力よりも大きく加速させる場合、防御側がより大きな利益を得る。しかし逆の結果も起こりうる。また、個々のケースで加速効果が防御側に有利であっても、攻撃頻度の増加は一定期間内の成功攻撃件数の増加につながる可能性がある。

防御側には速度をある程度制御する能力がある。ハードドライブの暗号化には時間がかかり、さらに遅くすることも可能だ。不審な動作は、パスワードの誤入力が多すぎた後に発生するロックアウトのような一時停止で遅延させられる。暗号化やマスターブートレコードの削除、大容量ファイル転送といった不審または危険な活動は、人間の承認を得るた

めに遅延させられる。こうした遅延は生産性を多少妨げるかもしれないが、セキュリティを優先するなら生産性は犠牲にしても構わない。

拡張性

あるいは、多層防御の目的は、攻撃者が単一のキャンペーンで全ての障壁を突破できないよう、十分な数の障壁を設けることにある（A. Lohn, 2019）。その場合、AI の規模は、標的が実装・監視できる防御手段の数を増やすのに役立つかもしれない。しかし攻撃側の観点からは、標的が直面する独立した攻撃者の数も増加させる可能性がある。

同じ時間内に多様な攻撃を仕掛ける能力は、防御側が侵入率全体を一定に保つためには、個々の攻撃の成功確率を低下させねばならないことを意味する。 同時に、攻撃者が試行回数を増やすと、発見される可能性も高まる。こうした発見は、防御側が攻撃をその発信源まで追跡し、攻撃者の攻撃実行能力を妨害するのに役立つ。攻撃者のインフラや運用を妨害することは、現在の攻撃を阻止するだけでなく、将来の攻撃も防ぐことができる。

7.1.6 防御側は攻撃側よりも数が多い

ほぼ全ての者がシステムを防御する必要がある。自らシステムを防御するか、サービスに代行させるかのいずれかだ。比較すると、攻撃者は比較的少数である。防御側は遍在性を活用し、情報を共有する。あるシステムで発見された脆弱性や攻撃手法は、YARA ルールやアンチウイルスシグネチャの更新といった自動化された方法で、迅速に他の防御側に共有される。

攻撃者でさえ、集団防御に貢献している。おそらく多くの防御者以上に。攻撃者は国家主体の高度な諜報活動や侵入キャンペーンの標的となることが多い。攻撃者のマシン上の防御サービスは、新規の手法を発見する可能性が高く、その情報は全ての防御者と共有される。攻撃者同士も、様々なハッキングフォーラムや Metasploit のようなツールスイートを通じて攻撃手法やツールを共有している。 しかし、その共有は防御側にも利益をもたらす。防御側はそれらのフォーラムを監視したりツールスイートにアクセスしたりして、新たな情報開示に適応できるからだ。

防御側の存在が普遍的であることは、防御技術革新の巨大な市場を生み出し、規模の経済を可能にし、急速な進歩を促す競争圧力を生み出す。防御が非常に高コストであることは、サイバー空間では攻撃側が有利だという主張の根拠として用いられてきた。しかし、防御側がより大きな予算を有することは、損害効果を引き起こす難度を高めることや、危機や紛争時の安定性を提供することなど、攻撃と防御のバランスを測る他の指標において優位性をもたらす可能性がある。

規模

信頼性が高く独立した AI エージェントは、標的あたりの防御者数を増加させる可能性がある。AI は参入障壁を下げることで攻撃者数を増やし、攻撃者あたりのエージェント数も増加させるかもしれない。AI がソフトウェア開発の速度を加速させる場合、標的の数やそれらの攻撃対象領域も拡大する可能性がある。

防御側は、AI の規模拡大が続く中で情報共有の恩恵を引き続き享受するだろう。AI 防御システムは、発見した脆弱性や効果的な技術を相互に共有し、全ての防御側に利益をもたらすことができる。一方、AI を活用した攻撃側は、エクスプロイトや技術の共有が防御側に知れ渡るリスクに依然として直面する。ただし、このリスクが攻撃側にとって障害となりにくいケースも存在する。 例えば、ハードウェアの脆弱性は 修正が困難な場合があり、防御側に情報を提供してしまうリスクはそれほど顕著ではない（A. Lohn, 2018）。

7.1.7 攻撃には時間がかかる

攻撃は、「サイバーキルチェーン」や Mitre の ATT&CK 枠組み（Lockheed Martin、2025 年、MITRE、2025 年）で概説されているような、いくつかの段階を経て進行することが多い。 段階は並行して完了することもあるが、多くの場合、段階は順番に相互に依存しており、段階が数回繰り返されることもある。例えば、初期アクセス前に偵察が必要となり、その後、リソース開発のために再び偵察が必要になる場合がある。リソース開発は、初期アクセスに必要となり、権限昇格や横方向の移動のために再び必要になる場合がある。

攻撃的なキャンペーンを実施するには長い時間を要するため、国家が平時に重要なインフラにサイバー能力を事前配置することを選択する理由を説明できる。 攻撃者はまた、防御側に攻撃手法を観察・適応させる機会を与えることで能力を「消耗」させるのを避けるため、攻撃の特定段階を遅らせる選択をする場合もある。2022 年のウクライナ侵攻においてロシアのサイバー攻撃が限定的だったのは、攻撃側に十分な準備時間が与えられなかったためだと指摘されている（C. Lee, 2023）。

スピード

攻撃タイムラインの加速は、この防御的優位性を弱める。たとえ防御側が攻撃側よりも加速したとしても、攻撃側が数か月ではなく数日や数週間で侵入を試みる可能性が高まれば、攻撃側の戦略的立場は強化される。彼らはより短い準備期間で攻撃を実行でき、防御側に与える事前警告時間も短縮される。数日間の地上侵攻や数分・数時間のミサイル攻撃には準備が間に合わないほど遅かったサイバー攻撃も、実行可能になる可能性がある。

しかし攻撃キャンペーン全体のタイムラインを加速させるのは困難な課題だ。なぜなら、多くの領域（全てではない）を加速させても効果が限定的である可能性があり、サイバー領域の多くの側面は既に極めて高速だからだ。サイバーキャンペーンの一部が低速なままなら、キャンペーン全体も低速なままとなる。

さらに、非国家主体の攻撃者にとっては、攻撃の速度はそれほど重要ではないかもしれない。彼らは、サイバー攻撃を現実世界のタイムラインと連動させる明確な理由が少ないからだ。とはいえ、その逆の例もある。例えば、株式市場の変動、自然災害、エクスプロイトや効果的な攻撃戦術の漏洩など、様々な機会に合わせて攻撃を調整する必要がある機会主義的なサイバー犯罪者などがそれにあたる。

規模

標的が特定されていない場合、規模そのものが一種の速度を提供する。業界や地域に混乱をもたらすことが目的であれば、警察署、近隣のダム、電力網、消防署を同時に攻撃する能力は、必要なタイムライン内で少なくとも一つが侵害される確率を高める。 同様に、AI が攻撃対象領域を拡大するような開発規模の拡大をもたらす場合、攻撃者はより迅速に標的を侵害できる可能性がある。

7.1.8 攻撃がエスカレーションリスクを抱える

攻撃側に不利な官僚的な理屈が存在する。例えば、攻撃的行動は外交的反発や報復を招くリスクがあり、その報復はデジタル、物理的、法的形態をとりうる（Valeriano & Jensen, 2019）。 被害を受けた国家は、重要インフラへのサイバー攻撃に対して攻撃国を爆撃する形で対応する可能性がある。あるいは犯罪ハッカーは逮捕されたり罰金を科されたりする。結果として、サイバー攻撃者はある程度行動を制限されることが多い。外国の領土内で防御されているサイバー犯罪者でさえ、保護者の寛容さや被害者の自制には限界があることを認識し、それに応じて攻撃を抑制するだろう。

反論として、この防御的優位性が攻撃者の破壊的・広範な攻撃を常に阻止してきたわけではない。ワナクライやノットペティアワームのような無差別攻撃も発生している（NATO, 2017）。 コロニアル・パイプラインや電力網、病院といった重要インフラへの攻撃も発生している（DNI, 2024）。さらに原子力施設への攻撃事例もある（Zetter, 2014）。サイバー攻撃はエスカレーションリスクが低いという評判が、逆にエスカレーションが起きやすい状況での使用を助長する可能性があるのだ。

誤動作率の低減

攻撃用 AI が誤作動率を低減できるなら、攻撃者は人間には任せられないサイバー攻撃を AI エージェントに実行させることを容認するかもしれない。現状では誤作動が頻発しているため、少なくとも信頼性が高く自律的な AI が求められる。重要な作戦上の決定、さらには戦術的な判断さえも専門家が自ら行い、信頼できる人間にすら委ねないケースが多いことから、専門家レベルではなく超人的な信頼性が求められる可能性すらある。

透明性

攻撃的サイバー戦の使用決定は、エラー率だけでなく透明性にも基づく信頼に依存する。AI システム自体が不透明であることは周知の事実であり、その行動や判断も不透明になり得る。しかし AI が人間よりも自らの計画や行動を説明できるなら、意思決定者の信頼を得るのに十分な透明性を備えていると言える。 その場合、AI 攻撃者は人間には許可されない攻撃を実行できるかもしれない。そのためには、現状では AI が人間より透明性が低いと見なされている状況から、大幅な進歩が必要となる。

攻撃的判断やエスカレーションリスクに影響する透明性の別の側面は、帰属問題だ。高度な AI が帰属を助けるか妨げるかは不明だ。攻撃者が識別されにくければ、エスカレーションリスクは低くなる。多くの攻撃者が同様の AI 技術を使えば、帰属はより困難になる。AI は偽旗作戦も容易に実行できるかもしれない。 さらに、攻撃のより多くの部分を AI が担う場合、帰属判断に利用可能な人間オペレーターの痕跡は減少する。しかしインフラ、動機、傍受コミュニケーションなど多くの手がかりは残る。AI は多様な情報源を統合し、帰属判断をより正確かつ迅速に行う助けにもなり得る。また、様々な攻撃者が独自の AI システムを使用する場合、それらのシステム固有の特徴が攻撃者特定に利用される可能性がある。

人間の嗜好への適合

AI は人間の感情や行動への反応を理解するのに苦労しており、今後もその傾向は続くだろう。 どの行動がエスカレーションにつながるかを判断する能力は、人間にとっても困難であり（ ）、AI にとってはさらに困難であると予想される。この問題は、訓練データが標的の文化ではなく攻撃者の文化に基づいて構築された場合に生じうるバイアスによって悪化する可能性がある。この欠点は、攻撃者が AI エージェントの使用を躊躇させるか、使用した場合に意図しないエスカレーションのリスクを高める可能性がある。

7.1.9 攻撃者は非合法である

国家による諜報活動に限定される場合、攻撃は正当な手段として認められるが、それ以外のほぼ全てのケースでは違法である。諜報活動においても、正当なサイバー攻撃は国際的・国内的な支持を失うリスクを伴う。国家以下のレベルでは、企業・組織・個人が自己防衛や報復のために攻撃的技術を正当に利用する能力は極めて限定的である。

これは技術的問題というより社会学的問題だ。AI がこの力学を直接変える可能性は低い。技術進歩の結果、社会的規範や期待が変化する可能性はある。例えば、標的環境ではなく攻撃源に近い場所で妨害できない場合、AI 攻撃への防御が著しく困難になれば、企業はより攻撃的な手法の使用を容認されるかもしれない。

時間の経過と攻撃の増加に伴い、社会はサイバー攻撃に対して感覚が麻痺するかもしれない。これは攻撃を公式に正当化するものではないが、次第に容認されるようになる。あるいは、社会は「もう十分だ」と判断し、以前は容認されていたサイバー攻撃に対してより厳しい姿勢を取るようになる可能性もある。

7.2 攻撃的非対称性

7.2.1 攻撃者は一度の成功で十分

攻撃者は成功するまで攻撃キャンペーンを繰り返すことができる。キャンペーン全体で防御網を突破もしくは迂回する必要はあるが、目標の多くを達成するには単一のキャンペーンが成功すれば十分だ。個々の失敗による損失は小さく、防御側に攻撃を察知されても、その後のキャンペーンを困難にする形で攻撃手法やインフラが露呈しない限り問題にならない。

攻撃者は短期間での成功を必要とする場合があるため、防御側は、攻撃を危機が収束するまで、あるいはデータやコードが最終化され改ざんや窃取が不可能になるまで遅らせることができれば、わずかな成功で十分である。さらに、攻撃者の目的が標的内部への持続的な潜伏である場合、繰り返し発見・排除されるため、多くの成功が必要となる可能性がある。

速度と規模

AI によって攻撃と防御が同じ時間内により多く発生するようになる場合、つまり攻撃速度が速くなるか、攻撃者の数が増えるかによって、防御側は侵入率を一定に保つために防御成功確率を高める必要がある。

専門知識へのアクセス

たった一度の成功したキャンペーンで十分だという事実は、新たな戦術・技術・手順（TTP）や脆弱性の発見が、防御側よりも攻撃側にとって容易な問題であることを示唆している。 劣った攻撃 AI エージェントが優れた防御 AI エージェントと対峙する状況を想像せよ。防御エージェントは 100 万の攻撃 TTP や脆弱性を識別し緩和できるが、劣った攻撃エージェントは試行可能な TTP や脆弱性をわずか 1000 個しか識別できないと仮定する。その 1000 個のうち、防御エージェントの 100 万個に含まれないものが一つでもあれば、攻撃者は有効な手法を発見したことになる。

このシナリオは、双方がランダムな TTP と脆弱性のプールから選択する状況を想定している。別のシナリオとして、優れた AI エージェントが劣ったエージェントが発見した TTP や脆弱性を全て把握し、さらに多くの情報を得る場合も考えられる。その場合、攻撃側には同等の能力を持つエージェントが必要となる。さらに、キャンペーンでは複数の TTP や脆弱性が要求される場合もある。その場合、攻撃側は失敗した試みが独自の TTP や脆弱性を露呈する可能性を懸念するだろう。

7.2.2 攻撃対象の選択

場合によっては、攻撃者は特定の標的を想定している。例えば、地上部隊が特定の建物を突破しようとしており、ドアのロックを解除する必要がある場合などだ。 多くの場合、攻撃者は標的により柔軟に対応できる。身代金を支払う可能性のある個人や組織なら誰でもランサムウェアを仕掛けられる。デジタル防御が最も脆弱な病院、学校、政治家を選ぶこともできる。あるいは、悪意のあるアプリをダウンロードした砲兵部隊を追跡することも可能だ（CrowdStrike, 2017）。

規模

攻撃の規模とは別に、AI 強化型デジタル開発は攻撃者が選択できる標的の数を増やす可能性がある。拡大するコードベースと増加するデジタルシステム・アプリケーションは、防御すべき対象の量を増加させる。

ただし攻撃そのものに関しては、初期のサイバー攻撃でさえ自動化され大規模に複製可能だった（Nguyen et al., 2024; Orman, 2003）。こうした自動化攻撃では標的間の差別化はほとんど行われなかった。例えばランサムウェアワームは到達した標的を全て暗号化するが、攻撃対象の選択には単純なルールが適用される場合がある。 例えばロシアの犯罪組織は、特定国やロシア語キーボードがインストールされたターゲットへの攻撃を回避する（Krebs, 2021）。信頼性が高く自律的な AI 攻撃エージェントは、攻撃者がこうした単純な標的選定ルールを超え、そもそも標的を攻撃するかどうかのより微細な判断を可能にするかもしれない。

こうした知能の向上は、実際には攻撃の規模を縮小したり、発動されないインストール数を減らしたりする可能性がある。こうした広範囲にわたる、あるいは偶発的な攻撃は、そうでなければ発見されないマルウェアや攻撃キャンペーンを識別するのに有用である（Kushner, 2013）。

複雑性とサイロの打破

AI は、攻撃者がさまざまなデータソースを活用して、特に脆弱で収益性の高いターゲットを識別するのに役立つ可能性があるが、現在のアプローチに比べてどの程度改善されるかは明らかではない。現在のアプローチは、スプレー＆プレイフィッシングなど、攻撃を広く拡散させるか、ワームに脆弱性の悪用を試みさせ、失敗または成功させるものである。また、Shodan などを通じて、脆弱性について広範囲にわたるスキャンを行うことも簡単である（フランクリン、2018）。 このような単純な手法は、攻撃者がインターネット上で脆弱性や収益性の高い標的を特定するのにすでに役立っている。

専門知識へのアクセス

AI が防御の強化に役立つならば、攻撃者が選択できるソフトターゲットの数を減らすことができる。現状では、AI は人間の防御者がコードベースを強化し、ネットワーク内のアラートや不審な活動を理解するのに役立っている（Cruz、2025 年、Google、2024 年）。 信頼性が高く独立したレベルでは、防御側はソフトウェアやデジタルエコシステム全体をさらに強化できると同時に、リソースの少ない組織や個人に有能なサイバー防御者をプロバイダできる。

7.2.3 攻撃者は自ら攻撃時期を選択する

攻撃者は自らの意思で攻撃するか否かを選択する。これは常に真実だが、攻撃のタイミング選択に関する主張を弱める可能性のある、攻撃決定に影響を与える圧力も存在する。 例えば、地上侵攻などキャンペーンの他の側面と連携するため、サイバー攻撃を特定の時期に実施する必要がある場合がある。この場合、サイバー指揮官が決定したものではなくても、攻撃者は攻撃時期を選択したことになる。別の例として、敵の地上侵攻を撃退するためにサイバー攻撃を必要とするなど、サイバーチームが反撃を余儀なくされる場合もある。

より一般的なケースでは、サイバー攻撃者は攻撃のタイミングを自由に選択できる。防御が最も脆弱に見える時や、影響が最大となる時がそれにあたる。例えば、冬の寒さが電力網の停止を特に深刻な事態にする場合や、新たな脆弱性が公表され多くの標的が通常より脆弱になる場合などだ。

スピード

AI エージェントは、攻撃者が機会が生じた瞬間に即応することを可能にするかもしれない。攻撃手法の開発は既に脅威的な速さで進んでいる。NotPetya や WannaCry が流出した EternalBlue 脆弱性を流用した事例がそれを示している（Hern, 2017）。 しかし、この速度向上は攻撃者よりも、防御側がこうした好機を捉えた標的を迅速に強化する助けとなる可能性が高い。攻撃者は既に高速であるため、防御側は技術進歩から得る機会が増えるかもしれない。特に防御策の開発過程よりも、広範な防御策の導入過程においてその恩恵を受けやすい（A. Lohn & Jackson, 2022b）。

7.2.4 攻撃者は目標を選択する

攻撃側の目標は防御側よりも単純だと主張する者もいる（Slayton, 2017b）。目標が単純かどうかに関わらず、攻撃側は目標を選択できる。攻撃者はデータを流出させたり、モデルを改ざんしたり、回路ブレーカーを切り替えたり、ウェブサイトを改竄したり、あるいはネットワークに侵入した後に最も都合の良い手段を後で選ぶこともできる。防御側はこれら全てに対して同時に防御する必要がある。

可用性

信頼性が高く自律的な AI 意思決定が可能になれば、この攻撃側の優位性はさらに高まる。攻撃主体に認可が与えられれば、明示的な指揮統制なしに迅速に判断を下せる可能性がある。一瞬の機会をより効果的に活用できるのだ。

機会がどれほど一瞬で、人間の意思決定が真の制約要因となるか、あるいは攻撃者にとって指揮統制がどれほど困難かは不明だ。例としては、エアギャップされた産業制御システムネットワークや機密ネットワークが標的となる場合が挙げられる。指揮統制へのアクセスがなければ、信頼性が高く自律的な攻撃エージェントは、標的へのアクセス後にそこで観察した内容に基づいて行動を決定できる可能性がある。これは自動化が困難な方法である。

専門知識へのアクセス

攻撃エージェントは、人間では入手困難な標的システムに関する専門知識を持つ場合もある。産業用制御システムなど特殊なシステムは多く、組織構造は外部者にとって複雑で不透明だ。標的理解を支援する AI は、攻撃者が影響力の大きい効果や行動を特定する能力を高める可能性がある。

同じ AI は防御側エージェントにも多少役立つ可能性があるが、防御側は攻撃者よりも自組織のシステムや運用を深く理解している場合が多いため、得られる利益は限定的だろう。

7.2.5 標的は容易に見つけられる

標的は通常、サービスを提供するために見つけやすい必要がある。サービス拒否攻撃のように見つかりにくくなること自体が攻撃の目的となる場合もある。しかし全ての標的が見つけやすいわけではない。個人や小規模組織の中には、身元を偽装したりデジタル痕跡を最小限に抑えたりして、見つかりにくくしようとする者もいる。

標的が一般的に見つけやすい一方で、攻撃者は通常、発見されず見つけられないよう努める。例外として、サービス販売のために発見可能である必要がある犯罪者による攻撃者が挙げられる。それでも、サービスの宣伝はサービス自体よりも見つけやすく、攻撃者は通常、身元を特定されないよう多くの予防策を講じる。

規模

AI が攻撃の数を増やすなら、防御側が攻撃者を発見する機会も増える。防御側はこうした発見を攻撃の起源まで遡り、攻撃者やそのツール・インフラを特定するために利用することが多い。

複雑性とサイロの打破

現在の攻撃者発見の困難さは、防御側に一定の均衡をもたらす機会を提供している。AI が防御側を支援し、インターネット全体で不審な行動を特定・調査できるなら、既に発見が容易な標的を扱う防御側よりも、攻撃者識別においてより大きな効果を発揮する可能性がある。

7.2.6 防御には信頼性が求められる

防御側はシステムの稼働と使用可能性を維持する必要がある。セキュリティ更新のインストール中など、システムを頻繁に停止または再起動することは許されない。また、それらの更新はシステムの使い勝手や性能を過度に妨げてはならない。例えば、2024 年の Crowdstrike 更新で発生したように、航空会社を運航停止に追い込み、多くの業界で事業を混乱させたようなシステムクラッシュを引き起こすパッチは許されない（Allyn, Mann, Chappell & Al-Kassab, 2024）。

攻撃側は信頼性の低さによる影響が比較的少ない。攻撃者が成功を一度だけ達成すればよい以上、障害に対する許容度は高い。しかし、そうした障害は攻撃者の意図やインフラ、能力を露呈させるなど、悪影響を及ぼす可能性がある。

速度

稼働時間と信頼性は、防御側の適応速度と頻度に制約を課す可能性がある。ユーザーが 1 日 1 回の再起動しか許容しない場合や、サービスレベル契約がネットワークのダウンタイムを制限する場合、防御側は攻撃の速度に対応できないかもしれない。システムやプロセスを再起動せずにパッチ適用できるライブパッチング技術の進歩は、AI が約束する速度向上を防御側が活用するために必要となるかもしれない（RedHat, 2024）。

一方で、防御側のパッチ適用は必要な速度を大幅に下回る場合が多く、AI はパッチ導入の加速化に寄与する可能性がある（A. Lohn & Jackson, 2022b）。防御側が新たな修正に確信を持てるよう支援する AI 技術、あるいは修正実装プロセスを支援する AI 技術は、ネットワーク強化に極めて大きな効果をもたらす。優先事項の長いリストに対応できる防御担当者が不足している場合に生じる実装遅延については、信頼性が高く独立した AI 防御システムが防御強化の加速に寄与し得る（Jacobs, Romanosky, Suciu, Edwards & Sarabi, 2023）。主な課題が単に優先事項リストの実装ではない場合には、専門家レベルの AI 以上の能力が必要となる可能性がある。例えば、更新がネットワークや組織全体に及ぼす可能性のある悪影響を理解し評価するには、高度な専門知識が必要となる場合がある。

精度の向上

更新が有害な影響をもたらすかどうかの判断など、防御担当者が直面する信頼性の課題は、AI がより信頼性の高いコードやサービスの生成を支援することで軽減される可能性がある。これは、専門家レベルの AI がより広く利用可能に

なれば実現しうる。ただし、多くの更新や防御サービスは既に人間の専門家チームを擁する主要組織によって開発されているため、改善の余地は限られているかもしれない。それらの更新の信頼性を大幅に向上させるには、超人的な AI が必要となる可能性がある。

規模

AI は、更新やサービスの候補をスキャン、テスト、評価するのにも有用かもしれない。AI が防御側に、どの変更が安全で、どれが安全でないかを判断する手助けができるなら、信頼性への要求から防御側が直面する不利な点をさらに減らせるだろう。

7.2.7 防御側にはより多くの官僚主義が存在する

攻撃者はしばしば自律性の高い小規模組織である。国家、特に国際法や行動規範の遵守を優先する西側民主主義国家の場合は、この傾向が弱いかもしれない。サイバー犯罪者やテロリストなどの他の脅威アクターも何らかの官僚的障害に直面する可能性はあるが、標的に比べて制約は少ないと考えられる（シュナイアー、2012 年）。

防衛的な官僚主義は政府や産業界全体で深刻な問題となり得る。例えばサイバーや AI ツールの調達手続きは長期の遅延を招き、採用決定や侵害された端末の消去といった防御措置の承認も同様だ。言うまでもなく、攻撃者は防衛側が機密防御ツールをアップロードする際の官僚的プロセスを経ずに、AI 搭載ツールを機密ネットワークに投入できる。

防衛には戦術的な官僚主義も存在する。プロセスの強制終了、エンドポイントの消去、スパムやフィッシングフィルターの閾値変更といった決定には、ユーザーや上層部からの承認が必要となる場合がある。

スピード

官僚主義は防御側の速度制限要因となり得る。そうなると、進歩の速度は防御側よりも攻撃側に有利に働く。官僚的なプロセスは、技術的進歩を迅速に可能にするよう設計できる。 戦術的決定権限は委譲可能だ。承認済み製品ラインの新規リリースを自動承認する調達認可プロセスを構築すれば、最先端技術を防御側に迅速に提供できる。ただし、官僚主義の削減は技術的リスクや社会学的・法的リスクを伴い、常に許容できるとは限らない。そのため、高度で高速な AI システムを導入しても、官僚的遅延は残存する可能性がある。

専門知識へのアクセス

AI は、他の方法では入手困難な専門知識の提供が可能である可能性がある。例えば、新技術を迅速に評価する技術的スキルを持つ官僚や調達担当官を見つけるのは難しい。AI は、こうした担当官が抱える技術的疑問の解決を支援できるかもしれない。技術専門家が官僚の質問に答えるために新技術を検査する必要があった場合、代わりに AI が支援できる可能性がある。

7.2.8 防衛には体系的な脆弱性がある

体系的な脆弱性とは、繰り返される手法やアプローチに起因するものである。これは、単なる構成要素ではなくシステム全体に関連する「システム的脆弱性」とは異なる。体系的な脆弱性は、システム全体に影響を与えずに多くの構成要素で再現される可能性がある。一方、単一のシステム的脆弱性は、多くの構成要素に影響を及ぼす連鎖反応を引き起こす可能性がある。

開発者と防御者はしばしば同じ技術やコードベースを使用するため、サイバーエコシステム全体に同じ脆弱性の事例が多数生じる。例えば脆弱なライブラリが多くのプロジェクトで使用されれば、それらのプロジェクト全てが脆弱になる可能性がある。あるいは脆弱なコードが共通の指導サイトでプロバイダされれば、多くの開発者が自らのプロジェクトやコードベースで体系的にそれらの脆弱性を再現するかもしれない（F. Fischer et al., 2017）。

バイアス

コードの脆弱性と防御的傾向は、一種のバイアスと見なせる。共通の指示サイトからの脆弱性は、AI エージェントも人間と同様に容易に学習する。しかし、その AI エージェントがより多くのコードを書く場合、人間よりも体系的に脆弱性を再現する可能性がある。安全なコードベースや脆弱なコードベースを開発することに加え、同等の選択肢の中から防御構成や戦術の 1 つを選択する傾向もバイアスの一形態だ。大規模に行われると、このバイアスは体系的な決定となり、その決定が脆弱であれば体系的に悪用される可能性がある。

AI 開発者や防御担当者が設計や決定を完全に繰り返さないよう求めるだけで、AI はこのリスクを低減できる可能性がある。防御側は、人間の開発者や防御担当者に実装が困難な方法で、体系的な 脆弱性のリスクを低減するため、意図的に設計や決定に変動性を取り入れることができる。また、リポジトリからコードをコピー＆ペーストすることが多い人間の開発者よりも、AI 開発者はコードを正確に再現しないかもしれない。

セキュリティ

AI システム自体には体系的な脆弱性がある。つまり、どのシステムも同じ方法で操作されやすい傾向にあるのだ。こうした普遍的な弱点は、その動作を制御するために利用される可能性がある。これらの脆弱性は防御側エージェントにも攻撃側エージェントにも影響するため、サイバー戦闘において攻撃側と防御側のどちらに有利かは明らかではない。しかし、脆弱な AI コンポーネントをより多くのシステムに組み込むことは、防御側の困難さを増す可能性が高い。

7.2.9 防御には体系的な脆弱性がある

方法論的に再現される弱点とは別に、防御は複雑で相互に関連するシステムを保護する必要がある。これらのシステムはしばしば、侵害される可能性のある主要コンポーネントに依存している。 デジタルエコシステムの大部分は、脆弱な単一ライブラリに依存している可能性がある。悪名高い Log4j、Heartbleed、SolarWinds の脆弱性（Geer et al., 2003）がその例だ。あるいは、重要なサプライヤーが脆弱であり、そのサプライヤーを混乱させるデジタル攻撃が、それに依存する複雑な連鎖を機能不全に陥らせる可能性がある。

規模

AI が構成要素の数を増やし、組織やシステムの複雑性を高めるならば、システム的な脆弱性の可能性も増大する。それらの構成要素やサービスの上に構築されたデジタルシステムや組織は、侵害される可能性のある構成要素やサービスをより多く持つことになる。

複雑性とサイロの打破

こうしたシステム的脆弱性が存在する理由の一部は、システムや組織が人間が十分に理解・評価するには複雑すぎる点にある。AI は、組織のサプライチェーン全体にわたる全てのソフトウェアコンポーネントや取引を検証するなど、複雑なシステムの評価を支援できる可能性がある。AI が潜在的なシステム的脆弱性の識別を支援できれば、その知見は防御側が脆弱性を緩和するか、少なくともリスクをより適切に評価するのに役立つだろう。

8. サイバーの特性

ジェイソン・ヒーリー、ロバート・ジャービス、ディヴィヤム・ナンドラジョグによる近日発表予定の論文は、サイバー領域をサイバー領域たらしめている一連の特性を捉えようとしている（ジェイソン・ヒーリー、n.d.）。 彼らは、さまざまな公式の政策文書や学術出版物を調査し、サイバーセキュリティやサイバー戦に関する注目すべき発言を記録している。また、彼らはそれらの発言の妥当性確認をしようとはせず、単にそれらをまとめただけである。これらの発言は、その論文の著者や本論文の著者がこれまでに見てきた、サイバーセキュリティの特徴を規定するとされるものを網羅したリストとなっている。

AI がこれらの声明のそれぞれにどのような影響を与えるかを評価することで、AI がサイバーセキュリティとサイバー戦の特性をどのように変化させるかについて考察を始めたいと思う。

8.1 戦術的にネットワーク速度で動く

デジタル技術の速度のため、多くの個々のサイバー行動や活動はすでに迅速に行われている。ネットワークスキャン、マルウェアのダウンロード、ファイル検査、権限変更などの個々のツールや行動は、AI によってそれほど高速化される可能性は低い。しかし、ヒーリー、ジャービス、ナンドラジョグが指摘しているように、他の分野での戦術的タスクも迅速に行われることがある。 砲撃は、大容量ファイルの転送やディスクの暗号化よりも遅いとは限らない。また、脆弱性の発見、バイナリの逆エンジニアリング、エクスプロイトやパッチの作成など、一部のサイバー戦術は、他の分野における一般的な戦術よりも時間がかかる場合がある。

AI は通常、高速なサイバー戦術をさらに高速化することを約束するものではない。実際、デジタル標準では、大規模なニューラルネットワークを備えた AI は、しばしば低速であると見なされる。おそらく、エキスパートレベル以上の AI は、一部のサイバーツールやアルゴリズムを再設計して、より効率的に実行し、その方法で戦術の速度を向上させることができるだろう。しかし、一部のタスクをどれだけ高速化できるかには、厳しい限界がある。

例えば、情報は光速を超えて伝達できず、回路内の配線や記憶装置は RC 時定数で充電される。さらに、証明可能な最適アルゴリズムはこれ以上改善できない。特に攻撃者は検知を避けるため、速度閾値を下回る必要があるかもしれない。

AI は確かに、いくつかの遅い戦術を高速化する可能性を秘めている。脆弱性発見、リバースエンジニアリング、パッチ作成、コードリファクタリング、ハニーポット構築は、いずれも複雑だが戦術的な活動であり、AI が加速すると約束している。

8.2 作戦・戦略レベルでは遅くなる

サイバー作戦は、高速な戦術要素が示唆するよりも大幅に遅くなる可能性がある。人間や官僚機構が、複雑になり得る様々な作戦上・戦略上の懸念事項を処理する必要があるためだ。

人間が、その指示に従う主体（人間であれ機械であれ）に対する制御を維持しようとする限り、この傾向は続くだろう。人間が意思決定をより多く機械に委ねれば、運用上の遅延の一部は縮小するかもしれない。しかし、超人的な機械に意思決定を完全に委ねても、遅延が完全に解消されるとは限らない。 サイバー領域における作戦・戦略上の意思決定は、技術的にも社会的にも複雑だ。人間関係者が多様な、あるいは明示されていない選好や目的を秘めており、意思決定プロセスにおいてそれらを探り出し考慮する必要があるため、困難を伴う。

作戦・戦略のペースは、むしろ大胆さと委任への意欲によって決まる。その意味で、敵対者が AI システムに作戦・戦略決定を委任することで優位を得るかもしれないという恐れが、組織自身の作戦・戦略決定を AI に委任する動機となる可能性がある。敵対者の AI 利用への懸念は、実際の能力の進歩以上に、作戦・戦略的意思決定における AI 利用を加速させるかもしれない。

8.3 サイバー空間はスケールフリーネットワークである

「スケールフリー」とは、少数のノードが多数の接続を持つ一方、大半のノードはごく少数の接続しか持たないネットワーク構造を指す専門用語である（Barabási, Albert & Jeong, 1999）。こうしたネットワークは、大半のノードへの影響がネットワーク全体に深刻な打撃を与えない点でレジリエンスを持つ。しかし、ネットワーク全体に過大な影響を及ぼす単一障害点となるノードが比較的少数存在する。

AI は様々な形でインターネットやデジタルネットワークの構造を覆す可能性がある。AI が普及し、それが寡占または独占によってプロバイダされる場合、インターネットやデジタル技術は今日よりもさらに中央集権化されるかもしれない。必要な情報がモデル内に含まれるため、中小規模のノード同士が相互にやり取りする必要性が減り、トラフィックの多くが 1 つまたは少数の組織（すなわちモデル開発者）との間で流れるようになる。 ユーザーは直接ウィキペディアや天気ネットワーク、ニュースサイト、政府ページにアクセスする代わりに、AI サービスに情報を要求し、そのサービスが代わりにそれらのページにアクセスするようになるかもしれない。この構成は「スケールフリーネットワーク」というより、「ハブアンドスポークネットワーク」に近いものとなるだろう。

別の可能性として、AI が普及しつつも広く拡散し効率的に稼働するケースが考えられる。信頼性が高く自律的なオープンソースの低リソースモデルやエージェントをシステムに直接組み込むことで、ネットワークアクセスへの依存を減らすか排除できる。稼働に必要な情報やユーザーが求める情報のほとんどを、それ自体が完結させられるのだ。

将来のインターネットは、この二つのシナリオの組み合わせとなる可能性がある。この複合的な未来を予兆する例として、Apple と OpenAI の合意（Kerner, 2024）が挙げられる。 Apple の iPhone には小型で効率的なモデルが搭載されており、ネットワーク接続なしで可能な限り多くの処理を端末内で行う。より複雑なタスクのみが、OpenAI の高度なモデルへ送信される。従来 iPhone がインターネットから取得していた情報の一部は、現在では端末内部の AI モデル内に格納されている。さらに困難な課題に必要な追加情報は、iPhone からの直接リクエストではなく、応答プロセスの一環として OpenAI サーバーへ要求される場合がある。

これらの変化は、ほぼ全ての対象が常時接続・到達可能な現状のインターネット相互依存構造を大きく覆す可能性がある。効率的で信頼性が高く独立したモデルやエージェントの普及により、多くのシステムが不要となったネットワークから切り離されるか、機能に必要なネットワーク接続数を削減するだろう。こうした削減は、それらのシステムをより防御しにくい対象へと変える可能性がある。

8.4 サイバー空間は人工的で適応可能である

陸・海・空・宇宙といった他の領域とは異なり、サイバー空間は人間によって創られ、比較的容易に変更可能だ。

AI が既に貢献したコードの量から、サイバー空間はもはや「人間のみが作り出した」ものではないと主張できる。しかし重要なのは、人間か機械のどちらがサイバー空間を作るかではない。重要なのは、サイバー空間が人間の選好を反映して作られ、その選好をより良く反映するように変更可能だということだ。

もし AI が独自の嗜好を発達させ、それに従ってデジタルシステムを設計するならば、この点は再考が必要となる。限界を超える AI システムでさえ、独自の嗜好を持つ必要はなく、おそらく人間の操作者や設計者からの指示に従ってタスクを実行するだろう。知覚能力がなければ、サイバー空間が人間によって作られたという本質は、その仕組みが AI によって行われるからといって変わるものではない。

変化するのはサイバー空間の適応性だ。AI がコードや設定をより速い速度で変更可能にすれば、サイバー空間はより適応性を高める。これは現状の AI レベルでもある程度進行中であり、AI のさらなる進展に伴い加速すると予想される。

ある時点で、適応性は使いやすさと衝突するかもしれない。 人間のプログラマーがライブラリやサービスの更新が速すぎると苛立つように、AI プログラマーもシステム内のコンポーネントや基盤となるサービスが急速に変化すれば構築に苦労するだろう。この影響で、ライブラリやサービスの変化についていけなくなった人間開発者はさらに排除される可能性がある。

さらに開発者レベルではなくユーザーレベルでは、更新や変更にダウンタイムやユーザー行動の変更が必要となる場合、過度な変更は使いやすさやサービスレベル契約を妨げる恐れがある。

8.5 サイバー空間は計り知れないほど複雑である

サイバー空間内のコード量、接続数、可能な構成の多さは、人間が完全に理解することを不可能にしている。人間は様々な抽象化レベルを通じてその多くを理解している。デジタル機器がシリコンでできており、それが回路パターン化され、ファームウェアによって制御され、アプリケーションから信号を受け取り、インターネットとそのプロバイダ群、合意されたプロトコルを通じて他のアプリケーションと接続されていることは理解している。 人々はこれらの要素が組み合わさってサイバー空間を形成する仕組みを理解している。個々の人間は、そのうちのいくつかの詳細な複雑さを理解できるが、全てを同時に理解することはできない。

AI は複雑性の管理を専門とする。大量の情報を取り込み、パターンを発見したり、重要な要素を抽出したりできる。既に AI は、コードベースやデジタルバイナリの分析・説明、ネットワークトラフィックパターンの観察・説明など、この複雑性の側面を理解する上で人間を支援している。

さらなる進歩は、サイバー空間の計り知れない複雑性をより深く理解する助けとなるかもしれない。とはいえ、サイバー空間全体に存在するデータと情報の量は膨大であり、AI でさえその大部分を理解するのは想像しがたい。また、それほど詳細な分析が必要だとも考えにくい。人間が持つ抽象的な理解は、多くの目的において十分である。 しかし、より高度な知能は、今日の我々には想像し難い方法で、サイバー空間の複雑性をより詳細に理解することに価値を見出すかもしれない。

同時に、AI が膨大な新規コードを生成する可能性は、サイバー空間の複雑性を大幅に増大させる恐れがある。超人的レベルの AI、あるいはそれより能力の低い AI によって書かれたコードの一部は、人間にとって解釈が容易でないかもしれない。 そのコードやシステムは設計が複雑であるか、あるいは単に解釈困難な言語で記述されている可能性がある。AI は標準やプロトコルにも貢献しうるため、デジタル生態系の構成要素だけでなく、その構造全体にさらなる複雑性を付加するだろう。

現状レベルの AI でさえサイバー空間の複雑性を理解する助けとなるが、限界を超える AI でさえその不可解さを影響力のある程度まで軽減できない可能性がある。サイバー空間は抽象的な概念や個々の構成要素としては理解可能であり理解され続けるだろうが、全体としては計り知れないままである可能性が高い。

8.6 サイバー空間には固有の脆弱性が存在する

脆弱性はサイバー空間に遍在する。無限ではないにせよ、人間が脆弱性を発見する速度は、発見すべき脆弱性が無限に供給されている状況と一致している（A. Lohn & Jackson, 2022b）。

AI はいくつかの方法で脆弱性の蔓延を減らせる可能性がある。一つには、大規模なコードベースのレビューやリファクタリングに活用できる。現状ではある程度これは行われているが、AI は現在、脆弱性を減らすどころか増やしているかもしれない（Md Fazle Rabbi, 2024）。 信頼性と独立性を備えたレベルでは、AI エージェントが広大なコードベース全体で基本的なクリーンアップを実行し、基本的なセキュリティ対策を実施する信頼性を得られる可能性がある（DARPA, 2023）。

より高度な課題は脆弱性の発見だ。現状レベルの AI は既に人間の脆弱性発見者を支援しているが、専門家や超人的な AI は人間が発見できない脆弱性を見つけ出す可能性がある（Google, 2024）。 ただし、この利点がサイバーセキュリティに有効なのは、攻撃者が防御側が何らかの理由で修正できない脆弱性を見つけるのをツールが助けない場合に限られる。その理由としては、防御側が全てを発見するには脆弱性が多すぎる場合や、修正が困難・不可能・望ましくない場合（ハードウェア脆弱性のように性能を阻害する場合など）が挙げられる（A. Lohn, 2018）。

防御側はまた、最初からより安全なソフトウェア、場合によっては証明可能な安全性を備えたソフトウェアを設計できる AI システムからも恩恵を受ける可能性がある。メモリー安全言語でのコーディングは、デジタルエコシステムを強化する

近期的アプローチの一つだ（Rohlf, 2023）。証明可能な安全設計を大幅に前進させるには超人的レベルの AI が必要かもしれないが、専門家レベルの AI ソフトウェア設計者を多数投入すれば可能かもしれない（Woodcock, Gorm Larsen, Bicarregui & Fitzgerald, 2009）。

信頼性が高く独立した AI 防御システムは、大量のコードを強化できる可能性がある。その作業は特に難しくないが、人間には多すぎる、あるいは人間が退屈に感じる、あるいは魅力的でないものばかりだ。これによりデジタルエコシステムから多くの既知の脆弱性を排除できる。さらに専門家のレベルに近づく進歩が必要かもしれない。まだ発見されていない多くの脆弱性を発見するためには。 脆弱性の発見は防御側が製品を強化するのに役立つが、AI が未発見の脆弱性を大量に生み出すなら、攻撃側が防御側よりも大きな利益を得る可能性が高い（A. Lohn & Jackson, 2022a）。な超人的 AI へのさらなる進展は、証明可能な安全性を備えたソフトウェアを創出できれば、再び防御側に有利な状況をもたらすかもしれない。そうすれば、それらの未発見の脆弱性は再び発見されにくくなるからだ。

8.7 能力の低影響かつ可逆的な効果

サイバー攻撃は従来、災害というより迷惑行為のカテゴリーに属していたが、その傾向は薄れつつある。 NotPetya による数十億ドルの損失（McQuade, 2018）のように、災害レベルの経済的影響をもたらしたサイバー攻撃も存在する。コロニアル・パイプラインや医療業界全体への攻撃も増加傾向にある（Easterly & Fanning, 2023; Riggi, 2020）。

サイバー大国間の比較的平穏な時期には高影響度の攻撃が少なかったという前例から、今後もサイバー攻撃の影響は低いままだと結論付けるのは賢明ではない。 しかし影響が限定的だったのは、攻撃者の意図が限定的だったためだけではない。防衛側も、軍事装備や重要インフラの一部など特にリスクの高いシステムを過度にデジタル化したり接続したりすることを躊躇している可能性がある。AI の進歩によってデジタルシステムの脆弱性が増せば、防衛側はシステムをデジタル攻撃に晒さない選択ができる。しかし人間が AI 搭載システムへの依存度を高めれば、サイバー攻撃発生時およびその後のレジリエンスは低下するだろう。

サイバー攻撃の影響が歴史的に低かったのは、最も能力のある脅威アクターが一定の抑制を受けていたためだ。しかし AI の進歩は、サイバー攻撃から予想される影響の度合いを変える可能性がある。ならず者国家やサイバー犯罪者でさえある程度躊躇してきた領域でも、サイバーテロリストは躊躇しないかもしれない。 テロリストは従来、破壊的な攻撃を実行する技術力に欠けていた（Rid & McBurney, 2012）。しかし AI 攻撃ツールが十分な能力を持ち、広く普及すれば状況は変わる。ただし防御ツールが防御側によるシステム強化やデジタルエコシステム全体の堅牢化を可能にする場合、攻撃ツールの拡散はテロリストにとってそれほど有用ではないかもしれない。

8.8 攻撃はスパイ活動の包含関係における下位概念である

破壊的効果をもたらすことを目的としたサイバーキャンペーンは、正当なスパイ活動と見なされるキャンペーンと区別するのが難しい。ある意味、正当なスパイ活動と違法な効果の両方を撃退するのは防御側の責務だ。したがって、この区別（あるいは区別がないこと）は戦略的・外交的には真実であり重要かもしれないが、技術的レベルではそれほど重要ではないかもしれない。

AI がサイバーキャンペーンの所要時間（数時間から数日）を短縮すれば、破壊的な攻撃能力を諜報活動の偽装で事前配置する必要性は低下する。攻撃者は必要な時だけアクセス権を得て攻撃を実行できる。 同時に、AI がさらに進歩すれば、持続的な諜報活動の利点は維持されるか増大する可能性が高い。したがって、諜報キャンペーンは AI の進歩レベルに関わらず、おそらく人気を保ち続けるだろう。そしてそれらの諜報キャンペーンは、ほとんどの場合、違法な効果キャンペーンへと容易に転用され続ける。

ただし、AI の場合、現行のサイバーツールよりも、捕捉した AI インプラントから意図を推測しやすいケースもある。防御側が侵入エージェントに問い合わせたり、安全なサンドボックスで評価したりできれば、正当な活動ではなく違法活動をいつ行うか、あるいは行うかどうかを判断できる可能性がある。 しかし、サイバーエージェントが破壊ではなく正当な諜

報活動のために設計されているという確認は、攻撃者が エージェントを更新・交換して攻撃性を高めることが困難な場合にのみ安心材料となる。エージェントが更新を受け取れない場合もあり、その場合は意図を推測できるが、今後の基本前提として、そうした更新は容易であるべきだ。

8.9 技術変化の急速なペース

デジタル技術は急速に進化し変化する。攻撃者と防御者が用いる手法も同様だ。

AI は既に、人間が新しいコードやコンポーネントを設計・記述する能力、そして攻撃者と防御者が用いる手法の種類において、イノベーションの速度を加速させている。この加速は今後も続くと予想されるが、おそらく無限には続かないだろう。なぜなら、相互に関連するコンポーネントやサービスをあまりにも急速に変更することで、性能や相互運用性の課題が生じる可能性があるからだ。 基盤となるサービスが変更または更新されると、サービスが機能しなくなる可能性がある。また、グローバルなエコシステム全体の安定性に対する必要性は、その構成要素間の急速な変化に対する圧力となるかもしれない。

8.10 普遍的な相互接続と依存性

複雑なデジタルサプライチェーン上に構築されたグローバルインターネットと統一されたデジタルエコシステムは、人々、組織、国家を相互依存させている。その依存関係は評価や理解が困難な側面を持つ。しかし、輸出管理や検閲などによる国家間の分離が進み、企業がサプライチェーンリスクへのエクスポージャーを減らそうとする中で、この相互依存は幾分減少している可能性がある。

もし AI がインターネットをスケールフリーネットワークからハブ・アンド・スポーク型ネットワークへと再構築すれば、ネットワークレベルの相互接続性はさらに低下する可能性がある。信頼できる独立したエージェントがネットワークから切断される場合にも同様の事態が生じうる。

一方で、AI がより多くのコンポーネントを生み出し、より多様で豊富なリソースを活用するようになれば、デジタルエコシステムは相互接続性と相互依存性を高める可能性がある。ソフトウェアは、もともとニッチな目的で開発されたり異なる言語で記述されたりしたため、人間の開発者が認識していないライブラリやコーディング手法を取り込むかもしれない。

8.11 許可不要の革新と接続

インターネットやデジタル環境の多くの側面は、容易にアクセスでき、容易に貢献できるように設計されている。これはインターネット探索の自由やオープンソース貢献の普遍性からも明らかだ。

ロボット.txt や CAPTCHA といった一連の技術や手順は、非人間的なデジタルエージェントの役割をある程度制御するのに役立ってきた。これらの手法の一部は、現状の AI に対しても効果が低下しつつある。一部のアプリケーションでは、より堅牢な人間性証明（Sedlmeir, Smethurst, Rieger & Fridgen, 2021）の開発が可能あるいは必要となるかもしれない。

AI の重要性が高まるにつれ、オープン性の争点は、トレーニングデータセットへのデータ包含や、モデル・システム・サービスの流通に移行する可能性がある。一部の開発者や国家は、様々なデータソースの包含・排除を選択するだろう。検閲を含む、より積極的なデータクリーニング手順を採用する可能性もある。 中国のモデルは、アメリカのモデルと同じ方法で全ての国際的な 問題に対処しないだろう。輸出管理や顧客確認（KYC）ポリシーも、どの主体や国家がどの機能にアクセスできるかを制限する可能性がある。

8.12 曖昧な境界線

デジタル信号は地理的境界に容易に拘束されない。確かに国家や企業は、ファイアウォールによるデジタル境界や物理インフラに基づく地理的境界を設定できる。ゆえにサイバー空間は境界なき領域ではないが、境界は曖昧である。

AI がサイバー攻防を機械速度で加速させれば、遠隔国間の信号伝達時間が重大な要素となる可能性がある。通信遅延は高速取引における長年の制約要因であり、レース条件脆弱性などでは既に重大な影響を及ぼしている（C. N. Fischer, Netzer & Miller, 1992）。高速化が進む中、境界は依然として曖昧だが、地理的要因が現在以上に影響力を増すだろう。

より近い将来においては、AI はすでに外交上の争点となっている。特定の地理的地域では、最高性能のシステムへのアクセスや利用が制限される可能性がある。このシナリオは、最高性能の AI が広く普及するよりも、ごく少数の組織にアクセスが限定される場合に起こりやすい。

8.13 物理世界との結びつき

デジタル技術やデータは儚いように見えるかもしれないが、それらは回路、光ファイバーケーブル、ネットワークルーターといった物理的な構成要素の中に存在する。

これは AI であるか否かを問わず全てのデジタルシステムに当てはまるが、デジタル部品の重要性は技術によって異なる。 現在、最も高性能な AI システムは、単一プロバイダが支配する高価なコンピュータチップに依存している。それらのチップは、別の支配的プロバイダが製造するトランジスタを用いて作られ、さらに別の企業が設計・構築した部品を必要とするファブで生産される。そして最先端モデルは、大規模で膨大な電力・冷却需要を伴う物理的なデータセンターで訓練・運用されている。

AI の物理的側面は、これまでのデジタル技術の大半と比べ、AI に関連するサイバー空間においてより重要な要素となってきた。AI が今後も極めてリソース集約的であるならば、この傾向は継続あるいは拡大する可能性がある。逆に、十分な能力を持つ AI がリソース集約性を低下させるならば、その物理的依存度は従来のデジタル技術と同水準まで低下するだろう。

8.14 情報（および能力）の複製容易性

デジタル技術、少なくともソフトウェアや情報（デジタルハードウェアでなくとも）は、急速に複製できる。情報のコピーは既に非常に効率的であり、AI がそれをさらに容易にする可能性は低い。しかし、AI が本来なら人間に内包されていたり、多くの異なる情報源に分散していたりするデータや情報を符号化し集約する限り、AI はデジタル情報の複製容易性をより有害なものにするかもしれない。

例えば、大量の機密情報で訓練された AI モデルを想像してみよう。現在、様々な の許可制限で分離された区画に保管されている情報を集約するのは難しい。もし将来、その情報が単一の AI モデルに凝縮されたら、そのモデルをコピーすることは、訓練に使われた全ての情報を盗むことに等しい。

別の例として、個人の行動や関心を学習し、正確に代行するため、あらゆるやり取りや行動を観察するパーソナル AI アシスタントを想像してみよう。このモデルが存在すれば、攻撃者は現在収集・保持されていない情報を容易に複製できる。スマートフォンやウェアラブル端末がデータを収集する現状でも既に発生しているが、AI によって、あるいは AI をより機能させるために、この問題はさらに悪化する可能性がある。

8.15 民間セクターによる支配

デジタル世界のインフラと専門知識の多くは民間セクターに存在する。唯一の例外は、正当な攻撃権限を持つ唯一の組織である政府が、最も能力の高い攻撃的サイバーチーム、ツール、インフラを保有している点だ。

AI システムも同様に民間セクターが支配している。民間企業は政府が追いつけない予算と専門知識を持ち、AI システムを開発できる。政府はこうした企業に一定の影響力を行使し、協力関係を築く可能性が高い。政府は脅威インテリジェンスなどの有用な情報を企業に提供できる一方、不審なアカウントに関する情報なども受け取れる。政府は必要に応じて、規制を緩めたり厳しくしたり、場合によっては企業を閉鎖することさえできるかもしれない。

たとえ政府が最先端の AI システムを生産・取得できなくとも、攻撃的サイバー活動に関連する専門知識、ツール、認可を活用することで、サイバーセキュリティの攻撃面において民間企業に対する優位性を維持できる可能性がある。

8.16 全体的な攻撃者の優位性

サイバーセキュリティ分野では「攻撃者が優位にある」という通説がある。本稿全体の目的は AI がこの通説に与える影響を解明することであるため、その他の各節の議論を参照されたい。

8.17 迅速かつ正確な帰属の困難さ

デジタル上の身元や行動は実際の犯人とは切り離されているため、帰属の特定はしばしば困難または遅延する。

人間のオペレーターがサイバー攻撃プロセスから遠ざかるほど、その人間と攻撃を結びつける手がかりは少なくなる。ロシアと中国が同じ AI を使って攻撃スクリプトを作成した場合、そのスクリプトは両者を区別する上で有用性が低下する。また AI は、偽旗作戦のために他の脅威アクターが用いる手法でスクリプトを作成したり攻撃を実行したりするよう指示される可能性もある。

しかし防御の観点では、AI はより複雑なソースからの大量データを横断的に分析し、手がかりを統合することで、攻撃の帰属を特定犯行者に割り当てる根拠を構築できる可能性がある。攻撃全体が AI によって実行された場合でも、動機、攻撃インフラの所在地や所有権、指導者や指揮官のコミュニケーションなど、多くの手がかりは残るだろう。

AI が帰属特定のプロセスを加速できれば、数か月や数年待つよりも有用なタイムラインで結論を出せる。その間、世論の怒りは収まり、法的・技術的な報復は身柄引き渡しなどの追加的な課題に直面する可能性がある。

8.18 直接観察が困難

デジタル情報とデジタル効果は観察が難しい。侵入者はステルス性を持ち、攻撃コードは衛星画像には映らない。遠心分離機の設定変更さえも気づかれないことがある（Zetter, 2014）。

AI は以前から、システムやファイル内の潜在的な悪意あるコードや活動をスキャンするために使用されてきた。防御側の AI が進化する攻撃戦術に追いつくのは歴史的に困難であったが（Dissanayake, Jayatilaka, Zahedi & Babar, 2021）、AI は今後も改善を続けると推測される。 高度な AI が、既に広く普及している自律スキャンによる悪意のあるファイルや行動の識別能力を向上させるのか、あるいは高度な AI が、本物ユーザーのような振る舞いやポリモーフィックマルウェアの作成によって、それらのファイルや行動の検知を困難にするのかは不明だ。

攻撃者の活動を監視する能力は、サイバーセキュリティ分野における AI へのアクセス制御の程度にもある程度依存する。攻撃者が AI 企業のサーバー上で実行するリクエストを提出する必要がある場合、それらの企業は攻撃前または攻撃中に脅威アクターを監視・制限し、フォレンジック調査のための活動を記録できる。たとえより寛容な企業や敵対国がプロバイダとしてサービスを提供する場合でも、テロリストなどの特定の攻撃タイプや脅威アクターを抑制することで合意する可能性がある。

高性能な AI が広く普及し容易にアクセス可能になれば、監視や干渉はより困難になる。モデルに組み込まれた制限やガードレールは、現時点では悪意あるユーザーを制限する上で期待できない。さらに、脆弱性発見や侵入テストといった防御に必要な能力は、攻撃的な利用に必要な能力と類似している。

8.19 能力は一時的であり、予測困難な影響をもたらす

攻撃ツールや手法は、単純な防御策を一つや二つ加えるだけで無効化されることが多い。結果として、攻撃能力は長く有用ではない可能性がある。これは攻撃側に、将来使用できなくなるかもしれない能力を急いで使うよう圧力をかける。

攻撃的ツールや技術を迅速に無効化するこの特性と、サイバー空間の複雑性・相互依存性が相まって、攻撃効果の予測を困難にしている。キャンペーンや攻撃が予期せず失敗する可能性がある。標的に与える損害が意図した量より多くなったり少なくなったりする可能性があり、意図しない標的に付随的損害を与える可能性もある。

AI が新たな脆弱性を迅速に発見しパッチ適用プロセスを加速させれば、攻撃能力はさらに一時的なものとなる。単に標的システムの通常更新や変更を加速させるだけでも、意図せず攻撃を無力化できる（Ablon & Bogart, 2017）。 一方で、攻撃側が新たな脆弱性を発見し、新たなツールや技術を容易に開発できるようになれば、攻撃側は自らのツールや技術の一時的な性質をそれほど問題視しなくなるだろう。攻撃者は、事前に用意した選択肢のレポジトリを維持する必要はなく、必要な時に必要な能力を創出できると期待できる。

効果の不確実性に関して、攻撃者は攻撃回数を増やすことで不確実性を低減できる可能性がある。各攻撃が意図した効果を生むか不確実であっても、同じ時間や週数でより多くの攻撃を試みられるなら、攻撃者はその期間内に少なくとも一度は成功する確信を持てるようになる。

しかし、攻撃回数の増加は、意図した以上のエスカレーションや損害をもたらす過剰な効果を偶発的に引き起こす確率も高める。また、巻き添え被害の確率も増加する。一方で、攻撃回数の増加は、防御側が攻撃者の作戦を識別し妨害する機会も増やす可能性がある。

8.20 作戦上の休止がほとんどない常時接触状態の敵対勢力

攻撃者と防御者、特に国家や大企業、犯罪組織といった大規模な主体は、絶え間なくサイバー攻撃を仕掛けたり防いだりしている。これは、交戦が稀な他の領域における物理的紛争とは異なる。

この絶え間ない接触の一部は、迅速な損害・破壊効果を発生させるための事前配置である。攻撃者が AI によって攻撃の開始と終結を迅速かつ確実に実行できると確信すれば、この事前配置の必要性は低下するだろう。しかし破壊効果のための事前配置は、大半のサイバー交戦を動機づけるものではない。情報収集目的での継続的な情報収集欲求は、高度な AI によっても減少しない可能性が高い。サイバー犯罪者も定期的な攻撃を継続する動機を保ち続けるだろう。

ただし、AI によって攻撃者の数や攻撃者一人当たりのエージェント数が増加すれば、接触の頻度は高まる可能性がある。今日「絶え間ない接触」と感じられる状況も、信頼性が高く自律的に動作するサイバー攻撃エージェントが広く普及した時代には、そうは見えなくなるかもしれない。現在、DDoS 攻撃のような単純な自動メッセージ送信を主に実行するボットネットが多数存在するが、将来はそれらのボットがそれぞれ創造的な行動を取れるようになり、防御側は個々にそれに対応する必要が生じるだろう。

8.21 国家の権力源への即時的な大陸間接近性

デジタル信号は光速で世界を駆け巡り、地上部隊やミサイルよりもはるかに速く標的に到達する。さらにデジタル世界では、国家の境界線と首都や産業拠点との到達時間にほとんど差がない。

デジタル戦闘が十分に加速し、意思決定の速度が決定的要因となった場合、地理的場所間で情報を移動させる光速では不十分となる可能性がある。信号が目的地を隠すために難読化ネットワークを通過する必要があったり、反撃を防ぐためにファイアウォールで内容をアセスメント・フィルタリングする必要があったりする場合、その時間遅延は光速よりもはるかに長くなるだろう。

さらに、AI がインターネットを再構築すれば、重要な標的の一部はデジタル的に到達困難になる可能性がある。標的は独自の信頼性の高い独立した AI システムやエージェントを使用し、インターネットから切断するか、接続数を減らすことで、攻撃者による到達を困難にできる。

8.22 優位性は能力の保有ではなく使用から生まれる

サイバー攻撃は、例えば核ミサイルとは異なり、使用の脅威が実際の使用よりも効果的である場合がある。サイバー脅威に関する情報を過剰に提供すると、防御側はそれを無効化できる。

この見方はやや誇張されている。国家や企業は、敵対するサイバーチームの具体的な能力を知らなくても、その潜在能力を考慮しているからだ。防御側は、十分に高度なハッキングチームなら、時間と労力を費やせばどんなデジタルシステムにも侵入できると想定している。また、各国が保有するそのようなチームの推定数や、様々な犯罪グループの能力と洗練度の推定値も存在する。

高度な AI システムを用いれば、敵対者の攻撃能力をより正確に評価できるかもしれない。AI 攻撃ツールがエキスパートレベル以上になれば、侵入テスト担当者は敵対者が保有するツールや専門知識と同等か類似の AI ツールを使用できる。ただし攻撃者が能力を大幅に向上させる追加データ・ツール・専門知識を有する場合、テスト結果は敵対能力を正確に反映しない可能性がある。敵対者が重要なデータやツールを欠いている場合もある。また、最高性能の AI システムへのアクセスが制御可能であれば、攻撃者や敵対者への提供を拒否できる。

レッドチーム活動の結果が、標的が強化されているか脆弱性を有しているかを明確に示さない場合、AI が大幅に進化しても、能力の保有ではなくその活用から優位性が得られる可能性がある。

8.23 敵対者は通常、武力紛争レベルを下回る能力を日常的に使用する

攻撃的サイバー活動は、被害者に妨害効果をもたらすものの、事態のエスカレーションを招くほどではない限定的な影響を与える手段として好まれるようになった。サイバー攻撃は、罰則のない侵略手段としての評判を得ている。

AI がデジタル諜報活動の価値や犯罪者への経済的利益を大きく変える可能性は低いため、サイバー攻撃は影響力の低いレベルで継続的に使用され続けるだろう。攻撃者は過剰な報復を受けずに可能な限り多くの成果を得るため、サイバー攻撃を使い続ける。

ただし AI は武力紛争への閾値を超えるリスクを高める可能性がある。 攻撃者の数が増え、制御や監視が弱まる可能性のある状況では、偶発的なエスカレーションのリスクが高まる。また AI は、訓練データとは文化的に異なる人間の被害者の反応を理解するのが苦手かもしれない。たとえ AI が人間の反応をより正確に予測し、個々の攻撃がエスカレーションを引き起こす可能性が低くなったとしても、攻撃の量が増えればエスカレーションリスクは依然として高まる。特に攻撃主体が誤りを犯しやすい場合、その傾向は顕著だ。

8.24 参入障壁の低さ

攻撃的サイバーツールは既に、ハクティビスト、スリルを求める者、内部脅威、低レベルの犯罪組織など、技術的知識やスキルがほとんどない個人でも入手・使用可能だ。より高度な持続的脅威にはスキル、知識、チーム、インフラが必要だが、それでも国家予算や軍事予算に比べれば通常はわずかな投資で済む。

AI は専門知識を普及させることで、低スキルの攻撃者の能力をさらに高める可能性がある。AI は彼らに高度な攻撃を実行する専門知識をプロバイダしうる。同時に AI はコード基盤や潜在的な被害者を強化し、それらの攻撃能力を無力化する可能性もある。

考えられる進展として、信頼性が高く自律的な AI 防御システムが普及し、それらがシステムを十分に強化することで、信頼性が高く自律的な AI 攻撃者が苦戦する可能性がある。さらなる進歩により、専門的な AI 攻撃者の普及が、

専門的な AI 防御者に対する攻撃者に有利に働くかもしれない。しかし、超人的な防御者が構築する防御は、超人的な攻撃者でさえも突破が困難なものとなる可能性がある。

8.25 優位性は一時的である

デジタル世界の変化の速さ、再構成可能性、参入障壁の低さは、いかなるサイバー上の優位性も長くは続かないことを示唆している。

AI はデジタル世界の変化速度と再構成可能性をさらに加速させ、サイバーアクターの参入障壁をさらに低下させる可能性がある。加えて、AI 技術自体が急速に変化しており、今後も急速な変化が続くかもしれない。ある時点で最先端の AI を保有する国や企業も、他国や企業が革新を続ける中で、わずか数回の技術革新で自らの技術が陳腐化する可能性がある。

AI は専門知識を急速に普及させる仕組みでもある。今日、多くの卓越した専門家チームによって先進的な立場にある国家や組織も、他の組織が同等の専門知識を持つ AI を利用できるようになれば、それほど特別ではなくなるだろう。AI はサイバーセキュリティの攻防双方において、競争条件を平準化する可能性がある。

8.26 能力開発コストは他分野より大幅に安い

サイバー能力の開発コストは、通常、航空機や戦車に比べて低い。一度開発された能力の複製を多数生産するコストはさらに低くなる。

AI の生産や運用にはコストがかかるかもしれないが、サイバー能力のコストをさらに削減する可能性が高い。AI 開発コストはほぼ全産業に分散されるため、サイバー応用分野は規模の経済効果により比較的低コストの AI システムを利用できる。ただし、高性能なサイバーシステムには膨大なデータセットの生成が必要となり、その開発や学習に多額の費用がかかる可能性がある（A. Lohn et al., 2023）。その場合、サイバー能力は現在よりも高価になる可能性がある。

8.27 能力は迅速に再構築可能

防御側がサイバー攻撃を撃退または阻止しても、それは一時的な勝利に過ぎないことが多い。なぜならサイバー能力は迅速に再構築できるからだ。例を挙げれば、潜水艦の再建造には比較的長い時間を要し、戦車製造基地が爆撃された後の産業基盤の再構築にもより多くの時間を要する。一方、防御側がサイバー攻撃を撃退する場合、防御態勢を持続的に改善する方法でそれを達成できる可能性がある。

AI がサイバー能力の開発を加速させれば、その再生は今日よりもさらに速く起こるだろう。 ただし、AI が加速しにくいボトルネックが存在する場合、攻撃キャンペーンの遅延に占める再生時間の割合はより重要になる可能性がある。例えば、攻撃用インフラの構築にはアカウント作成が必要であり、 身元確認は、特に人間のフォロワーによる正当性が求められる場合、確立に時間がかかる。また、インフラプロバイダが広範なサイバー防御の一要素である限り、例えば顧客確認方針を通じて、こうした再構築のタイムラインをある程度制御する能力を持つ。

8.28 攻撃は壊滅的な影響をもたらす可能性がある

サイバー攻撃は小規模で一時的な効果しか持たず、エスカレーションの閾値を下回ると見なされる傾向があるが、より重大な影響をもたらす可能性を秘めている。理論上、水道供給の汚染、発電設備の破壊、物理的攻撃の開始などに利用されうる。また、小規模な攻撃の集積が、民主主義の後退といった壊滅的な結果を招くこともある。

攻撃量の増加は偶発的な壊滅的災害の可能性を高める。さらに、高度な攻撃ツールや技術がサイバーテロリストの手に渡れば、意図的な壊滅的サイバー攻撃のリスクも増大する。おそらく、サイバーテロリストのスキル向上は防御側

のスキル向上と並行して進むだろう。最も恩恵を受けるのは、専門知識と予算に制約のある中小組織であり、これには重要インフラの多くが含まれる。

より重要なシステムへの AI 統合も壊滅的影響の可能性を高める。現在、人間のオペレーターは攻撃者と影響の間に非デジタルの障壁を提供している。人間は攻撃中および攻撃後のレジリエンスプロバイダも提供する。AI は攻撃者に対して脆弱であることが知られているため、より多くの AI コンポーネントを統合すると、攻撃に対する脆弱性が増大する一方で、レジリエンスは低下する可能性がある。

8.29 機動力と主導権に依存する戦術的成功

状況の変化に応じて進路を変更し、機会に適応する能力、ならびに攻撃者や防御者の行動・戦術への調整能力は、個々の攻撃やキャンペーンの成功に重要である。

信頼性が高く自律的なレベル以上の AI は、追加の指示なしにリアルタイムで戦術的判断を下す認可を与えられ得る。これは攻撃側・防御側双方に当てはまる。攻撃と防御の数が単純に増加するにつれ、ミスが生じる可能性もある。特に、正確性や信頼性ではなく、速度や規模を理由に判断を AI に委任した場合に顕著だろう。ミスに対するペナルティは、利害関係者の利便性や稼働時間に悪影響を与えないよう回避する必要がある防御側の方が高くなる可能性がある。攻撃側は主に、発見されることや意図しないエスカレーション、巻き添え被害を避ける必要がある。

8.30 戦略的成功は大胆さと主導性に依存する可能性が高い

サイバーセキュリティにおける戦略的利益は、より攻撃的な主体に蓄積される傾向がある。これはサイバー脅威アクターが依然として境界線を踏み越えず、自らの攻撃がどの程度で高コストな報復を招くかを見極められていないためかもしれない。攻撃が小さなペナルティしか生み出さない間、より大胆な攻撃はより大きな戦略的利益をもたらす。

おそらく攻撃件数が増えれば、どの程度の深刻度や頻度の攻撃がコストのかかる報復を招くかが明らかになるだろう。今日の典型的な攻撃よりも大胆ではない攻撃が数多く発生することで、国家が軍事的対応に追い込まれる可能性もある。 その結果生じる破壊は、そうしたサイバー攻撃を戦略的失策に容易に変えうる。サイバー空間におけるレッドラインが明確になれば、攻撃者はより大胆な攻撃を選択するかもしれないが、さらなる大胆さと主導性が戦略的成功につながるという確信は弱まるだろう。しかし AI がレッドラインを明確化しない可能性もあり、大胆な AI 攻撃が戦略的利益をもたらし続けるかもしれない。

AI 駆動型攻撃に対し、一般市民や国家がどう対応するかは未だ未知数だ。AI エージェントは従来のボットやワームより抑制され制御されているかもしれない。しかし人々は、人間発の攻撃よりも AI 主導の攻撃に脅威を感じ、反発する可能性がある。その人工性、あるいは様々な理由から、AI 攻撃における大胆さは利益をもたらさなくなるかもしれない。

8.31 抑止が困難

サイバー攻撃は抑止が難しいと主張する者もいる。その理由として、攻撃の帰属が困難なため攻撃者が攻撃後に無実を主張できることや、サイバー能力に関する秘密主義が抑止側を阻み、抑止に必要な能力を把握できないことが挙げられる。

AI が意思決定に用いられている場合、抑止側は人間の攻撃者よりもその意思決定の計算式を理解しやすいかもしれない。防御側が同様の AI システムを保有しているか、実際の攻撃エージェントを捕獲している場合、それを検査して自軍の防御や挑発に対する反応を予測できる。しかし攻撃側が攻撃エージェントやその指令を更新できる場合、こうした防御側の調査はほとんど価値を持たないかもしれない。

防御側は相手の対応を理解するよりも、自らの攻撃エージェントを用いて脆弱性を把握すべきだ。現在のレッドチームは敵対者の能力を模倣しようとするが、熟練した敵対者を模倣する人間の専門知識には限界があるため、適切なレ

ッドチーム活動を実施できる組織は少ない。AI 攻撃ツールが広く普及すれば、多くの組織が抑止すべき敵対者の能力をより深く理解できるようになるだろう。

サイバー攻撃の抑止が困難なもう一つの理由は、攻撃者が攻撃後に無実を主張できると想定している点にある。AI がサイバー攻撃の帰属判定にどう影響するかは不明だ。AI が大量の複雑な情報を処理する能力は帰属判定を向上させる可能性がある。しかし、多くの攻撃者が類似の AI ツールを使用する場合、あるいは AI が他の攻撃者のスタイルで攻撃を実行できる場合、将来的に帰属判定はより困難になるかもしれない。

8.32 防御の成功は攻撃者を阻まない

攻撃者はたった一度の成功さえあればよいので、成功するまで攻撃を続ける。これに対する反論として、防御側の成功は単に攻撃を阻止するだけでなく、防御側が攻撃者のインフラを破壊したり、防御コミュニティ全体で共有できる警告を発して攻撃者を識別し活動を阻止したりする可能性もある。

人間の攻撃者は繰り返しの失敗で確かに挫かれる。一方、信頼性が高く自律的な AI 攻撃者は、本当に挫かれることはない。防御が不完全であれば、24 時間 365 日、疲労や挫折なく高頻度で再攻撃を仕掛ける大量の攻撃者が、最終的にその防御の不備を見つけ出すだろう。しかし、そうした繰り返しの失敗は、防御側が攻撃者のインフラを破壊したり、成功した防御戦術・技術・手順を広範な防御コミュニティと共有したりするリスクを攻撃者に依然として課すことになる。

8.33 攻撃の警告が困難であること

サイバー攻撃は突発的であり、準備段階を観察するのは困難だ。マルウェアが敵対者のシステム上で開発され、分析が敵対者のネットワーク内のみで実施される場合、攻撃開始前にその兆候を察知できない可能性がある。これが国家が前線防衛と持続的関与を行う理由の一部だ。攻撃開始前に敵対者の準備や計画を観察する必要性を感じているのだ。しかし、標的ネットワーク内で準備が行われる場合、例えばマルウェアや攻撃ツールの事前配置時や、ネットワークおよび標的の偵察中に、何らかの警告は可能である。

AI が侵入、ツール開発、攻撃計画のタイムラインを短縮すれば、指標が検知可能であっても警告のタイムラインはさらに短縮される可能性がある。 今日、警戒レベルは攻撃の技術的指標よりも、政治的緊張の高まりや事態の悪化といった要素に結びつけられることが多い。こうした非技術的指標は、警戒レベルを導く指針として今後も利用可能だろう。主に非デジタル領域におけるこうした政治活動や事態の悪化は、公的声明や軍隊の配置といった社会的・物理的なタイムラインに依存しているため、AI によってそれほど加速される可能性は低い。

8.34 意図の示唆は問題を抱える

攻撃的サイバーツールや侵入の目的を解釈するのは難しい場合がある。侵入やマルウェアは正当な諜報目的である可能性もあれば、不正な破壊を意図している可能性もある。さらに複雑なのは、攻撃者が明確な目的を持たず、両方の選択肢を残すことを好む場合だ。 敵対ネットワークへの侵入は、一歩踏み込んだ防衛や脅威能力の無力化を目的とする場合もあり、攻撃と防御の境界は曖昧だ。また個人や組織が企業に侵入し、防御強化やバグ報奨金獲得を図るケースもある。

意思決定を AI に委ね、その AI を捕獲・尋問できるなら、現在のサイバーインプラントよりも意図を特定できる可能性がある。これは AI がサイバー紛争の意図問題を解決する希望を示す。 しかし、AI が迅速に更新可能な場合（おそらく頻繁に起こりうる状況だ）、人間の攻撃者の実際の戦略的意図は現在と同様に不明のままである。これらの尋問技術は、インプラントを事前評価するよりも、事後的に意図を法医学的に発見する方が有用かもしれない。

8.35 先制攻撃兵器となる可能性

サイバー能力は、警告を受けた防御側によって迅速に打ち破られる可能性があり、またサイバー攻撃はエスカレーション化しない場合が多いため、サイバー攻撃は先制攻撃兵器として好まれると予想する者もいる（Healey & Jervis, 2020）。危機的状況においてサイバー攻撃を行う動機は、紛争へのエスカレーションのリスクを高める可能性があるが、サイバー攻撃は、エスカレーションを緩和する圧力解放弁としての役割も果たす可能性がある。

AI によって攻撃者がより短期間でアクセスと効果を得られることが期待できる場合、攻撃者は「使わなければ失う」というプレッシャーをあまり感じなくなる。危機的状況においてエスカレーションにつながる可能性のあるサイバー準備や事前配置を行うインセンティブは少なくなる。

しかし、サイバー攻撃は、リスクが低く戦争に至らない行動としては、依然として魅力的な選択肢であり続けるだろう。影響や破壊の程度には様々なレベルがあり、サイバー攻撃は物理的手段と比較してリスクが比較的低いと見なされるようになってきている。

8.36 驚異性がより重要である

防御側は十分な事前警告があれば通常は攻撃を阻止できるため、攻撃側は奇襲に依存する。ただし、奇襲に頼らず持続的な攻撃を好む攻撃例からも明らかなように、これは常に当てはまるわけではない。さらに、ロシアがウクライナの電力供給を二度遮断したように、過去に攻撃を実施した事実があっても、防御側が将来同様の攻撃を確実に防げるとは限らない。

もし AI が攻撃側に戦術的優位をもたらすなら、攻撃側は標的や効果を公表することを気にせず、奇襲効果も低下する。防御側は攻撃側が脅威を実行可能だと確信するだろう。標的やデジタル構成要素を完全に切り離さない限り、防御側はそれらを防御できない。その切り離しだけで、実際にサイバー攻撃を実行せずとも、攻撃側の作戦目標や戦略目標を達成するのに十分かもしれない。

しかし攻撃側がそれほど戦術的優位性を確立しない場合、奇襲要素は依然として攻撃側に大きな利益をもたらす。AI が攻撃側にこれほど大きな戦術的優位性をプロバイダするかは全く明らかではない。

8.37 国家による代理組織の活用は容易である

国家は攻撃を他国や組織を経由させることで、検知や帰属の特定を回避する。また犯罪者や個人に、自国の目的や利益に沿った攻撃を実行させることも可能だ。そうした犯罪者や個人は、日中は国家に雇用されていたり、国家から指示や任務を受けている場合すらある。

国家や犯罪者によって信頼性が高く独立した AI 攻撃者が作成・任務付与され、世界中に配置される可能性がある。これらのエージェントは設計者や展開者と結びつく観察可能な特徴がほとんどないため、既存のプロキシに代わる拡張性の高い選択肢となり得る。

攻撃者は他者の攻撃エージェントや善意のソフトウェア・エージェントを乗っ取ることも可能だ。AI システムは悪意ある改ざんや乗っ取りに極めて脆弱性である。例えば攻撃者は共通データソースにプロンプト・インジェクション攻撃を仕掛け、無害なプログラミングボットにコード内で権限昇格経路を生成させたり、取引ボットに特定株を吊り上げさせたりできる。AI のさらなる統合は、無自覚な代理攻撃の機会を増大させるだろう。

8.38 攻撃と防御は類似し、互いに情報を提供する

攻撃と防御に必要なツールや技術は多くが共通している（Buchanan, 2017）。 例えば脆弱性の発見は攻撃・防御双方にとって重要だ。侵入は攻撃の主要目的であり、防御では頻繁なテスト手法となる。特に防御側が攻撃者のネ

ットワークに侵入し攻撃能力を妨害する「一歩踏み込んだ防衛」を行うにつれ、サイバー防御は攻撃とますます類似してきた。は攻撃と防御の類似性が攻防均衡の前提を無効化すると主張する（Valeriano, 2022）。

攻撃と防御の類似性を過大評価するのは容易だ。脆弱性を除去するためにコードを書き換えたり、メモリー安全な言語へ変換したりすることは攻撃行為ではない。二要素認証の導入や、ファイル・フォルダ・アカウントへのアクセス権制限も同様だ。一方で、マスターブートレコードの消去、暗号化メカニズムの弱体化、否認可能なインフラの開発、コードへの脆弱性導入は防御行為ではない。

攻撃と防御の類似点は、その行為が人間か AI かによって大きく左右されない。ゆえに AI は攻撃と防御の類似性に大きな変化をもたらさない。変化する可能性があるのは、それらの行動の普及度と重要性だ。 脆弱性発見、侵入テスト、一歩踏み込んだ防衛がサイバーセキュリティの中核となるならば、攻撃と防御はより類似して見えるだろう。一方、防御側がコード基盤や設定を強化しようとするのに対し、攻撃側が脆弱性や脆弱な設定を導入しようとする競争が中心となるならば、攻撃と防御の類似性は低下するだろう。

8.39 概念の混乱と明確な定義の欠如

人々はしばしば、サイバー空間の構成要素やその内部活動について理解し、伝達するのに苦労する。例えば専門家は「効果」の定義を巡って議論する。別の例として、正当な活動と不正な活動の区別は、同じ考えを持つ学者や政策立案者の間でも不明確だ。対立する者同士では、その差は劇的である。

AI は複雑なシステムを検査し、理解しやすい方法やユーザーにとって馴染み深い用語で説明できる可能性がある。これは現状の AI システムにおける強みだ。正確で合意された用語が欠如していても、AI が用語使用時の意図を理解し、別の読者向けに適切な用語へ変換できれば問題は軽減される。

ただし、一部の混乱は不可避か意図的なものかもしれない。「効果」を具体的に構成する要素が不明確であることは、攻撃者に「効果」を引き起こしたと否定しつつ作戦を遂行する柔軟性を与える。不正確さは被害者にも攻撃が不正だったと主張する余地を与える。AI が不正確さを減らす手段のプロバイダであったとしても、その利点ゆえに不正確さは残存する可能性がある。

AI は混乱を増大させる形でソフトウェアやライブラリ、ネットワークを構築するためにも利用されうる。AI は人間にとって解釈が困難なことで有名だが、その生成物も同様である。特に、人間よりも機械にとって便利な形式や言語で記述されている場合だ。インターネットとその構成要素も、人間の意図を超えた速度や方向で進化すれば、人間にとってより不透明になる可能性がある。

8.40 不十分かつ競合する認可

サイバーセキュリティの官僚主義は重大な問題となり得る。攻撃者にとって、官僚主義はサイバー攻撃を実行する認可を持つ組織の数や種類、認可される攻撃の種類、許容される攻撃の範囲を制限する。 防御側では、プロセスを強制終了したり、ユーザーを削除したり、サービスを停止したりといった防御行動の多くを実行する前に、人間でさえ許可を申請する必要がある。その他の防御 権限には、どの政府機関がどのシステムを担当するか、防御企業が守るシステムに対してどれほどの可視性が認められるか、チームがどこまで一歩踏み込んだ防衛を許されるかといった点も含まれる。

AI システムへの権限委譲は、AI の技術的能力に関わらず課題であり続ける。組織内で許可されるソフトウェアの選定や重要システムのエアギャップ化といった判断は、人間が AI に委ねたくない重大な経営判断となり得る。より戦術的な観点では、信頼性の懸念や法的責任の所在などから、防御・攻撃いずれの文脈においても、AI エージェントが能力の全てを行使する完全な自律性を与えられない場合がある。 人間の関与がなくても、AI システムは他の防御・攻撃エージェントと協議し、自らの行動がそれらに与える影響を評価する必要があるだろう。そして人間は、超人的な AI エージェントに対しても、官僚的な統制や指導権を保持したいと考える可能性が高い。

利害関係者の間で対立する利益は根本的に解決不能であり、全ての利害関係者を満足させる権限構造は存在しないかもしれない。ある利害関係者は性能の最大化を優先する一方、別の利害関係者は性能を犠牲にしてでも

安全性や信頼性を求めるだろう。競合し不十分な権限構造は、AI にどれほどの知能があろうとも解決できない現実かもしれない。

8.41 困難な指揮統制

指揮統制は攻撃者にとって困難な場合がある。不正なオペレーターが指揮官の監視外で行動する可能性があるからだ（Libicki, 2009）。また技術的にも困難である。サイバー攻撃は通常、防御側のネットワーク内で発生するため、攻撃エージェントやインプラントとのコミュニケーションが多すぎると防御側に警戒される恐れがある。そのコミュニケーションは攻撃者の身元やデジタル位置を保護するため、難読化ネットワークを経由する必要もある。

信頼性が高く自律的な AI エージェントを活用すれば、遠隔からの指揮統制の必要性を低減できる。攻撃エージェントは戦術的判断、場合によっては作戦的・戦略的判断すら自律的に下せるようになる。 標的システム上、あるいはボットネットの侵害ホスト上に存在する自律エージェントは、ニューラルネットワークの規模、メモリ、演算能力といったリソースが著しく制限される可能性がある。こうした制約により、標的に近い高性能ハードウェア上でより高度なモデルを運用できる防御エージェントと比べて、十分な性能を発揮できない恐れがある。

たとえそれらの指揮統制エージェントが十分なリソースを持ち、標的ネットワークではなく大規模な遠隔サーバーにホストされていたとしても、性能が低ければリスクを増大させる可能性がある。指揮統制を AI システムに委任するのは、効果性よりも速度や規模の理由による場合がある。それは信頼性の低い決定や、より多くの誤りや誤算につながる可能性がある。

8.42 極秘扱い

サイバーツールや技術、特に攻撃的なものは、しばしば高度に分類される。この分類が、各国の能力格差を維持している。全ての国や組織が、実装の専門知識はともかく、同様の防御技術を持っている一方で、脅威アクターが利用できる攻撃技術には大きな差がある。

最も高性能な汎用 AI システムが、分類可能な政府によって管理される可能性は低い。むしろ、機密指定された攻撃的ツール・技術・データセットこそが、汎用 AI システムを攻撃的サイバーセキュリティに応用する上で重要となるだろう。その場合、分類による脅威アクターの階層化は継続する可能性がある。

開発者が最高性能の AI へのアクセスを制御する取り組みには、分類との類似点がある。データやツール自体は技術的に機密指定されていないが、知的財産保護、輸出管理、あるいは単に技術が高価な場合の経済格差を通じて、アクセスが制限される可能性がある。アクセス制御は脅威アクター間の階層化を生み出すだけでなく、様々な顧客やユーザーに強弱の異なる防御ツールをプロバイダすることで、防御能力の階層化も招きかねない。

AI は機密防御を複雑化する可能性もある。機密ネットワークへのデータ・ソフトウェアの持ち込みと持ち出しには保護策が存在するが、高機密から低機密への転送を防ぐ保護策は、低機密から高機密への転送を防ぐものより多い（Arnold, 2016）。機密ネットワークへの AI エージェント導入は困難だが、自律型攻撃エージェントは防御側にネットワーク内での情報流出や改ざん防止への注力を強いる可能性がある。信頼性が高く独立したエージェントは、指揮統制なしに機密ネットワークを移動し、データの流出や改ざんをより効果的に実行できるかもしれない。しかし、こうした攻撃エージェントは効果を発揮するために膨大な計算資源を必要とし、機密ネットワークの防御者から隠蔽するのが困難かもしれない。

8.43 戦術的関与は分析の基本単位である

サイバーセキュリティの成否を戦術レベルで検討する方が分析者にとって容易である。個々のハッキングの影響は、しばしば単独で計算・評価される。しかし時間の経過とともに、この傾向は弱まりつつある。単一の脅威アクターと防御者が継

続的に相互作用する持続的関与においても、この傾向は弱まる可能性がある。さらに、サイバー攻撃が広範な軍事作戦や影響力行使キャンペーンの支援手段として使用される場合、この傾向はより顕著に弱まる。

戦術的意思決定が AI システムに委任される場合、人間はそれらのシステムに非戦術的な指針を提供する必要がある。その場合、人間のオペレーターは指針プロバイダに注力せざるを得なくなる。その結果、アナリストも指針のアセスメントに重点を移し、サイバーオペレーターと監視者の双方において、戦略的・長期的な影響への関心がより重要視されるようになる可能性がある。

作戦上あるいは戦略上の権限を委譲しなくとも、AI は小規模攻撃の頻度と規模を増加させ、個々の攻撃は注目に値しなくなるが戦略的影響は重大となる水準に達しうる。分析は、日常的なハッキングに追いつけず、あるいは関心を失い、より大きな枠組みでの影響、目的、傾向へと移行する可能性がある。この行動は、log4shell 脆弱性（TrendMicro, 2021）のような広範に分布した脆弱性において発生する。 さらに、ランサムウェア攻撃は個々の影響は小さいものの、その膨大な数が蓄積されることで国家・地球規模の影響を生み出し、戦術的関与から戦略的考察への移行を促している。

8.44 サイバー空間内では水平的・垂直的に紛争がエスカレートするが、まだ外部には波及していない

サイバー攻撃への対応は、主に抑制的であり、報復的なサイバー攻撃に限定されている。逮捕や制裁などの例外はある（Rusinova & Martynova, 2024）が、物理的・軍事的対応へのエスカレーションは限定的だ。

サイバー空間外へのエスカレーションは、単独または累積的により破壊的な効果を目的としたサイバー攻撃が増加した場合に起こり得る。これは AI の関与拡大による可能性のある結果だが、必然ではない。深刻化よりも、AI が攻撃件数を増加させ、国家や標的がサイバー空間外で増加した攻撃量に対応する可能性が高い。 しかし、AI がサイバー攻撃の容易さを増すならば、それはサイバー攻撃を報復手段としてさらに魅力的な選択肢とし、エスカレーションがサイバー領域内に留まる可能性を高めるだろう。

AI はまた、サイバー攻撃と非サイバー対応を分けるエスカレーション閾値を誤って越えるリスクを高める可能性がある。これは、AI がサイバー攻撃の数を増やすことで、1 つまたは数件の攻撃が閾値を越える確率が高まることによって生じうる。また、より高速または拡張性が高いという理由で、能力の低い AI エージェントに意思決定権限を委譲した結果としても起こりうる。

8.45 サイバー紛争はエスカレーション、誤算、不安定化を招く恐れがある

サイバー攻撃は絶えず発生している。各攻撃は標的、手法、状況において独自性を持つ。同じ標的を二度攻撃する場合でも、被害者が単発攻撃は許容できるが繰り返しの攻撃は容認できないと判断する可能性があるため、二度目は異なる性質を持つ。

攻撃の量を増やすことは、不適切な AI システムに認可を委譲することと同様に、エスカレーションや誤算のリスクを高める。AI は人間の心理を理解するのが苦手であり、自らの行動に対する人間の反応を確実に予測するのに苦労する可能性がある。将来、AI が人間を理解する能力を改善すれば、特に被害者とその文化や信念を代表する訓練データにアクセスできる場合、この傾向は弱まるかもしれない。

AI システムがエスカレーション経路の提案、あるいは攻撃者や報復者の選択肢決定を支援する場合、設計者はサイバー攻撃からのエスカレーション傾向や抑制性を決定できる。意思決定者はまた、行動実行前に AI システムから推奨対応策を得て、その提案が敵対者の AI が提供するものと類似すると予測できる。 これにより、エスカレーションを招く誤解や誤算を減らせる。敵対者はエスカレーション手前で線引きを維持できるが、モデルには依然としてランダム性が含まれるため、反応を完全に予測することは不可能だ。この線引き能力は危険を伴う。

しかし現実的には、AI モデルの操作可能性や、意図的な破壊行為下での信頼性の低さから、誤解を他のリスクより恐れる主体間であっても、AI に過度の意思決定権限を委ねることは賢明ではない。戦略的意思決定には専門家集団が存在する以上、AI が戦略決定に大きく貢献するには極めて高度な能力が求められる。

8.46 インターネットには共通モード障害がある

インターネット上の技術の多くは共有されている。企業は自社製品やサービスを他の多くの企業や組織に販売している。さらに、デバイスが効果的に相互運用するためには、同じコミュニケーションプロトコルや手順が必要だ。そしてオープンソースのコードベースやライブラリは、システムやソフトウェア全体に統合されており、その程度は観察するのが難しい。

AI は技術や手法をさらに集中化させ、共通モード障害を増大させる可能性がある。同一または類似の AI システムがほとんどのソフトウェアを記述する場合、異なる組織で開発されたソフトウェアさえも類似性を増すだろう。特定のコーディング手法や嗜好に起因する脆弱性は、その手法や嗜好を採用した作者がいる場所ならどこでも存在する。インターネット上のコードの大部分を、1 つまたは少数の異なる AI 作者が記述するようになれば、それらの手法や嗜好はより広く共有されることになる。

同様のリスクは、防御サービスを提供する AI システムが一つか数個に集中した場合にも生じうる。AI 防御システムが依存する技術や、そのシステムが好む設定に弱点があれば、多くの標的が同時に危険に晒される可能性がある。

AI がより多くのコンポーネントを生成し、人間の開発者が扱わない言語で設計されたものなど、より多様なリソースを活用するようになれば、サイバー環境は相互接続性と相互依存性を増す可能性がある。

AI が共通モード障害を増大させるこれらの各経路は、サイバーセキュリティが既に共有コードと防御サービスプロバイダの寡占に大きく依存している文脈で考慮すべきだ。主要な OS 開発者は数社、多くの企業は同じ少数のクラウドサービスにホストされ、密接に連携する主要なアンチウイルスプロバイダも数社に過ぎない。AI モデルやシステムに活動に多様性を提供するよう単純に指示するだけで、共通モード障害を悪化させるどころか軽減できる可能性がある。

8.47 多数のデバイスが攻撃者に有利に働く

多数のデバイスを保有することは、攻撃者が攻撃規模を拡大することを可能にする。特に DDoS 攻撃のように直接的なボリューム依存型攻撃において顕著だ。また、防御側が攻撃源を特定しようとする際のステルス性や、防御側が攻撃デバイスを無効化しても効果を維持する点でも規模は利点となり得る。ただしこの前提は誇張されがちだ。特に破壊的な攻撃の多くはノート PC1 台で実行可能であり、防御側に多数のデバイスが存在すれば攻撃対象領域が増大するからだ。

現在、AI は計算処理に多くのデバイスを必要とするリソース集約型である。効率が急速に向上し続ければ、将来はそうではなくなるかもしれない。しかし、AI システムの運用がリソース集約型であり続けるなら、高性能 AI システムを稼働させる可用デバイス数は、AI エージェントやシステムの数量・速度・能力を大きく左右する要因となり得る。攻撃者と防御者の競争は、攻撃・防御キャンペーンを遂行するためのデバイス数によって左右される可能性がある。

攻撃者にとって、AI の規模を活用するには、複数の攻撃エージェントを運用するため、さらに広範に分散したデバイスが必要となるだろう。単一のデバイスから過剰な攻撃が発生すれば、防御側は容易に警戒態勢に入り、それらの接続を遮断したりデバイスを妨害したりする措置を講じられる。

8.48 持続的に活動する者のサイバー空間上の優位性

攻撃者は散発的な活動よりも持続的な活動からより多くの利益を得るとする見方がある。これはある程度、奇襲が重要であるという前提と矛盾する。 その論拠は「国家がサイバー戦略環境において、またそれを通じて行う主要な［既成事実化］と二次的な［直接的サイバー関与］行動は…持続するという構造的必然性と、武力攻撃に相当しないサイバー搾取による利益追求という構造的に導かれた戦略的インセンティブの結果である」（Fischerkeller, Goldman & Harknett, 2022）というものだ。 言い換えれば、サイバー攻撃者は戦争の閾値を下回る「千の紙切れによる死」を引き起こすことを好む（Harknett, 2024）。さらに侵入者は、敵対者と持続的に関与することでより多くの情報を収集でき、初期アクセスや偵察といった攻撃の一部を既に実施していれば、突発的な行動を取る準備が整っている可能性が高い。

AI は情報収集の価値を低下させないため、国家レベルの攻撃者は依然として持続的関与を望むだろう。エンドツーエンド攻撃の実施期間を AI が短縮する場合、将来の標的攻撃オプションを維持することを目的とする攻撃者にとって、持続的関与の必要性は低下する可能性がある。しかし犯罪者や一部の国家・テロリストは持続性をほとんど必要とせず、可能な限り早期に効果を発生させることを好む。AI がこうした動機を変える可能性は低い。

攻撃者が脆弱性や新規手法を入手できるが量が限られている場合、持続性を維持・再確立するためにそれらを使い果たすのは賢明ではない。攻撃者は保持したい能力を消耗するリスクを避けるため、持続的関与を好まない可能性がある。 しかしより可能性が高いのは、脆弱性や手法が限られている場合、攻撃者は防御側が独自にそれらを発見するのを懸念するかもしれないということだ。攻撃者はまた、他の攻撃者が先に新たな脆弱性や手法を発見し、使用して消耗させることを恐れるかもしれない。したがって、能力が消耗される前に使用しようとする動機が、持続的な関与をさらに促す可能性がある。

AI を活用した攻撃と防御が迅速に行われ、防御側の成功が比較的長期間持続する場合、攻撃者にとって持続的な関与は魅力が薄れる可能性がある。例えば、防御側の成功が攻撃者のインフラ破壊を伴う場合、そのインフラの再構築には多大な費用や時間を要する。その場合、攻撃と防御のサイクルが加速することは攻撃者にとって苛立たしく、あるいはコスト高となり、攻撃者がアクセス可能な時間の割合が全体的に減少する結果を招くかもしれない。 一方で、頻繁ではあるが持続的でないアクセスも、依然として持続的関与である。

9. AI がサイバーに与える影響

本節では、第 7 節でサイバー空間の攻防優位性に関する議論への AI の影響を評価した際、また第 8 節でサイバー紛争の性質への AI の影響を評価した際に提起された様々な論理展開をまとめる。これらは AI が将来のサイバー空間に与える可能性のある影響に関する論理である。

これらの論理の多くは複数の議論で繰り返し登場したり類似していたが、本節では各論理を 1 回のみ掲載する。

AI がサイバーに影響を与えうる全手法を収集した後、類似性に基づいて分類を試みた。ここではそれらを以下のグループに分類して提示する：デジタルエコシステムの変容、デジタル環境の強化、デジタル交戦の戦術的側面、インセンティブと機会、紛争・危機への戦略的影響。

9.1 デジタルエコシステムの変化

攻撃者と防御者の双方にツールや能力を提供するだけでなく、AI はより多くのシステムやサービスに統合され、開発者による利用が増えるにつれて、すでにデジタル環境を変えつつある（JetBrains, 2024）。このセクションでは、AI が攻撃対象や防御対象、そして攻撃と防御が行われる場所を今後も変え続ける可能性について概説する。

デジタルエコシステム	
コーディング、設計、サービスへの影響	
	防御対象となるコードとサービスの増加
	コード、サービス、アーキテクチャの複雑化を促進する
	複雑なシステムを理解しやすくする
	体系的な脆弱性を増減させる
	システム的な脆弱性を増減させる
AI コンポーネントの統合	
	脆弱性のある AI コンポーネントを組み込む
	人間の弱点を排除する
	手動制御やバックアップを排除し、人間のスキルを低下させる
	AI モデルという形で新たな標的を生成する
ネットワークとインターネットの構造	
	ネットワークをハブ・アンド・スポーク構造へ再編成する
	より多くの分離とエアギャップを可能にする

AI がコーディング、設計、防衛に与える影響

第一の帰結として、AI がソフトウェア開発を加速させるにつれ、防御側は保護すべきコードやサービスが増えることになる。デジタル要素の増加に加え、 AI によって生成されたコードやサービスはより複雑化したり不透明になったりする可能性がある。AI は人間にとって不便な手法や言語でコンポーネントを設計するかもしれない。一方で、AI は複雑なコードベースやシステム・サービスを分析・説明し、人間が迅速に理解する手助けとなる点で有用である。

AI システムがサービスやコードをプロバイダとして提供・適応させる速度は、使いやすさを維持する必要性によって制限される可能性がある。特に AI の利用がより複雑で相互に関連するデジタルエコシステムを生み出す場合、変化が速すぎると、その価値以上に問題を引き起こす恐れがある。

AI はデジタルエコシステム全体で共有・反復される脆弱性の程度にも影響する可能性がある。広範なシステムに影響する脆弱性や体系的に再発する脆弱性は、既に今日、多くの支配的なサービス・製品プロバイダ、広範なコード共有やライブラリ利用、アンチウイルスや監視・スキャンといった共通防御サービスによって蔓延している。 少数の支配的な AI 製品、あるいは同じ訓練データを共有する類似の AI 製品は、こうした体系的・システム的なリスクを増大させる可能性がある。しかし逆に、AI はそれらを軽減するのに適しているかもしれない。単一の AI 開発者や防御者は、ランダム

性を組み込んだり、アプローチを変化させたりするよう指示できる。出力の変動性は、AI 開発者が容易に調整できるパラメータである。

AI コンポーネントの統合

AI モデルやエージェントをより多くのシステムやサービスに組み込む動きは、デジタルエコシステムにも大きな変化をもたらす。サイバー防衛用 AI を含む AI システムは、欺瞞や操作に対して脆弱性を持つ可能性がある。防御側は、より脆弱なコンポーネントで構築されたシステムを管理しつつ、それ自体が脆弱性を持つ可能性のある AI 防御ツールを使用する必要に迫られる。標的となるシステムや組織の人間要素が、デジタル攻撃可能な AI エージェントに置き換わるにつれ、防御上の課題は増大する。 人間はサイバーセキュリティにおける共通の弱点となり得るが、デジタルで制御するのは難しい。

AI の統合は攻撃へのレジリエンスも低下させる可能性がある。時間の経過とともに、AI システムは故障やサイバー攻撃時に重要な手動制御を置き換えるかもしれない。例えば、ハンドルのない自動運転車やコンソールのない産業制御システムは、AI オペレーターが侵害された場合、人間が操作できない。インターフェースが残っていても、人間はそれらを操作するスキルを失うかもしれない（H.-P. Lee et al., 2025）。

AI は新たな標的を生み出す可能性もある。AI の目的は膨大な情報や能力をコンパクトなデジタル形式に集約することだ。大量の機密報告を吸収した AI モデルは、個々の報告が分散している状態よりもはるかに容易に盗難や漏洩の危険に晒される。また所有者の発言・行動・嗜好を追跡する AI モデルは、金銭的価値が高いか、あるいは致命的なサイバー攻撃の標的となり得る。

ネットワークとインターネットの構造

AI はインターネットのアーキテクチャも変える可能性がある。少数の AI 開発者が市場を支配し、AI が情報への主要経路となれば、インターネットは現在の「スケールフリー」構造から「ハブ・アンド・スポーク」モデルへと再構築されるかもしれない。 現在、情報は様々なユーザーと情報プロバイダの間で流れる。しかし別の未来では、情報は主に情報生成者から AI 企業へと流れるようになる。ユーザーは AI 企業にアクセスを求め、その企業がユーザーに代わって情報プロバイダとやり取りするモデルを利用することになる。 ハブ・アンド・スポーク型のインターネットは単一障害点のリスクが高まるが、防御側は重要なノードに集中して対策を講じられる。

少数のコア AI プロバイダへの中央集権化がなくても、AI の普及はインターネットの再構築をもたらす可能性がある。AI システムが機能実行に必要な要素を多く内包する場合、ネットワーク接続の必要性は低減し、場合によっては不要になる。これによりエアギャップシステムの増加が予想される。Apple と OpenAI の提携はこの手法を予兆しているかもしれない。 iPhone に組み込まれた AI は、可能な限りネットワーク接続なしで処理を行い、必要な場合にのみ OpenAI にアクセスする。OpenAI はその後、必要に応じて他の情報プロバイダにアクセスし、iPhone ユーザーに情報を返す。この隔離性の向上により、AI 搭載デバイスの防御は容易になる。ただし、AI が現在の人間よりも多様な情報源から情報を収集するために使用される場合は例外だ。その場合、逆の結果になる可能性がある。

攻撃者は分散型ネットワークやインターネット上の否認可能なインフラに依存して攻撃を仕掛けるようになるかもしれない。攻撃規模が拡大すれば、防御策による遮断や妨害を回避するため、より多くの攻撃発信源が必要となる。攻撃インフラの構築・保守は、攻撃の他の側面ほど加速しない場合、あるいは攻撃増加により防御側がインフラの追跡・妨害を容易に行えるようになった場合、ボトルネックとなる可能性がある。

9.2 デジタル環境の強化

サイバー防御の大半は攻撃発生のずっと前に行われる。デジタルシステムの設計や機能において防御は実現される。製品リリース前のコーディングやテスト段階でも行われる。さらに、アクセス管理、多要素認証、ファイアウォール規則などの決定を含むリスクアセスメントや構成管理の過程でも実施される。

デジタル環境の強化

コードベース	
	より安全なコードを書く
	安全でないコードをリファクタリングする
設定	
	標準的な防御策とベストプラクティスを実装する
更新とパッチ	
	脆弱性を発見する
	パッチと更新のテストと評価
	パッチをより迅速に生成する
	修正が難しい脆弱性をより多く悪用する
リスクの受諾または拒否	
	デジタル在庫を評価する
	リスクシナリオをブレインストーミングする
	レッドチームによるセキュリティ態勢やリスクの評価

コードベース

あらゆる業界において、サイバーセキュリティ専門家の数が大幅に不足している。また、重要インフラプロバイダを含む小規模組織のサイバー防衛予算も不足している。 AI が防御担当者を支援したり、防御タスクを自律的に実行する能力は、人材不足によるセキュリティホールの解消に寄与し得る。この強化策は、高度なスキルを持つチームを既に擁し、製品の開発・テスト・更新を安全に行っている大手プロバイダよりも、小規模で知名度の低い組織や貢献者にとって有益だろう。こうした熟練チームは全ての欠陥に対処できるわけではないが、優先度の高いものは迅速に修正できる。

AI はまた、小規模な貢献者によるコード、特にオープンソースコードのセキュリティ強化においてより大きな利益をもたらすプロバイダとなる可能性がある。こうしたコードはそれ自体が重要であるか、重要な製品に組み込まれている。コードベースの規模と作業の単調さを除けば、そのコードの評価および保護は特に難しい作業ではない。比較的控えめな進歩で、AI は膨大なコードベースを強化できるかもしれない。これは人間には時間も興味もない作業だ。

設定

コードベースを超えて、組織の防御体制を構築することは膨大な作業となる。国立標準技術研究所の情報システム保護ガイドライン 800-53 だけでも 500 ページに及ぶ。規定されたタスクの実行は必ずしも困難ではないが、特に小規模組織にとっては大きな負担だ。 こうしたベストプラクティスの加速化、あるいは自律的な実施により、重大な侵害や影響を招いた多くのセキュリティホールを塞ぐことができる。AI システムはより多くの防御ツールを実装・監視し、攻撃者が直面する防御の深さを増すことも可能だろう。

更新とパッチ

AI は新たな脆弱性の発見速度を加速させ（Google，2024）、パッチ開発時間を短縮する可能性が高い。しかし歴史的に見て、更新やパッチの開発時間は主要な防御課題ではなかった。むしろパッチや更新を実装する方がはるかに長い時間を要するケースが多い。これらの更新はユーザー体験を妨げたり、デバイスの再起動などのダウンタイムを必要としたり、更新に不具合があれば広範囲なサービス停止を引き起こす可能性がある。 AI が防御に影響を与える可能性は、更新そのものの開発を加速させるよりも、更新のテストと評価プロセスを加速させることにある。新しいテストツールがなくても、AI はシステム管理者が、複雑な更新が自組織やシステムの複雑さとどう相互作用するかを迅速に評価する助けとなるかもしれない。

更新実装における課題は信頼性だけではない。更新頻度が高すぎると、使い勝手が悪くなったり稼働時間に影響が出たりする。AI がサービスを頻繁に更新すると、そのサービスに依存する他のシステムが不安定になったり不便になったりする可能性がある。

また、パッチの作成が遅れるケースもある。例えばハードウェアの脆弱性には、必ずしも容易なパッチが存在しない。仮にソフトウェアパッチで対応可能でも、パフォーマンスや使い勝手の低下を伴う場合がある。システム管理者はこうした犠牲を受け入れるか拒否するか判断せねばならず、パフォーマンスと引き換えに一定の脆弱性を選択する可能性もある。

リスクの受容または拒否

おそらく、どんな戦術的防御よりも重要なのは、防御側が受け入れ可能なリスクのレベルを選択することだ。防御が提供するセキュリティレベルに不満がある場合、システムを切断したり、サービス・機能・提供を停止したりする選択ができる。防御側にとっての大きな課題は、そうした判断を行うためにリスクを正確に評価することにある。

AI は複数の方法でリスク評価を支援できる。複雑な組織が自社の運用、サプライチェーン、シャドーIT を含むデジタル資産を理解する手助けとなる。 AI はリスクシナリオのブレインストーミングを支援し、想定されるリスクシナリオをより網羅的にカバーできるようにする。さらに、攻撃的サイバーエージェントの普及により、防御側はより優れたレッドチーム評価を実施できるようになる。将来的に攻撃的エージェントが広く利用可能で高性能になったシナリオでは、防御側はそれらのツールを自ら活用してリスクを正確に評価できる。製品やサービスを提供する前に評価を実施し、リスクを許容範囲に合わせて調整することが可能となる。

9.3 デジタルエンゲージメントの戦術的側面

デジタルエンゲージメントは、脆弱なサービスへの単一クエリで情報が漏洩する SQL インジェクション攻撃のようなケースではほぼ瞬時に行われる。より一般的なのは、攻撃側と防御側の戦術的行動と対応が長期化するプロセスだ。持続的エンゲージメントの場合、それは継続的なプロセスとなる。防御側は複数の同時攻撃者、あるいは分散型サービス拒否攻撃時のように単一攻撃者の複数インスタンスに直面する可能性がある。

攻撃の規模

自律型攻撃はデジタル戦闘主体が存在して以来ほぼ常に存在してきたため、AI によって攻撃ボットの規模が増大するとは限らない（Orman, 2003）。しかし高度な AI は、これらの自律型攻撃者がより独立して行動することを意味し、単一のソフトウェアの複製であるワームや従来型ボットネットよりも個別化された防御対応を必要とする可能性がある。より信頼性が高く自律的な攻撃は参入障壁をさらに低下させ、新たな脅威アクターを許容するだろう。

デジタル戦闘の戦術的側面	
規模の拡大	
	攻撃数の増加
	防御手段の数を増やす
	より多くのアラートを検査する
速度の向上	
	攻撃戦術を加速する
	防御戦術を加速させる
リソースの利用可能性	
	攻撃能力は防御能力より低く設定する
委任	
	指揮統制を縮小する
	エラーや事故の発生確率を増減させる

サイバー攻撃を実施する。例えば、テロリストは高度な AI がプロバイダできる専門知識へのアクセスが限られているため、これまで少数のサイバー作戦しか実施していない可能性がある。

同時に、AI の進歩は防御側に専門知識を提供する。比較的控えめなレベルの AI 進歩でも、国内外のサイバー防衛要員不足の解消に役立つ可能性がある。また、より自律的に動作する防御システムを実装できるかもしれない。

AI が警報の調査を支援できれば、それらの防御システムはより積極的に作動または警報を発するよう設定でき、誤検知の調査に追われる防衛要員の負担を軽減できる。

スピード

AI の利点としてよく挙げられるのがその速度だが、サイバー活動の一切を AI で加速できるわけではなく、加速できる範囲にも限界がある。AI モデルは計算負荷が高く、既に高速な他のデジタルプロセスと比べて遅く、さらなる加速が難しい場合が多い。例えばデータ転送、ディレクトリスキャン、ハードドライブ暗号化などは、AI によるさらなる加速が容易ではない。

帯域幅や回路設計、情報理論などに基づく物理的な限界により、こうした処理の速度向上には限界がある。防御側はさらに特定のプロセスを遅延させる選択も可能だ。例えばパスワードロックアウト後の手動承認要求や、重要ファイルの暗号化・削除前の承認手続きなどがそれにあたる。攻撃側が処理を加速できる場合でも、悪意ある活動と検知されるのを避けるため、あえて加速しない選択をするケースすら存在する。

ただしデジタル攻撃の一部要素は加速可能だ。バイナリの逆コンパイルや脆弱性発見など、現在遅いタスクが存在する。遅いタスクが作戦のボトルネックなら、それらを加速することで全体的な大幅なスピードアップが得られる。

デジタル攻撃の一部が AI によって加速される場合、加速が困難な他の側面がより重要になる可能性がある。サイバー空間は遠隔地へのアクセスを可能にするが、光速でさえ遅すぎて、特に標的から地理的に離れた攻撃者にとっては意味のある遅延を引き起こす。攻撃者はまた、防御側が攻撃中に破壊または遮断する可能性のあるインフラや否認可能なアカウントを再構築する時間も必要とするかもしれない。

資源の利用可能性

おそらく、戦闘中の地理的遅延よりも重要なのは、防御側がより高性能な AI システムやエージェントを利用できる場合が多いという点だ。 防御側は AI 防御システムに任意の資源を割り当てられ、AI エージェントとの間で必要な情報を自由にやり取りできる。一方、攻撃側は被害システム上で防御側に警戒を招かないよう、限られた資源の使用に制約されることが多い。あるいは遠隔サーバーとの間で情報をやり取りする選択肢もあるが、その場合でもデータ転送で防御側に警戒を招かないよう注意が必要だ。

とはいえ、防御が攻撃より容易になるとは限らない。交戦時の両者の任務が根本的に異なるからだ。パスワードの解読と生成は別物であり、コードへの悪意ある改変と改変の悪意識別も異なる。AI が特定の任務を支援する一方で他の任務を進歩させないため、デジタル交戦における好ましい戦術は変化するだろう。

委任

戦術的、あるいは作戦レベルの決定は、指揮統制の最小化や速度向上といった様々な利点、あるいは単に人間が個別に管理できるエージェント数を超過しているという理由から、AI サイバーエージェントに委任される可能性がある。攻撃エージェントが指揮統制なしに動作する能力は、機密ネットワークやエアギャップシステムといった堅牢な標的に対してより有用となる可能性がある。

信頼性以外の理由で権限が委譲された場合、エージェントは適切な判断を下せず、ミスが増える可能性がある。これは攻撃側・防御側双方に当てはまるが、攻撃組織はリスク許容度が高い傾向にある。

攻撃側では、こうしたミスが意図せぬエスカレーションや巻き添え被害となる可能性がある。防御側では、エージェントがプロセスを強制終了させたり、システムを切断したり、エンドポイントを消去したりする恐れがある。これらは、信頼性の高いサービスを提供しようとするユーザーや組織にとって、非常に大きな混乱を招く恐れがある。AI を活用した防御側も操作に脆弱性がある場合、このリスクは特に深刻だ。攻撃者は、防御側エージェントに混乱や損害を引き起こさせることができれば、自ら効果を発動させる必要がなくなるかもしれない。

9.4 インセンティブと機会

攻撃的サイバー活動は、非対称的でアクセスしやすい攻撃形態と見なされることが多い。他の領域に比べて投資、人口、産業基盤が少なくて済むからだ。しかしサイバー脅威環境は依然として大きく分化している。スクリプトキディ、テロリスト、ハクティビストは国家よりもはるかに能力が限定されている。国家間や国家内部でも、様々なチームの能力には大きな差がある。

インセンティブと機会	
攻撃者にとっての入手可能性	
	志願する攻撃者の参入障壁を下げる
帰属	
	膨大で多様な情報源を分析する
	攻撃における人的関与を減らす
	偽旗作戦を実行する

攻撃者への入手可能性

AI が普及し攻撃能力を高めれば高めるほど、下位レベルの脅威アクターもより強力になる。攻撃ツールやスクリプトは既に個人、テロリスト、犯罪者に広く入手可能だ。AI はこれらのツールを次第に強力にし、 より使いやすくするだろう（Singer et al., 2025）。これらのツールと、それらが可能にする攻撃者がより効果的で脅威的になるかどうかは、防御側が攻撃者の進歩以上に強化されるかどうかにかかっている。

防御用 AI エージェントの拡張性と入手可能性により、防御側は既知の脅威に対する防御を強化できるかもしれない。 その場合、既知の脆弱性を悪用する既知の攻撃ツールが広く入手可能であることは、たとえそれらのツールが AI 対応であっても、将来的に問題が小さくなる可能性がある。もし AI 対応攻撃ツールが新たな脆弱性や手法を発見するのに有用であれば、防御側は苦戦するかもしれない。そのような能力は、今日では最も先進的な国家レベルの脅威アクターだけが持つものに相当するが、おそらくより大規模になるだろう。もしそれが広まれば、最も能力の高い脅威アクターと最も低い脅威アクター間の格差は縮小する。

ただし脅威アクター間の格差が依然として大きいシナリオも存在する。例えば、高性能な攻撃エージェントが少数のプロバイダからしか入手できない場合、それらのプロバイダはアクセスを制限できる可能性がある。たとえ競合国が対抗製品を提供していても、テロリストや特定の犯罪組織への使用を禁止することで合意するかもしれない。

別の例として、攻撃的エージェントには機密扱いのツールやデータセット、専門知識が必要となる場合がある。その場合、攻撃的 AI エージェントは能力を持ちながらも広く普及しない可能性がある。こうしたツールは制限付きで展開される必要があるだろう。もし保護されていないネットワークや標的環境に展開された場合、敵対者や様々な脅威アクターがそれらのツールやその一部を複製・流用する可能性が高い。

帰属

一部のサイバー攻撃では、攻撃者は身元を特定されることを気にせず、むしろ攻撃の功績を主張したい場合さえある。しかし多くの場合、攻撃者の抑止や処罰には帰属の特定が重要であり、AI は帰属特定プロセスを加速させてインシデント発生時刻に近づけたり、帰属特定をより困難にしたりする可能性がある。

AI はより多様な情報を統合し、点と点を結びつけたり、特定の容疑者に対する証拠を強化したりできる可能性がある。AI はパターンマッチングに優れており、人間が見逃す攻撃間の類似性を特定できるかもしれない。 しかし AI が攻撃の大部分を担えば、人間の犯行者からの手がかりは減少する。異なる攻撃者が同じ AI ツールを使用すれば、それらの区別は困難になる。さらに AI は偽旗作戦の実行に長けている可能性がある。例えば、既知のハッカー集団の文体でスクリプトを書くよう指示されるかもしれない。

とはいえ、攻撃の帰属判断は AI が影響する技術的側面だけに依存するわけではない。緊張関係や動機といった他の手がかりは残るし、資金の流れや流出データの痕跡も残る可能性がある。また、エージェントが依然として指揮統制下にある限り、コマンドアンドコントロールサーバーの所在地や所有者も依然として手がかりとなり得る。

9.5 紛争と危機への戦略的影響

サイバー攻撃は戦略的な効果に限界がある。確かに広範な経済的損害や混乱、死さえも引き起こしてきたが、他の領域と比較すると、サイバー領域は戦略的影響力が低く、直接達成した戦略目標も少ない。しかしこれは主に、サイバー領域が発展した時期に、主要なライバル国の戦略的関心事がエスカレーション回避にあったためかもしれない。

意思決定プロセス

攻撃的サイバー能力の使用可否や使用方法、あるいは防御態勢の大幅な変更に関する戦略的決定は遅延しがちだ。これらは往々にして多面的な決定であり、達成すべき多様な目標とリスクが同時に存在する。 政府や組織の各部門の責任者は、所属部門のニーズに基づいて互いに異なる視点を持つ可能性があり、それらの視点は複雑で明示されない場合もある。AI はこうした意思決定を加速または改善できるかもしれないが、その効果が限定的であると考えられる理由もある。

紛争と危機への戦略的影響	
意思決定プロセス	
	意思決定の迅速化
	意思決定の質を向上または低下させる
サイバーの戦略的価値	
	戦略的要請に応えるため、作戦や戦術を加速させる
	高リスクシステムの強制的な切断または強化
確実性	
	少なくとも 1 つの攻撃が成功する確実性を提供する
	偶発的なエスカレーションや付随的損害のリスクを高める
意思疎通の誤りと誤算	
	敵対者の行動や反応に関する理解を深めるか、あるいは損なう
	フォレンジック分析において攻撃的意図を露呈する
使わなければ失う	
	サイバー能力や脆弱性の儚さを増大させる
	能力や脆弱性の可用性と代替可能性を高める
	攻撃目標をオンデマンドで達成する能力を増減させる

AI が極めて高性能になっても、人間は意思決定プロセスに関与し続けたい、あるいは関与を求められるだろう。AI はプロセスの大部分を加速できるかもしれないが、他のボトルネックが解消されても人間のボトルネックは残り、プロセス全体の加速度合いを制限する可能性がある。そして、あらゆる知能を駆使しても、戦略的意思決定から対立やトレードオフを排除し、全ての関係者の成果の質を向上させることは不可能かもしれない。

サイバーの戦略的価値

サイバーの戦略的価値が歴史的に低かった理由の一部は、防御側が大きなリスクを受け入れないことにある。サービスプロバイダがサービスを十分に保護できない場合、そのサービスを提供しない選択をするかもしれない。 例えば、インターネット経由で遠隔投票（ ）を行うのは便利で効率的だが、その防御は困難だ。防御側は特定のサービスや機能を提供しない選択肢を持ち続けるだろう。しかし攻撃側がより能力を高めれば、実際のサイバー攻撃がなくても、防御側は既存サービスの提供中止を余儀なくされるかもしれない。このように、攻撃と防御のバランス変化は、実際の攻撃がなくても戦略的影響をもたらし得るのだ。

サイバー攻撃の戦略的価値が歴史的に限定的だった別の説明として、戦術的な速さにもかかわらず作戦レベルでは遅い可能性がある。ミサイル攻撃は数分～数時間で、地上侵攻は数日～数週間で実行できる。一方、サイバー作戦は標的に合わせた攻撃計画の策定に多大な準備とカスタマイズを要する。AI がこれらのプロセスを加速させれば、意思決定者が必要とする時間軸でサイバー攻撃の選択肢が増えるだろう。

確実性

サイバー攻撃は不確実性も極めて高く、戦略的重要性を帯びた作戦において指導者が依存するのは困難だ。攻撃用 AI エージェントが、実行速度の向上や独立エージェント数の増加により、1 日あるいは 1 週間で実行可能な攻撃数を増やせば、少なくとも 1 つの攻撃が成功する確率は高まる。意思決定者は自らの作戦が戦略目標を達成する確信を強められる。

同時に、一定期間における攻撃数の増加は、ミス発生の可能性を高める。意図しないエスカレーションや付随的影響が生じる確率が増加する。より多くの戦略目標を追求する中で、こうした事故は過去に比べてより大きな影響をもたらす可能性がある。

意思疎通の誤りと誤算

意図せぬ結果としてエスカレーションや意図せぬ標的の被害が生じるだけでなく、意図した効果にも予期せぬ結果が伴う可能性がある。AI は人間の反応を理解する能力が特に弱く、不適切な決定を下したり助言したりする恐れがある。これは特に、モデルやエージェントが限定的なデータ、あるいは攻撃者の文化圏のデータのみで訓練されている場合に起こりやすい。

しかし適切に訓練されれば、AI は人間の攻撃者が予測しない被害者のレッドラインや対応策を提案できる可能性がある。被害側の意思決定者が AI の助言を受けている場合、攻撃者が同様の AI システムを入手し、攻撃実行前にそのシステムが推奨する対応策を評価することさえ可能になるかもしれない。

攻撃への対応には、攻撃の規模や意図に関する理解も必要だが、これらは必ずしも明確ではない。重要インフラに侵入した外国のマルウェアが、正当な諜報活動なのか、破壊攻撃のための違法な事前配置なのか、サイバーアナリストの間で議論が続いている。 攻撃用 AI エージェントを捕獲すれば、現在の攻撃ツール分析が明らかにしている能力だけでなく、その意図についてもより多くの情報が得られる可能性がある。セキュリティアナリストは、捕獲した AI エージェントをテストし、損害を与える行動やエスカレーション行動を実行する判断条件を特定できるかもしれない。攻撃後に行うフォレンジック分析では、攻撃前に予測する場合よりも明確な結果が得られる可能性がある。なぜなら、人間の攻撃者が無害なエージェントを容易に攻撃的なものへ更新できる場合、攻撃前の予測は困難だからだ。

使わなければ失う

意思決定者は、将来的に攻撃手段が利用できなくなる恐れから、サイバー攻撃の実行を迫られることもある。サイバーツールやアクセス権、エクスプロイトの短命性は、既に脅威アクターをより攻撃的に駆り立てている。防御が強化される前に、入手可能なエクスプロイトを急いで使用したり、危機や紛争の初期段階でより攻撃的になったりする可能性がある。

もし AI が、防御側による脆弱性の発見、他の攻撃者による暴露、あるいは通常のソフトウェア更新による単純な削除までの予想時間を短縮するなら、攻撃者はより迅速かつ攻撃的に行動する圧力を感じるかもしれない。一方、防御側が脆弱性をより早く発見できるようにする AI 技術が、攻撃者にとっても脆弱性をより豊富に生み出すなら、その圧力は緩和される可能性がある。

10.結論

AI が攻撃側か防御側に有利をもたらすかという問いに、唯一の答えは存在しない。コスト、破壊、強制、情報伝達など、利益を考慮する方法はあまりにも多い。そしてサイバー紛争や競争のあまりにも多くの側面が、相反する形で影響を受けるだろう。

AI は、防御対象となる製品をより複雑で数多くのものにする可能性が高い。これらの製品には、従来のデジタル部品よりも脆弱な AI コンポーネントが組み込まれる一方で、サイバー攻撃時およびその後のレジリエンスとして人間が提供する耐性は低下するだろう。しかし AI の利用は、インターネットと接続性を再構築し、接続性の必要性を減らすことで標的を強化したり、情報の流れを少数の重要ノードに集中させることで標的を強化したりする可能性もある。

AI がコードや組織慣行のセキュリティ脆弱性を自律的に検証し、ベストプラクティスに沿うよう適応させる段階まで進めば、デジタルエコシステム全体の防御力はさらに強化されるだろう。こうしたタスクは AI の学習に適したデータ豊富さを有する可能性が高い。しかし AI が脆弱性発見を加速させ、特に修正困難な脆弱性に対しては、AI がパッチ開発を迅速化できても防御側が追いつけなくなる恐れがある。 その場合、防御を維持するためには、AI がより脆弱性の少ないシステムの設計を支援する必要がある。

AI がなくても攻撃と防御の速度と量は既に高いため、追加の知能がもたらす変化は一部が予想するほど大きくないかもしれない。 ただし、脆弱性発見や偵察の一部など、サイバーキャンペーンの戦術的側面は AI によって加速される可能性がある。これらの工程がボトルネックとなっている場合、全体的な速度は向上するだろう。速度や規模以上に影響が大きいのは、戦術的判断の委任によって防御側が直面する攻撃の多様性が増し、同時に事故リスクが高まる点だ。特に、信頼性よりも速度や規模を優先して委任が行われた場合に顕著である。事故は攻撃側エージェントと防御側エージェントの双方から発生しうる。

より戦略的なレベルでは、AI はテロリストなど従来能力の低かった脅威アクターへの参入障壁を下げ得る。ただし AI が同時に防御強化を助ける場合、それらのアクターは攻撃実行能力が低下する可能性がある。より伝統的な脅威アクターにとっては、サイバー作戦の加速や単独攻撃の試行回数増加が、サイバー攻撃をより現実的な選択肢に変え得る。AI が攻撃を容易にする場合、攻撃者は実際に攻撃を実行する必要性が減る。脅威そのものが十分であり、防御側は攻撃されると壊滅的なサービスや運用を縮小・回避するだろう。それらのサービスを削減または廃止するだけで、攻撃なしに十分な壊滅的状況が生じうる。一方、AI が標的の防御強化に寄与する場合、防御側が対策を講じる前に攻撃能力を迅速に活用する圧力が高まる可能性がある。しかし強化された防御は攻撃者の成功を稀かつ限定的なものに留めるだろう。

著者

アンドルー・ローンは、セキュリティ・新興技術センターでサイバー AI プロジェクトに携わる上級研究員である。

謝辞

繰り返しフィードバックとサポートを提供してくれた、ジョン・バンセマー、レベッカ・ゲレス、コリン・シェイ・ブライマー、カイル・ミラー、ジェシカ・ジ、クリス・ロールフに感謝する。フォーマットと校正を担当したアレクサンドラ・フォールにも感謝する。また、有益な議論と意見交換をしてくれた、ジェイソン・ヒーリー、ブルース・シュナイアー、ブランドン・バレリアーノにも感謝する。

資金

この研究は、英国科学技術革新省からの助成金によって支援された。

著作権

© 2025 セキュリティ・新興技術センター。この著作物は、クリエイティブ・コモンズ表示-非営利 4.0 国際ライセンスに基づいて利用許諾されている。このライセンスの写しを見るには、https://creativecommons.org/licenses/by-nc/4.0/ を参照のこと。

参考文献

Ablon, L. & Bogart, A. (2017). ゼロデイ、数千の夜：ゼロデイ脆弱性とその悪用の歴史と現状（技術報告書）。RAND。

Aljamaan, F., Temsah, M.-H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., ... Malki, K. H. (2024). 医療用人工知能チャットボットの参照幻覚スコア：開発とユーザビリティ研究。*JMIR Medical Informatics*。https://doi.org/10.2196/54345 から取得。

Allyn, B., Mann, B., Chappell, B. & Al-Kassab, F. (2024).世界中のシステムを混乱させているコンピューターの更新の不具合について我々が知っていること。ナショナル・パブリック・ラジオ。

Alperovitch, D. & ヒーリー、J. (2020). 攻撃的なイノベーションを理解し、破壊する。https://www.rsaconference.com/Library/presentation/USA/2020/ understanding-and-disrupting-offensive-innovations.

Arnold, R. D. (2016). 機密ネットワークと非機密ネットワーク間のデータ転送戦略（技術報告書 No. ARWSE-TR-15037）。米国陸軍兵器研究開発技術センター。https://apps.dtic.mil/sti/tr/pdf/AD1005160.pdf から取得。

Barabási, A.-L., Albert, R. & Jeong, H. (1999). スケールフリーランダムネットワークの平均場理論. *Physica A*. 取得元: https://doi.org/10.1016/S0378-4371(99)00291-5

ビン・ラシッド、A. & カウシック、A. K. (2024). 世界中の産業を変革する AI：その多様な応用に関する包括的概観. *Hybrid Advances*.

Brumfiel, G. (2016). 米海軍、将校向けに星による航法訓練を復活させる. *National Public Radio*. https://www.npr.org/2016/02/22/467210492/u-s-navy-bringsback-navigation-by-the-stars-for-officers.

ブランデージ、M.、アヴィン、S.、ワン、J.、ベルフィールド、H.、クルーガー、G.、ハドフィールド、G.、... アンデルユング、M. (2020). 信頼できる AI 開発に向けて：検証可能な主張を支えるメカニズム. *arXiv* プレプリント *arXiv:2004.07213*. https://arxiv.org/abs/2004.07213 から取得。

ブキャナン、B.（2017）。『サイバーセキュリティのジレンマ：国家間のハッキング、信頼、恐怖』。オックスフォード大学出版局。doi: 10.1093/acprof:oso/9780190665012.001.0001

CISA. (n.d.). 重要インフラセクター. https://www.cisa.gov/topics/criticalinfrastructure-security-and-resilience/critical-infrastructuresectors.

CISA. (2022). 重要インフラに対するロシア国家支援型及び犯罪的サイバー脅威 (技術報告書). 米国サイバーセキュリティ・インフラストラクチャ庁. https://www.cisa.gov/news-events/cybersecurity-advisories/ aa22-110a から取得

CISA. (2024). *2023* 年に最も頻繁に悪用された脆弱性トップ（技術報告書）. サイバーセキュリティ・インフラセキュリティ庁.

Corsi, G., Kilian, K. & Mallah, R. (2024). 人工知能が攻撃と防御の力学に及ぼす影響に関する考察. *arXiv* プレプリント *arXiv:2412.04029*. doi: 10.48550/arXiv.2412.04029

CrowdStrike. (2017). ウクライナ野戦砲兵部隊追跡における *Fancy Bear Android* マルウェアの使用 (技術報告書). 著者. https://www.crowdstrike.com/wp-content/brochures/FancyBearTracksUkrainianArtillery.pdf.

Cruz, M. (2025). セキュリティおよび *IT* 役割におけるセキュリティコパイロットのユースケース (技術報告書). Microsoft. 取得元: https://learn.microsoft.com/en-us/ copilot/security/use-case-role-overview

サイバーセンター（2022）。サイバー脅威環境の序論（技術報告書）。カナダサイバーセキュリティセンター。

DARPA. (2023). *Aixcc: AI* サイバーチャレンジ. 取得元: https://www.darpa.mil/research/programs/ai-cyber

Ding, D., Han, Q.-L., Xiang, Y., Ge, X. & Zhang, X.-M. (2017). 産業用サイバーフィジカルシステムにおけるセキュリティ制御と攻撃検知に関する調査. *Neurocomputing*.

Dissanayake, N., Jayatilaka, A., Zahedi, M. & Babar, M. A. (2021). ソフトウェアセキュリティパッチ管理 - 課題、アプローチ、ツール、実践に関する系統的文献レビュー. *arXiv* プレプリント *arXiv:2012.00544*. https://arxiv.org/pdf/2012.00544 から取得。

DNI. (2024). 米国インフラに対する最近のサイバー攻撃は、米国の重要システムの脆弱性を浮き彫りにしている、*2023* 年 *11* 月～*2024* 年 *4* 月 (技術報告書). 国家情報長官室. https://www.dni.gov/files/CTIIC/documents/products/Recent_Cyber_Attacks_on_US_Infrastructure_Underscore_Vulnerability_of_Critical_US_Systems-June2024.pdf.

国防総省（2011）。国防総省サイバー空間政策報告書：*2011* 会計年度国防授権法第 *934* 条に基づく議会提出報告書（技術報告書）。米国国防総省。https://irp.fas.org/eprint/dod-cyber.pdf から取得。

Durumeric, Z., Li, F., Kasten, J., Amann, J., Beekman, J., Payer, M., ... Halderman, J. A. (2014). ハートブリードの問題点. インターネット計測会議 *2014* 論文集 (p. 475–488). 米国ニューヨーク州ニューヨーク: 計算機学会. https://doi.org/10.1145/2663716. 2663755 doi: 10.1145/2663716.2663755 から取得

イースターリー、J. & ファニング、T. (2023). コロニアル・パイプラインへの攻撃：過去 *2* 年間で学んだことと行ったこと。https://www.cisa.gov/newsevents/news/attack-colonial-pipeline-what-weve-learned-what-wevedone-over-past-two-years.

フィッシャー、C. N.、ネッツァー、R. H. B.、ミラー、B. P. (1992). レース条件とは何か？：いくつかの問題と形式化. *ACM Letters on Programming Languages and Systems*. https://dl.acm.org/doi/10.1145/130616.130623 から取得

Fischer, F., Bottinger, K., Xiao, H., Stransky, C., Acar, Y. & Backes, M. (2017). スタックオーバーフローは有害か？ Android アプリケーションセキュリティへのコピー＆ペーストの影響.

Fischerkeller, M. P., Goldman, E. O. & Harknett, R. J. (2022). サイバー持続性理論：サイバー空間における国家安全保障の再定義。オックスフォード大学出版局。

フランクリン、C. (2018)。*Shodan* での検索を始める *7* つのステップ。https:// から取得。www.darkreading.com/iot/7-steps-to-start-searching-with-shodan

Garfinkel, B. & Dafoe, A.(2021).攻撃と防御のバランスはどのように取られているのか？ラウトレッジ。https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003179917-10/offense-defense-balance-scale-ben-garfinkel-allan-dafoe。

Geer, D., Bace, R., Gutmann, P., Metzger, P., Pfleeger, C. P., Quarterman, J. S. & Schneider, B. (2003). サイバーインセキュリティー：独占の代償。

Google. (2024). 昼寝から深い眠りへ：大規模言語モデルを用いた実世界のコード脆弱性検出. https://googleprojectzero.blogspot.com/2024/10/ from-naptime-to-big-sleep.html.

グレイ博士、C. S. (2013). サイバーパワーの戦略的意義：なぜ天は落ちてこないのか。米国陸軍戦争大学出版局。

Greenfield, V. A., Welburn, J. W., Schwindt, K., Ish, D., Lohn, A. J. & Hartnett, G. S. (2023). サイバーセキュリティとサプライチェーンリスクマネジメントは単純な加算ではない（技術報告書）. RAND.

Harknett, R. J. (2024). アメリカの同盟国は変化している：サイバー空間は抑止力ではなく、持続性である。*Cyberscoop*.

ヒーリー、J. (2021a)。サイバー空間における攻撃側のシステム全体の優位性を理解する。*Lawfare*。https://www.lawfaremedia.org/article/understanding-offensessystemwide-advantage-cyberspace。

ヒーリー、J. (2021b). サイバー空間における攻撃側のシステム全体の優位性を理解する。*Lawfare*.

ヒーリー、J.、ジャービス、R. (2020)。エスカレーションの逆転と状況的サイバー安定性のその他の奇妙さ。テキサス国家安全保障レビュー、*3*(4)、? doi: 10.26153/tsw/10962

Hern, A. (2017). Wannacry、petya、notpetya：2017 年にランサムウェアが大きな成功を収めた経緯。*The Guardian*.

ハルム、G. V. (2024). サイバーセキュリティにおける人工知能と機械学習の現状（技術報告書）. ダーク・リーディング.

ハントリー、W. & シャイブス、T. (2024). サイバー空間における攻防のバランス. 第 *23* 回欧州サイバー戦・セキュリティ会議論文集.

IBM. (2023). 脅威アクターとは何か？ https://www.ibm.com/think/topics/threat-actor から取得

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. & Madry, A. (2019). 敵対的サンプルはバグではなく機能である. https://arxiv.org/abs/1905.02175 から取得.

Jacobs, J., Romanosky, S., Suciu, O., Edwards, B. & Sarabi, A. (2023). 脆弱性優先度の強化：コミュニティ主導の知見を用いたデータ駆動型エクスプロイト予測. *arXiv preprint arXiv:2302.14172*. doi: 10.48550/arXiv.2302.14172

Janow, M. E., Rattray, G. & Venables, P. (2017). 防御可能なサイバー空間の構築 (Tech.Rep.). ニューヨーク・サイバー・タスクフォース。

Jason Healey、D. N.、Robert Jervis。(n.d.)。サイバー紛争と競争のダイナミクス。

ジャービス、R. (1978)。安全保障のジレンマにおける協力。*World Politics*、*30*(2)、? doi:10.2307/2009958

JetBrains. (2024). 開発者エコシステムの現状に関するレポート *2024.* https:// から取得。www.jetbrains.com/lp/devecosystem-2024/

Ji, J., Goldstein, J. A. & Lohn, A. J. (2023). 大規模言語モデルの出力制御：入門編。Kerner, S. M. (2024). Openai と Apple の提携について。*TechTarget*。Kozen, D. C.(1977). オートマトンと計算可能性. (章: ライスの定理).

Springer.https://link.springer.com/chapter/10.1007/978-3-642-85706-5_42 から取得

クレブス、B.（2021）。ロシアのハッカーが嫌う奇妙なトリックを試してみろ。https://krebsonsecurity.com/2021/05/try-this-one-weird-trick-russianhackers-hate/。

Kushner, D. (2013). スタックスネットの実話：カスペルスキー・ラボがイランの核燃料濃縮計画を妨害したマルウェアを追跡した経緯。*IEEE Spectrum*.

https://spectrum.ieee.org/the-real-story-of-stuxnet から取得

Lee, C. (2023). ウクライナの重要インフラに対するロシアのサイバー作戦。スタンフォード国際政策レビュー。

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R. & Wilson, N. (2025). 批判的思考に対する生成的 AI の影響：ナレッジワーカーを対象とした調査による、認知的努力と自信効果の自己申告による減少。 https://doi.org/10.1145/3706598.3713778 から取得。

Libicki, M. C. (2009). サイバー抑止力とサイバー戦争 (技術報告書). RAND Corporation.

ロカテッリ、A. (2013)。サイバー空間における攻防のバランス (技術報告書)。戦略政策研究所。https://www.ispionline.it/sites/default/files/ pubblicazioni/analysis_203_2013.pdf。

ロッキード・マーティン (2025). サイバーキルチェーン. https://www.lockheedmartin.com/enus/capabilities/cyber/cyber-kill-chain.html.

Lohn, A. (2018). サイバーセキュリティの将来にとって、メルトダウン、スペクター、ライゼンフォールはどのような意味を持つのか？ *TechCrunch*.

Lohn, A. (2019). 深度防御：封鎖と遅延の基本。*arXiv* プレプリント *arXiv:1910.00111*. doi: 10.48550/arXiv.1910.00111

Lohn, A. & Jackson, K. A. (2022a). *AI* はサイバーの剣か盾か（技術報告書）. セキュリティ・新興技術センター. doi: 10.51593/2022CA002

Lohn, A. & Jackson, K. A. (2022b). *AI* はサイバーの剣か盾か：技術進歩の数理モデル（技術報告書）. セキュリティ・新興技術センター. doi: 10.48550/arXiv.2207.13825

ローン、A.、ナック、A.、バーク、A. & ジャクソン、K. (2023). 自動化されたサイバー防衛：研究室から運用へのロードマップ (技術報告書). セキュリティ・新興技術センター.

Lohn, A. J. (2020a). AI の脆弱性の推定：安全度水準と分布外性能のテスト必要性. *arXiv* プレプリント *arXiv:2009.00802*.

https://arxiv.org/abs/2009.00802 から取得

Lohn, A. J. (2020b). *AI* をハッキングする：政策立案者向け機械学習サイバーセキュリティ入門（技術報告書）. セキュリティ・新興技術センター.

Lohn, A. J. (2023). *AI* のスケール化：最先端 *AI* のコストと性能（技術報告書）. セキュリティ・新興技術センター. 取得元：https://cset.georgetown.edu/wp-content/uploads/Scaling-AI-Cost-and-Performance-of-AI-at-the-Leading-Edge.pdf

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D. & Hoy, D. E. (2024). 幻覚なし？主要 *AI* 法律研究ツールの信頼性評価. 取得元: https://arxiv.org/abs/2405.20362

McQuade, M. (2018). 史上最も破壊的なサイバー攻撃「NotPetya」の知られざる真実. *WIRED*.

Md Fazle Rabbi, M. Z., Arifa Champa. (2024). AI が書き、我々が分析する：ChatGPT Python コードの物語. In *ACM* 第 *21* 回ソフトウェアリポジトリマイニング国際会議.

https://dl.acm.org/doi/pdf/10.1145/3643991.3645076 から取得

ミアーズハイマー、J. J.（1989）。「通常戦力の均衡アセスメント：3 対 1 の法則とその批判」。『国際安全保障』13 巻 4 号、54-89 頁。doi: 10.2307/2538780

MITRE. (2025). エンタープライズ向け *ATT&CK* マトリックス. https://attack.mitre.org/ から取得

Moore, T., Kenneally, E., Collett, M. & Thapa, P. (2019). サイバーセキュリティ研究データセットの評価. 第 *18* 回情報セキュリティ経済学ワークショップ *(WEIS)* にて発表.

NATO. (2017). *Notpetya* と *WannaCry* は国際社会の共同対応を要請する. https://ccdcoe.org/news/2017/notpetyaand-wannacry-call-for-a-joint-response-from-international-community/ から取得.

Nguyen, D. D. A., Alain, P., Autrel, F., Bouabdallah, A., François, J. & Doyen, G. (2024). エターナルブルーを利用したマルウェアはどれほどの速さで拡散するのか？ワナクライとノットペティアの事例。*2024* 年 *IEEE* 第 *10* 回ネットワークソフトウェリゼーション国際会議（*NetSoft*）収録（p. 399-404）。doi: 10.1109/NetSoft60951.2024.10588886

NIST. (2024). *NIST* サイバーセキュリティ枠組み*(CSF) 2.0* (技術報告書). 米国国立標準技術研究所. 取得元: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

Orman, H. (2003). モリスワーム：15 年後の視点. *IEEE Security & Privacy*,*1*(5), 35-43. doi: 10.1109/MSECP.2003.1236233Pratt, M. K. (2024). AI の 24 の利点と欠点. *TechTarget*.

ラドイニ、A. & シディキ、M. (2021). 化学・生物・放射性物質・核（*CBRN*）施設に対するサイバー脅威（技術報告書）. 国連地域間犯罪司法研究所. 取得元 https://unicri.org/sites/default/files/2021-12/16_cyber_threat.pdf

RedHat. (2024). *Linux* カーネルのライブパッチングとは何か？ https://www.redhat.com/en/topics/linux/what-is-linux-kernel-live-patching から取得

Rid, T. & McBurney, P. (2012). サイバー兵器. *The RUSI Journal*. 取得元: https://doi.org/10.1080/03071847.2012.664354

リッギ、J.（2020 年）。病院に対するランサムウェア攻撃は変化した。*AHA* ヘルス・イノベーション・センター。

Rohlf, C. (2023). (技術報告書). セキュリティ・新興技術センター. 取得元: https://cset.georgetown.edu/article/memory-safety-an-explainer/

ルシノヴァ、V. & マルティノヴァ、E. (2024). 制裁によるサイバー攻撃対策：デジタル脅威と経済的対応策. *Israel Law Review*, *57*(1), 135–174. doi: 10.1017/ S0021223722000255

Salem, A. H., Azzam, S. M., Emam, O. E. & Abohany, A. A. (2024). サイバーセキュリティの進展：AI 駆動型検知技術の包括的レビュー。*Journal of Big Data*(105)。https://doi.org/10.1186/s40537-024-00957-y から取得。

シュナイアー、B. (2012). 技術の変化がセキュリティに与える影響. *IEEE* コンピュータ・信頼性学会.

シュナイアー, B. (2018). 人工知能と攻撃/防御のバランス. *IEEE* コンピュータ・信頼性学会.

Sedlmeir, J., Smethurst, R., Rieger, A. & Fridgen, G. (2021). デジタルアイデンティティと検証可能クレデンシャル. ビジネス情報システム工学. 取得元: https://doi.org/10.1007/s12599-021-00722-y

Shannon, C. (1949). 雑音存在下におけるコミュニケーション. *Proceedings of the IRE*, *37*(1), 10-21. doi: 10.1109/JRPROC.1949.232969

シルバー、L.（2022 年）。アメリカ国民は、2020 年 3 月と比べて現在自国が直面する世界的な脅威を異なって認識している。ピュー研究所。https://www.pewresearch.org/shortreads/2022/06/06/americans-see-different-global-threats-facing-thecountry-now-than-in-march-2020/ から取得。

シンガー、B.、ルーカス、K.、アディガ、L.、ジェイン、M.、バウアー、L. & セカール、V. (2025). 多段階ネットワーク攻撃の実行における *LLM* 利用の可能性について。https://arxiv.org/abs/2501.16466 から取得。

Slayton, R. (2017a). サイバー攻防の均衡とは何か？その概念、要因、アセスメント. *International Security*, *41*(3), 72-109. 取得元: https://www. jstor.org/stable/26777791?seq=1

Slayton, R. (2017b). サイバー作戦が常に攻撃側に有利とは限らない理由. 国際安全保障.

Smith, A. L., Greaves, F. & Panch, T. (2023). 幻覚か虚構か？大規模言語モデルにおける比喩としての神経解剖学. *PLoS Digital Health*, *2*. 取得元: https://doi.org/10.1371/journal.pdig.0000388

スミス、S.（2021）。サイバー脅威と大量破壊兵器。大量破壊兵器研究センター。https://wmdcenter.ndu.edu/Portals/68/Documents/wmd-proceedings/CSWMD%20Proceedings%20Jun%2021%20PQ.pdf から取得。

スミス, C. (2020). サイバー攻撃の崇拝：サイバー攻撃と防御のバランスに関する誤解. イエール国際問題ジャーナル. 取得元 https://www.yalejournal.org/publications/cult-of-the-cyberoffensive-misperceptions-of-the-cyber-offensedefense-balance

Tang, J., Saade, T. & Kelly, S. (2024). サイバーセキュリティにおける人工知能の影響：攻防バランスの変化（技術報告書）. Institute for Security + Technology.

寺島アキオ（2022 年）。ウクライナ戦争最新情報：ロシアが 10 月以降ウクライナエネルギー施設に「1000 発以上のミサイルとドローン」を発射、電力不足は依然「深刻」。キエフ・インディペンデント。

トレンドマイクロ（2021 年）。「*Apache Log4j*（*Log4Shell*）脆弱性とは何か？」 https://www.trendmicro.com/en_us/what-is/apache-log4j-vulnerability.html

Valeriano, B. (2022). サイバーセキュリティにおける攻撃と防御のバランス失敗。*The Cyber Defense Review*.

Valeriano, B. & Jensen, B. (2019). サイバー攻撃の神話：抑制の必要性について。*Policy Analysis*.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L. & Hobbhahn, M. (2024). データは枯渇するのか？ 人間生成データに基づく *LLM* スケーリングの限界.

フォン・エッシェンバッハ、W. J. (2021). 透明性とブラックボックス問題：なぜ我々は AI を信頼しないのか。哲学と技術、1607-1622。https://doi.org/10.1007/s13347-021-00477-0 から取得。

Woodcock, J., Gorm Larsen, P., Bicarregui, J. & Fitzgerald, J. (2009). 形式手法：実践と経験. *ACM Computing Surveys*, *41*(4). 取得元：https:// dl.acm.org/doi/abs/10.1145/1592434.1592436

ユーニス、M.（2023）。米国ではサイバー妨害が最も重大な脅威。ギャラップ。 https://news.gallup.com/poll/472544/cyber-disruption-criticalthreat.aspx から取得。

Zetter, K. (2014). 世界初のデジタル兵器スタックスネットを未曾有の視点で検証. *WIRED*.

Zetter, K. (2016). NSA のハッカー責任者が明かす、システム侵入を防ぐ方法. *WIRED*.