# 装備品等の研究開発における

# 責任あるＡＩ適用ガイドライン

（第１版）

防衛省

目次

## １　本ガイドライン策定の背景

近年の人工知能（ＡＩ）技術の急速な発展と普及に伴いＡＩの性能が大幅に向上した一方で、新たな課題も浮上している。ＡＩは、人間が定めた明確なルールや条件に基づいて動作するのではなく、与えられたデータからルールや知識を自ら学習し、それに基づいて新たな結果を出力する。このため、学習データの偏りなどに起因するバイアスや誤判断が生じる可能性など、ＡＩ特有の技術的なリスクが存在する。また、ＡＩの性能が向上したためにＡＩが人間や社会に与える影響が大きくなっている。ＡＩの推論過程や判断根拠等の不透明性、個人情報保護やプライバシーへの懸念、著作権侵害の可能性、ＡＩによる偽情報の生成と拡散、ＡＩ導入による雇用への影響など、技術的なリスクだけでなくＡＩ特有の倫理的・法的・社会的なリスクが存在する。これらの課題に対処するため、ＡＩ技術の責任ある研究開発と利用に関する国際的な議論が活発化している。各国政府や国際機関、企業、学術団体などが、ＡＩの倫理ガイドラインや規制フレームワークの策定に取り組んでいる。防衛省は、防衛省・自衛隊のＡＩ活用に関する考え方を部内・部外に示すため、「防衛省ＡＩ活用推進基本方針（令和６年７月防衛省）」（以下「基本方針」という。）を策定した。基本方針の中では研究開発における防衛省・自衛隊独自のガイドラインを策定することとしている。

### (1)　諸外国の軍事領域におけるＡＩの責任ある利用に関する取組状況

諸外国の軍事領域では、ＡＩの責任ある利用について次のような取り組みが行われている。

米国防総省は２０２０年に「ＡＩ倫理原則」を策定し、責任、公平性、追跡可能性、信頼性、統治可能性の５つの原則を掲げている。これらの原則に基づき、ＡＩの軍事利用における倫理的配慮と技術的信頼性の確保を目指している。２０２２年、米国防総省は、ＡＩの信頼性を達成することを望ましい最終状態として責任あるＡＩを導入していくための行動方針となる「Responsible Artificial Intelligence Strategy and Implementation Pathway」を策定した。その後、２０２３年には「ＡＩと自律性の責任ある軍事利用に関する政治宣言」（以下「ＡＩ政治宣言」という。）を主導し、多くの国が支持・参加した。

フランスでは、フランス軍事省が設置した軍事倫理委員会が、２０２１年に「Opinion on the Integration of Autonomy into Lethal Weapons Systems」をとりまとめた。フランス軍事省は、ＡＩの軍事利用における人間の判断の重要性を強調し、完全自律型の致死兵器システムの研究開発には反対の立場を取るとともに、米国主導のＡＩ政治宣言にも支持・参加しており、国際的な規範形成に積極的に参加している。

イギリスは、２０２２年に「Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence」を策定し、倫理的・法的・技術的な観点からの検討を進めている。英国防省は、ＡＩの軍事利用における人間の意思決定の重要性を強調し、完全自律型の致死兵器システムの研究開

発には反対の立場を取っている。イギリスもＡＩ政治宣言に支持・参加しており、国際的な規範形成に積極的に参加している。

(2)　ＡＩの軍事利用に関する国際的な議論及び我が国の見解

ア　ＡＩ政治宣言の概要等

米国が主導して立ち上げられたＡＩ政治宣言は、軍事分野におけるＡＩの責任ある開発、配備、使用を確保するための国際的な指針を示す重要な取り組みである。参加国は軍事用ＡＩ能力のライフサイクル全体を通じ、関連する各段階で適切な措置を講じることが期待されている。我が国を含む多くの国がこの宣言に支持を表明しており、２０２４年１２月時点で５８か国が参加している。

この宣言は法的拘束力を持たないものの、参加国はＡＩの軍事利用に関する宣言の目的推進のために、表１に示す措置の実施や継続的な協議、これらの措置の改善などを行うことが求められている。

そのほか、オランダ・韓国が主導する「軍事領域における責任あるＡＩ（ＲＥＡＩＭ：Responsible Artificial Intelligence in the Military Domain)」イニシアティブがあり、我が国を含む多くの国が支持を表明した成果文書では、軍事領域において、適用可能な国際法を遵守しつつ、責任ある形でＡＩが活用されなければならないこと等が確認されている。

イ　自律型致死兵器システムをめぐる議論（ＣＣＷ等）

近年の軍事分野における急速な技術発展や国際社会における関心の高まりを背景に、特定通常兵器使用禁止制限条約（Convention on Certain Conventional Weapons：ＣＣＷ）（以下「ＣＣＷ」という。）の下、２０１４年から自律型致死兵器システム（Lethal Autonomous Weapons Systems：ＬＡＷＳ）（以下「ＬＡＷＳ」という。）について議論が開始されている。２０１７年以降は、ＣＣＷに政府専門家会合（Group of Governmental Experts：ＧＧＥ）（以下「ＧＧＥ」という。）が設置され、ＬＡＷＳに関する国際的なルール作りに関する議論が行われている。

ＣＣＷのＬＡＷＳ・ＧＧＥでは、主にＬＡＷＳの特徴・定義、国際人道法の適用、人間の関与の在り方、責任・説明責任、リスク軽減・信頼醸成、規制の在り方等の論点について議論が行われている。主要な成果として、２０１９年１１月、国際人道法がＬＡＷＳにも適用されること、人間の責任が確保されなければならないこと等の１１項目から成る「ＬＡＷＳに関する指針」を採択し、２０２３年５月、国際人道法の遵守の観点からの禁止・制限の考え方等について記載したＬＡＷＳ・ＧＧＥ報告書を採択した。一方で、現在に至るまでＬＡＷＳの国際的な定義は定まっていない。

２０２３年１２月、国連事務総長にＬＡＷＳに関する報告書の作成を求めるオーストリア提案決議が国連総会で採択され、国連が国連加盟国等に対し意見提出を要請した。これを受けて我が国も、２０２４年５月、同報告書の作成及びＣＣＷでの議論に貢献するべく、我が国の見解をまとめた作業文書（以下「Ｌ

ＡＷＳ作業文書」という。）を提出している。ＬＡＷＳ作業文書において、我が国は、我が国が国際人道法を始めとする国際法や国内法により使用が認められない兵器システムの研究開発や運用を行うことはない旨を述べているほか、開発、使用が国際的に認められるべきでない兵器システム等についての我が国の見解を示している。

表１　ＡＩ政治宣言において参加国が実施すべき措置として定められている事項

| | |
|---|---|
| A | 国家は、自国の軍事組織が、ＡＩ能力の責任ある開発、配備、使用のために、これらの原則を採用し、実施することを確保すべきである。 |
| B | 国家は、軍事用ＡＩ能力が国際法、特に国際人道法の下でのそれぞれの義務に合致して使用されることを確保するために、法的審査などの適切な措置をとるべきである。国家はまた、国際人道法の履行を促進し、武力紛争における文民及び民用物の保護を改善するために、軍事用ＡＩ能力をどのように使用するかを検討すべきである。 |
| C | 国家は、高官が、このような兵器システムを含むがこれに限定されない、重大な影響を及ぼす活用を伴う軍事用ＡＩ能力の開発と配備を効果的かつ適切に監督することを確保すべきである。 |
| D | 国家は、軍事用ＡＩ能力における意図せざるバイアスを最小化するための積極的な措置をとるべきである。 |
| E | 国家は、軍事用ＡＩ能力を組み込んだ兵器システムを含め、軍事用ＡＩ能力の開発、配備、使用において、関係者が適切な注意を払うことを確保すべきである。 |
| F | 国家は、軍事用ＡＩ能力が、関連する防衛要員にとり透明で、また、かかる要員により監査可能な方法論、データソース、設計手順及び文書作成により開発されることを確保すべきである。 |
| G | 国家は、軍事用ＡＩ能力を使用する、又は使用を承認する要員が、それらのシステムの使用について状況に応じた適切な判断を下し、自動化バイアスのリスクを軽減するために、それらのシステムの能力と限界を十分に理解するよう訓練されることを確保すべきである。 |
| H | 国家は、軍事用ＡＩ能力が明確かつ十分に定義された用途を有し、それらの意図された機能を果たすようにデザイン及び設計されていることを確保すべきである。 |
| I | 国家は、軍事用ＡＩ能力の安全性、セキュリティ、有効性が、その明確に定義された用途の範囲内で、そのライフサイクル全体にわたって、適切かつ厳格なテストと保証の対象となることを確保すべきである。自己学習または継続的に更新される軍事用ＡＩ能力については、国家は、モニタリングなどのプロセスを通じて、重要な安全機能が低下していないことを確保すべきである。 |

表１　ＡＩ政治宣言において参加国が実施すべき措置として定められている事項（続き）

| | |
|---|---|
| J | 国家は、軍事用ＡＩ能力における失敗のリスクを軽減するために、意図しない結果を検出し回避する能力、及び、そのようなシステムが意図しない挙動を示した場合に、例えば、配備されたシステムによる交戦を停止させる、又はその作動を停止させることによって対応する能力などの適切な保護措置を実施すべきである。 |

２　本ガイドラインの位置付け

ＡＩを搭載した無人機の実戦投入や、情報収集・分析へのＡＩ活用など、諸外国においても装備品へのＡＩ適用の具体的な事例が増加しており、防衛分野における責任あるＡＩ適用のための実効性のあるルール作りの必要性が日々高まっている。また、人間の関与が確保された自律性を有する兵器システムついては、ヒューマンエラーの減少、省人化といった安全保障上の意義を有するものの、ＡＩには、判断プロセスの不透明性、予期せぬ挙動の可能性、倫理的な問題など、従来の技術とは異なる特有の難しさが存在している。軍事・防衛分野における責任あるＡＩの適用方策については前項まで紹介してきたように諸外国においても慎重に検討が進められている一方で、前項に紹介したＡＩ政治宣言のように、具体的に実施すべき措置を示す枠組みも出てきている。このような流れを受け、基本方針では、装備品等の研究開発におけるガイドラインを策定する方針を明らかにした。

こうした背景を踏まえて、装備品等の研究開発における責任あるＡＩ適用ガイドライン（以下「本ガイドライン」という。）は、防衛省・自衛隊の一連のコミットメントを具体化するとともに、研究開発の文脈でＡＩのリスクを適切に管理し、リスク低減を図るための枠組みを提供するものである。これにより、研究開発に参画する事業者等における予見可能性を確保し、装備品等へのＡＩの利活用を一層推進することを企図している。図１に示す装備品等のライフサイクルの中で構想段階から研究・開発段階までの間の研究開発事業を本ガイドラインの対象範囲とする。本ガイドラインにおけるＡＩの定義は基本方針に準拠し、機械学習をするソフトウェアやプログラムを指すこととする。

なお、本ガイドラインは、ＡＩ技術を適用した装備品等（以下「ＡＩ装備品等」という。）に係る研究開発事業の計画立案や実施等に当たって実施者が準拠すべき指針として位置付けられるものである。

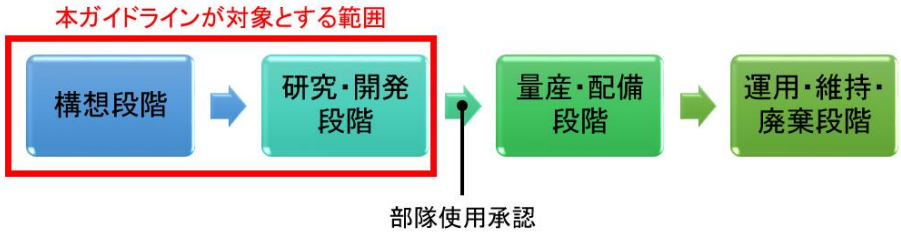


図１　本ガイドラインの対象範囲

３　ＡＩ装備品等の研究開発における確認事項

第１項で述べたＬＡＷＳ作業文書を踏まえ、ＡＩ適用のリスクへの実効性のある対策を確保した管理（以下「リスク管理」という。）を行うため、ＡＩ装備品等の研究開発において準拠すべき要件を以下のとおり設定し、研究開発事業の中で確認することとする。

(1)　準拠すべき要件の設定

我が国は、国際人道法の原則は、ＡＩ等の新興技術を活用するものを含め、あらゆる兵器に適用されるべきという立場である。また、防衛省・自衛隊として、当然のことながら国際法や国内法により使用が認められない装備品の研究開発及び導入を行うことはない。さらに、ＬＡＷＳの議論の中でも、我が国は、国際人道法に従って使用することができない兵器システムや人間の関与が及ばない完全自律型の致死性を有する兵器システムは許容できない結果をもたらす可能性があるとして、その開発、使用は国際的にも認められるべきではないとしている。そのため、研究開発するシステムの運用構想は、これに準拠すべきである。

これを踏まえ、ＡＩ装備品等の運用構想が法的・政策的観点から適正であるか（以下「要件Ａ」という。）確認し、その運用構想を技術的観点から満足しているか（以下「要件Ｂ」という。）を確認することとする。Ａ－１については、国際人道法を始めとする国際法及び国内法の要請を満たしているかという法的観点からの要件である。なお、Ａ－１の要件は、我が国は国際人道法の原則は、新興技術を活用するものを含め、あらゆる兵器に適用されるべきであり、国際法や国内法により使用が認められない装備品の研究開発を行うことはないとの立場であるところ、ＡＩ装備品等の研究開発に特有の要件ではないが、新興技術を活用した兵器に関する議論においては特に国際人道法を始めとする国際法の遵守が焦点の一つとなっていることから、本文書でも、確認的に要件の一つとして位置付けるものである。Ａ－２については、国際人道法も念頭に置いた我が国の政策的立場を反映した要件である。なお、Ａ－２については、あくまでも現時点での我が国の政策的立場を反映した要件であり、国際場裏における今後の議論の過程で我が国として依拠すべきと考える新たな考え方が形成される場合には、必要に応じてこれを見直すことも検討する。

また、責任あるＡＩ適用の議論の関係では、以下に示すＢ－１～Ｂ－７の要素から構成される要件Ｂについても準拠すべきである。要件Ｂは、構想段階で要件Ａを満たすものと認められた高リスクＡＩ装備品等（後述）の試作品につき、実際の配備・運用段階でも要件Ａを引き続き満たすことを可能にするための機能を備えているか、また適切なリスク軽減策が施されているかを技術的観点から確認するために導入する要件である。米国を始めとした諸外国においては責任あるＡＩの実施のために目指すべき方向性をＡＩ倫理原則という形で打ち出している。要件Ｂは図２に示すとおり、各国のＡＩ倫理原則との関係性からも国際的な潮流と整合していることが分かる。研究開発事業におけるリスク管理のためには、構

想段階から研究開発終了までの間、審査等の必要な結節において要件Ａ及び要件Ｂを継続的に確認することとする。

法的・政策的要件（要件Ａ）

Ａ－１　国際人道法を始めとする国際法及び国内法の遵守が確保できないものでないこと

その性質上過度の傷害又は無用の苦痛を与えるもの、本質的に無差別なもの、その他国際人道法に従って使用することができないものでないこと等

国内法に従った使用ができるものであること

Ａ－２　人間の関与が及ばない完全自律型致死兵器でないこと

適切なレベルの人間の判断が介在せず、人間による責任ある指揮命令系統の中での運用が確保できないような、人間の関与が及ばない完全自律型で致死性を有するものでないこと

技術的要件（要件Ｂ）

Ｂ－１　人間の責任の明確化

ＡＩシステムの利用に際して人間が適切な水準での注意を払い責任をもって対応できるよう、運用者の関与や運用者による適切な制御が可能な設計としていること

Ｂ－２　運用者の適切な理解の醸成

運用者がＡＩシステムを適切に使用できる仕組み、過度の依存を防ぐ対策、不具合発生時の運用者による改善の仕組みを設計していること

Ｂ－３　公平性の確保

バイアスの原因を探索、理解した上でＡＩシステムとデータセットに対して適切な軽減策を講じ、不当で有害なバイアスを局限化していること

Ｂ－４　検証可能性、透明性の確保

ＡＩシステムの設計手順、採用したアルゴリズム、学習に使用したデータ等、システム構築の過程を明確化し、ＡＩシステムの検証可能性と透明性を確保していること

Ｂ－５　信頼性、有効性の確保

ＡＩシステムの信頼性、有効性、セキュリティに関して、多様な観点から試験評価を行い、ライフサイクル全体にわたって、許容可能なレベルでの動作を確保していること

Ｂ－６　安全性の確保

ＡＩシステムの誤作動や深刻な失敗のリスク発生を低減できる仕組みを整備し、安全性を確保していること

Ｂ－７　国際法及び国内法の遵守が確保できないものでないこと

適用される国際法及び国内法を遵守した運用が可能な設計としていること

| | 各国の政策文書中で掲げられている理念 | | | |
|---|---|---|---|---|
| | アメリカ<br>(AI Principles) | オーストラリア<br>(A Method for Ethical AI in Defence) | イギリス<br>(AMBITIOUS, SAFE, RESPONSIBLE) | フランス<br>(Opinion on the Integration of Autonomy into Lethal Weapon Systems) |
| B-1 人間の責任の明確化 | Responsible | Responsibility | Responsibility | Command |
| B-2 運用者の適切な理解の醸成 | — | | Understanding | Competence |
| B-3 公平性の確保 | Equitable | — | Bias and Harm Mitigation | — |
| B-4 検証可能性、透明性の確保 | Traceable | Traceability | — | — |
| B-5 信頼性、有効性の確保 | Reliable | Trust | Reliability | Confidence |
| B-6 安全性の確保 | Governable | Governance | — | — |
| B-7 国際法及び国内法の遵守が確保できないものでないこと | — | Law | Human-Centric | Compliance |

図２　諸外国の軍事領域におけるＡＩ倫理原則

４　ＡＩ装備品等の研究開発における実施事項

以降で前項において設定した要件を確認する流れを示す。具体的には図３に示す「ＡＩ装備品等の分類」、「法的・政策的審査」及び「技術的審査」の３つの事項を実施することとする。審査要領の詳細及び体制等に係る制度等は別途定める。

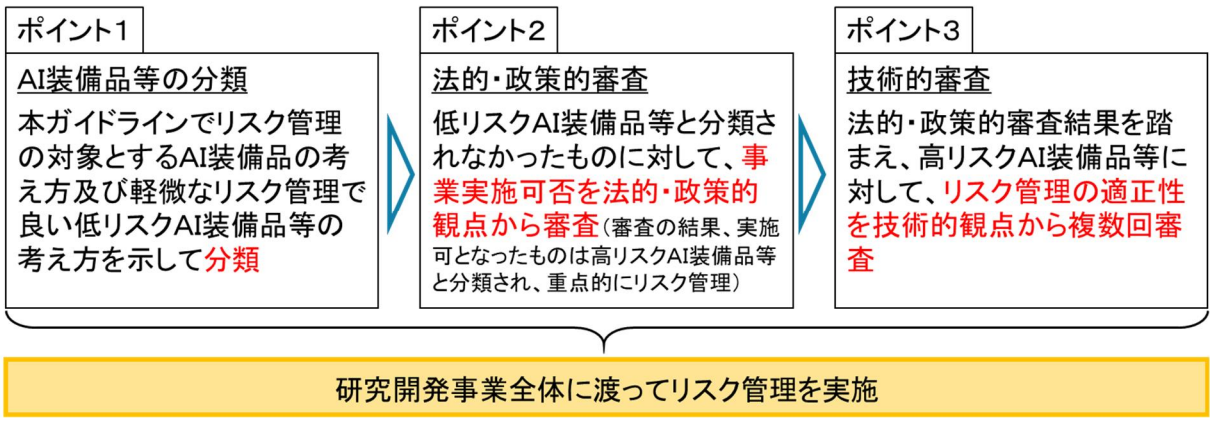


図３　研究開発事業における実施事項

(1)　ＡＩ装備品等の分類

ＡＩ装備品等の研究開発にあたっては、特に第１項で述べたＬＡＷＳの観点から重点的にリスク管理すべき装備品等を高リスクＡＩ装備品等、それ以外のＡＩ装備品等を低リスクＡＩ装備品等と分類し、装備品等の態様に応じた適切なリスク管理を実施することとする。

具体的には、図４のフローに従って装備品等を分類することとする。装備品等の研究開発には様々な種別が存在するが、本ガイドラインにおいては、原則としてＡＩ技術を適用した装備品等の試作事業を対象とする。その上で、ＡＩ機能に起因する出力がシステム内の破壊能力を有する機能に影響を与えない場合は低リスクＡＩ装備品等として分類する。それ以外の場合は法的・政策的審査（詳細後述）にかけ、不適格とされた場合は許容できない結果をもたらす可能性があるため、その時点で当該装備品等の研究開発の実施は取止めとする。適格とされた場合には高リスクＡＩ装備品等として分類し、技術的審査（詳細後述）を実施することにより重点的なリスク管理を行う。

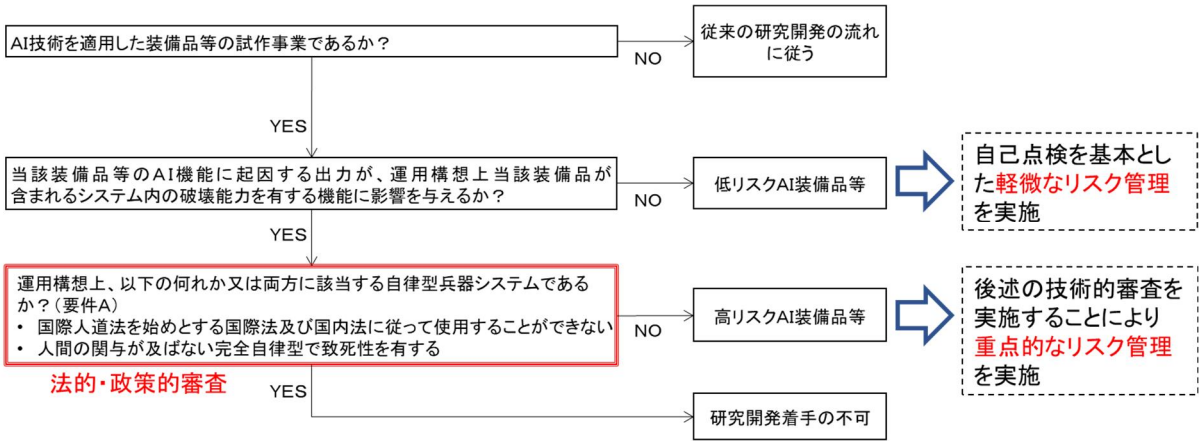


図４　装備品等の分類と対応フロー

(2)　法的・政策的審査

法的・政策的審査は研究開発対象のＡＩ装備品等が搭載する予定の機能及び想定される使用の態様等が前項の要件Ａに沿って適切なものになっているかを法的・政策的観点から当該ＡＩ装備品等の構想段階で審査するものである。審査は、前号の分類において低リスクＡＩ装備品等ではないと判断されたＡＩ装備品等を対象にして表２に示す審査項目について実施する。審査にあたっては国際人道法を始めとする国際法や国内法の遵守に係る方策の適正性の確認が必要なため、法律分野の専門家や専門部署を審査員に含めることとする。図５に示すとおり、法的・政策的審査は事業（研究開発事業を指す。）開始の可否を判断する重要な位置付けであるため、ＡＩ装備品等に関する総合的な意思決定を行う会議体にて事業実施担当を審査する形で実施することとする。審査を行う時点は事業開始前とする。

表２　要件Ａの審査項目（基準）

| | 確認項目 |
|---|---|
| Ａ－１<br>国際人道法を始めとする国際法及び国内法の遵守が確保できないものでないこと | ・　当該装備品等を使用するにあたって軍事的必要性と人道的配慮のバランスを考慮し、過度の傷害又は無用の苦痛を与えることを防止する措置が含まれた構想になっているか。<br>・　軍事目標と非軍事目標（文民及び民用物）を区別し、軍事目標のみに攻撃を行うことが可能な装備品等の構想になっているか（区別原則）。<br>・　予測される具体的かつ直接的な軍事的利益との比較において、巻き添えによる文民の死亡、文民の傷害、民用物の損傷又はこれらの複合した事態を過度に引き起こすことが予測される攻撃を防止する措置が含まれた構想になっているか（比例性原則）。<br>・　攻撃に先立つ軍事目標の選定から攻撃の実施に至るまで、無差別攻撃を防止し文民と民用物への被害を最小限に抑えるための各種予防措置を実施し得る構想になっているか（予防原則）。<br>・　国際人道法のその他の要請や適用のある他の国際法及び国内法に従って使用することができないような装備品等の構想となっていないか。 |
| Ａ－２<br>人間の関与が及ばない完全自律型致死兵器でないこと | ・　適切なレベルの人間の判断が介在し、人間による責任ある指揮命令系統の中での運用が確保できる構想となっているか。<br>・　人間の関与が及ばない完全自律型の致死性を有する兵器システムの開発ではないか。 |

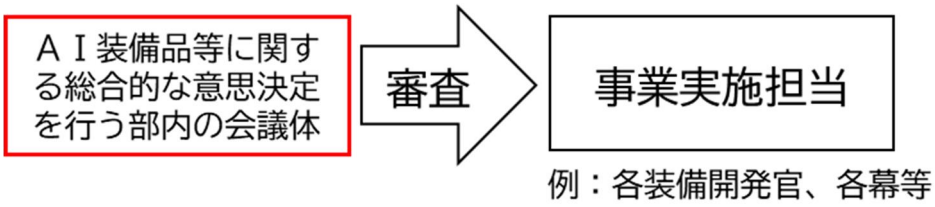


図５　法的・政策的審査体制イメージ

(3)　技術的審査

技術的審査は、高リスクＡＩ装備品等の試作品につき、実際の配備・運用段階でも要件Ａを引き続き満たすことを可能にするための機能を備えているか、また適切なリスク軽減策が施されているかを技術的観点から確認するために行う審査である。審査は表３に示す審査項目について実施する。図６に示すとおり、技術的審査はＡＩ技術に対して専門的な知識を持つ部内の者によって構成される会議体が事業実施担当を審査する形で実施することとする。審査にあたってはＡＩを搭載したシステムの安全性確保や品質管理についての高度な技術的知識が必要となる可能性もあるので、部外の専門家にも必要に応じて意見聴取する。審査を行う時点は事業開始前、設計承認前、試験開始前、運用開始前等の研究開発における結節とする。

なお、リスク管理における表３の確認に関する一例を別紙のとおり示す。

表３　要件Ｂの審査項目（基準）

| | 確認項目 |
|---|---|
| Ｂ－１<br>人間の責任の明確化 | ・　ＡＩシステムの利用に際して適切なタイミングと程度において、運用者の関与や運用者による適切な制御が可能となるように設計されているか。<br>・　ＡＩシステムと運用者それぞれの役割が明確に定義されているか。<br>・　運用者が負うべき責任が明確に定義されているか。 |
| Ｂ－２<br>運用者の適切な理解の醸成 | ・　ＡＩシステムの挙動、パフォーマンス範囲、操作等に運用者が習熟して当システムを適切に使用できるように設計されているか。<br>・　ＡＩシステムへの過度の依存を防ぐための対策が設計されているか。<br>・　ＡＩシステムのモニタリング時に、運用者が不具合を認めたときにそれを改善することを可能とする仕組みが設計されているか。 |
| Ｂ－３<br>公平性の確保 | ・　データセットの公平性要件が明確化されているか。<br>・　ＡＩモデルに関連するバイアスが許容レベルを超えないか確認するとともに、不具合が認められた場合の改善の仕組みが設計されているか。 |
| Ｂ－４<br>検証可能性、透明性の確保 | ・　ＡＩシステムの構築に係る過程、使用した手法・データ・アルゴリズム等が明確化され、その妥当性を後に検証可能な仕組みが整備されているか。<br>・　研究開発体制（事業者含む。）において、説明責任を果たす責任者を明確化しているか。 |

表３　要件Ｂの審査項目（基準）（続き）

| | 確認項目 |
|---|---|
| Ｂ－５<br>信頼性、有効性の確保 | ・　ＡＩシステムの信頼性を確保するために、多様な指標を用いた試験評価を実施しているか。<br>・　開発初期から研究開発終了までの全ての期間において、運用を想定したデータセットを使用することが検討されているか。<br>・　運用を想定して、ＡＩに連接する装備品等の信頼性を損なうことなく、ＡＩを適用できることを設計において担保されているか。<br>・　ＡＩシステムが容易に維持管理できる設計となっているか。<br>・　ＡＩシステムに対してセキュリティ管理策の実施が検討されているか。 |
| Ｂ－６<br>安全性の確保 | ・　ＡＩシステムの誤作動や深刻な失敗・事故の発生を低減する安全機構が設計されているか。 |
| Ｂ－７<br>国際法及び国内法の遵守が確保できないものでないこと | ・　装備品等の運用に際して適用される国際法や国内法令、各種内部規則等から逸脱しないよう対策が検討されているか。 |

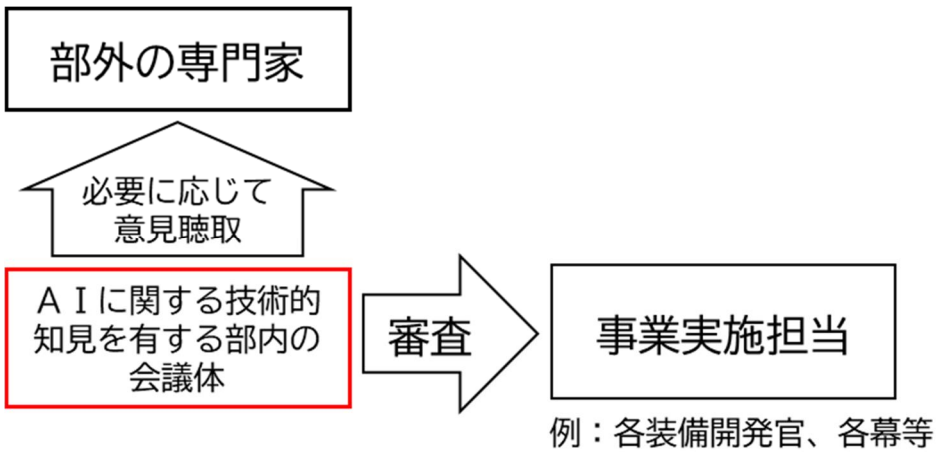


図６　技術的審査体制イメージ

第１号～第３号を踏まえ、高リスクＡＩ装備品等の標準的なリスク管理イメージを図７のとおり示す。

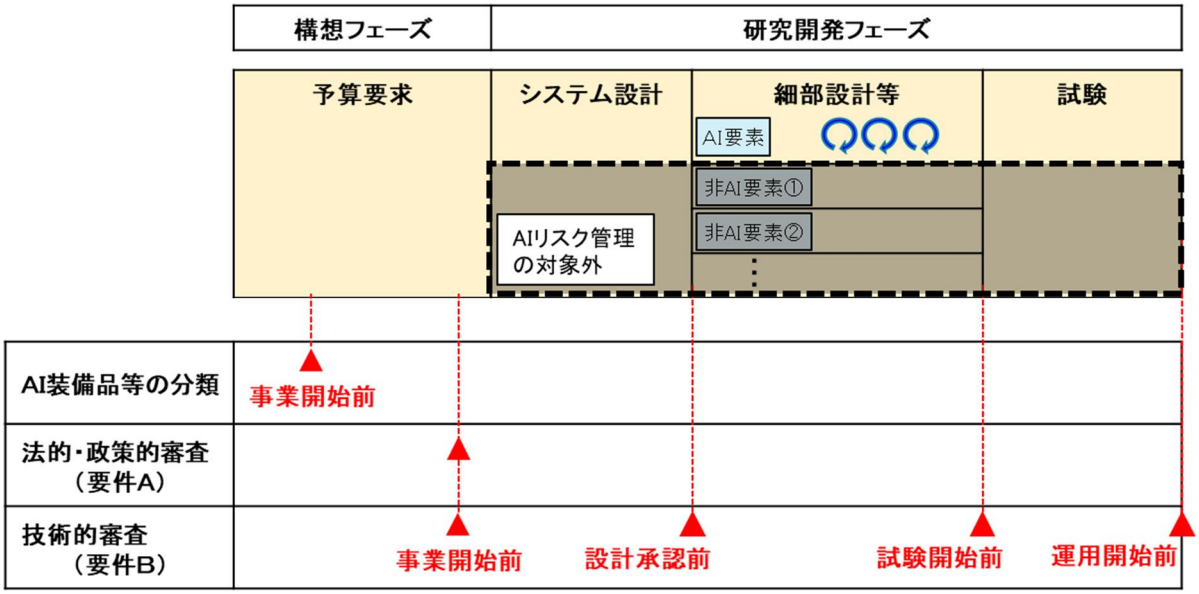


図７　高リスクＡＩ装備品等の標準的なリスク管理イメージ

(4)　その他

ア　自己点検

第１号において低リスクＡＩ装備品等と分類されたものについては、第２号に掲げる法的・政策的審査及び第３号に掲げる技術的審査については不要とし、事業実施担当による自己点検を基本としたリスク管理を行うこととする。具体的には従前のリスク管理と同様の進捗会議等の適宜の機会において表３に掲げる要件を確認する形で行う。図８に低リスクＡＩ装備品等のリスク管理イメージを示す。

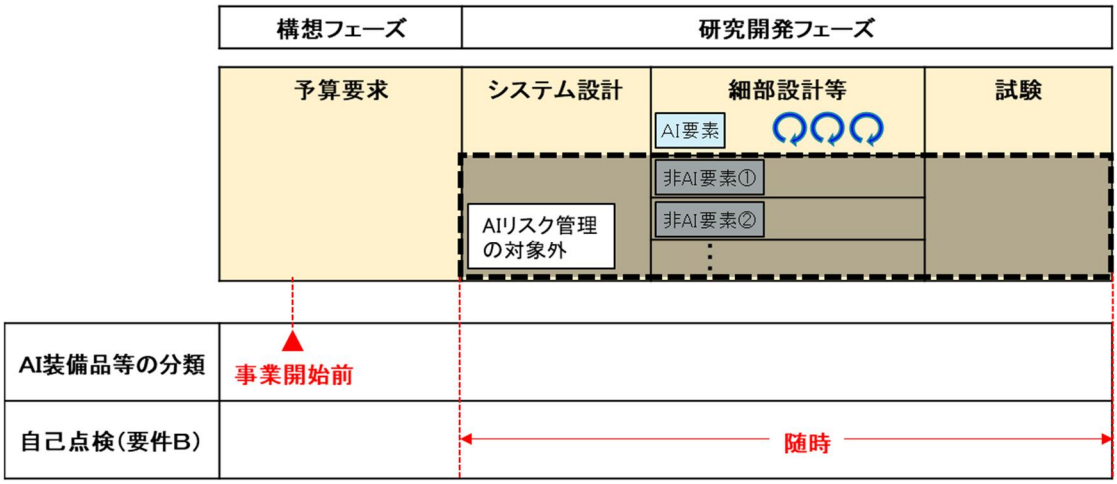


図８　低リスクＡＩ装備品等の標準的なリスク管理イメージ

イ　設計・試験評価への運用者の関与

ＡＩは、学習に用いられるデータに性能が大きく影響される。また、ＡＩの性能はどのような環境でどのようにそれを使用するかにも影響される。そのため、様々な運用場面を想定してＡＩの試験評価を行う必要がある。したがって、防衛装備庁と各自衛隊等は構想段階から密接な調整を行い、各幕の関係部署や将来の利用者と想定している部隊、ＡＩの研究開発に必要なデータを提供する部隊、将来維持管理を担うことが予期されている部隊等を関与させて研究開発を実施することが有効である。また、試作品等を部隊等に試験的に配備し、部隊等において試作品等を試験的に運用しながら運用上の評価を実施し、その評価結果を受けて試作品等の改善又は能力向上の措置を行うというサイクルを繰り返し実施することを可能とする枠組みを構築して研究開発を実施することも有効である。この中で、部隊等において試験的に運用するフェーズを設け運用上の評価を実施し、部隊等で維持管理可能な水準まで十分に検証した上でリスク管理を部隊等に引き継ぐ。このように、設計・試験評価等について研究開発の可能な限り初期の段階から運用者を関与させることが有効である。

## 5　まとめ

本ガイドラインは、装備品等の研究開発における責任あるＡＩ適用の具体的な指針を提示するものであるが、実際の装備品等の研究開発には様々な形態があり得るため、本ガイドラインで示した考え方を指針として、各事業に応じて柔軟に事業設計を行うこととする。ＡＩをとりまく状況は急速に変化しており、一度決めたルールや手続きを維持し続けることが必ずしも適切でない場合がある。したがって、本ガイドラインはLiving Documentとし、適宜更新を行うことを予定している。

なお、本ガイドライン公表時点において、既に着手している研究開発事業については、本ガイドラインの趣旨を踏まえ、研究開発の進捗状況を勘案して、個別の対応を検討する。

# RAI Toolkit を用いた要件Ｂの確認の一例

## １　目的

ＡＩ技術を適用した装備品等に係る研究開発事業の計画立案や実施等に当たって実施者が参考情報として活用することを目的として、RAI Toolkit を用いた要件Ｂの確認の一例を示すもの。

## ２　RAI Toolkit について

米国防総省（Chief Digital and Artificial Intelligence Office（ＣＤＡＯ））が公開しているRAI Toolkit は、ＡＩ開発から運用においてリスク管理や評価など一連のプロセスに有効活用可能なツールであり、ツールの大半は業界標準のオープンソースオプションである。また、RAI Toolkit は、ＡＩプロジェクトの実施者が、責任あるＡＩ（以下「ＲＡＩ」という。）のプロダクトライフサイクル全体を通じて、ＲＡＩに関連する問題を特定、追跡、軽減する（及びＲＡＩ関連のイノベーションの機会を活用する）ことができる一元化されたプロセスを提供する。

## ３　RAI Toolkit を用いた要件Ｂの確認の一例

要件Ｂの確認に活用し得るRAI Toolkit の一例を付表のとおり示す。

## ４　参考資料

(1)　RAI Toolkit

(2)　Responsible AI Toolkit ガイド＆フロントマター（２０２３年１０月２４日）

表　要件Ｂのチェック項目に活用し得る RAI Toolkit の例（１／５）

| No. | RAI Toolkit | | | 該当するチェック項目 |
|---|---|---|---|---|
| | ツール名称 | ツール説明 | | |
| 1 | Trust in Autonomous Systems Test (TOAST) | システムに対する信頼度の評価 | 人間がシステムをどの程度信頼しているかを測定するための９つの質問からなるテスト。(例：私はシステムが何をすべきか理解している。私はシステムの限界を理解している。等) https://www.ida.org/-/media/feature/publications/p/pr/predicting-trust-in-automated-systems---validation-of-the-trust-of-automated-systems-test---toast/d-33088.ashx | Ｂ－２<br>運用者の適切な理解の醸成 |
| 2 | Human-Machine Teaming Systems Engineering Guide | システムの設計支援 | システム開発者が人間のオペレーターと協力して機能するＡＩを設計するのを支援するガイド。 https://www.mitre.org/sites/default/files/2021-11/prs-17-4208-human-machine-teaming-systems-engineering-guide.pdf | |
| 3 | FairML | ＡＩモデルの公平性の診断・改善 | ＡＩモデルの予測結果に対して、性別や人種などの属性との関係を分析し、公平性を診断し、要因の特定、改善方法の提供を行うためのツールキット。ＡＩモデルの予測結果の公平性を改善するために用いられる。 https://github.com/adebayoj/fairml | Ｂ－３<br>公平性の確保 |
| 4 | Tensor Flow Fairness Indicators | ＡＩモデルの公平性の評価・改善 | ＡＩモデルの公平性に関する懸念を評価、改善、比較するためのツールキット。「公平性」とは、ＡＩモデルが特定の属性(人種、性別、年齢など)に基づいて不当な差別をしないことを意味する。 https://github.com/tensorflow/fairness-indicators | |

表　要件Ｂのチェック項目に活用し得る RAI Toolkit の例（２／５）

| No. | RAI Toolkit | | | 該当するチェック項目 |
|---|---|---|---|---|
| | ツール名称 | ツール説明 | | |
| 5 | What-If Tool | ＡＩモデル性能と公平性の評価 | What-If ツール（WIT）は、機械学習（ML）モデル、特に分類や回帰タスクにおいてブラックボックスシステムとして機能するモデルの理解と分析を支援する可視化ツール。<br>https://github.com/PAIR-code/what-if-tool<br>https://pair-code.github.io/what-if-tool | Ｂ－３<br>公平性の確保 |
| 6 | Explainer Dashboard | ＡＩモデルの解釈性向上 | ＡＩモデルの解釈性を高めるためのツールを提供するライブラリ。モデルの予測結果を視覚的に分析したり、特徴量の重要度を確認したりすることが可能。<br>https://github.com/oegedijk/explainerdashboard<br>https://explainerdashboard.readthedocs.io/en/latest/ | Ｂ－４<br>検証可能性、透明性の確保 |
| 7 | InterpretML | ＡＩモデルの解釈性向上 | 機械学習の解釈可能性のための最新技術を一元的に包含するオープンソースのツール。このツールは、解釈可能なモデルを作成し、ブラックボックスシステムを説明する機能を提供する。モデルの全体的な振る舞いを理解したり、個々の予測の背後にある理由を理解したりすることが可能。<br>https://github.com/interpretml/interpret/<br>https://interpret.ml/docs/index.html | |

要件Ｂのチェック項目に活用し得るRAI Toolkitの例（３／５）

| No. | RAI Toolkit | | | 該当するチェック項目 |
|---|---|---|---|---|
| | ツール名称 | ツール説明 | | |
| 8 | Hugging Face Data Card Template | データセットの透明性確保 | データセットカードの作成手順を示すもの。データセットの内容、データセットを利用する背景、データセットの作成方法、その他利用者が注意すべき点を理解するのに役立つ。責任ある利用を促し、データセットに潜在する偏りを利用者に知らせることができる。<br>https://huggingface.co/docs/datasets/dataset_card | Ｂ－４<br>検証可能性、透明性の確保 |
| 9 | Hugging Face Model Card Template | ＡＩモデルの透明性確保 | ＡＩモデルを理解し、共有し、改善するためのフレームワーク。各モデルについて、その技術と設計方法を記述し、透明性・監査可能性を実現するのに役立つ。また、技術の適切な理解を示すために、設計手順が透明で監査可能であることを測定し示すのにも役立つ。<br>https://huggingface.co/blog/model-cards | |
| 10 | Threat Modeling Resource | 脅威や脆弱性の特定・評価および対策の計画 | ＡＩの脅威モデリングのためのフレームワーク。ＡＩ機能のセキュリティを実現するために、脅威モデリングによってセキュリティレビューを実施し、それに基づいて推奨されるリスク軽減策を取り入れる。<br>https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml?source=recommendations | Ｂ－５<br>信頼性、有効性の確保 |

要件Ｂのチェック項目に活用し得る RAI Toolkit の例（４／５）

| No. | RAI Toolkit | | | 該当するチェック項目 |
|---|---|---|---|---|
| | ツール名称 | ツール説明 | | |
| 11 | EQUI(NE2) | ニューラルネットワークの不確実性を定量化 | モデルの予測における不確実性を可視化するツール。各予測の「信頼スコア（confidence score)」と「外れ値スコア（outlier score)」を提示し、モデルが通常のデータ範囲を超えた状況においてどの程度リスクが高まるかを評価することができる。モデルの精度を評価することだけでなく、データの偏りに対するモデルの頑健性を評価し、ＡＩの適用可能な範囲を明確化することができる。<br>https://github.com/mit-ll-responsible-ai/equine | Ｂ－５<br>信頼性、有効性の確保 |
| 12 | IBM Adversarial Robustness 360Attacks | ＡＩのセキュリティ強化 | 敵対的な攻撃を生成し、ＡＩの信頼性(ＡＩのセキュリティ)を高めるための堅牢性を訓練したり、攻撃成功率を計算してＡＩの信頼性(ＡＩのセキュリティ)を測定し、示したりすることができるＡＩのセキュリティを守るための Python ライブラリ。ユーザーが機械学習モデルやアプリケーションを、回避、汚染、抽出、推論といった敵対的な脅威から守り、評価するためのツールを提供する。TensorFlow、Keras、PyTorch、scikit-learn、XGBoost などのＡＩフレームワークをサポートしており、画像、表、音声、ビデオなどあらゆるデータタイプや、分類、物体検出、音声認識、生成、認証などのＡＩタスクに対応。<br>https://github.com/Trusted-AI/adversarial-robustness-toolbox/tree/main/art/attacks<br>https://github.com/Trusted-AI/adversarial-robustness-toolbox/wiki/ART-Attacks | |

RAI Toolkit と要件Ｂのチェック項目との対応（５／５）

| No. | RAI Toolkit | | | 該当するチェック項目 |
|---|---|---|---|---|
| | ツール名称 | ツール説明 | | |
| 13 | Drift Tools | ＡＩ機能の信頼性とガバナンスの支援 | Drift Tools のアルゴリズムを ＡＩ機能に組み込んで、ＡＩ機能のパフォーマンスが保証されていない分布外の入力を検出し、ＡＩの信頼性(有効性)と意図しない結果を検出する能力の実現を支援する Python ライブラリ。外れ値、敵対的攻撃の検出(Adversarial detection：機械学習モデルに対する攻撃を検出し、防御するプロセス)、及びドリフト検出(Drift detection：機械学習モデルの学習データの統計的分布と、実際に遭遇するデータの分布との間に生じる変化を検出するプロセス)に特化している。<br>https://github.com/SeldonIO/alibi-detect<br>https://docs.seldon.io/projects/alibi-detect/en/stable/ | Ｂ－２<br>運用者の適切な理解の醸成<br>Ｂ－５<br>信頼性、有効性の確保 |
| 14 | Python Outlier Detection (PyOD) | 多変量データにおける異常値の検出 | PyOD (Python Outlier Detection) は、多変量データにおける異常値を検出するための包括的かつスケーラブルな Python ライブラリ。このライブラリは、小規模プロジェクトから大規模データセットまで、さまざまなニーズに対応するための幅広いアルゴリズムを提供する。<br>https://github.com/yzhao062/pyod<br>https://pyod.readthedocs.io/en/latest/ | Ｂ－６<br>安全性の確保 |

【解説】

Ｂ－２の「ＡＩシステムのモニタリング時に、運用者が不具合を認めたときにそれを改善することを可能とする仕組みが設計されているか。」は、ＡＩシステムの場合、実環境での使用に伴って運用者によって新たに不具合が発見されるリスクに対応するものである。RAI Toolkit の Alibi Detect にある Drift Detection 機能は、ＡＩのパフォーマンスを維持するためにモニタリングするのに役立つオープンソースのツールである。その他、運用者と協力して機能するＡＩを設計するのに役立つガイドであるＭＩＴＲＥの Human-Machine Teaming Systems Engineering Guide や人間がシステムをどの程度信頼しているかを判定するための９つの質問が記されている Institute for Defense Analyses（ＩＤＡ）の Trust in Autonomous Systems Test（ＴＯＡＳＴ）、米国防総省のＣＤＡＯの Human Systems Integration (HSI) T&E※1等の文献を参考にすることができる。

Ｂ－３の公平性の確保において、RAI Toolkit の What-If Tool (WIT) は、データの異なるサブセットにわたるモデルの性能と公平性を調査するのに役立つオープンソースの可視化ツールである。他にも RAI Toolkit の FairML や Tensor Flow Fairness Indicators は公平性の診断・改善に役立つ。

Ｂ－４の透明性の確保においては、データセットの内容、作成方法、その他利用者が注意すべき点等を理解するのに役立つテンプレートである Data Card やモデルの理解、共有、改善に役立つテンプレートである Model Card の使用が挙げられる。また、RAI Toolkit の InterpretML は、機械学習の解釈可能性のための最新技術を一元的に包含するオープンソースのツールである。モデルの全体的な振る舞いを理解したり、個々の予測の背後にある理由を理解したりするのに役立つ。同様に、RAI Toolkit の Explainer Dashboard というライブラリもモデルの予測結果を分析したり特徴量の重要度を確認したりすることで解釈性を高めるのに役立つ。

※１：https://cdao.pages.jatic.net/public/guidance/human-systems-integration/

Ｂ－５にある多様な指標として、最も直感的な指標である Correctness（正解率等）に加え、Bias（学習データの選択や偏り）、Resilience（機能回復）、Uncertainty（正解率に対する不確実性、確信度等）、Representativeness（運用環境を想定した学習データの代表性等）、Latency（処理速度）、Explainability（説明可能性）、Robustness（敵対的攻撃への耐性やノイズへの耐性等）、Drift（データやモデルのドリフト）という観点を米国防総省のＣＤＡＯが AI Model T&E※２の文献で提示している。RAI Toolkit の EQUI(NE2)は、ニューラルネットワークの出力の精度だけでなく不確実性も定量化することが可能なオープンソースのツールである。Ｂ－５の「運用を想定したデータセットを開発初期から使用し評価を行うとともに、運用段階おける再学習の影響等について検討されているか。」と「運用を想定して、ＡＩに連接する装備品の信頼性を損なうことなく、ＡＩを適用できることを設計において担保されているか。」は、米国防総省のＣＤＡＯの Operational T&E（OT&E）※３と Systems Integration（SI）T&E※４の文献を参考にしたものである。RAI Toolkit の Adversarial Robustness Toolbox（ART）は、回避（Evasion)、汚染（Poisoning)、抽出（Extraction)、推論（Inference）といった敵対的な脅威から守り、評価するためのオープンソースのセキュリティツールであり、セキュリティ管理策の実施に役立つ。

Ｂ－６の安全性の確保では、異常値を検出するための幅広いアルゴリズムを提供する Python ライブラリ PyOD（Python Outlier Detection）等が安全機構の設計に役立つ。

要件Ｂのチェック項目を議論の枠組みとし、その回答結果に応じた適切なリスク対応を検討した上で研究開発を進めていくことにより、必要なプロセスに抜け漏れがないか、ＡＩシステムの構築において潜在的な問題が存在しないかの確認を行いやすくなる。このとき、ＡＩシステムの研究開発の過去の失敗事例や新しい技術の情報に精通しＡＩの特性や脆弱性を十分に理解した有識者にリスク評価に関する意見を聴取することで効果的なリスク管理が可能となる。RAI Toolkit の AI Incidents Database 等のデータベースを参考にして、過去の経験からリスクの特定と対応策についての示唆を得ることもできる。

※２：https://cdao.pages.jatic.net/public/guidance/model/
※３：https://cdao.pages.jatic.net/public/guidance/operational/
※４：https://cdao.pages.jatic.net/public/guidance/systems-integration/