講演７．AIとシステムセーフティとの関係

－規格適合性において何が課題か－

鉄道認証室　　※森　崇

１．はじめに

鉄道分野においても、技術開発において、AIによる検知をもとにした安全関連装置の開発が進められつつある。例えば、踏切における滞留者検知装置や、ホームからの転落を検知する装置など、実用化事例から現在試験中のものまで、各種装置が提案されている。

本稿においては、国際規格上安全についてどのようなことが要求され、その要求とAI側からのリスクマネジメントのガイドラインを対比し、どのようにAIを活用できるか、その可能性についてまとめることを目的とする。

２．国際規格から見たAI

まずこの章では、機能安全国際規格及び鉄道関係の国際規格が、AIについてどのように要求を行っているか概説する。

２．１．機能安全規格における安全の考え方

機能安全規格においては、規格ごとに段階の細分化は少し異なっているが、図１のようなアプローチが一般的である。

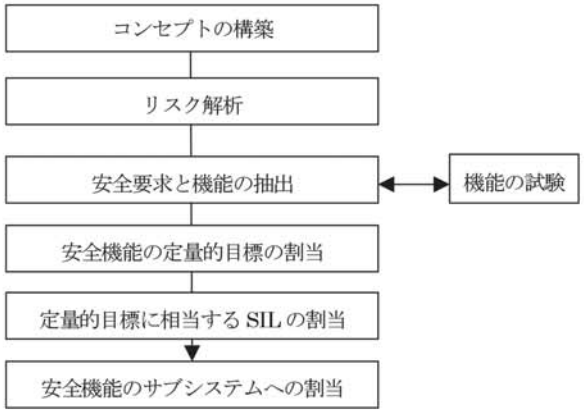

図１　一般的な機能安全における設計の進め方

ここで重視されるのは、上位の段階の事項が下位にもれなく引き継がれるか(フォーワードトレース)、また下位の段階で決定した内容が、上位の段階から入力されたものであるか(バックトレース)どうかである。フォーワードトレースは、要求が実装にもれなく展開されることを担保し、バックトレースは、上位仕様の漏れを検出する。安全機能において非常に重要な事項とされており、仕様については、明確に動作定義が要求されており、決定論的な状態遷移が前提となっている。学習によって状態遷移確率が変化し、改善性が高く状態遷移が決定するような考え方には立っていない。

図２　駅構内の進路設定

例えば、駅構内の信号設備設計を行うことを考える。信号装置は、列車の競合を防止し、衝突及び追突を防ぐという機能がある。競合するか、競合しないかは、列車の建築限界、線路の配置、列車の停止位置、列車進路により決定できる。これを網羅的に決定論的に定義したものが、「連動図表」といわれるもので、この定義表により、駅構内の信号装置の動作定義を行う。

また、列車の建築限界、線路の配置、列車の停止位置、列車進路を仮想的に構築し、数多くの列車走行試行を行うことで、競合するか否かを学習し、システムを構築することもできる。但しこの方法は、試行が網羅的であるかどうかが担保できればこの方法も採用できるが、そもそも試行が網羅的であるとの判断がで

きる場合は、設計上最初から網羅的に定義すればよいはずなので、鉄道信号における試行と学習による保安装置構築は一般的ではなかった現状がある。

但し、画像処理による安全確保など、網羅的な試験を計画できないような分野には、学習におけるアルゴリズム構築のニーズがあることは確かである。

２．２．機能安全規格である IEC 61508

IEC 61508 (Functional safety of electrical/electronic / programmable electronic safety-related systems)においては、Part 3 -Software requirements 及び Part 7-Overview of techniques and measures に AI に関する記述が存在する。当該規格において、Artificial intelligence fault correction の是非について述べられており、安全性インテグリティ(SIL)1 以外の安全性を要求される場合は、Not Recommended (NR)となっている。NR とされた技術を使用する場合、規格に付属している付録資料を参照して、仕様の根拠を詳細に説明する必要がある。また、ソフトウェアアーキテクチャ設計の段階において、ソフトウェア安全仕様についての正確性、単純でわかりやすい構築、動作に対する想定のしやすさ、設計に対する試験と確認については、実現が困難ではないかという表明が IEC 61508-3 Table C.2 でなされている。

２．３．鉄道保安装置ソフトウェア規格 IEC 62279

IEC 62279 においては、IEC 61508 と同様に、Artificial intelligence fault correction の是非について述べられており、ソフトウェア安全性インテグリティ(SSIL)0 以外の安全性を要求される場合は、Not Recommended (NR)となっている。Annex D.1 に解説があるが、故障予測、故障修正、保守とその行為支援について例示されている。安全にかかわる機能について NR である点からして、この規格において、稼働率の向上及び保守支援についての活用は許容されているが、他の目的においては、適切な正当化が必要であると考えられる。

３．AI 推進側から見た安全

次に AI を活用、推進することを研究している側からどのような見解が示されているかを述べる。今回はその研究成果のうち、現在最新であると思われる、Trustworthy & Responsible AI Resource Center, The National Institute of Standards and Technology (NIST)の AI Risk Management Framework (AI RMF)及びその解説である RMF Playbook をもとにまとめていく。

３．１．AI 技術者側から見た既存ソフトウェアとの違い

AI RMF の Appendix B において、伝統的なソフトウェアと AI のリスクの差異が述べられている。この項目で、鉄道をはじめとする安全関連系に適用した場合、関係の深いと思われるものを挙げておく。

表 1 AI RMF に挙げられているリスク差異の概要

リスク概要	具体的な鉄道への影響
用意したデータの品質の問題による安全性の影響	学習させるデータの品質の見極めと安全性が直結する
学習の複雑さの増加と再現性の問題	不具合が起こった場合、再現試験が求められるが、再現されるとは限らない
状況の変化による学習の正当性の欠如	線路切り替え、車種の変更など学習をどの程度生かせるか不明
決定ポイントの複雑性による説明の難しさ	Validation についてどのような方法で行うか決定する必要あり
事前学習をされている場合の影響の不確実性	Pre-existing software のルールを準用できにくい
試験方法の標準化不足	規格上試験のカバレッジと要求とのトレースが求められる
計算リソースが多くなることによる環境影響	CO2 排出の増大

この表で問題とされていることは、大きく区分すると 5 つあると考えられる。学習データの問題、説明の難しさ、Pre-existing data の活用、試験方法の確立、改修方法の確立である。換言すれば、Verification と Validation およびトレーサビリティの確保をどのように行えばよいか、また Generic Application software における役割と、Application data における役割の明確な分離が規格上では求められるのであるが、その観点を AI としてどのように作りこむか、Generic Application に相当する部分をどのように Pre-existing として設定し、学習させるかどうかの観点を問われているように私は考えている。

3．2．AI 技術者側から見たリスクマネジメントの切りロ

AI RMF において、AI RMF Core として 4 つの機能が定義されている。Govern、Map、Measure そして Manage である。

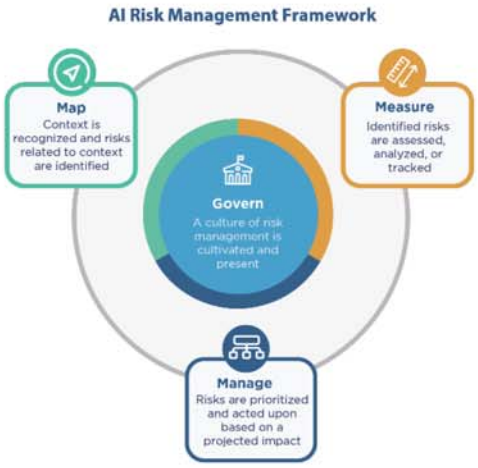

図 3　NIST RMF Core の概念図

4 つを非常に簡単に要約すれば、Govern は、組織やルールにおける機能を定め、Map は対象システムの目指すものや背景について、Measure はリスク評価方法、Test, Evaluation, Validation and Verification(TEVV)の手法、リスクの同定などがあげられ、Manage はリスク対策と文書化、第三者によるモニタリングがあげられている。

それほど安全におけるリスクマネジメントと考え方に大きな差異はないと思われるが、しかしながらある程度スタイルが定まっている機能安全のリスクアセスメントやマネジメントとは異なり、現状においてどのような方法で test や V&V を実施するかについては、今後規格においても記述されることが期待される。

3．3．3．安全にかかわる RMF Playbook の対応

AI RMF において定義された 4 つの要求は、カテゴリーに分けられ、さらにサブカテゴリー化されている。

その解説として、AI RMF Playbook が発行されており、AI RMF Playbook において、System safety についてはどのような観点で記述が行われているのかを調査する。

サブカテゴリーの中で、Safety という言葉が記載されている項目は全体で 72 項目あるうちの 9 項目であり、TEVV についての記載されている項目は 16 項目である。

この中で、鉄道保安システムで、関係が深そうな事象が、どのように AI RMF Playbook において紹介されているかを概説する。また、出来る限り鉄道での国際規格ではどのような方針であるかを比較する。

1) 他の組織や会社が作成したデータやシステムについて

AI のようなシステムの場合、鉄道保安システムを製作している会社が、一からすべてソフトウェアを構築することは現実的ではないと考える。

AI RMF Playbook Govern 6 及び Manage 3.1 には、他社のソフトウェアを使用するポリシーの構築、評価のポリシーの確立、学習データやアルゴリズム、インターフェース及び制限事項の透明化が求められており、完全なマニュアルの整備、また安全にかかわるものにおいての故障時の冗長構成の考慮などがあげられている。

これは、鉄道制御・保安システムソフトウェア規格 IEC 62279 においても、Pre-existing software を活用する際に、ほぼ同じ事項が 7.3.4.7 項に定められている。満たすべき要求、制約条件およびインターフェースの明確化、試験方法の確立、SIL3,4 の機能を含む場合の故障時の状態の明確化と対応方針など、大きな差異はない。しかしながら、現実的に AI を活用した場合の、インターフェースの明確化、評価ポリシーの具体化などは課題がある。

2) 運用中に安全度は変化するかもしれない

機能安全のシステムは、仕様から実装、試験に至るまでのトレースを取ることにより、動作定義とその実現の担保を行う強い原則がある。このため、仕様を変えない限りは、実装が変わることはないということで、安全を担保している。

しかし AI においては運用中の学習において、状態遷移が変化するという今までのシステムにはない特徴を備えている。人間に当てはめると、トライアンドエラーで成長し、その時々の経験でより適切な選択肢を選択するということになる

が、そのあいまいさを防御する安全関連システムが、柔軟であってもよいかという根源的な疑念はある。但し、今後の少子高齢化を見越した時には、限定された条件での決定論的制御だけではなく、数多くの状態を入力し、出来るだけ適切な制御を行うということが避けられないかもしれない。

AI RMF Playbook Manage 2.2 には、入力されるデータの管理及びルール化、出力の制限（おかしい出力は抑止するなど）、ライフサイクル全体の TEVV の実施など、変化に対応するようなルールを構築する必要性が生じる。鉄道関係規格において、安全性については、メンテナンスが規定通り行われ、改修がない場合は変化しないという立場で、RAMS ライフサイクルにおいてもそのような考え方であり、まったく考え方が異なる。

3) データと試験について

鉄道の安全関連系のシステムにおいては、網羅的な試験を行うため、ソフトウェアコンポーネントの段階での試験カバレッジを重視し、次にソフトウェアを実機で動作させた際には、アプリケーションデータとの組み合わせ、外部インターフェースとの試験、パフォーマンス試験などを行うことが通例であり、IEC 62279 の 7 章において考え方が述べられている。

AI の場合において、このような考え方で試験を遂行するには、大きな障壁がある。動作定義は学習で行われていき、従来のコンポーネントテストからの積み上げ方式とは全く異なる TEVV が必要である。AI RMF Playbook Map 2.3 においてその観点での記述があるが、注意すべき項目として、TEVV の方法をサブコンポーネント、設置時、運用時各々決定することや、テストモジュールの特定、独立した評価者の検証などがあげられている。これは今までの V&V の方法とかなり異なっており、AI 技術者と鉄道技術者が協力して実施する必要があると考える。

4) 説明責任等について

鉄道事業者は、アクシデントがあった際、その内容について説明を求められる。これは単に「説明責任」とされているが、AI RMF Playbook Measure 2.9 において、通常説明責任と言われている事項は Transparency(透明性), explainability(説明可能性), Interpretability(解釈可能性)の 3 つの要素があるとされている。

透明性は、何が起こったかの観点であり、説明可能性は、どのように決定がなされたのか、解釈可能性は決定が行われた理由とその意味付けについて返答できることとされている。

このような観点は非常に大事であるが、これは制御装置だけではなく、監査ログの重要性も非常に問われているように考える。

5) 運輸業界について

AI RMF Playbook Measure 2.6 に重篤性の高いものとしての筆頭に運輸があげられている。網羅的な試験は行えない場合も数多くあることから、シナリオの工夫による十分な試験、余裕を持ったシステムパフォーマンス、改善の継続と運用中における試験などが求められている。この分野でも研究が各種あるようであるので、注目していきたい。

4．おわりに

今回国際規格での要求と AI の活用の指針との比較を行うことで、どのような点にギャップがあり、注力しなければならない点を紹介した。鉄道認証室においても、新しい技術に果敢にチャレンジしていきたい。

参考文献

1) NIST, Artificial Intelligence Risk Management, Framework (AI RMF 1.0) (2024)
2) NIST, AI RMF Playbook (2023)
3) IEC, IEC 62279, Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems (2015)
4) IEC, IEC 61508-3, Functional safety of electrical/electronic/programmable electronic safety-related systems –Part 3: Software requirements