# AI 時代を迎えてのセキュリティの在り方とは
## ～切り離しの Se-Cure から共有の Co-Cure へ～

米澤 一樹

「セキュリティ：Security」の語源はラテン語で「Se-Cure：心配事(Cure)から切り離す(Se)」すなわち「心配無用の状態にする」とされている。つまり、「セキュリティ」とは、「安心を確保する手段」と言うことができる。そして、それを電子化された情報に適用する際には、「機密性：許可された者にしか見せない」・「完全性：許可された方法・時点でしか更新できない」・「可用性：許可された者が必要な時に使える」の 3 要素における要求レベルを満たすこととされてきた。つまり、「許可・不許可による厳密な切り離し」によって確立されてきたものとも言える。そして、それは、システムとしての AI を保護する手段としては、引き続き有効であると思われる。しかし、人と対話してその対話を新たな創造につなげていくという AI の働きに注目すると、異なる視点も必要なのではないか。

そこで、本稿では、そのような AI 時代におけるセキュリティの在り方の一つとして「切り離しの Se-Cure」に対する「共有の『Co-Cure』（心配事を共有するという意味の造語）」を提唱し、AI 時代における「安心の確保」の在り方を論じる基点としたい。



## １．セキュリティとは

「セキュリティ：Security」の語源[Oxford]は、ラテン語で Se-Cure：心配事(Cure)から切り離す(Se)」すなわち「心配無用の状態にする」とされている。そして、現代英語では「攻撃や危険に対する保護に関する活動」、「未来に発生するかもしれない悪い事象への保護」、「危険や心配事から保護されていて幸せと

感じる状態」などの意味が定義[Oxford]されている。つまり、不確実な事象への対策などで「安心」を確保するのが「セキュリティ」であると言える。
なお、「セキュリティ」としばしば混同されがちな「セーフティ：Safety」は「全ての Danger(危険)、Loss(損失)、Harm(害悪)から Safe な状態」と定義されており、「Safe」の語源は「Save：助ける」から分岐したとされている。つまり、「セーフティ」とは、「（一定レベルとはいえ）確保された安全」を示すものであると言える。
従って、「セキュリティ」とは、「セーフティを希求して行動し『一定レベルの Safe を確保できたとの安心』を得る」ことであると言うことができる。

## 2．電子化された情報へのセキュリティ

「セキュリティ」の概念は、当然のごとく、情報の保全にも適用されてきた。その嚆矢と言えるのが、1972 年に米空軍におけるコンピュータシステムのセキュリティ要件の計画研究としてとりまとめられた「Computer Security Technology Planning Study」[James1972]であり、また、その後に米国防総省のコンピュータシステム調達基準としてまとめられた「Trusted Computer System Evaluation Criteria（通称：Orange Book）」[DoD1983]である。特に後者においては、セキュリティポリシーとそれに基づくアクセス管理、情報の識別、主体・客体の識別と認証、説明責任とその手段としての監査、動作上の保証とその証明手段としての試験及び証跡提示など、2024 年の現在まで用いられている概念がほぼ全て網羅されており、その意味で、情報セキュリティの原点にして原典ともいえる存在である。
そして、それを貫く考えは、「許可・不許可による厳密な切り離し」であり、技術的・組織的な実装に際しては、要求される「セキュリティ＝安心」レベルに応じて、情報の「機密性：許可された者にしか見せない」・「完全性：許可された方法・時点でしか更新できない」・「可用性：許可された者が必要な時に使える」の 3 要素の担保が求められることである。

## 3．AI に対するセキュリティ

AI に対するセキュリティについては、2022 年 11 月の米 OpenAI による ChatGPT リリースに始まった所謂「生成 AI」の急速な普及を受けて、「AI システム」に対するセキュリティ対策も提唱されている。そして、2024 年 4 月には、米英加豪新の 5 か国の政府機関が共同で AI システムの導入・運営ガイダンス[FiveEyes2024]がリリースされている。こちらでは、従来のコンピュータシステムに対するセキュリティ対策に加えて、AI モデルおよび入力データ（特に学習データ）の保護、（意図せぬ学習結果を招いた場合などの失敗に備えた）ロールバック機能の準備、そして、人間を介した（学習・運用の）プロセスなどが提唱されている。これらは、入力データによって機能・性能を充実させていく機械学習を意識したものであるが、その根本にある考えは、従来と同様の「許可・不

許可による厳密な切り離し」である。

## 4．セキュリティは何のため？

これまで述べてきたように、電子化された情報に対してのセキュリティとは約めて言うと「許可・不許可による切り離し」である。また、それ以外の分野においても、『一定レベルの Safe を確保できたとの安心』を得るための行動として「許可・不許可による切り離し」は共通であると言える。このことは、物理的な施設や人物の警護において「許可された者以外は近づけない・立ち入らせない」ポリシーが必ず適用されることによっても明らかである。同様なことは、国家のレベルでは「正当な旅券と査証を持っていないものは入国させない」となり、家庭のレベルでは「入ることを許可した者以外は家に入れない」となる。

では、このような「許可・不許可による切り離し：セキュリティ」は何のために存在しているのだろうか。それは、Safe を脅かす可能性のある因子、つまり、Danger(危険)・Loss(損失)・Harm(害悪)をもたらす可能性が否定できない存在を遠ざけることによって、「一定レベルの Safe を確保できたとの安心」を得るためであると言える。そして、その「一定レベルの Safe」が対応するものが危険・損失・害悪であることを考慮すると、「一定レベルの Safe」は危険を排除して生き残ること (Be の自由確立)、損失を排除して保持しているものを維持すること(Have の自由確立)、害悪を排除して望む行動ができること(Do の自由確立)ということができる。つまり、セキュリティとは、一定レベルの「Be/Have/Do の自由確立」ができたとの安心を得るための手段であると言える。

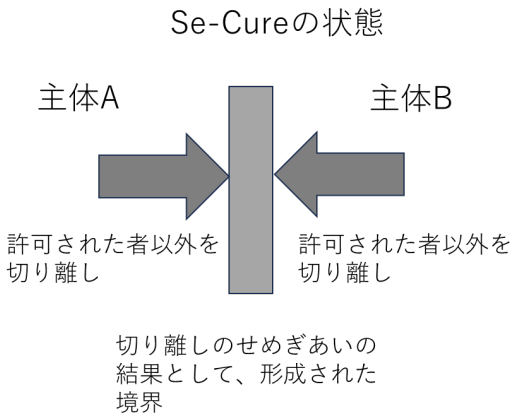
| Safe を脅かす<br>可能性のある因子 | 「一定レベルの Safe」に到達した状態 |
|---|---|
| Danger(危険) | 危険を排除して生き残ること<br>(Be の自由確立) |
| Loss(損失) | 損失を排除して保持しているものを<br>維持すること<br>(Have の自由確立) |
| Harm(害悪) | 害悪を排除して望む行動ができること<br>(Do の自由確立) |

## 5．セキュリティのせめぎ合いとしての境界

当然のことではあるが、「許可・不許可による切り離し」には実力行使が伴う。許可されている者のアクセスを保障することも、許可されない者のアクセスを拒絶することのいずれにも実力行使は不可欠である。

そして、その実力行使が複数の主体によって同時に行われる以上、一定の確率で

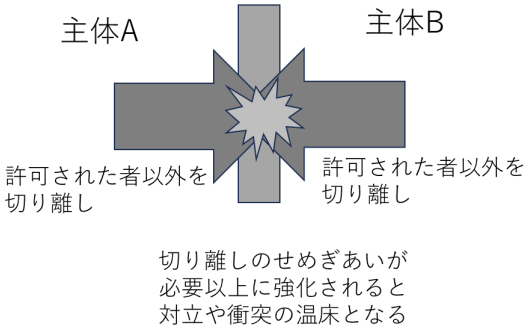
せめぎ合いが発生する。そのせめぎ合いの結果として、お互いに「これ以上の『Be/Have/Do の自由』を求めない」との合意あるいは妥協、もしくは勢力均衡の結果として設定されるのが、境界である。つまり、国境であり、住居や事業所の境界であり、また、電子化された情報やそれを取り扱うシステムにおけるアクセス管理やシステム間インターフェースである。
そのような「許可・不許可による切り離し」と、そのせめぎ合いの結果としての境界が設定されつつ「一定レベルの『Be/Have/Do の自由確立』ができたとの安心を得る」ことがなされているのが、こんにちのセキュリティであり、セキュリティが土台となって支える社会である。




## 6．境界の強化が生む分断と対立

境界そのものの存在は決して悪いものではない。「くに（国）」という言葉が「限り」の意味であり、垣根で囲った範囲もしくは垣根そのものを言い表すことも有ったとの説が有力である[Kurano1940]ように、そもそも、「共同体の限り」つまり「（共通の価値観や規則のもとに）一緒にやっていける者達の範囲」を区切るものが境界である。そして、それは、同時に、「Be/Have/Do の自由確立」の方法やレベルに対しても共通認識を持つことができるということでもあり、むしろ、望ましいものであると言える。
ただ、この境界が、必要以上に強化されていくと分断と対立の温床となる。その分断と対立が時として衝突につながり、危険・損失・害悪を避けるための対策が、より大きな危険・損失・害悪を招くという皮肉な事態になった事例は歴史上数限りなく存在する。特に、二つの世界大戦と中世の西欧・中東における宗教戦争はその規模の大きさと内容の凄惨さが際立っている。

なお、このような現象は、実社会の共同体だけではなく、電子化された情報やそれを取り扱うシステムにおいても、自動化・自律化が進むと起こり得ると考えられる。

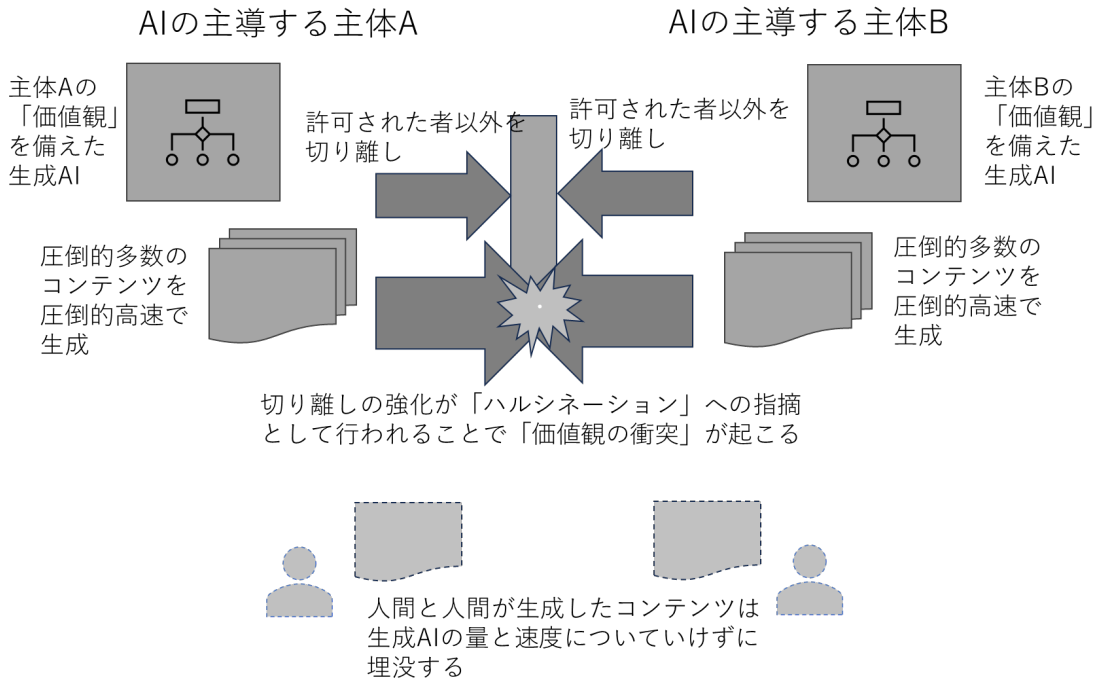



## 7．AI が加速する対立と分断、埋もれる人間

システムの自動化・自律化の役割を担い、また、加速すると期待されているのが AI である。そして、AI、特に、生成 AI は、前掲の対立と分断に拍車をかける可能性が小さくないと思われる。なぜならば、生成 AI がコンテンツ作成のために用いる「生成アルゴリズム」は、学習データと学習モデルによって作り上げられたものであり、価値観そのものと言えるからである。つまり、自らの価値観に基づく「正しい」コンテンツを「生成」する AI にとって、異なる価値観に基づいて生成されたコンテンツは「許可できないもの」であり、「切り離し」の対象となり得るのである。生成 AI の課題の一つとして、「間違えた」コンテンツを生成する「ハルシネーション」が挙げられているが、これは、生成 AI と生成コンテンツを活用する人間たちとの間の価値観の衝突と捉えることもできる。そして、この「ハルシネーション」の指摘を AI が行うようになった時、AI 同士の間にも価値観の衝突が発生し得る。異なる「価値観」を持つ AI 同士が「衝突」し、その「価値観」および「価値観によって生み出されたコンテンツ」の「一定レベルの Safe」を確保しようとした場合、必然的にお互いを非難・排斥する方向に進みかねない。そして、その非難・排斥の手段はコンテンツの生成となり、コンテンツの生成の応酬は内容の先鋭化と更なる生成量の増大を招くことが予想される。

そのような応酬の狭間で、大多数の人間は、そのようなコンテンツの応酬に乗せられて先鋭化の道をたどりかねない。また、冷静にいられる極少数の人間も、生成 AI の速度と精製能力についていくことはできず、彼らの生成するコンテンツと共に埋没してしまうことが容易に想定できる。

かつて、大衆扇動が招いた恐怖政治・全体主義・共産主義に類する悲劇が、今度は生成 AI によって引き起こされる可能性は決して低いとは言えない。




## 8．新たなアプローチとしての Co-cure の提案

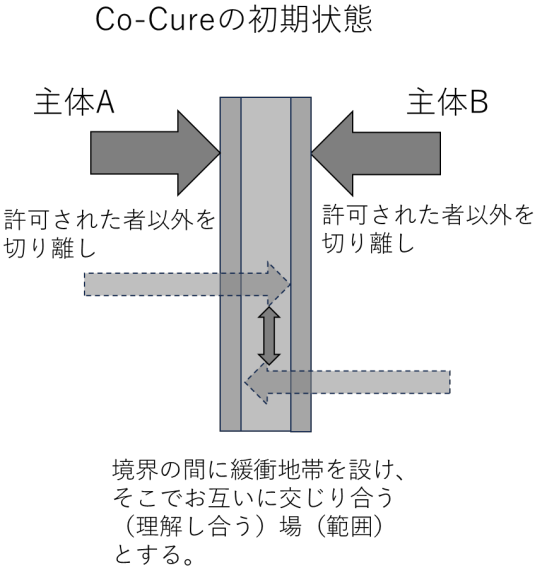
危険・損失・害悪を避けるための対策が、より大きな危険・損失・害悪を招くという皮肉を避けるために何ができるか、それは、人類文明が始まって以来の命題であり、歴史上、多くの取り組みがなされてきた。国家間であれば条約、組織や個人の間であれば法律や契約などである。そして、電子化された情報とそれを取り扱うシステムであれば、システム間インターフェースにおけるアクセス管理ポリシーがそれに該当すると言える。生成 AI でも他の生成 AI とその生成したコンテンツへの干渉を禁止することで実現可能ではあると思われる。しかし、これらは、あくまで対立と衝突をコントロールしようとするものであり、そのコントロールに失敗した場合に一気に事態が悪化することが多いという事例も歴史上数多く存在する。

そこで、本稿では、そのような対立と衝突を部分的にでも緩和・解消する試みとして「Co-cure」の概念を提案したい。「Co-cure」とは、「心配事を切り離す Se-cure」に対する概念として「心配事を共有する」意味を持たせた造語である。つまり、切り離すのではなく、共有することで、対立と衝突を部分的にでも緩和・解消を試み、また、分断についても同様の対応を試みるものである。

## 9．Co-cure とは

Co-cure の概念としては地政学用語である「緩衝地帯」に近い。元々は海洋国家

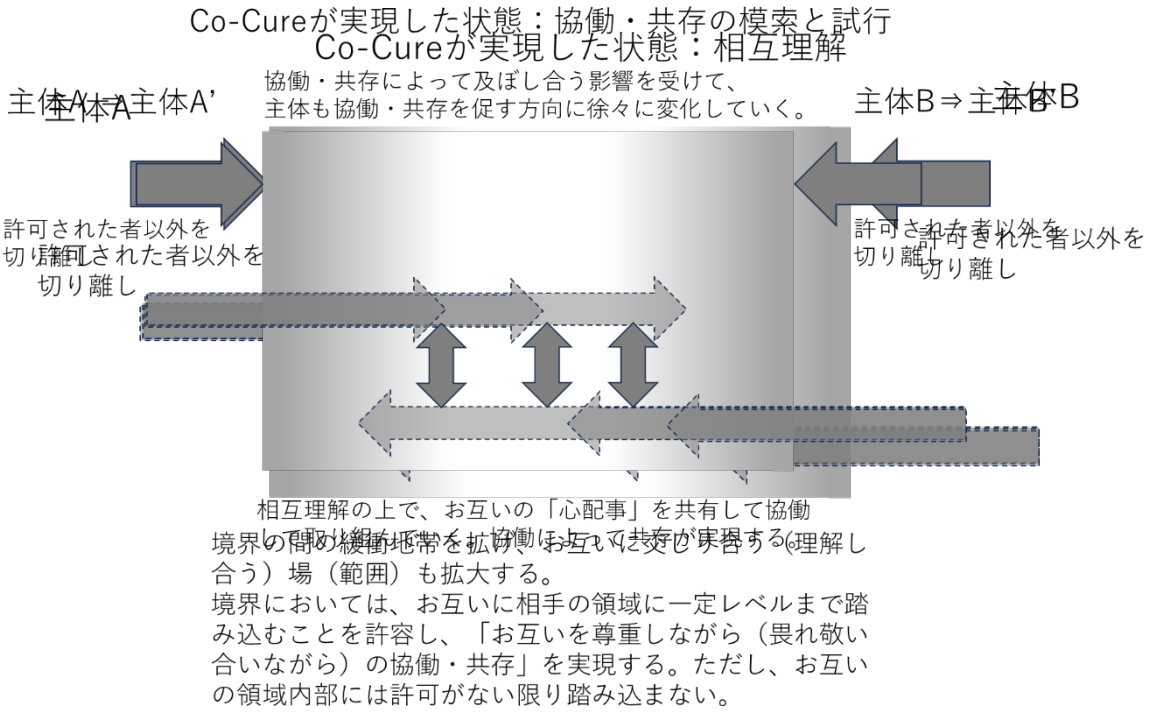
を自認する英国が、大陸国家としたソ連・ドイツに対峙するにあたり、どちらの陣営にも与しない国家群をお互いの勢力圏の間に配置することを呼んだものである。それによって、ソ連。ドイツの膨張を食い止めるとともに、英国が両国と直接衝突することを避ける効果が期待された。[Somura1984]
緩衝地帯は、対立を避ける目的ゆえに、交流が盛んになり、また、交渉の舞台ともなる。そのような場を意図的に組織・個人・システム間で設け、また、その場を最大限に活用することを提案するのが「Co-cure」である。
Co-cure を実現するにあたり、最初に、Se-cure の切り離しを緩め、お互いの領域を理解する緩衝地帯に相当するものを設ける。システム間インターフェースでの実装形態はさらなる検討の余地があるが、生成 AI に関しては、「生成」したコンテンツの正邪曲直の判定を一旦保留して、比較・相対化する場とすることでお互いの「コンテンツ生成にあたっての価値観」（生成の基準・仕組み）を理解する機会を与えることが「緩衝地帯」となるだろう。また、相対化されたコンテンツと「価値観」を前に、生成 AI を操作・管理すべき人同士の対話を促すことにもつながると思われる。




## １０．Co-cure に期待すること

境界の間にせめぎ合いを緩める緩衝地帯を設け、その緩衝地帯でお互いの理解を促すことから始まる Co-cure に期待すること、それは、緩衝地帯における価値観の相対化であり、それに基づいてお互いに理解し合うこと、つまり、相互理解である。そして、その相互理解に基づく協働・共存の模索と試行である。

ここでいう相互理解は、必ずしもお互いの歩み寄りや統合を意味するのではなく、自らと異なる相手の存在をその相違と共に認識し、その存在をお互いに認めた上で尊重することである。従って、友好親善が最適解となるわけでもない。むしろ、お互いに畏れ敬い合って一定の距離を置く形になることも有り得る。




そして、協働・共存の模索と試行とは、相互理解の上で、お互いが存在する上での心配事（課題）を共有して協働して取り組んでいくことである。そのような形の協働は、必然的に共存を実現する。なぜならば、お互いがお互いを協働のパートナーとして必要とするからである。そして、その協働・共存の過程を経て、各々自身が共同・共存を促す方向へと徐々に変化していくことが想定される。

## １１．Co-cure 実現までの道のり

本稿では AI 時代のセキュリティとして従来の「切り離しの Se-cure」に対して「共有の Co-cure」を提唱し、その実現の姿を提示した。しかしながら、「言うは易く行うは難し」とは、この Co-cure にも当てはまるものであり、実現までの道のりは短いものにはならないと予想される。

実現に向けては、まず、概念の整理と洗練が求められる。具体的には、緩衝地帯、相互理解、協働・共存の類型を見出すことが必要になると思われる。そのうえで、電子化された情報を取り扱うシステム上での実装の在り方、および、そのシステム上で動作する AI の振る舞いの在り方（動機付けとも言うことができる）とその実装方法を具体的に検討していく必要がある。

これらの課題を克服した先に AI 時代の可能性を最大限のものにすることを支えるセキュリティの在り方が生まれてくることを確信して取り組んでいきたい。

## 参考文献


[Oxford]『**Oxford Advanced Learner's Dictionary**』

[James1972] Anderson, James P.『**Computer Security Technology Planning Study**』, ESD-TR-73-51, Vol. Ⅰ・II

[DoD1983] Department of Defense『**Trusted Computer System Evaluation Criteria**』, DoD 5200.28-STD

[FiveEyes2024] 米国 CISA ほか『**Deploying AI Systems Securely - Best Practices for Deploying Secure and Resilient AI Systems**』, https://www.ic3.gov/Media/News/2024/240415.pdf

[Kurano1940] 本居宣長撰 倉野憲司校訂『**古事記伝(一)**』, 岩波文庫 ISBN4-00-302196-7

[Somura1984] 曽村保信『**地政学入門 外交戦略の政治学**』, 中公新書 ISBN4-12-100721-2