【限定なし】

CAA-2024038

2024年度 宇宙航空安全・ミッション保証シンポジウム

～宇宙機のソフトウェア品質保証に係る活動状況と今後の取り組み～
「自律化技術・AI」×「Assurance(アシュアランス)」

開催結果報告

2025年2月6日

宇宙航空研究開発機構
安全・信頼性推進部

2024年度 宇宙航空安全・ミッション保証（S&MA）シンポジウム

1. 概要

- テーマ：「自律化技術・AI」×「Assurance（アシュアランス）」
- 日程：2025年1月15日（水）13:00 – 17:45
- 場所：御茶ノ水ソラシティカンファレンスセンター＋オンライン
- 参加登録者数(講演者・事務局含む)：311名（対面77名／オンライン234名、詳細はp.13参照）
- 参加者数実績(講演者・事務局含む)：239名（対面72名／オンライン167名、詳細はp.13参照）

2. 背景

- 特殊な環境に置かれる宇宙機は、独自の自律化技術・データ処理/制御技術の進化を遂げており、今後は月面着陸等、更に高度なAI技術を利用したシステム開発が計画されている。
- 他産業、例えば自動車業界では衝突回避等の運転支援システムなどの実用化が進んでおり、この分野にもAI技術が導入されている。
- これらのAI技術は、その品質確保や検証方法に困難さが伴っており、それぞれの工夫により、目標が達成されている。

3. 趣旨

本シンポジウムでは「ソフトウェア品質保証」に着目し、システムへの高度な自律化・AI技術の導入の動機・ニーズを参加者と共有し、更にシステムに組み込む際の開発や信頼性・安全性確保に係る課題や留意点を研究者の知見や他業界の事例を通じて意見交換を行うことで、JAXA内および国内宇宙関連企業に示唆を与える。

アジェンダ（1）

時間	演題	発表者
13:00 - 13:05(05)	開催挨拶	JAXA 髙畑 博樹 理事補佐
13:05 - 13:45(40)	1. 宇宙機へのAIソフトウェア搭載に向けた課題と対策 ～宇宙機搭載開発ハンドブックの概要	JAXA 研究開発部門 第三研究ユニット 石濱 直樹 研究領域主幹
13:45 - 14:25(40)	2. 自動運転におけるAIの安全性に向けた取り組み ～探索的アプローチとLLMの活用	国立情報学研究所 石川 冬樹 准教授
14:25 - 14:35(10)	3. 宇宙機へのAI搭載に向けた ソフトウェア品質保証の取り組み	JAXA 安全・信頼性推進部 神戸 大輔 主任研究開発員
14:35 – 14:45(10)	休憩	
14:45 - 15:20(35)	4. 小型月着陸実証機SLIMの成果： 自律的な航法誘導制御系の事前検証と 運用結果を中心に	JAXA 宇宙科学研究所 宇宙機応用工学研究系 福田 盛介 教授 JAXA 研究開発部門 第一研究ユニット 植田 聡史 研究領域主幹

※講演資料を本資料p.15以降に示す。

アジェンダ（2）

時間	演題	発表者
15:20 - 15:55(35)	5.自動車の自動運転・運転支援システムにおける AI導入とアシュアランス	株式会社本田技術研究所 先進技術研究所 知能化研究ドメイン 杉本 洋一　フェロー
15:55 - 16:30(35)	6. デンソーにおけるAI品質保証の仕組み	株式会社デンソー ソフト生産革新部 先端ソフト開発室 中神　徹也　様
16:30 - 17:05(35)	7. 変化し続けるLLMモデルをプロダクトに組み込む際のテストの考え方	株式会社ベリサーブ 研究開発部 須原　秀敏 部長
17:05 - 17:15(10)	休憩	
17:15 - 17:40(25)	8. 講演者によるフリーディスカッション	JAXA 研究開発部門 第三研究ユニット 安全・信頼性推進部 片平　真史　研究領域総括 （コーディネータ）
17:40- 17:45(05)	閉会挨拶	JAXA 安全・信頼性推進部 杢野　正明　部長

※講演資料を本資料p.15以降に示す。

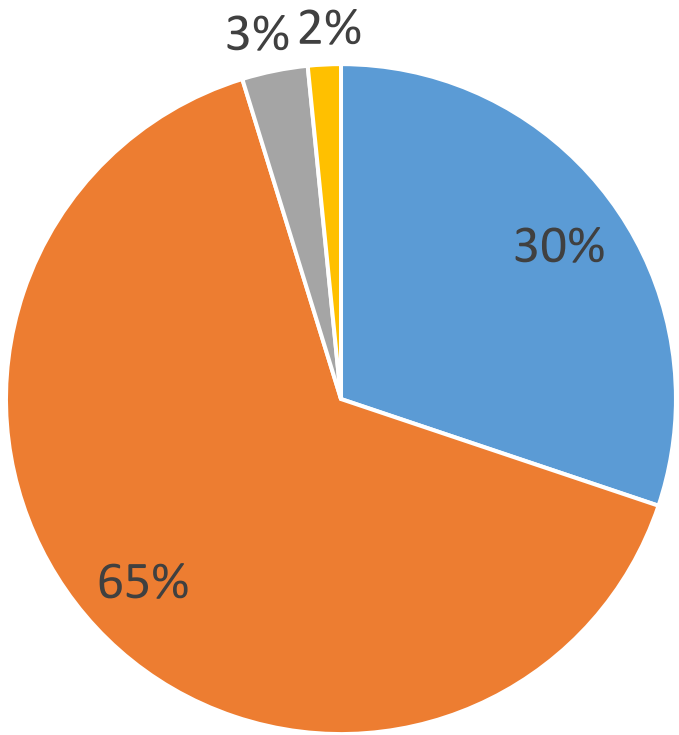
会場内からの質問と回答

演題	質問	回答
演題1	**宇宙機搭載開発ハンドブックにおけるAIに関して、宇宙ならではの事項**はあるか？	**安全性の担保に関しては自動車業界も宇宙も基本的に変わりはなく**、現時点で宇宙ならではの概念は無いと考えている。
	資料P38の誤推論に対する対策とP39のAIシステム開発フローの関係性について、P38の各要素はP39左側のフロー及び右側フローの確認及び評価において考慮する必要があるとの理解で良いか？	ご理解のとおりである。P38はシステムレベルで考慮すべき事項について記述しており、図の右下に示すシステム要求/設計フェーズへというのは、最終的に軌道上で問題が発生した際には、改めてP39に示すシステム要求や設計フェーズに戻って考え直す必要があるという意味で記述している。
	（資料P22において）**十分な学習・テストデータを準備するとの記述があったが、どのような形でデータを準備し、どのような基準に達したら問題ないと判断できる**といった内容はハンドブックに記載されているか？	そこまでは記載していない。**宇宙の場合には十分なデータが無い**ということもあり、如何に効率よくデータを集めるかといった観点から分析の考え方を記載している。
	最終的な方向性として、AI開発のプロセスが変わっていくと考えてよいか？また、現在の宇宙機開発のプロセスに対してどのように対応していくのか？	（資料P18表において）左側が従来の開発プロセスであり、センサーの抽出アルゴリズムのようなものを開発する場合は、現在でも左側のレガシーなループで開発を行うことになり、新しいものと変わりはないと考えている。ハンドブックに関しても、ターゲットとするところは変わりないとご理解いただきたい。
演題2	AIは知らないものは知らないと回答することができないとの説明があったが、検出対象の検討不足といった話にも関連して、例えば、自転車や車以外は知らないと学習させれば良いと考えるが、そのような単純なものではないのか？	その他の特徴がバラバラであり、かつ消去法的なやり方は適用できないため、相当なデータが無い限り検出精度を高めることは極めて難しいと考える。
演題3	**ソフトウェアの品質保証について、従来のレガシープロセスに加えてAIの品質保証を行っていくのか**、又は従来の品質保証の部分をAI特有のものに置き換えていくのか、どのような方向性で検討しているのか？	**基本的にAIを搭載することによって品質保証面で考えるべきことが増える**と認識している。一方で、AIの使用による開発も増えていけば減る要素もあると考えている。
演題4	**画像航法に関して苦労した点**は何か？	**CG画像を実際の月表面の様相にどのくらい合わせられるか**といった点である。結果として、そっくりに見えるCG画像を作成できことは大きな成果であり、月画像のシミュレーション技術向上につながったと考える。
	検証計画の作成において、その適否はどのように決めていったのか？	ノミナル検証では千本ノックのようにとにかく精度を達成するまで実施し、End-to-End検証では、アイデアベースで仕様の範囲外で起こり得る事象を想定しながら決定した。
演題5	日本の道幅は狭く、また交通ルールを守らない人がいる中で、**どのように自動運転を実現していくのか**？	ご指摘のとおり、交通弱者や交通ルールを守らない人がいて、安全性を担保するのが非常に難しいことから、交通環境が比較的整っている高速道路での実証を行うこととした。End-to-Endの観点から、**ルール外でどのような振る舞いをするかといった点はAIを導入した方が対応する力が広がり、また安全性の担保は従来のルールベースで作り込みを行い、これらを組み合わせていくことが必要**と考える。
	バックしてくる車に対してクラクションを鳴らすことがあって、交通ルール的に推奨されていないが、このような状況についても基本的には交通ルールを重視して取り入れていくことになるのか？	基本的には交通ルールを順守することになる。ただし、交通ルール自体も概念的で曖昧な面があり、警察庁と状況を共有して決めていくことになると考える。
演題6	例えば、車載系と非車載系の違いがある中でどのように統合をしていったのか？	（資料P14において）例えば、自動車系では起こり得るリスクを極力減らすという考え方で、IT系についてはその逆で、こういったことを守っていこうという考え方になり、それぞれ求めることが少しずつ異なることから、達成事項を抽象的に考えていくことで進めた。
	AI人材育成に向けた取り組みについて、実際に効果が得られているか？	新入社員及び新任管理者に対する教育は浸透できていることを確認しているが、現時点で導入の段階であり、全社的にAIとしてリテラシーが高まっているかどうか不明確なところである。
演題7	MBTモデル作成について、インフォーマルな自然言語の仕様書から形式モデルを構築する作業をAI支援で行ったとの認識であり、これ自体が注目すべき事項だと考えるが如何か？	100%できたということではなく、ベースモデルを作成することができたといった状況である。
	テストに関して、まったく学習させて無い部分に対しても影響が及ぶ場合、原因を特定するための方法やツールはあるか？	現時点では無い。
	同じくテストに関して、ChatGPTとかAI同士が議論を重ねて最終的な結論を出すような方式が主流になっているが、AIの出力のわずかな違いが後段まで波及する場合があり、この点を画一的に試験するようなツール又は方法はあるか？	現時点では無い。なお、興味深いテーマであり、分割処置において比較的良い推論が得られることがあり、途中の思考過程を評価する必要がなく、結果しか見なくてよい世界が訪れると考えている。

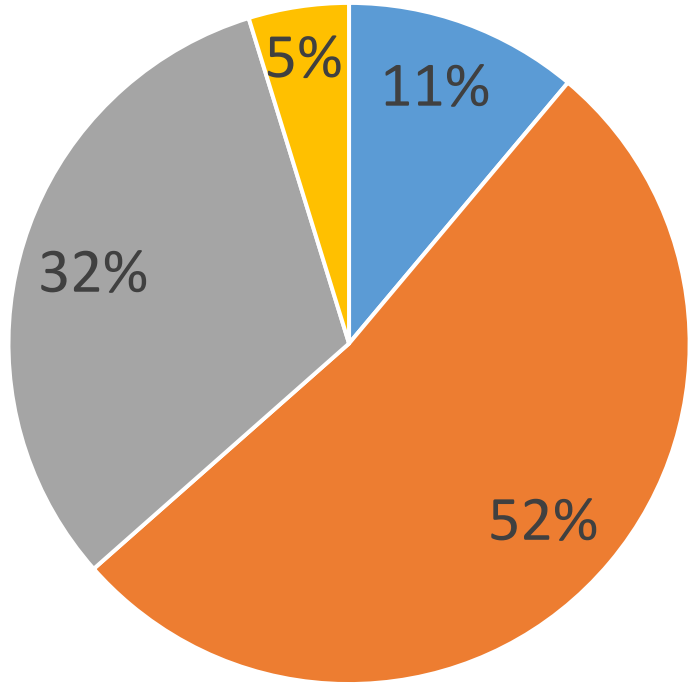
事後アンケート結果に対する考察（1）

- 事後アンケートへの回答者数は63名で参加者全体の約30%、回答者の約70%は企業。
- 「期待に沿うものであった」との意見が95%であるものの、「AI導入にはまだハードルがある」といった意見があり、具体的なアクションに繋がる気づきを得られたのは63%程度。

Q2：本シンポジウムはご期待に沿うものでしたか。



Q3：本シンポジウムで得られた「気づき」によって、ご自身のすべき／したいことがはっきりしましたか。

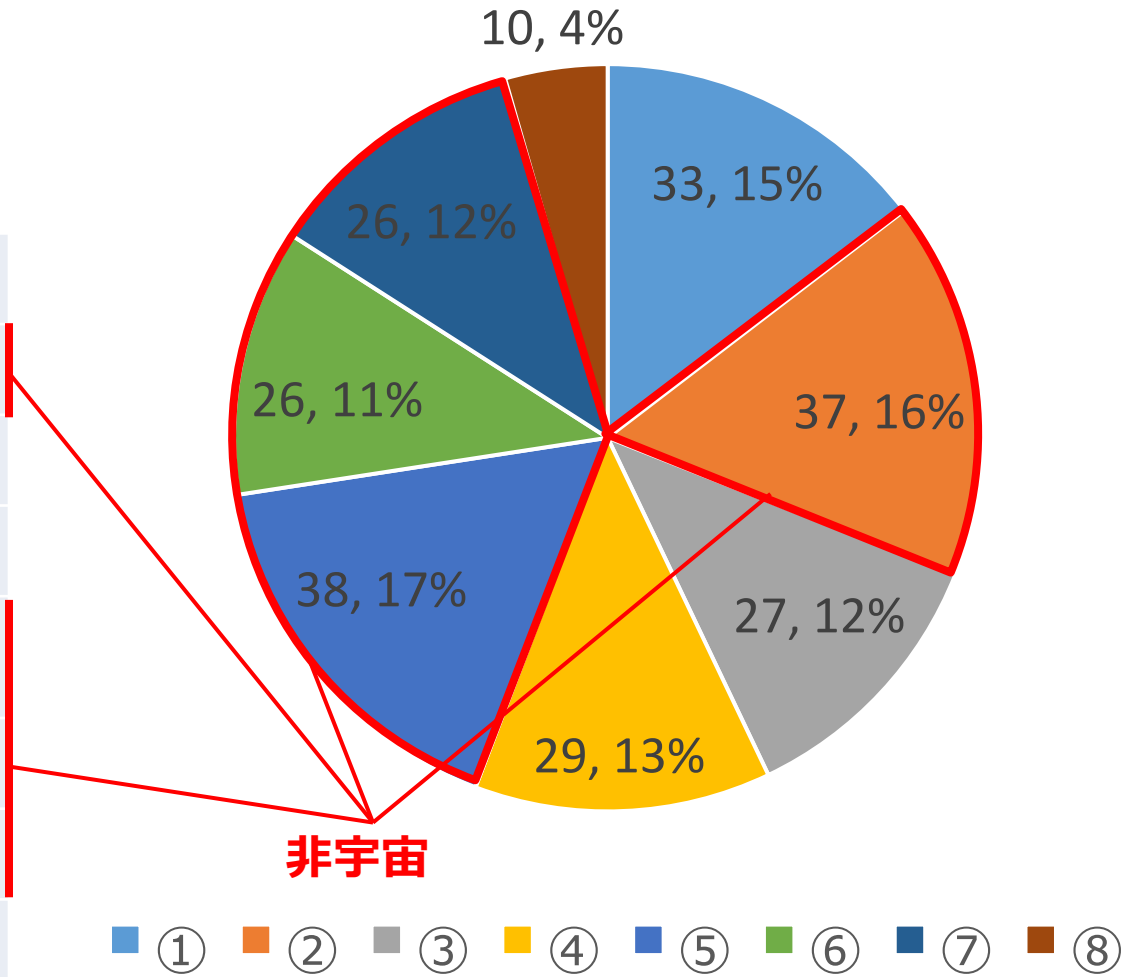


事後アンケート結果に対する考察（2）

- 自動車業界の取組み事例が印象に残ったとの回答が相対的に多い。
- 宇宙に限定せず、自動車を始めとした他産業からも講演いただいたことで、様々な業界の参加者にとって価値のある情報を発信できたのではないか。

Q4：特に印象に残った（学び・気づきを得た）講演を教えてください。（複数回答可）

① 宇宙機へのAIソフトウェア搭載に向けた課題と対策
② 自動運転におけるAIの安全性に向けた取り組み
③ 宇宙機へのAI搭載に向けたソフトウェア品質保証の取り組み
④ 小型月着陸実証機SLIMの成果
⑤ 自動車の自動運転・運転支援システム運転支援システムにおけるAI導入とアシュアランス
⑥ デンソーにおけるAI品質保証の仕組み
⑦ 変化し続けるLLMモデルをプロダクトに組み込む際のテストの考え方
⑧ 講演者によるフリーディスカッション



事後アンケート結果に対する考察（3）

- JAXAに対して「自律化・AIのアシュアランス」の観点では、AIの品質保証に対する具体的な取組み、他産業との連携、実際の活用事例の共有などが求められている。

Q7：自律化・AIのアシュアランスの観点から、JAXAに望むことがございましたら、お聞かせください。

(回答の一部)

- ✓ JAXA S&MAに関するAI（ChatGPT）の活用事例を紹介してほしい。
- ✓ 現状、実機にAIを持ち込むには課題があるので、AIを設計プロセスのツール、製造時のツールとして使うときについて、もっと掘り下げられないかという観点で検討できないか。
- ✓ 宇宙探査だけでなく他の分野（無人航空機の運用、衛星データの解析）でのAIの信頼性について知りたいです。
- ✓ JERGにAI品質保証を盛り込む際には、実現性も考慮したものにしてほしい。
- ✓ 宇宙ミッションは現状では一点もの開発がメインであるため、開発プロセスもそれに対する品質保証もなかなか標準化しにくいというか、標準化した場合に抽象度を高めざるを得ないと思われます。一点ものの開発に対し、抽象化された品質保証の仕組みをどの様に適用するのかを発表していただきたいと思いました。
- ✓ 実開発適用事例等があれば、是非共有いただきたい。
- ✓ 宇宙開発におけるガイドラインおよびプロセスの明確化

事後アンケート結果に対する考察（4）

- 今後の「S&MAシンポジウム」「S&MA活動」に対して、継続的な情報発信、他産業との交流、国内外の連携が期待されている。

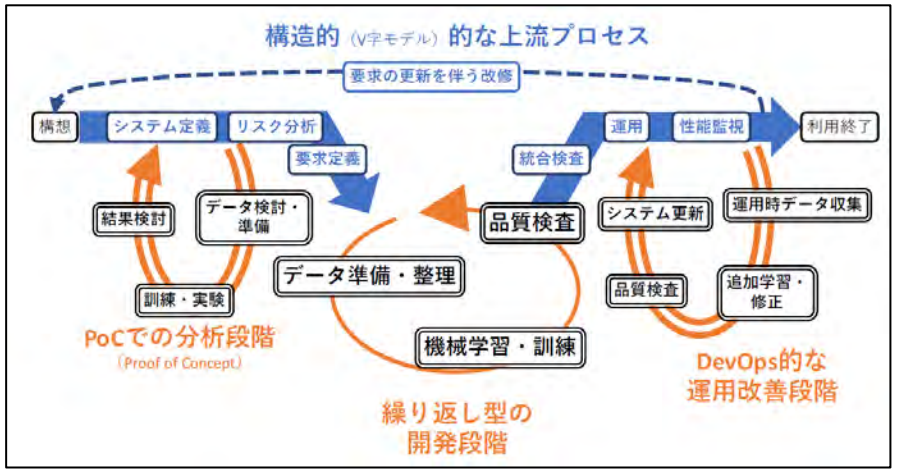
Q8：今後の企画、JAXAの安全・信頼性・品質活動へのご意見・ご要望等ございましたら、お聞かせください。

(回答の一部)

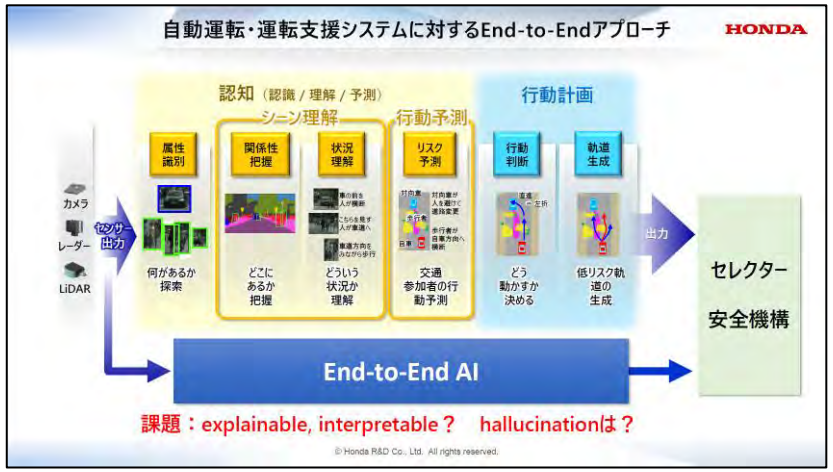
- ✓ AIと同様に、ロボティクス工学関連もかなり技術動向が進んでいると思いますので、今回のAIと同じように、一般的な技術動向、適用状況、必要とされるS&MA活動などについてシンポジウムの開催はいかがでしょうか。
- ✓ JERGや標準策定に関わる、国内業界全体でのシンポジウム情報をもっと開催・周知頂けますとありがたいです。
- ✓ 講演会参加させて頂き、大変ありがとうございました。当社内の取り組みに活かしていきたいと思います。本講演会のテーマに対するアップデートや、先進的な取り組みについて、講演会等の企画をして頂けると幸甚です。
- ✓ S&MA分野でJAXA殿が蓄積された、加えて、獲得される情報の積極的発信
- ✓ ニュースペースの方々が考えるS&MAにかかわる作業と、実際に必要な作業の間には大きなギャップがあり、このことが製品等の信頼性や健全性に大きな影響を与えているものと考える。この分野で、先を行く米国や欧州の実態、（可能であれば）中国やインドでの実態についてご教示いただきたい。
- ✓ 今回のような業界を横断した発表は気付きが多いと感じましたので、定期的なシンポジウムの開催をお願いします。

AI搭載宇宙機のソフトウェア品質保証に向けての今後の取り組み（1）

- 不完全性・不確実性をもち、ブラックボックスであるAIに対し、機械学習型AIの品質保証については国内外で近年ガイドラインが整備されつつあり、他産業含め、以下のようなアプローチで品質保証の取組みが行われていることが確認できた。AIが搭載された宇宙機のソフトウェア品質保証のためのガイドライン検討の際に、これらの知見を活用できる。
  - AIモデルの学習/開発にあたり、Iterative（反復的）なプロセスを導入し、AI品質を向上
  - 実利用環境に則した膨大な数のシミュレーションや実証実験により、実製品の品質を担保
  - AIで実現する領域の外側に安全機構を設け、システムとしてリスクを受容可能な範囲に限定



出典：機械学習品質マネジメントガイドライン 第4版
（国立研究開発法人 産業技術総合研究所）

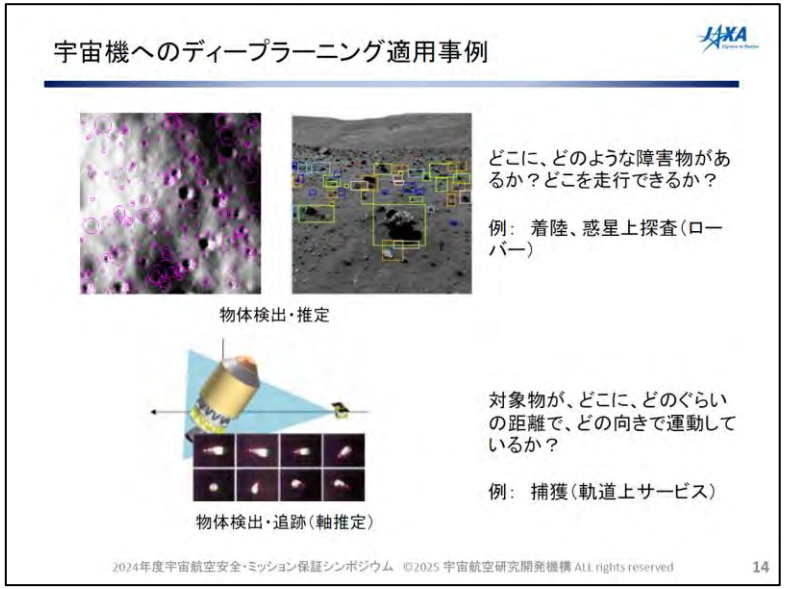


出典：自動車の自動運転・運転支援システムにおけるAI導入とアシュアランス
（株式会社本田技術研究所）

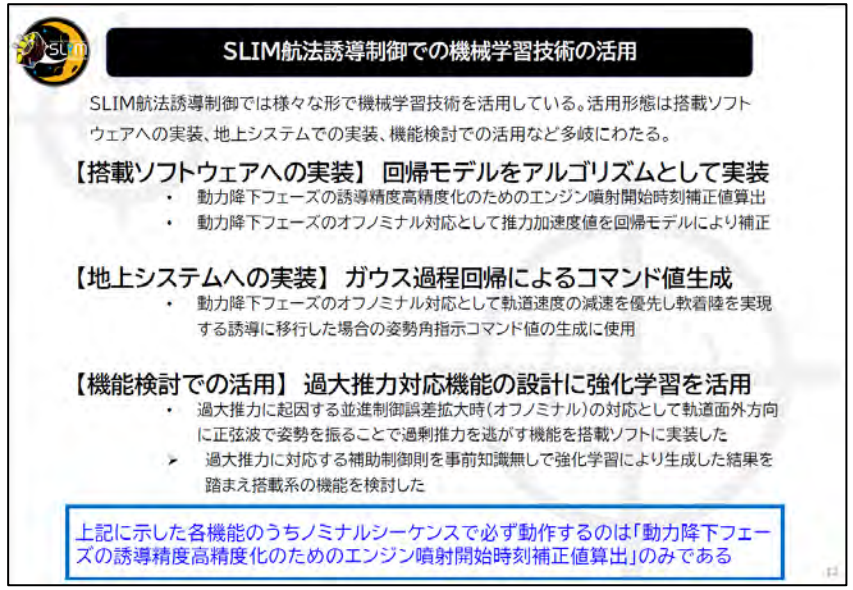
AI搭載宇宙機のソフトウェア品質保証に向けての今後の取り組み（2）

- AIの宇宙機搭載に際し、品質保証を行う上での課題に対して具体的対策の検討が必要
  - 利用環境の未知性/多様性によるAI推論精度確保の困難さ
  - AI開発に必要なデータセットの量・質（十分性/被覆性/均一性/妥当性）確保の困難さ
  - AI推論精度が低い場合にも安全が求められる安全要求の高さ

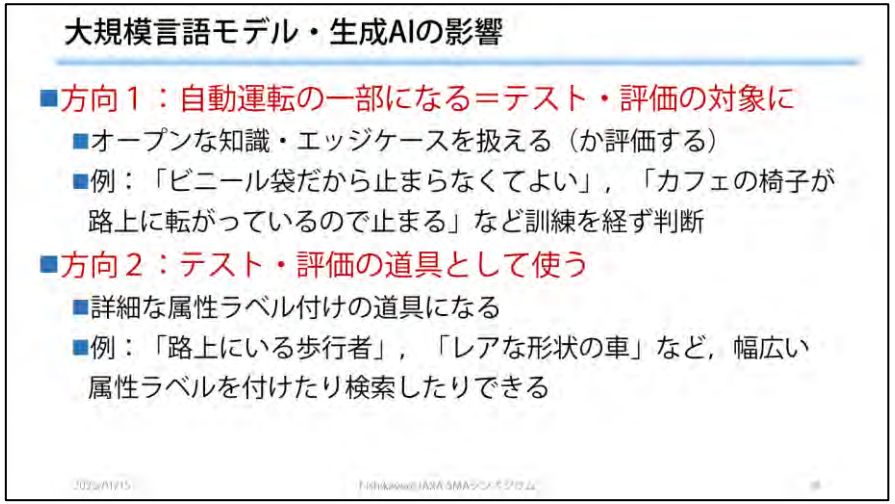
- 生成AI普及に伴い、AI利用拡大が期待される一方で、品質保証のさらなる課題が想定される



出典：宇宙機へのAIソフトウェア搭載に向けた課題と対策
（JAXA 研究開発部門 第三研究ユニット）



出典：小型月着陸実証機SLIMの成果
（JAXA 宇宙科学研究所/研究開発部門 第一研究ユニット）



出典：自動運転におけるAIの安全性に向けた取り組み
（国立情報学研究所）

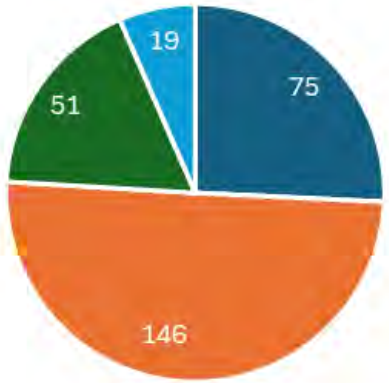
まとめ

- システムへAIを組み込む際の安全性・信頼性確保における課題や技術動向を、宇宙業界のみならず、アカデミアや自動車業界の事例を通じて、JAXA内および国内の宇宙関連企業に広く共有でき、シンポジウム開催の趣旨を達成できた。

- S&MAの文脈で、他産業との交流の機会や継続的な情報発信の場を提供することが期待されているため、その時の状況に合わせた適切なテーマでの定期的な開催を計画していく。

【参考】参加登録者数／参加者数実績(講演者・事務局除く)の内訳

登録者の内訳

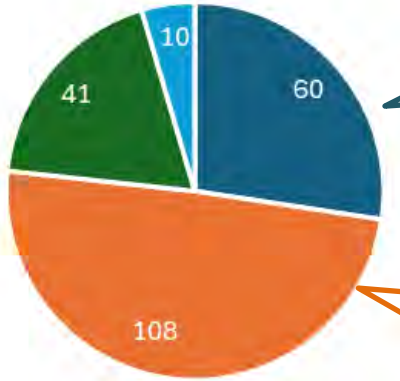


■ JAXA ■ 宇宙関連企業 ■ 非宇宙 ■ 大学・研究機関等

JAXA登録者の所属内訳

■ 安信部 ■ 研究開発 ■ 有人宇宙
■ 宇宙輸送 ■ 第二宇宙 ■ 航空
■ 追跡ネットワーク ■ S&MA総括 ■ その他

参加者の内訳（実績）



約1/4がJAXA

約半数が宇宙関連企業

■ JAXA ■ 宇宙関連企業 ■ 非宇宙 ■ 大学・研究機関等

JAXA参加者の所属内訳（実績）

JAXA内の各部門/事業共通組織から参加

■ 安信部 ■ 研究開発 ■ 有人宇宙
■ 宇宙輸送 ■ 第二宇宙 ■ 航空
■ 追跡ネットワーク ■ S&MA総括 ■ その他

【参考】フリーディスカッションでの質問と回答

【会場参加者からの質問1】

（会場参加者）：AIの開発においては、従来のソフトウェア開発とは異なるAI固有の活動、課題、プロセス等があり、この点をどのように考えているか？
（JAXA片平）：AI固有の活動等について、その対策を含めて伺いたい。
（JAXA石濱）：設計と開発がほぼ一体化する中で要件を満たす形になっていて、ウォーターフォール開発と異なる点があると考える。対策としては、ブラックボックスのテスト方法についてもきちんと考えていく必要がある。また、これまでは要件に対して必要なデータによってテストを行い、最終的にバリエーションによるテストという流れになっていたが、学習の段階で評価しなければならないという点にも違いであると考える。
（NII石川様）：不確実な点であり、完成したものが上手くいくかやってみないと分からない、上手くいっても何故上手くいっているのか分からない、といったことだと考える。基本的には、試して評価して少しでも良くなるように継続的な試行錯誤を行っていくしかないと考える。
（JAXA神戸）：プロセスだけでなく、扱う学習データや評価データによってAIシステムの品質が決まってくることから、この点は大きな違いであり、データ管理や準備を含めたプロセスがAI搭載システムに求められると考える。不確実性に対する対策に関しては、確立された方法論があるわけではなく、説明可能性を意識しつつ、その水準を上げて品質向上を図っていくことが目指すべき方向であると考える。
（JAXA福田）：今後、AIシステムを使用していく場合には、どのようなデータを使用でき学習できるのかといった点が重要になると考える。特に太陽系探査においてはデータが少ない中で使用していくことになり、AIに対して何を任せるかといった点はデータを見ながら検討しないとAIの適切な活用はできないと考えている。
（JAXA植田）：現状としては機械学習の活用となっているが、実際の検証では機械学習を搭載ソフトに実装し検証を行っており、AIや機械学習のモデルは現状に対しては問題ないが、今後より高度になっていく際に検証プロセスに与える影響がどうなるか関心がある。
（ホンダ杉本様）：ブラックボックスに対して検証を如何に効率的かつ短期間に行えるかが重要だと考える。自動車の場合には十分にデータを保有しており、そのデータのフィードバックによる検証やシミュレーションを活用していく必要がある。また、これからは顧客の同意のもとデータ収集の仕組みが出来てくることから、実装したAIが狙い通り作動していない状況を集めることができる可能性や、実際には機能させないアルゴリズムを組み込んで検証を行う仕組みが検討されていることから、如何にこれらを組み合わせて高速に回すことができるかといった点が重要になる。
（デンソー中神様）：ゆらぎが生じる点が大きな違いであり、これまでは演繹的に機能を実装してきたが、AIの場合には帰納的に決まってくることになると考える。対応については、四つのゆらぎを押さえる活動を繰り返し行っていて、どのようなケースに関してどこまで確認したか明示的に示すことをできるようにすることが重要であると考えている。
（ベリサーブ須原様）：これまでのルールベースで作ることに比べて、要件定義がこれまで以上に必要になってくることから、データ空間を如何に狭くすることができるかという点と、実物で如何にテストを行いそれを如何に早くフィードバックさせていくかといった点が重要になると考える。

【会場参加者からの質問2】

（会場参加者）：今後も実証実験の重要性や必要性が残るかどうか伺いたい。
（ホンダ杉本様）：想定外についてはある程度読み切ったと認識される中で、それでも想定外は起こるのではないかとの観点から、実証実験は必要だと考えている。自動運転の実証実験では、もともと100万kmをターゲットとしていたが、想定外が出切ったことを確認するために、更に30万km走行した。
（会場参加者）：宇宙において実証実験は難しく、またデータも少なくシミュレータも使えない状況であり、どのように確認を行っていくのか？
（JAXA石濱）：ご指摘のとおりであり、打ち上げ前に可能な限りのシミュレーションを行うことが必須と考える。とは言え、想定外は必ず発生し、次の機会に活かすために取得したデータを地上において検証しながら最終的に確認していくことが必要だと考える。また、安全性を担保するといった観点も必要と考える。

【参加者から事前に提出された質問】

（JAXA片平）：SLIMや自動車の場合でもシミュレータを回して検証されているが、どこまで実施すればよいのかポリシーがあれば教えていただきたい。
（JAXA福田）：SLIMではアルゴリズムの検証の際に画像航法を行う7箇所を対象とし、1箇所当たり1万枚から数万枚のノミナルケースの画像を使用して検証を行った。たまに失敗するケースが出てくるが、それを必ず吟味することが重要であり、このようなケースの場合にこのようなことが起こると分かった時には、画像処理においてシステム的にそれが起こらないようにしていくという考え方をとっていた。したがって、SLIMの画像航法に関しては、これまで失敗は見たことがないといった状況にしていくことが重要と考える。
（ホンダ杉本様）：シミュレーションに関して1日200万km相当と説明したが、それにはAIアルゴリズムの検証は含まれていない。難しいのはAIによる画像認識であり、量産システムであれば弱点シーンはかなり分かっているため、これを極力再現してシミュレーションを行い、自動運転の場合には、ISO34502のシナリオベースで様々なバリエーションを組み合わせて網羅していく考え方で進めようとしている。
（デンソー中神様）：ここまで実施すればよいといった理論的な基準はないため、極力ゆらぎを抑える活動を行うことにしている。演繹的なシーンを予め認識し、苦手なシーンについては要因を把握して、品質を高めていく必要があると考えている。
（ベリサーブ須原様）：シミュレータではカバーできるデータ領域があり、その範囲内のサンプリングでは画一的に一定以下のバグ率等を設定して判断していくことになるが、データ空間の範囲外については、いわゆる想定外としてシミュレーションではカバーできないと考えている。

（JAXA片平）：例えば、自動運転のLevel2や3においてドライバーが対処するとか、SLIMの場合にはAIシステムに任せることに不安があって、人を含めてシステム上で対処するといった考えがあると思うが、この点についてご意見があれば伺いたい。
（JAXA石濱）：人が操作して止められるのであれば、AIによって最後まで進められるが、宇宙の場合には通信遅延があって直ちに止めることはできないことから、安全的な対策と条件の中で担保するといったアーキテクチャになってくると考える。
（NII石川様）：システム依存になると思うが、例えば、これだけ近づいた場合にはブレーキを踏むしかないといったルールベースで決められることは部分的にあり、このような対策の組み合わせしかないと考える。それが人間なのかルールやアルゴリズムなのかはシステムによるが、代替手段が無いケースもあり、リスクがゼロにならないケースがあっても当たり前として、それにどう対応していくか別途検討していく必要があると考える。

（JAXA片平）：SLIMのピンポイント着陸の際に、月面の画像を地上で見ながら万が一のことに備えていたと思うが、それ以外に人がモニターしながら運用したことはあるか？
（JAXA植田）：万が一の場合にはピンポイント着陸は諦めるが、軟着陸だけは実現することと、また次の段階として小型ローバーだけでも落として生き残らせることを最低限のミッションとして成功させることを考えていた。その場合には地上において人間が判断することになるが、複雑な情報は一瞬で判断することができないため、ある範囲を超えそうになった時には人間の判断に移行することについて、事前に運用訓練を繰り返して実際に機能するか確かめた上で本番に臨んでいた。システムとして複雑な場合でも、人間が判断することは分かり易く、加えて判断結果に納得できるといった点が重要だと考える。

（JAXA片平）：自動車の場合は、人が介在しないケースが増えてくると思われるが、AIの不安定さをカバーするためにどのような対策をとるのか？
（ホンダ杉本様）：自動運転Level3ではドライバーが乗っているが、万が一のことを考えてリスク最小化制御機能を搭載している。Level4ではシステムが安全な行動をとるしかなく、基本的には安全に停止するところまでは車が自律的に判断するように設計されていると聞いている。
また、AIの不確かさに関して、認識については画像認識だけでなくLiDAR等の異種冗長センサーを組み合わせることによって信頼性を向上させている。

宇宙機へのAIソフトウェア搭載に向けた課題と対策
～宇宙機搭載開発ハンドブックの概要

2025/1/15

宇宙航空研究開発機構
石濱　直樹

はじめに

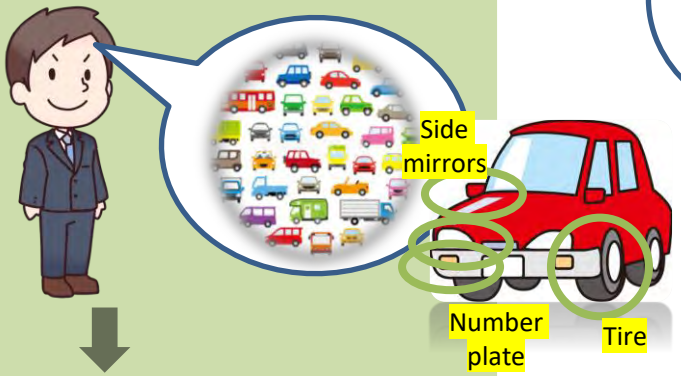
- AIについて
- AIで起きること
- 宇宙機システムに適用とその時に起こりえる問題
- AI搭載ソフトウェア開発ハンドブック

AIの種類

演繹的推論と帰納的推論　～AI視点で見ると～

（例）自動車を例として考えてみた場合・・・

演繹的アプローチ

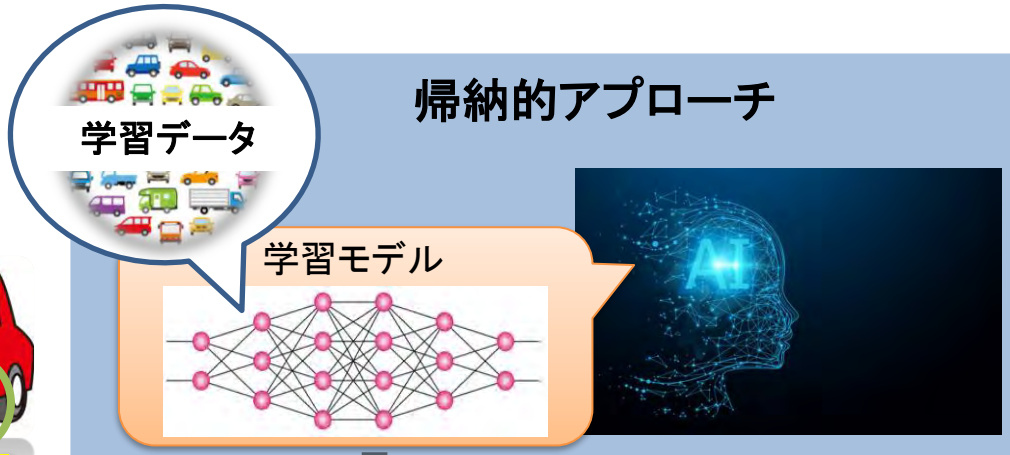


人間が設定するルール

- ・四輪
- ・形の特徴
- ・色の特徴
- ・ナンバープレート
- ・高さ　　・・・

このルールで推論

帰納的アプローチ



AIが見つけ出した特徴

- ・特徴α
- ・特徴β
- ・特徴γ
- ・・・

知識・概念を獲得

この特徴で推論

学習データにないことは
知識・概念を獲得できない
（知らない）

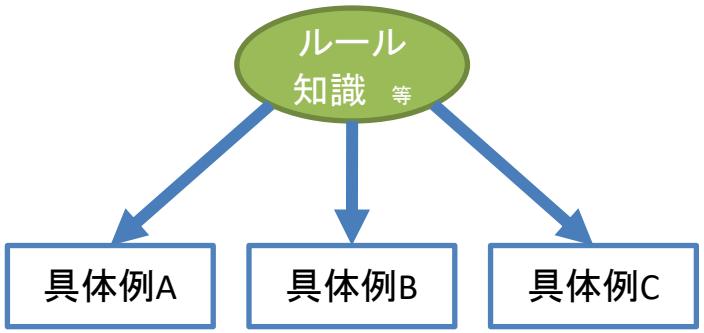
AIの種類

演繹的推論と帰納的推論　～AI視点で見ると～

Legacy System

演繹的推論

**「決められた
ルール・知識・原則」
をもとにした推論**

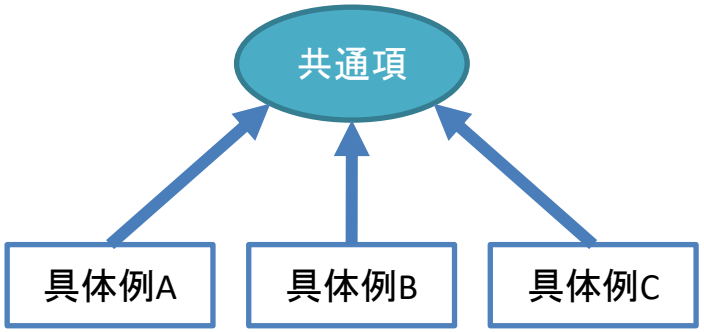


人間がルールを与える
（ルールベース）

AI machine learning System

帰納的推論

**「多くの具体的事例」から
抽出した「共通項」
をもとに推論**

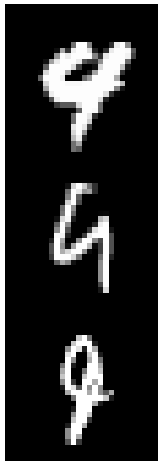


AIがルールを見つける
（機械学習）

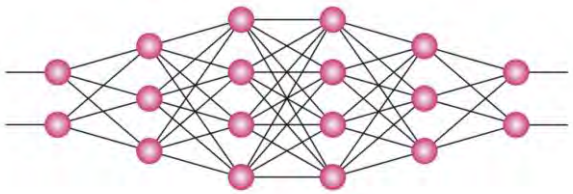
Deep Learningによる知識獲得と推論

学習フェーズ

学習データ

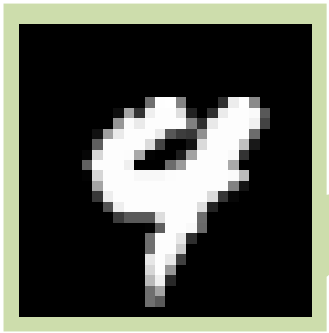


学習器
（ネットワーク）



知識・概念を獲得

推論フェーズ



Deep Learning

・・・
“4”の確率　60.6%
・・・
“9”の確率　39.4%

答えは・・・“4”

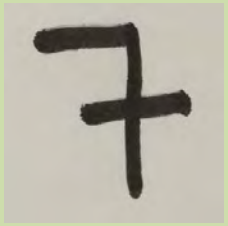
確率の一番高い値が答え!!

Deep Learningの推論結果

自信をもって答えている
（これだ！）

迷った答えをしている
（これかな・・・？？）

これ初めてのことだよ
（知らないことだよ）

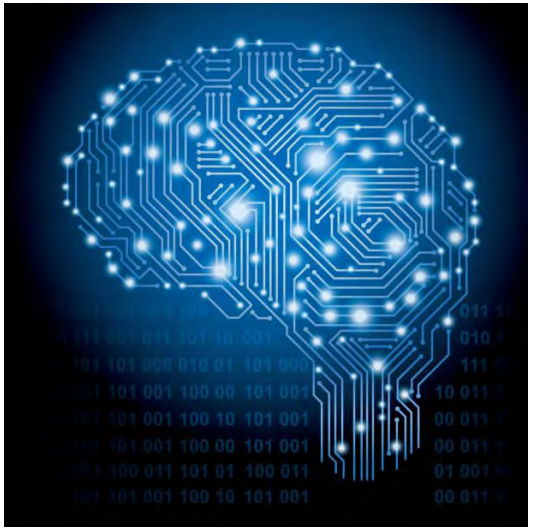


Deep Learningの推論結果

推論結果
（Deep Learningは…）

自信をもって答えている
（これだ！）

まよった答えをしている
（これかな。。。??）



Deep Learningの推論結果と答え合わせ

推論結果
（Deep Learningは…）

正解
（答え合わせ結果…）

自信をもって答えている
（これだ！）

○ 自信をもって答えた結果
正しかった！

× 自信をもって答えた結果
誤った答えだった…

まよった答えをしている
（これかな。。。??）

まよっていたが
たまたま正しかった！

？？ 

まよっていたので
誤った答えだった…

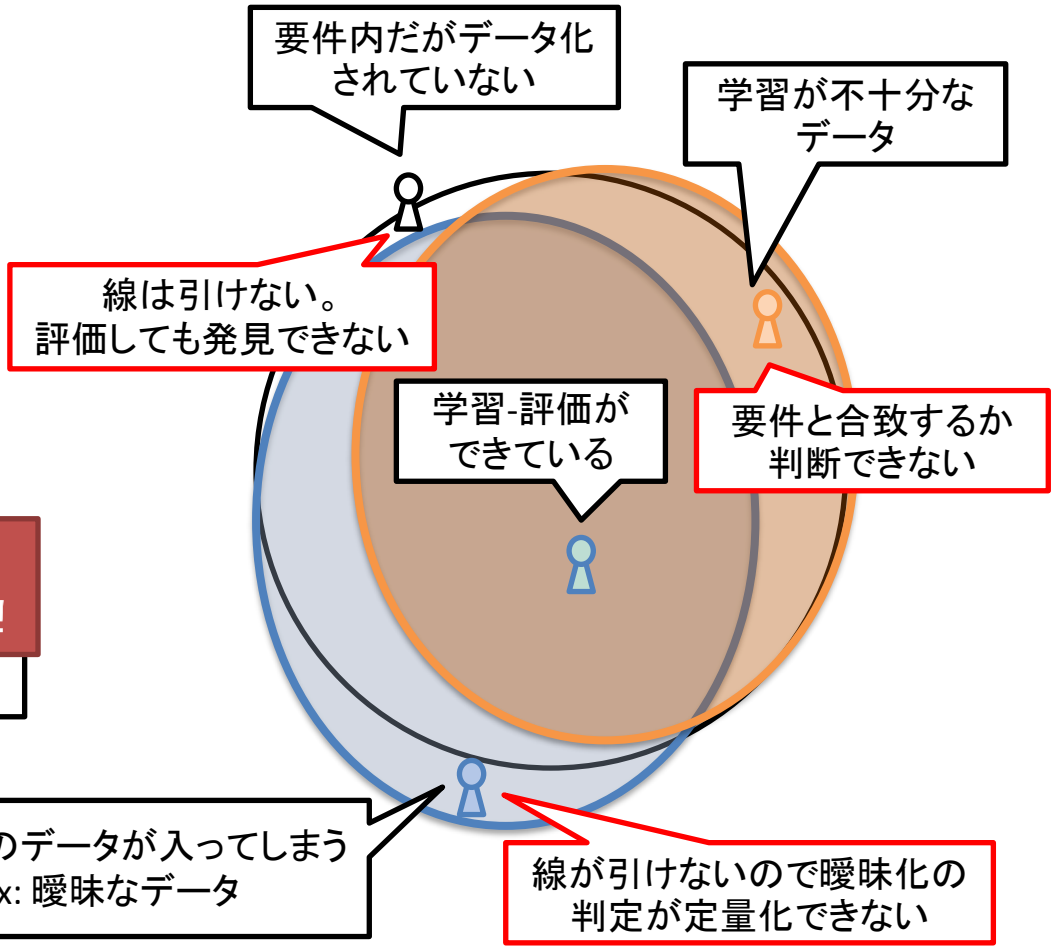
このケースは、無理やり
答える必要があるか？

要件 ― 学習データ ― テストデータの関係

要件で決めた空間
（実のデータ空間）

学習データで収集した空間

テストデータで収集した空間



この3つのデータ空間が重要
3つが完全に重なることは難しい‼

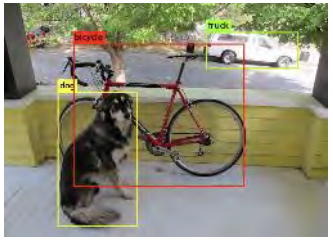
実際に右図は描けない

Deep Learningを用いた外界認識

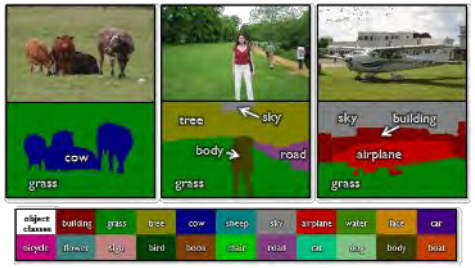
1）分類（classification）…”**その画像が何なのか**”を識別

2）検出（detection）…”**その画像のどこに何があるのか**”を識別

例：R-CNN,YOLO,SSD系



3）セグメンテーション(segmentation)…”**その画像領域の意味**”を識別



http://jamie.shotton.org/work/research.html

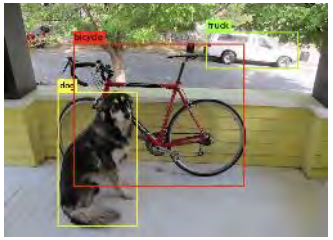
Deep Learningの物体識別で起こりえる問題

1）分類（classification）…”**その画像が何なのか**”を識別

ありえる問題：
識別対象を他のモノと答える

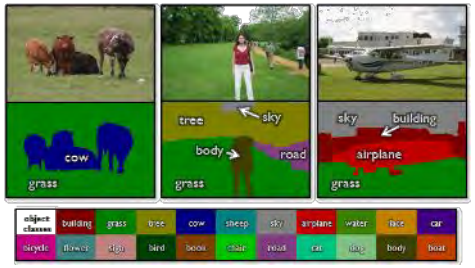
2）検出（detection）…”**その画像のどこに何があるのか**”を識別

例：R-CNN,YOLO,SSD系



ありえる問題：
識別対象を見落とす（あるのにないと答える）
識別対象がないのに識別する（ないのにあると答える）
識別対象を他のモノと答える

3）セグメンテーション(segmentation)…”**その画像領域の意味**”を識別



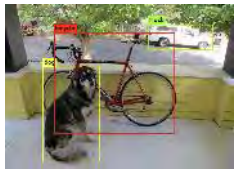
ありえる問題：
識別対象を他のモノと答える

http://jamie.shotton.org/work/research.html

Deep Learningの物体識別で起こりえる問題と具体例

2）検出（detection）…”**その画像のどこに何があるのか**”を識別

ありえる問題：
識別対象を見落とす（あるのにないと答える）
識別対象がないのに識別する（ないのにあると答える）
識別対象を他のモノと答える

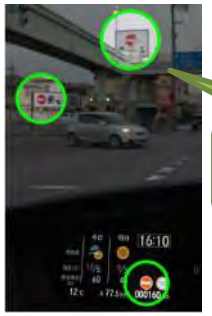


識別を誤ることで<u>ハザード（非安全）に至る</u>具体例：

- 障害物を識別出来ず、障害物がないと判断し走ってしまい、事故につながる。

識別を誤ることで<u>信頼性への影響</u>がでる具体例：

- 住宅街を走行中に、絵を道路交通標識（進入禁止）と識別してしまい、迂回してしまう。
  - この迂回は非安全ではないが、走れるはずなのに、迂回してしまうので、乗車している人の**<u>信頼感を失う</u>**。

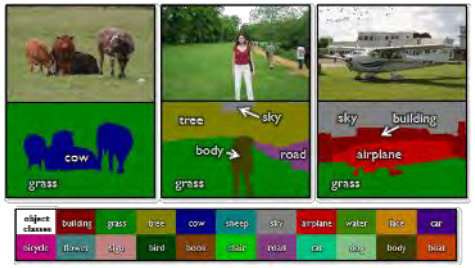


https://twitter.com/Bleu_kakeru727/status/937680760491753473

Deep Learningの物体識別で起こりえる問題と具体例

3）画像セグメンテーション(segmentation)…”**その画像領域の意味**”を識別

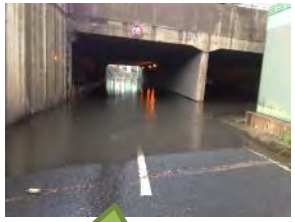
ありえる問題：
識別対象を他のモノと答える



http://jamie.shotton.org/work/research.html

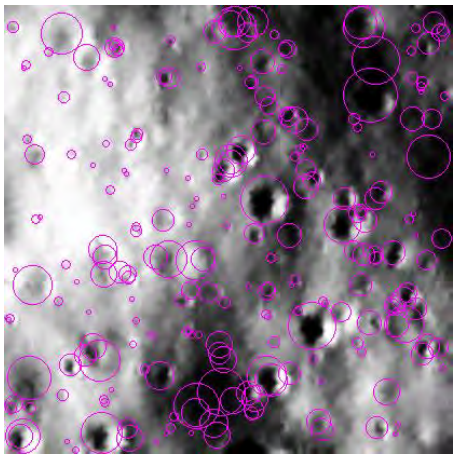
識別を誤ることで<u>ハザード（非安全）に至る</u>具体例：

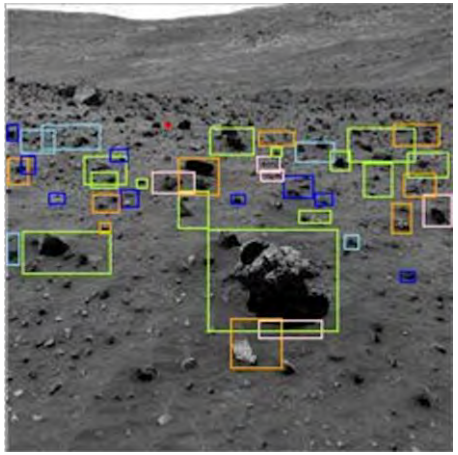
- 走れない領域（例えば、深い水たまり）を、乾いた道路と識別してしまい、走行してしまうことにより、水没してしまい事故につながる。



走行可能と
思ってしまう

宇宙機へのディープラーニング適用事例

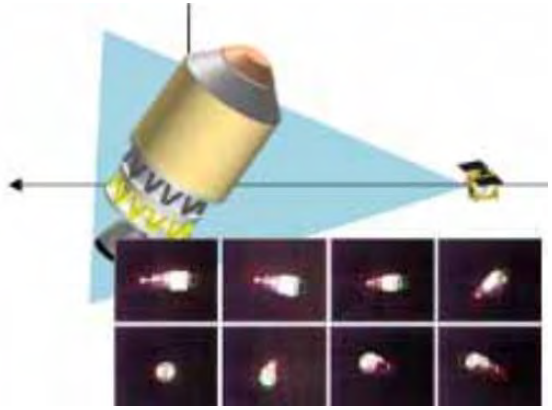




物体検出・推定

どこに、どのような障害物があるか？どこを走行できるか？

例： 着陸、惑星上探査（ローバー）

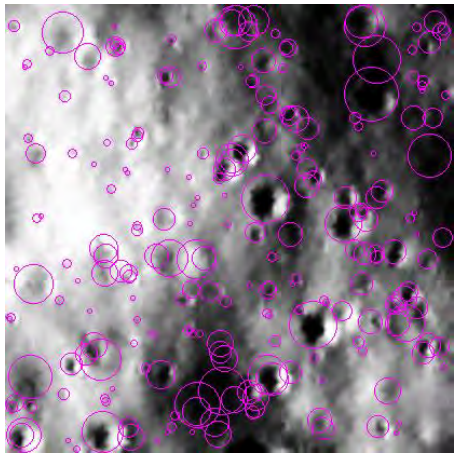


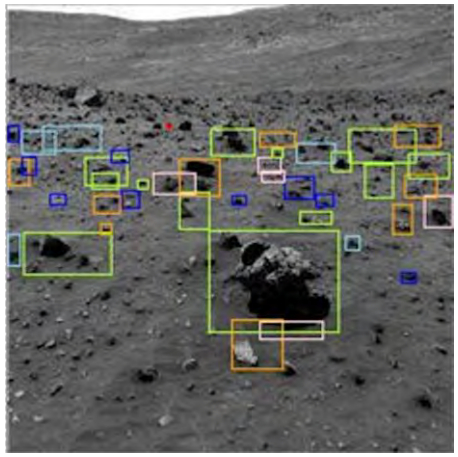
物体検出・追跡（軸推定）

対象物が、どこに、どのぐらいの距離で、どの向きで運動しているか？

例： 捕獲（軌道上サービス）

宇宙機へのディープラーニング適用事例で起こりえる問題





物体検出・推定

どこに、どのような障害物があるか？どこを走行できるか？

例： 着陸、惑星上探査（ローバー）

ありえる問題：
識別対象を見落とす（あるのにないと答える）
識別対象がないのに識別する（ないのにあると答える）
識別対象を他のモノと答える

識別を誤ることでハザード（非安全）に至る具体例：

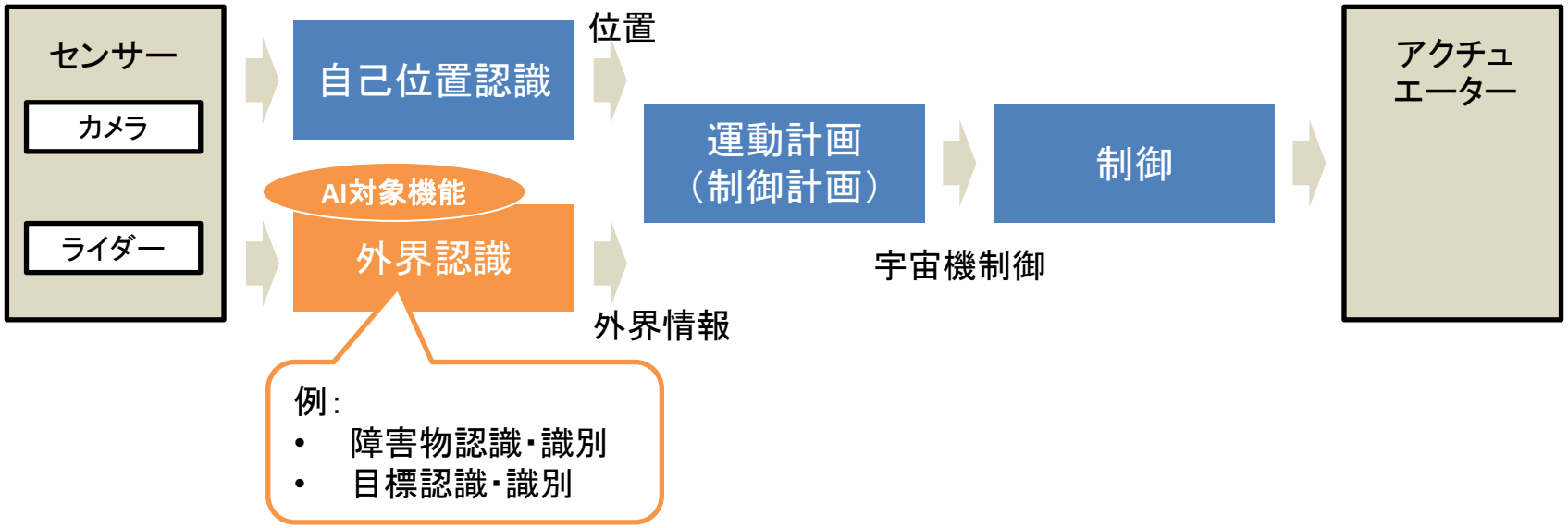
- 障害物を識別出来ず、障害物がないと判断し走ってしまい、クレータに突っ込んでしまい、事故につながる。
- レゴリスに突っ込んでしまい、脱出できなくなる。

識別を誤ることで信頼性への影響がでる具体例：

- 誤って障害物があると判断し、迂回（遠回り）してしまう。場合によっては、走行できないと判断してしまう。

AI対象機能

宇宙機システムとAI対象機能の例



AI搭載ソフトウェア開発ハンドブック　全体像

本ハンドブックの位置づけ:

システムにとって目となる外界認識機能に対し、AI技術を適用した研究結果をもとに、AI推論結果におこるAI特有の問題点に着目し、それが発生する要因、原因とそれに対する対策例と要求事項を解説した資料である。

章	内容
2章　基本的な考え方	• AIの種類 • AIの課題（誤推論）
3章　開発プロセスごとの要件	• レガシーシステムとAI搭載システムの開発プロセスの比較 • AI搭載システムの開発プロセスの概要
4章　誤推論の要因	• AIが誤推論を行ってしまう要因（フェーズ毎に作りこむ要因） • 誤推論の原因
5章　誤推論に対する対策	• 誤推論の各原因に対する具体的な対策例 • 誤推論しないためのフェーズ毎の対策要求
6章　開発プロセス詳細	• 誤推論したいための対策を含めたAIソフトウェア開発フロー全体像とフェーズ毎のプロセス要求

レガシーシステム開発プロセスとAI搭載システム開発プロセスの比較

レガシーシステム開発とAI搭載システム開発をプロセス視点で対比。
AI搭載システムの開発プロセスのフェーズ毎のアクティビティ要件を明確化。

レガシーシステム開発プロセス	AIを搭載したシステム開発プロセス
システム要求分析フェーズ	
システム設計フェーズ	
ソフトウェア要求フェーズ	AIソフトウェア要求分析フェーズ/ AI学習準備フェーズ
ソフトウェア設計フェーズ	AI学習フェーズ 
ソフトウェア製作フェーズ	
ソフトウェア試験フェーズ	AIテストフェーズ
ソフトリリース	AIソフトリリース
システムテストフェーズ	
システム出荷	
システム運用フェーズ	

AIシステム開発プロセス概要

	フェーズ	概要
	システム要求分析フェーズ	システムの利用シーン・ユースケース・外部要因を分析し、要求を明確化する。
	システム設計フェーズ	ハードウェア・ソフトウェアのアーキテクチャ設計を行う。ソフトウェアのなかでAIで実現する機能を明確化する。
AIソフトウェア開発フェーズ	AIソフトウェア要求分析フェーズ	AIソフトウェアの機能・非機能要求を識別し、明確化する。
AIソフトウェア開発フェーズ	AI学習準備フェーズ	ユースケースシナリオ、要求等をもとに学習データの準備・前処理を行う。
AIソフトウェア開発フェーズ	AI学習フェーズ	AIモデルを作成し、準備したデータをもとに、要求を満たすまで学習を繰り返す。
AIソフトウェア開発フェーズ	AIテストフェーズ	学習後モデルに対し、準備したテストデータを用い推論結果の妥当性検証を行う。
AIソフトウェア開発フェーズ	AIソフトリリース	AIソフトウェア単体での検証は完了
	システムテストフェーズ	他ソフトウェアを含めたシステム全体での妥当性検証を行う。
	システム出荷	製品を出荷
	システム運用フェーズ	運用中の状態をモニターするとともに、想定外の状況が発生したことを検知し、データを保存し、事後解析を行う。

誤推論の原因

- フェース毎に作りこむ要因と具体例を整理

4.1.1 システム要求分析フェーズに作りこむ要因

- 「ユースケース分析」「システム要求の識別」を行う中で、以下の要因で誤推論をおこす可能性がある。
- **（原因1）推論対象となる学習データがそもそもない**
  - 状況・状態の変化を考慮した走行シナリオ、外部要因（走行環境）、推論対象の検討漏れや、法規等の制約検討漏れにより、推論対象となる学習データを抽出することが出来なかった
- **（原因2）推論対象となる学習データのバリエーションが足りない**
  - 状況・状態の変化を考慮した走行シナリオ、外部要因（走行環境）、推論対象の検討漏れや、法規等の制約検討漏れにより、推論対象となる学習データのバリエーションの考慮が不足した

□ 上記であげた原因の具体例を以下に示す。（ここでは、自動車を例に示す）

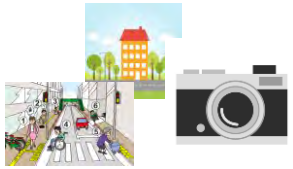
- **利用シーンの検討漏れ**
  - ✓ 全世界を対象にしたいのに、日本国内シーンのみでの分析を行っていたため、左車線走行のみのデータとなっていた
  - ✓ 走行シーンで、逆走車がいるとは思っていなかった
- **推論対象の検討漏れ**
  - ✓ 高速道路で人/動物が歩いているとは思わなかった
  - ✓ 市街地・欧米での利用シーンが抜けていたので、走行している標識の違いを考慮できなかった
  - ✓ 自動車全体が見えているとは限らないことを考慮できていなかった

誤推論の原因まとめ

原因A:　十分な学習データがない	原因A1:　推論対象となる**学習データがそもそもない**
	原因A2:　推論対象となる**学習データのバリエーションが足りない**
	原因A3:　推論対象データへの**外的・内的要因に対する学習データが足りない**
	原因A4:　推論対象となる**学習データの量・質が足りない/悪い**
原因B:　誤った学習データが含まれている	
原因C:　不十分なモデルである	
原因D:　誤った実装を行っている	

誤推論に対する対策

原因	対策
原因A:　十分な学習データがない	**対策1**:　体系的な手法を用いて**必要十分な学習・テストデータを準備する**



- 対象物に対し影響を与える外部環境
  - 撮影時の環境条件
    - 温度
    - 光
      - 天候（雨、曇り）
      - 時間帯（逆光、夜）
    - 外界情報（道路状態）
- 対象物の種類（バリエーション）
  - 車種
- 対象物を捉える条件
  - 撮影条件
    - センサ性能（カメラの性能）
    - 撮影角度（上下・左右撮影角度）
    - 障害物
      - 直接（対向車、外壁）
      - 間接（反射（鏡、窓））

誤推論に対する対策

原因	対策
原因A:　十分な学習データがない	**対策1**:　体系的な手法を用いて**必要十分な学習・テストデータを準備する**

….. 必要十分な学習・テストデータを100%準備することは難しい（困難）

誤推論に対する対策

原因	対策
原因A:　十分な学習データがない	**対策1**:　体系的な手法を用いて**必要十分な学習・テストデータを準備する**

….. 必要十分な学習・テストデータを100%準備することは難しい（困難）

原因	対策
原因A:　十分な学習データがない	**対策2:　誤推論していることを識別する**

誤推論に対する対策

原因	対策
原因A:　十分な学習データがない	**対策2:　誤推論していることを識別する**

詳細化した原因に対する対策

原因	対策
原因A1:　推論対象となる**学習データがそもそもない**	**対策2-2:　入力データ傾向の変化を識別**
原因A2:　推論対象となる**学習データのバリエーションが足りない**	**対策2-2:　入力データ傾向の変化を識別** **対策2-3:　正しい特徴を捉えていないことを識別** **対策2-4:　あいまいな推論結果を識別**
原因A3:　推論対象データへの**外的・内的要因に対する学習データが足りない**	**対策2-2:　入力データ傾向の変化を識別** **対策2-4:　あいまいな推論結果を識別**
原因A4:　推論対象となる**学習データの量・質が足りない/悪い**	**対策2-1:　データ分布のズレ/不整合を識別** **対策2-2:　入力データ傾向の変化を識別** **対策2-3:　正しい特徴を捉えていないことを識別** **対策2-4:　あいまいな推論結果を識別**

誤推論に対する対策

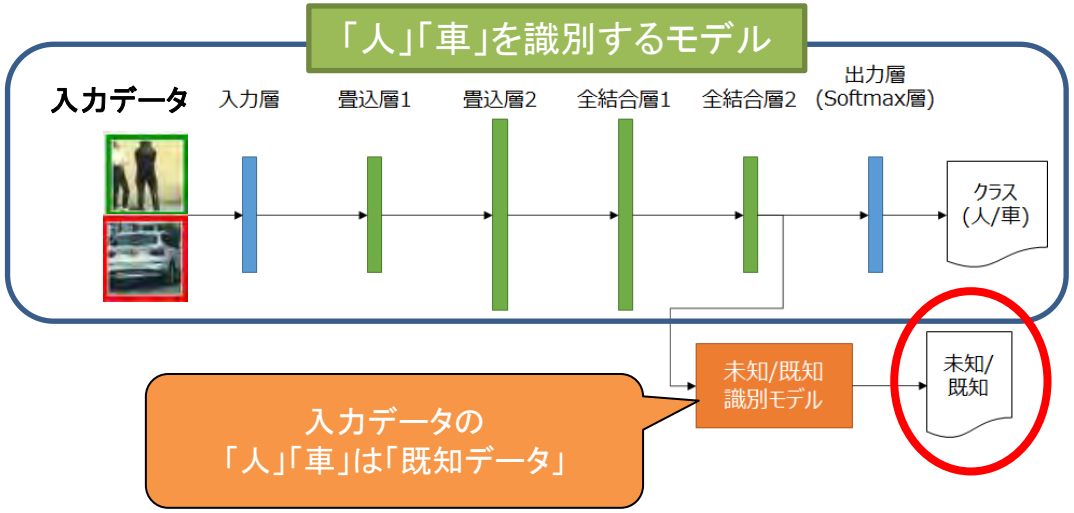
対策
対策2:　誤推論していることを識別する
対策2-1:　データ分布のズレ/不整合を識別
対策2-2:　入力データ傾向の変化を識別
対策2-3:　正しい特徴を捉えていないことを識別
対策2-4:　あいまいな推論結果を識別

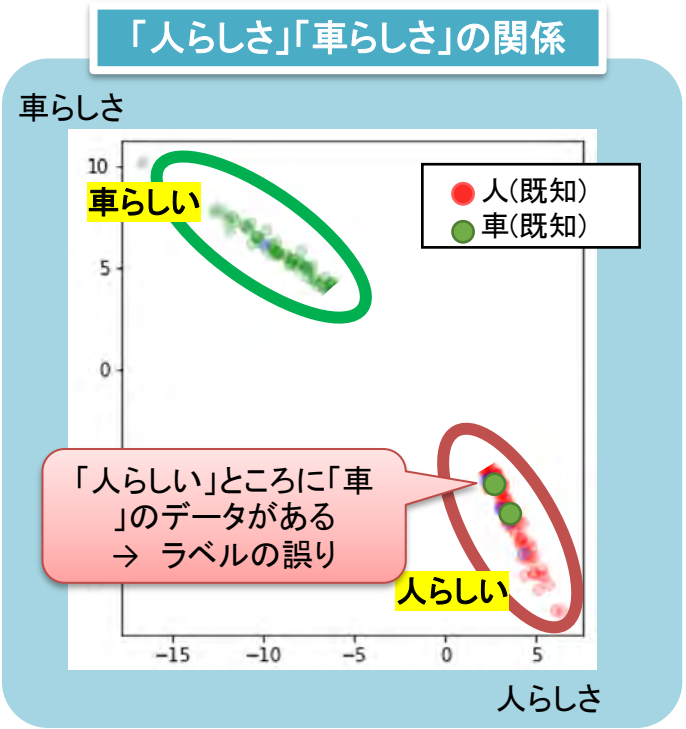
誤推論に対する対策毎の具体例

対策2-1: データ分布のズレ/不整合を識別

整理された知識マップから矛盾している（学習）データを見抜きたい。
例：　異なるラベルをつけていた

方法例:





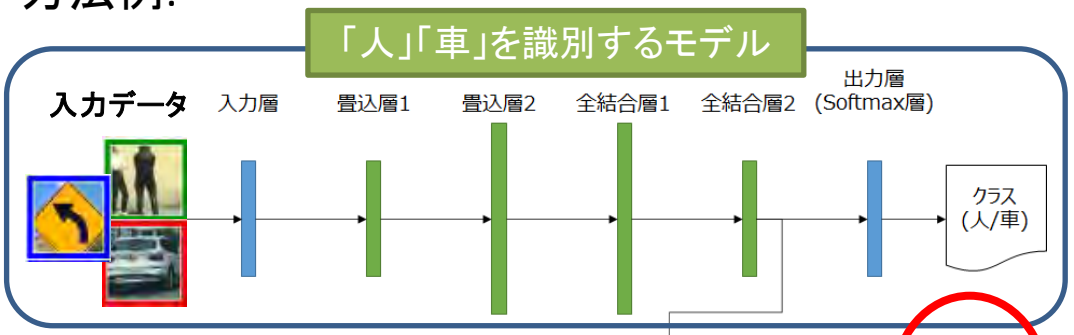
誤推論に対する対策毎の具体例

対策2-2:　入力データ傾向の変化を識別

『教わっていないために誤る』ことを見抜きたい
学習データとしてなかったデータ（識別対象とされていない）なので、
知識獲得が出来ていなく、誤った回答をした
例：　まさか高速度道路を自転車が走るとは思わなかった

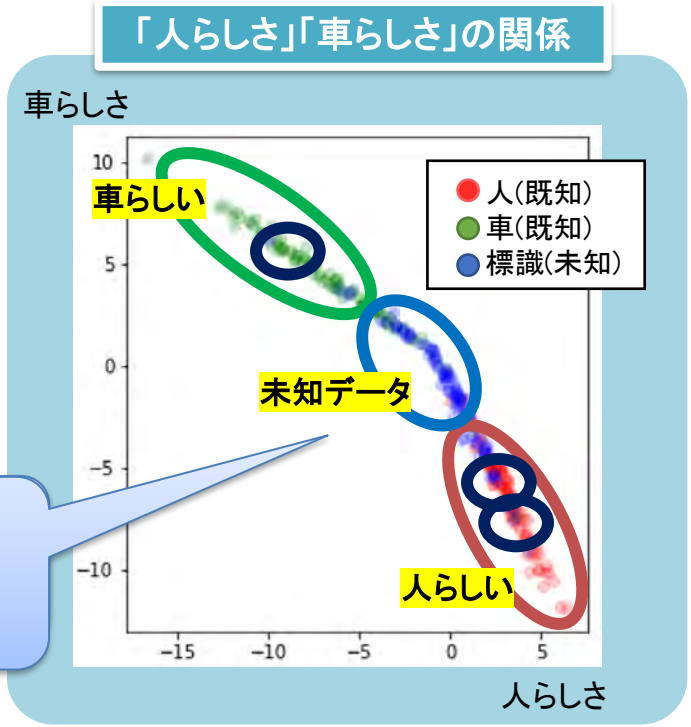
答えようがないものを
無理やり答えている

方法例:



入力データの
「人」「車」は「既知データ」、
「標識」は「未知データ」

未知データは、
「車らしさもなく」
「人らしさもない」
ようになってほしい



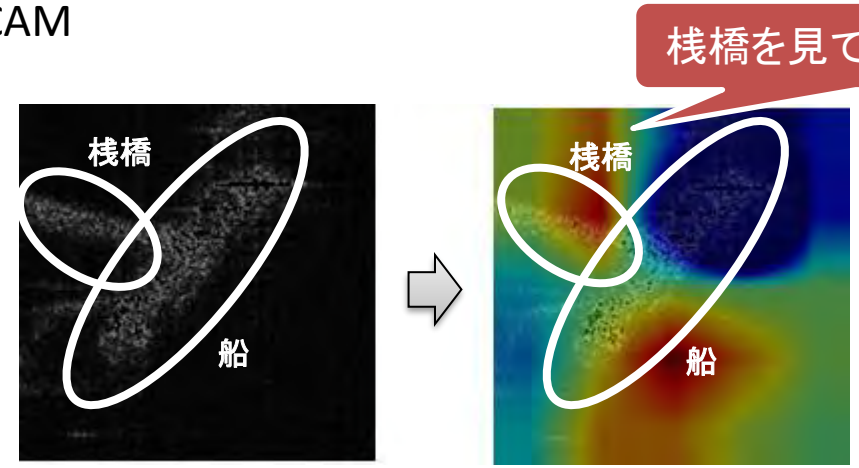
誤推論に対する対策毎の具体例

対策2-3:　正しい特徴を捉えていないことを識別

『誤った覚え方をしていたので誤る』ことを見抜きたい
- 識別対象物ではなく、**周辺の環境情報の特徴をみて回答**をしていたため、誤った回答をした
  - 例：車の特徴をとらえて車と回答をしたいところなのに、道路の特徴をとらえて車と回答
    - ⇒ 考えられる理由：道路の上にある車の写真を使い学習したため、共通的特徴として道路を抽出した

自信はあるが間違った覚え方をしている

識別手法例：　Grad-CAM



Input Image　　Grad-CAM

人による確認が必要

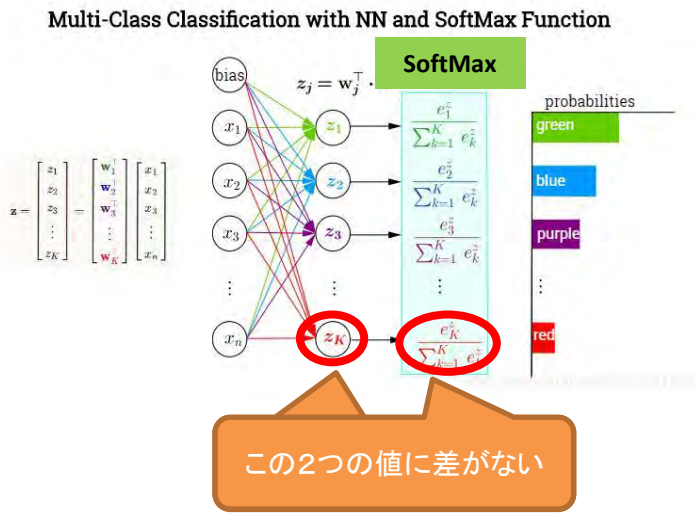
AIが誤った推論したケース

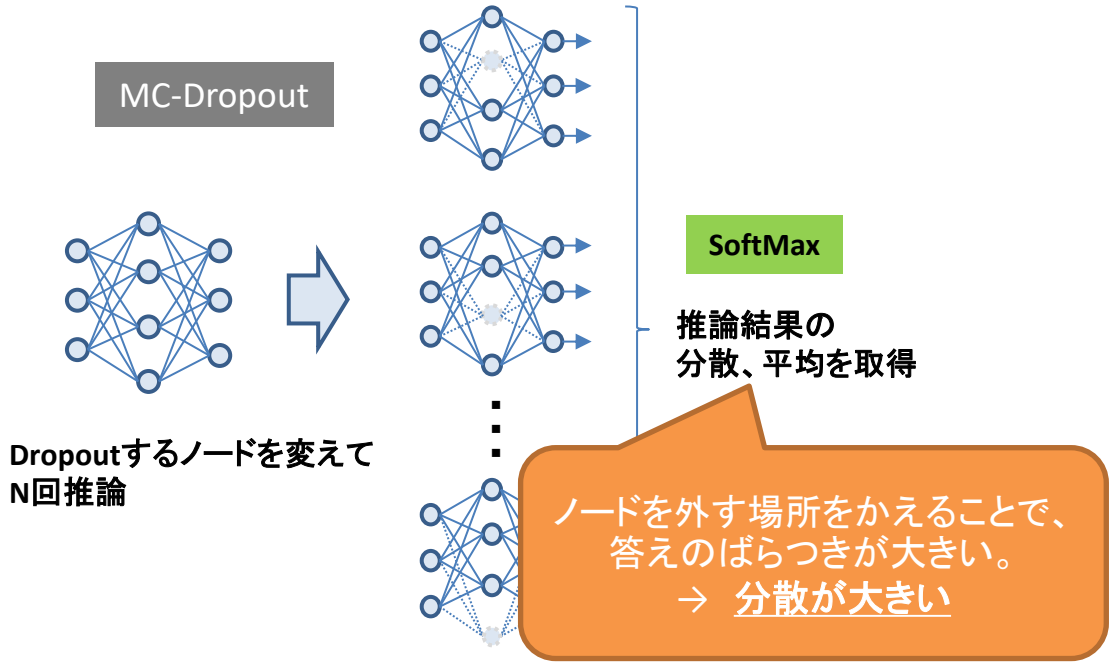
誤推論に対する対策毎の具体例

対策2-4:　あいまいな推論結果を識別

『中途半端に覚えていたので誤る』ことを見抜きたい
識別対象物が、獲得された知識に対して変化があり、一意に判断することが難しく、
悩んだ末に誤った回答をした

識別手法例：　MD-Dropout



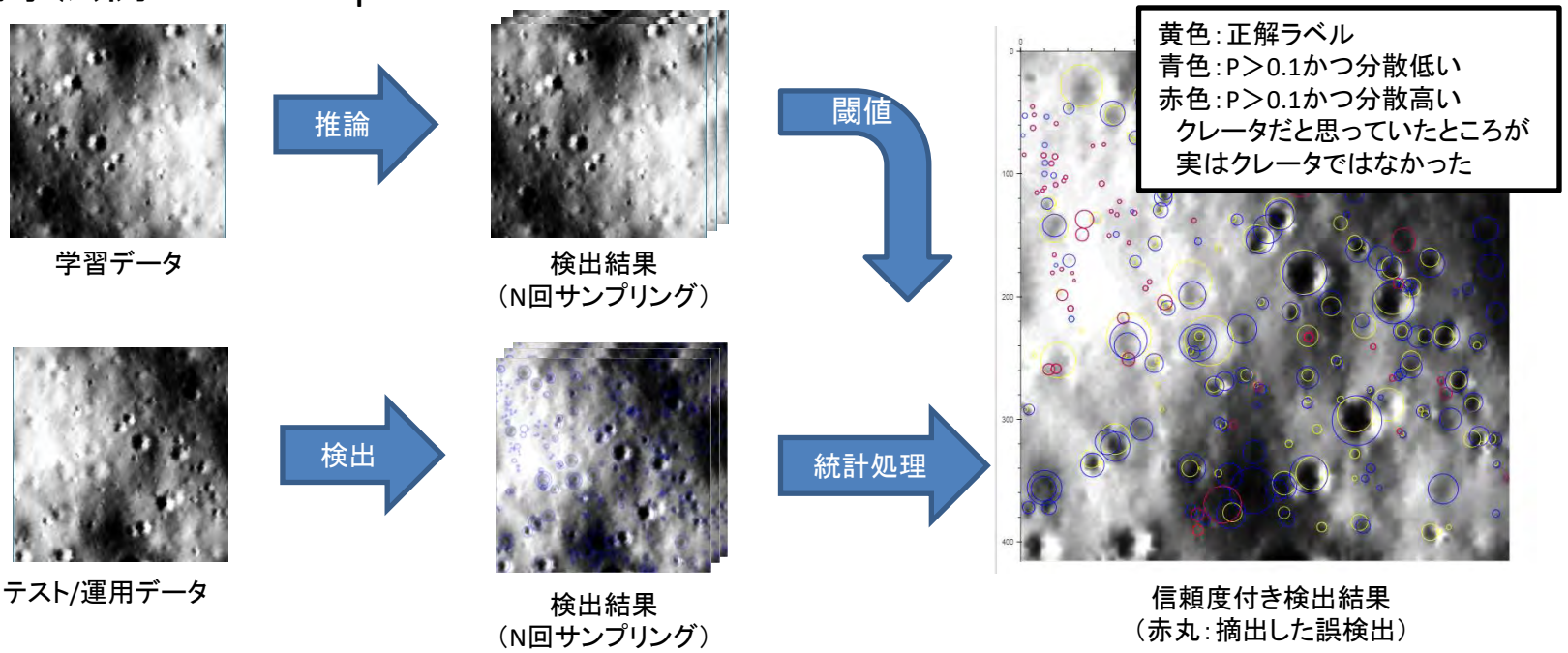


誤推論に対する対策毎の具体例

対策2-4:　あいまいな推論結果を識別

『中途半端に覚えていたので誤る』ことを見抜きたい
識別対象物が、獲得された知識に対して変化があり、一意に判断することが難しく、
悩んだ末に誤った回答をした

識別手法例:　MD-Dropout



誤推論に対する対策

原因	対策
原因A:　十分な学習データがない	**対策1**:　体系的な手法を用いて**必要十分な学習・テストデータを準備する**

….. 必要十分な学習・テストデータを100%準備することは難しい（困難）

原因	対策
原因A:　十分な学習データがない	**対策2:　誤推論していることを識別する**

….. 誤推論結果を100%識別することは難しい（困難）

誤推論に対する対策

原因	対策
原因A:　十分な学習データがない	**対策1**:　体系的な手法を用いて**必要十分な学習・テストデータを準備する**

….. 必要十分な学習・テストデータを100%準備することは難しい（困難）

原因	対策
原因A:　十分な学習データがない	**対策2:　誤推論していることを識別する**

….. 誤推論結果を100%識別することは難しい（困難）

原因	対策
原因A:　十分な学習データがない	**対策3:　システムレベルで対応策をとる**

誤推論に対する対策

原因	対策
原因B:　誤った学習データが含まれている	対策4:　ラベル誤りデータがないこと、非推論対象データが含まれていないことを識別する **=>対策2-1:　データ分布のズレ/不整合を識別**

誤推論に対する対策

原因	対策
原因B:　誤った学習データが含まれている	対策4:　ラベル誤りデータがないこと、非推論対象データが含まれていないことを識別する **=>対策2-1:　データ分布のズレ/不整合を識別**

….. 誤推論結果を100%識別数ることは難しい（困難）

原因	対策
原因B:　誤った学習データが含まれている	**対策3:　システムレベルで対応策をとる**

誤推論に対する対策

原因	対策
原因B:　誤った学習データが含まれている	対策4:　ラベル誤りデータがないこと、非推論対象データが含まれていないことを識別する **=>対策2-1:　データ分布のズレ/不整合を識別**

….. 誤推論結果を100%識別数ることは難しい（困難）

原因	対策
原因B:　誤った学習データが含まれている	**対策3:　システムレベルで対応策をとる**

原因	対策
原因C:　不十分なモデルである	**対策5:　（通常と同様な）性能要求の未達によりモデルを再構築する**

原因	対策
原因D:　誤った実装を行っている	**対策6:　（通常と同様な）レビュー・テストを行う**

誤推論に対する対策

原因	対策
原因B:　誤った学習データが含まれている	**対策3:　システムレベルで対応策をとる**

重要になるのはシステムレベルでの設計

Point1：推論結果の正しさを（点でなく空間情報をふくめた時系列情報の）因果関係で判断

- 意味論としてあり得ない推論結果は除く。
  - あり得ない例：　場所、速度、方向

例：　自動運転走行の場合
ユースケースシナリオから、走行している場所・速度、また、識別対象物の属性、行動/運動能力が明確化出来る。
その情報と、推論結果の因果関係が成立しないのであれば、推論結果が誤っているので推論結果を使わない。

STEP2：推論結果の信頼性・確実性評価を行う

- 推論結果確信をもって使えないデータを抽出する。

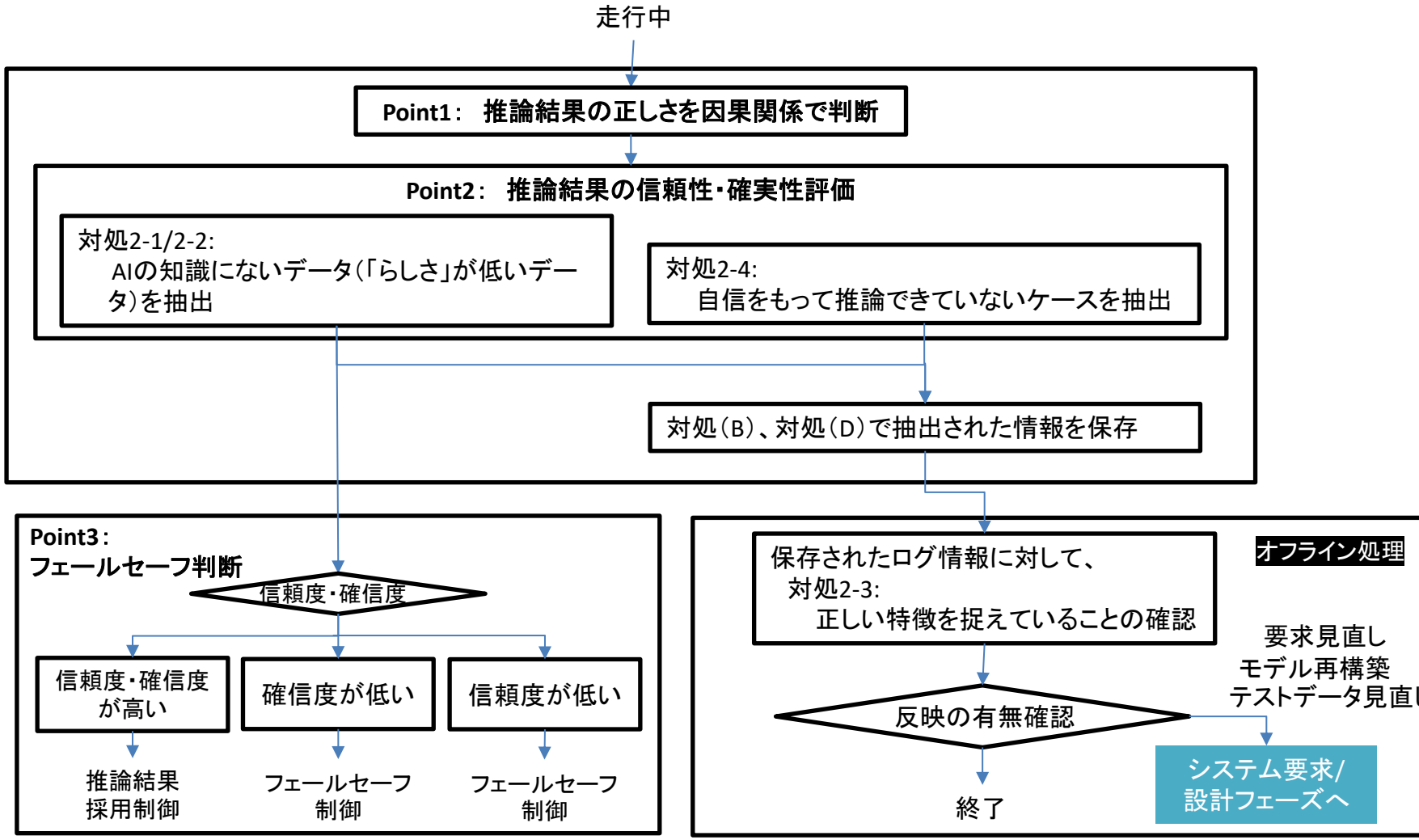
対処2-1/2-2:
AIの知識にないデータ（「らしさ」が低いデータ）を抽出

対処2-4:
自信をもって推論できていないケースを抽出

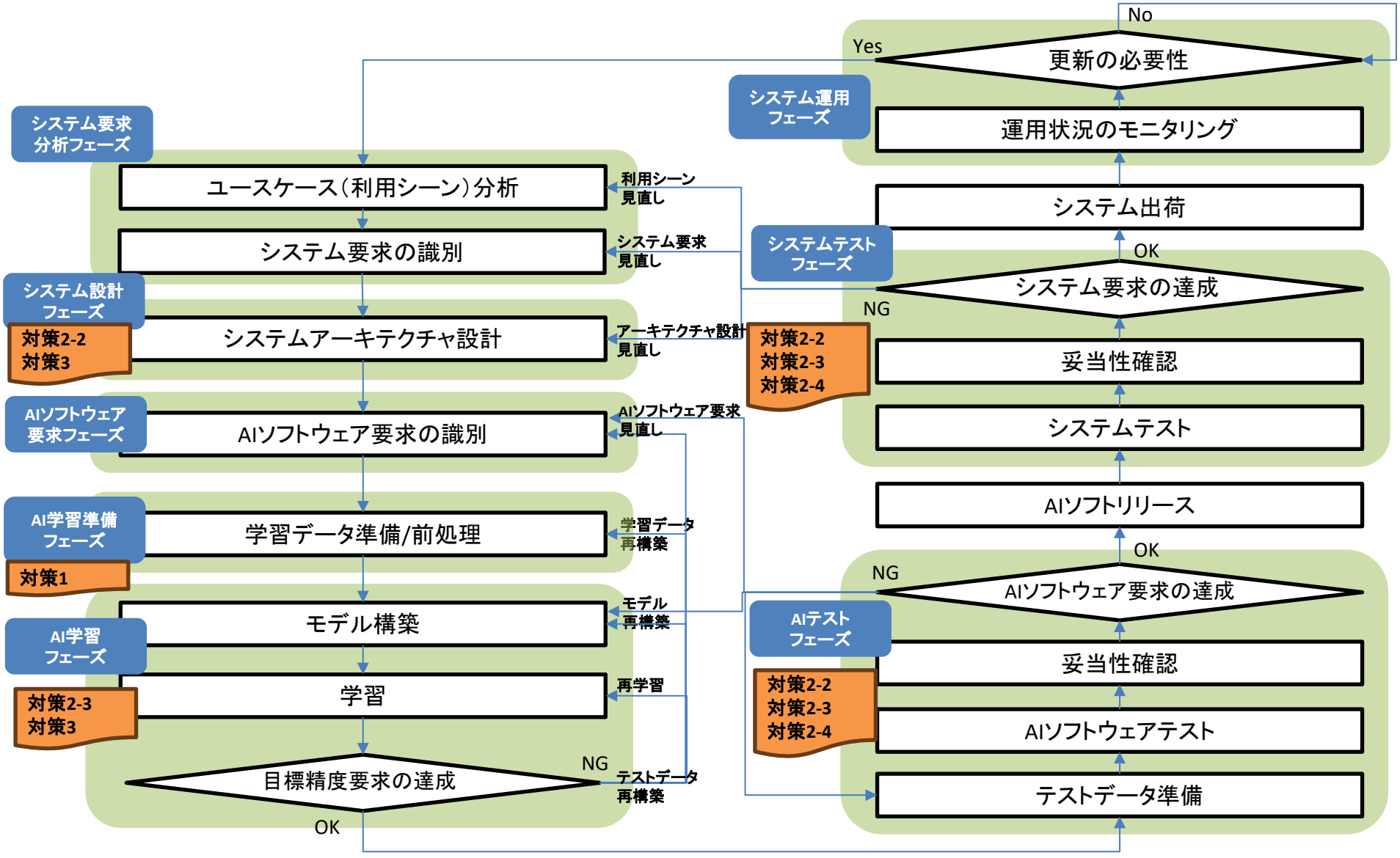
STEP3：推論結果の信頼度・確信度によった制御選択（フェールセーフ判断）

誤推論に対する対策

重要になるのはシステムレベルでの設計



誤推論しないための対策要求を含めたAIシステム開発フロー（全体像）



誤推論しないための対策要求を含めたAIソフトウェア開発プロセス要求

アクティビティ要求：
- （一般アクティビティ要求）一般的な各フェーズで実施する要求
- （誤推論対策アクティビティ要求）AIが誤推論したいための各フェーズ毎の対策要求

インプット（入力）：
- 各フェーズのアクティビティへの入力情報

アウトプット（成果物）：
- 各フェーズのアクティビティ実施後の成果物

誤推論しないための対策要求を含めたAIソフトウェア開発プロセス要求

（例）システム設計フェーズ

一般アクティビティ要求:

- システム要求・制約を基に、アーキテクチャ設計を行うこと。
- 利用シーン（例：走行シナリオ）を基に、各センサー・アクチュエータ・計算機システム等ハードウェアへの機能・非機能要求を明確化し、各センサー・アクチュエータの配置設計を行うこと。
- 学習・テストデータ準備に必要な、以下の要件を漏れなく識別すること。
  - アーキテクチャ設計情報(配置など)
  - ハードウェア特性
  - ハードウェアへ影響を与える外部環境条件
- 安全要求を満たすアーキテクチャ設計を行うこと。

誤推論対策アクティビティ要求:

- （必要に応じて）出荷後、利用シーンの変化、識別対象の変化、環境条件の変化等が生じる可能性があるため、入力データに対し、これらの変化を検知できる機能を設けること。
- 出荷後、運用中の状態をモニターして、必要に応じて、想定外の状況が発生したデータを保存する機能を設けること。
- AIソフトウェアは、想定外の推論結果を出力する可能性があるため、それを考慮した設計を行うこと。

インプット:

- 利用シーン（例：走行シナリオ）
- 識別対象分析結果
- システム要求仕様

アウトプット:

- アーキテクチャ設計仕様
- センサー/ハードウェア等の設計仕様

まとめ

- 本発表では、JAXAがまとめた、クリティカルシステムへのAIソフトウェア搭載に向けた『AI搭載ソフトウェア開発ハンドブック』の概要について紹介させていただいた。

- 今後は、利用者の声を反映（コンテンツの改良）していくとともに、AI対象機能の拡張に伴う本ハンドブックのアップデートを行っていく予定。

- ハンドブックにご興味がある方は、以下問い合わせ先までご連絡ください。

自動運転におけるAIの安全性に向けた取り組み～探索的アプローチとLLMの活用

国立情報学研究所　石川　冬樹
f-ishikawa@nii.ac.jp / @fyufyu
http://research.nii.ac.jp/~f-ishikawa/

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

自己紹介

- NII・総研大・電通大
  - ソフトウェア工学，特にディペンダビリティ：
    形式手法，自動テスト生成，安全性論証など
  - SE for AI & AI for SE
- JST MIRAI-eAI：機械学習型AIのエンジニアリング支援
- 産業界向け教育・実践研究
  - 日科技連SQiP，トップエスイー
  - 機械学習工学コミュニティ（MLSE研究会，QA4AI）
  - ソフトウェアテストコミュニティ（ASTER）

AI性能アラインメント技術
→お試し・実証実験大歓迎

年70名の実務者が受講
20期生募集

eAI

TOP SOFTWARE ENGINEERS EDUCATION PROGRAM FOR TOP SOFTWARE ENGINEERS

MLSE 機械学習工学研究会
MACHINE LEARNING SYSTEMS ENGINEERING

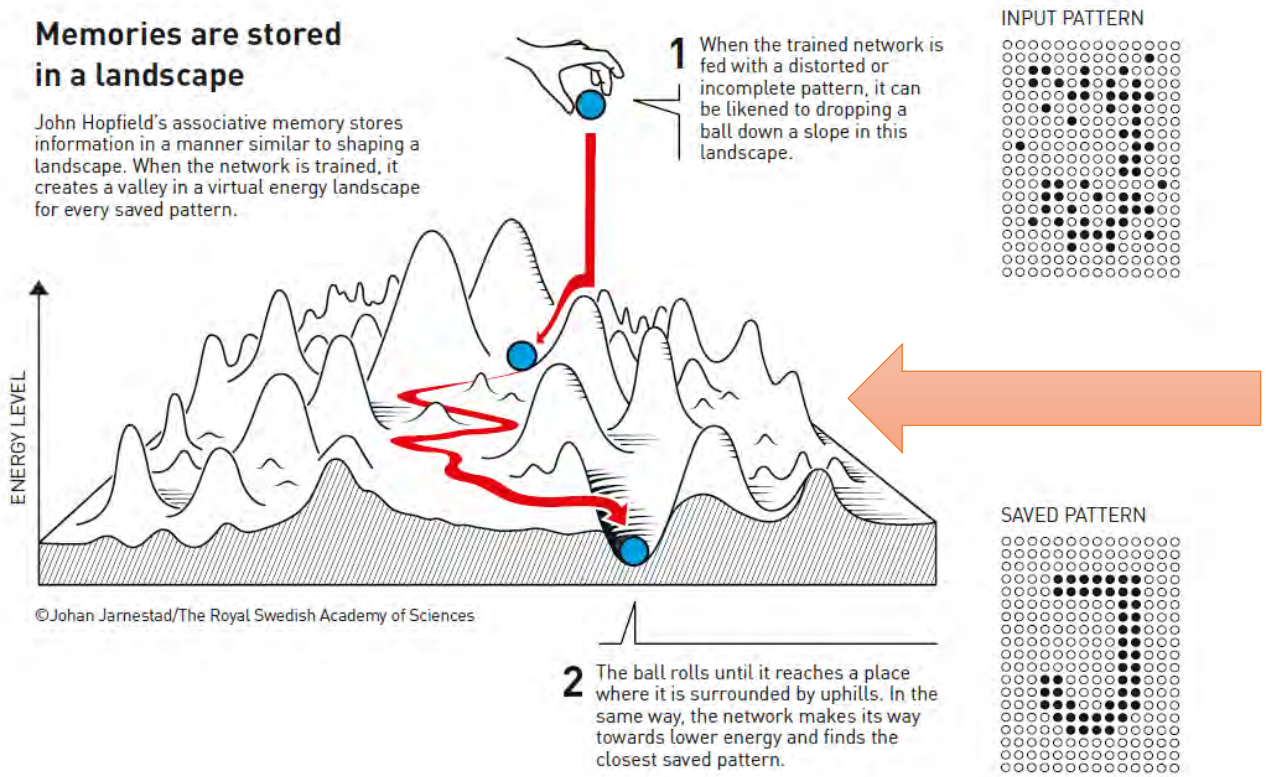
QAI
AIプロダクト
品質保証
コンソーシアム

目次

- 機械学習型AIの品質・トラスト
- 探索アプローチによる技術例
- 大規模言語モデル・生成AIによる変化

AIまたは「機械学習技術を用い構築した部品・システム」

- Hopfield先生・Hinton先生
  2024年度ノーベル物理学賞受賞！（2024/10/08）

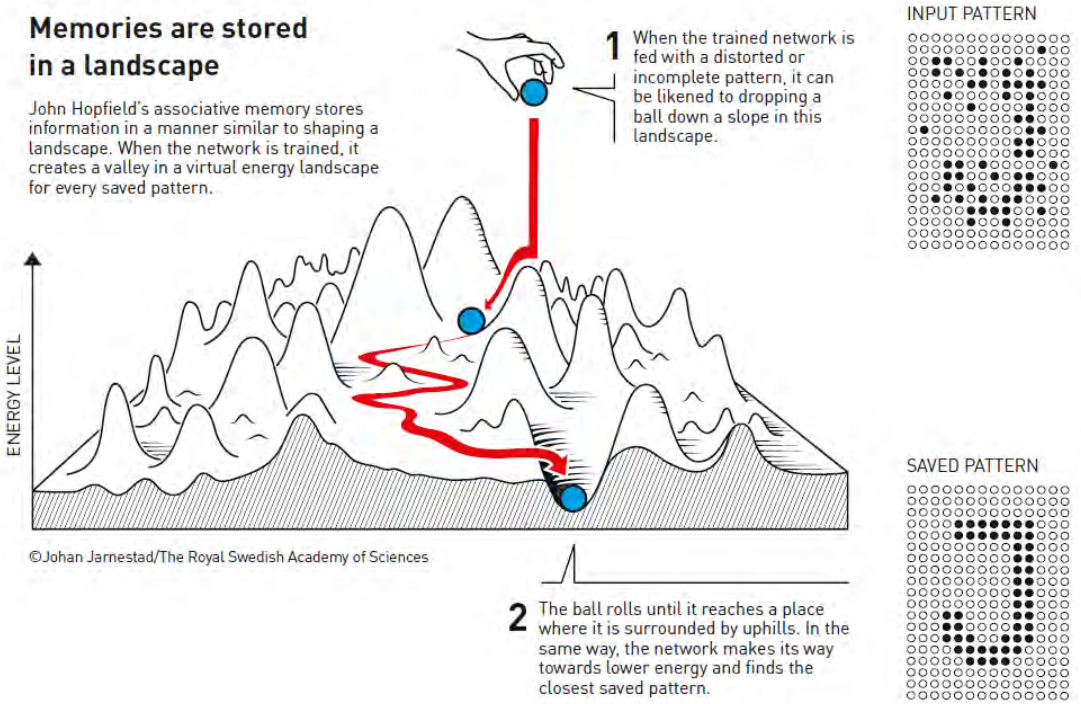


（仮想的な）エネルギー場で，
高低をうまく組み上げると，
入力に対する出力の求め方を
「覚える」ような仕組みになる

仮にこの山々がAIシステムだと
イメージしてみると，
その品質・安全性はどうする？？

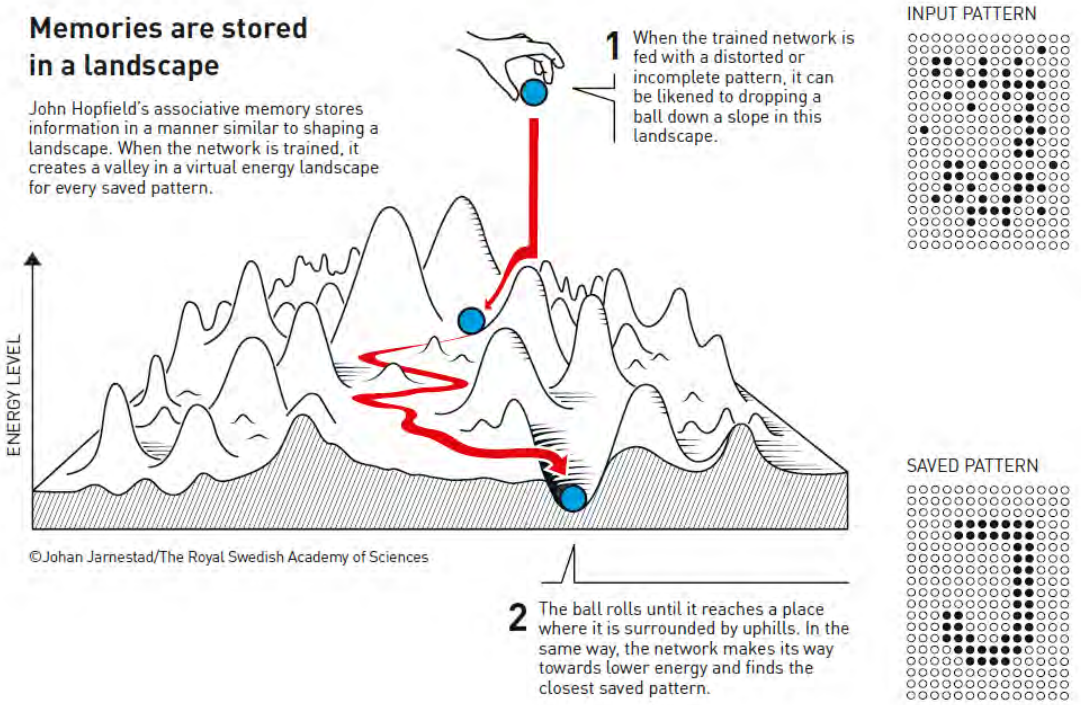
©Johan Jarnestad/The Royal Swedish Academy of Sciences
[https://www.nobelprize.org/prizes/physics/2024/press-release/]

AIのテスト・品質評価で議論されたこと（の一例）



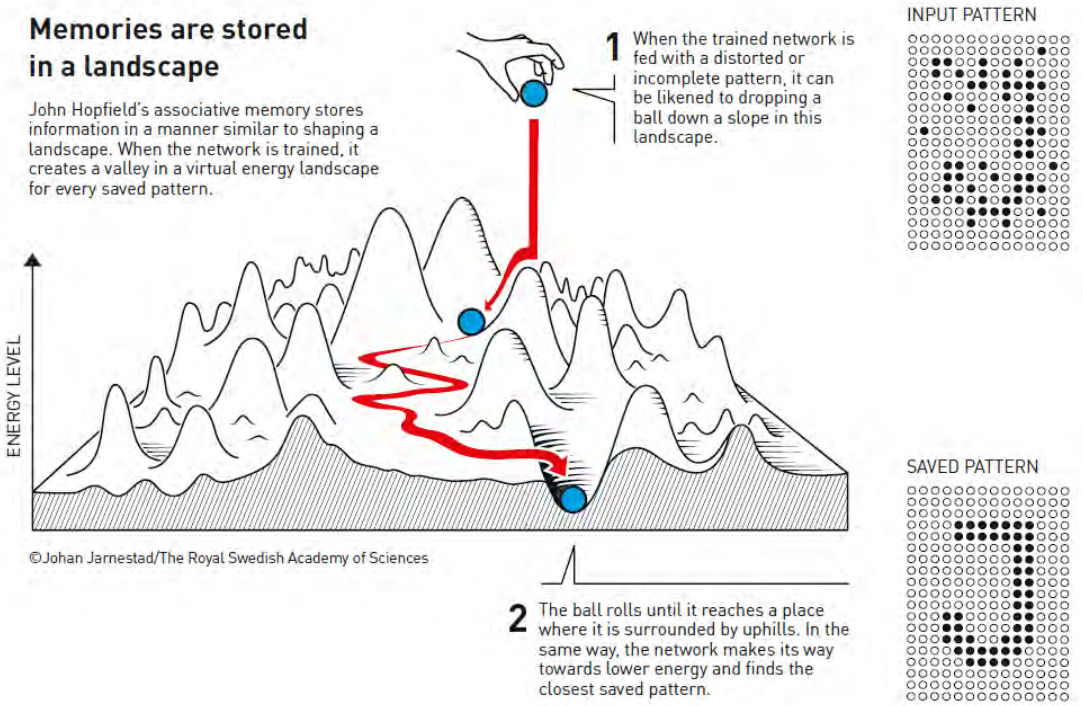
問題領域・要求に対する
テスト入力データのカバレッジ：
運用を想定したときに，
どこから球を落としてみるべき？

AIのテスト・品質評価で議論されたこと（の一例）



「実装を十分に検査した」って何だ？
多様な「動きパターン」は試したいが，
網羅を求める意味はなさそう
（ニューロンカバレッジなど）

AIのテスト・品質評価で議論されたこと（の一例）



予測性能（精度）の評価が基本だが,
正解が一意に決まらないケースや
用意するコスト・時間がとれない場合もある
（オラクル問題）

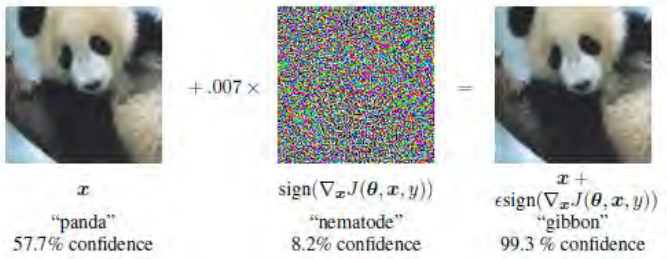
「この入力がここに落ちるというならば,
入力をこうずらすと落ち方は
こう変わるはず」というテストの作り方も
（メタモルフィックテスティング）

振り返り：機械学習型AIシステムの「品質」

- 2010年代初期：深層学習技術の発展を受けAI構築技術が進展
  - 広告推薦事例などを基にしたGoogleの指針解説など
- 2010年代後半：AIの品質や倫理に関する多くの動き
  - XAI（説明できるAI）の潮流
  - 国内では「AI品質」という観点から二つのガイドライン（AIQM, QA4AI）
  - 欧州では倫理という観点からのガイドライン
- 2020年以降
  - 倫理・トラスト・アライメントといった用語で人間・社会観点に焦点
  - ISO標準や政府レベルのガイドラインなどが発行フェーズに
  - ChatGPTなど対話型生成AIに対する議論に焦点が急速に移行

追及されたAI固有の品質特性（例）

頑健性
（入力ノイズで出力が変わる
敵対的サンプルの問題）



[Goodfellow et al., Explaining and Harnessing Adversarial Examples, 2015]

[Ackerman, Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms, IEEE Spectrum'17]

公平性
（過程や結果に社会的に
不適切な偏りがある問題）

[https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN] (access: 2024/07/19)

Google広告における推薦偏り

[L. Sweeney, Discrimination in Online Ad Delivery, ACM Queue'13]

説明可能性・解釈性
（出力の意味・根拠がわからない
と信頼・活用できない問題）

入力画像における
注目領域を出力させる技術の例

[Ribeiro et. al., " Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD'16]

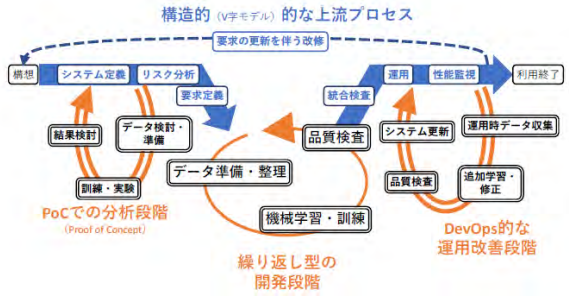
AIセキュリティ
（不正アクセスを要さないような
固有の攻撃がある問題）

訓練データ内の画像を推測した例

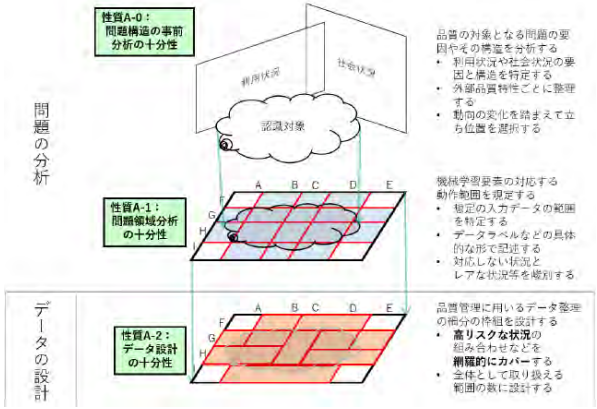
[Fredrikson et al., Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, 2015]

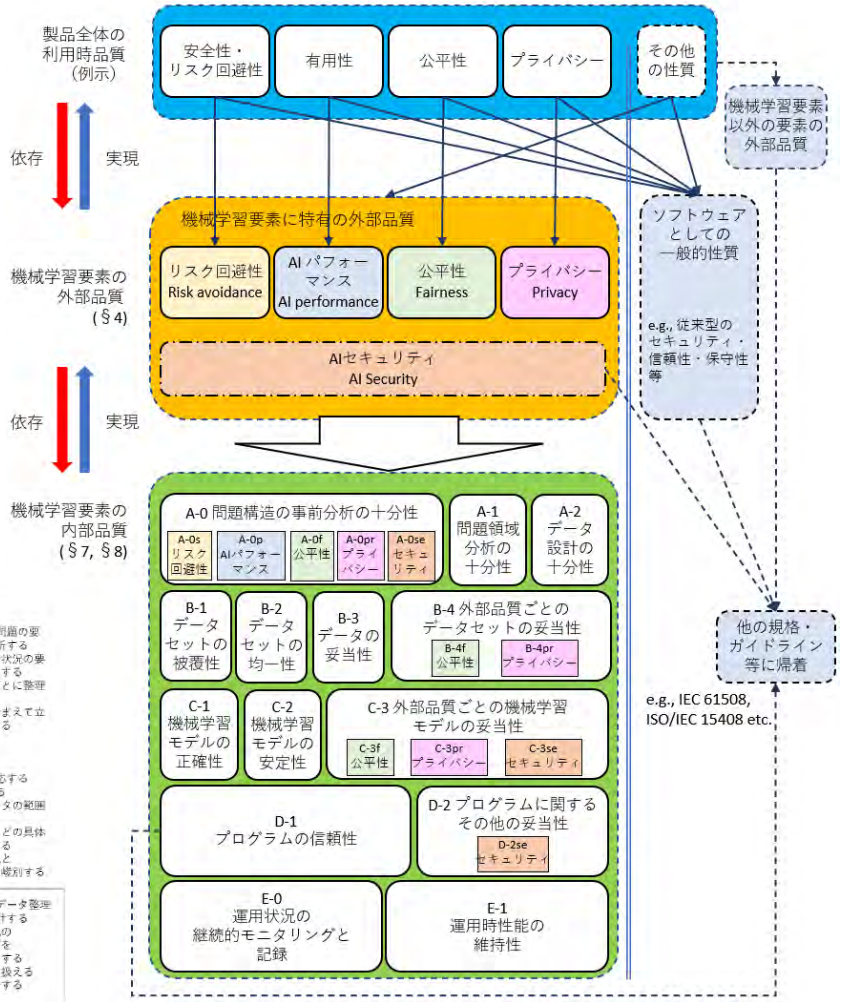
国内での指針（例）：AIQMガイドライン（2020~）

- 「品質モデル」を軸に規範・指針を提供
  - 品質の属性（観点）を定義
  - 認識対象に関する問題領域での「場合分け」に基づくデータ，モデルの品質評価が重要な軸
  - ISO/IEC TR 5469:2024 に反映



[AIQMガイドライン4.2.0より引用
https://www.cpsec.aist.go.jp/achievements/aiqm/]



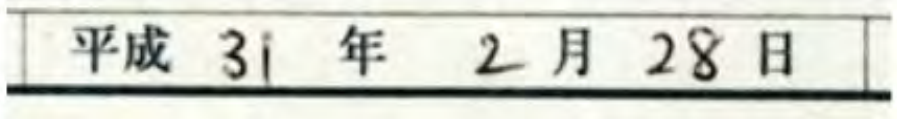


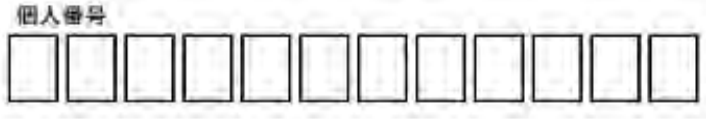
国内での指針（例）：QA4AIガイドライン（2019～）

- 具体的な取り組み指針の例：文字読み取り（OCR）

  どれだけの認識対象を考えていくか？

  - プレ印字の有無・色
  - 株式会社や年号の異なる表記
  - 印鑑かぶり
  - ボックス区切り・網掛け
  - …





[QA4AI 2024.01版より引用
https://www.qa4ai.jp]

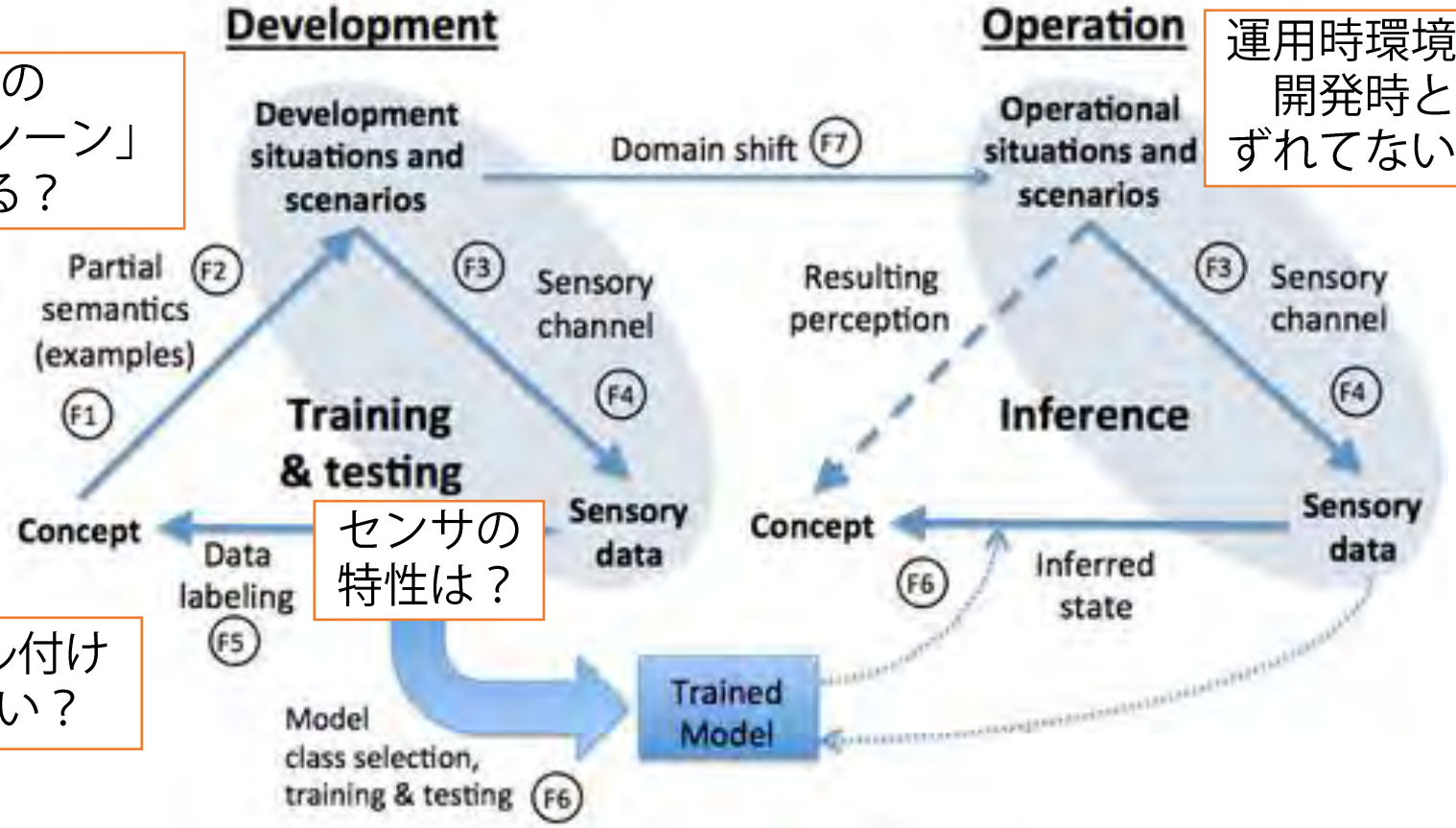
最近のガイドライン・標準（一例）

- 経産省・総務省 AI事業者ガイドライン（2024）
  - 包括的にワークシートを整理：「・・・しているか？」
  - 技術的な点も含むが，組織文化や利用者リテラシー確保まで言及

- EU AI Act（2024理事会承認）
  - 高リスクなAIに関する明示的な禁止を含む：
    人間の社会スコア，不特定の顔画像データベース構築，
    リアルタイムな生体情報収集の活用（特定目的に制限）など

自動運転に関する当初の課題提起の例 (1)

- 様々な不確かさを見積もり低減する必要性



どれだけの「歩行者」「シーン」を想定する？

運用時環境は開発時とずれてない？

「歩行者」の定義は？

センサの特性は？

正解ラベル付けぶれてない？

実装が何をやっているのか不確か

[https://uwaterloo.ca/waterloo-intelligent-systems-engineering-lab/projects/assuredai-safety-assurance-ai-based-automated-driving]

自動運転に関する当初の課題提起の例 (2)

- 自動運転レベル3～4における V&V の挑戦の一つ：
  「機械学習が入っていること」
- 例：運用・監視で意識する必要がある点
  - Unknown unknownsは検出できない
  - 不明確な状況でも高い確信度で結果を出すことがある
  - 意味のある特徴量で判断を下しているとは限らない
  - 入力分布の変化により性能変化が起きる

SaFAD:
ダイムラーなど10社以上による
自動運転に対する安全性原則の整理

[https://www.daimler.com/innovation/case/autonomous/safety-first-for-automated-driving-2.html]

その後 ISO TR 4804 としても発行

関連する標準の例 (1) ANSI/UL 4600

- ANSI/UL 4600：自動運転をはじめとした自律プロダクトのためのセーフティケースに対する要件
  - 入力空間のうち，安全性に関わる部分集合に対して性能を論じる
  - 運用時における入力の分布変化について監視する
  - 運行設計領域での多様な環境条件での予測性能評価を行い，弱点となる条件を明らかにする
  - 新しい入力に対して既知かのように誤りを犯すことに留意する
  - …

関連する標準の例 (2) ISO/PAS 8800

- ISO PAS 8800:2024： 26262 FuSAおよび21448 SOTIFに対し，機械学習型AI固有の性質や考慮すべきリスク要因を定義
  - AIの誤りが安全性に影響するかどうかの分析が必要
  - Systematic Error / Systematic Fault の考え方を機械学習でも：「特定の入力領域に対する訓練不足が原因となり，特定の状況で誤りが多発，性能が不十分になる」など
  - 更新の難しさ（破滅的忘却，後述）などの留意点も列挙

まとめ：これまでのAIの品質・トラスト

- 「従来AI」： 特定タスクを想定しデータから機能を構築
  - 正解率など予測性能（精度）がやはり主流
  - 訓練・評価双方においてデータ品質が最重要
  - 公平性や説明可能性など，技術だけの問題ではない固有の観点
- 2019年前後に方向性は確立・広く合意
  - AIQM・QA4AIガイドラインなどで，開発者の考え方はよく整理
  - 事業者ガイドライン，EU AI Act などは，組織全体の話題も含むが，根本的には考え方はそのまま（EUでの特定の禁止事項はあるが）

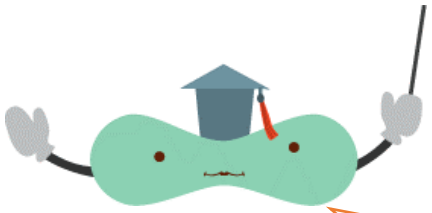
理想・規範を踏まえ「どこまでどうやるか？」が実際の問題

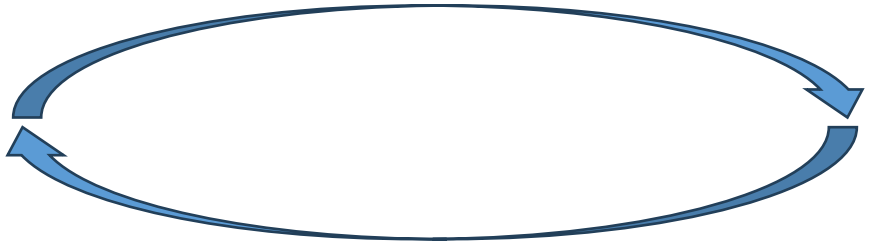
目次

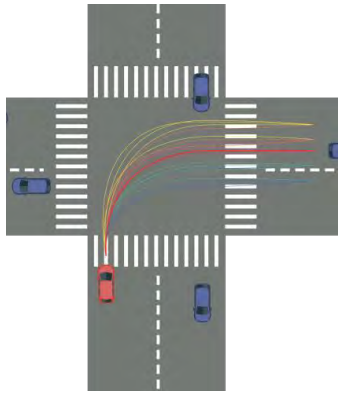
- 機械学習型AIの品質・トラスト
- 探索アプローチによる技術例
- 大規模言語モデル・生成AIによる変化

ベースとなる技術：探索・最適化によるエンジニアリング

シミュレーター上で他車の配置や動きをうまく決め，
不要な急加速でぶつかってしまうような
テストシナリオを見つけたい！





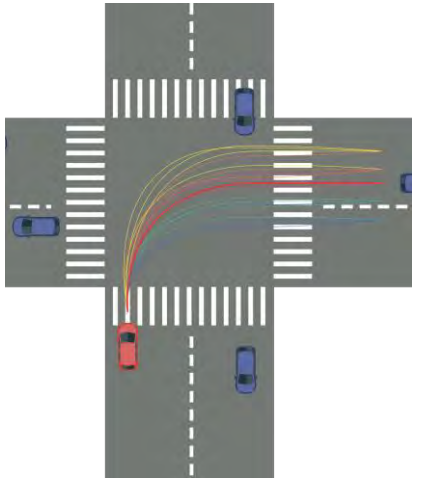


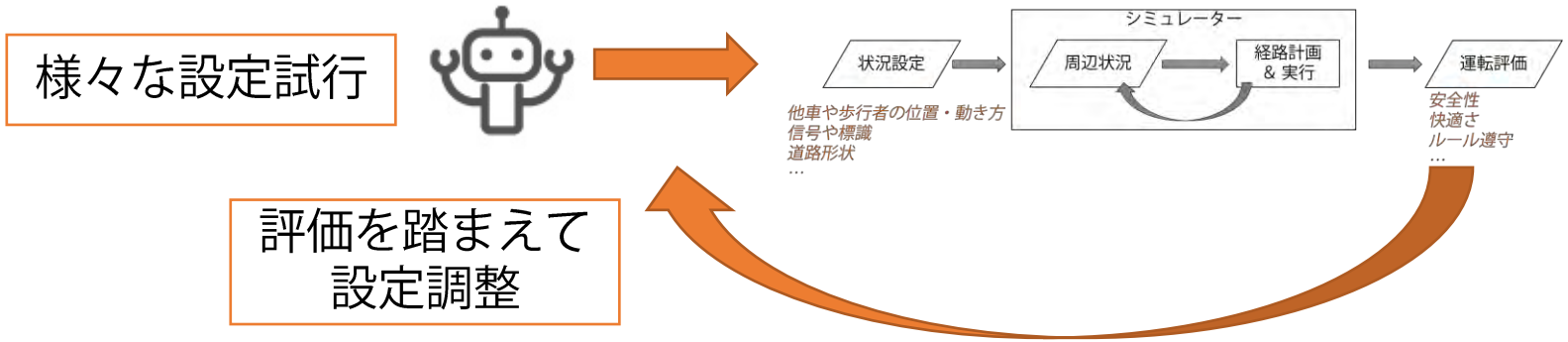
こういう設定でやってみたら
急加速レベル50点，危険度75点
次はここを変えて試してみるか！

技術用語としては Search-based Software Engineering という一大分野があり，
進化計算などのメタヒューリスティック最適化を活用

事例 (1) 自動運転の経路計画機能のテスト・デバッグ

- ハンドル角やアクセル・ブレーキ量を決定
- 「右折のケースをテストしよう」の中に膨大な可能性が含まれる！
  - 特定の配置やタイミングで事故が起きるかも



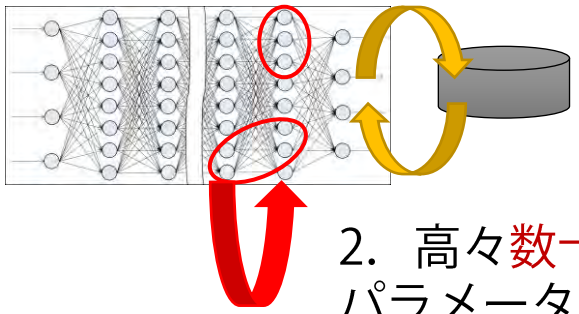


安全性だけでなく快適さなど
多数の要求違反を効率よく検出

パラメータ空間や振る舞いの
観点からのカバレッジも追求可能

事例 (2) 認知AIにおけるリスク低減

- カメラに写った物体を分類するなどの認知AI
  - 「90%正解」のような全体評価ではなく，リスクの分析が重要
    - 例：「歩行者をバイクに乗っていると間違える」
      → ブレーキをせずに衝突するのはとても危険
    - 特定の誤り（複数）を減らすような訓練は深層学習では困難
      - 数百万超のパラメーターの調整，「破滅的忘却」も発生



1. 訓練データや運用データから
致命的な誤りの要因を分析

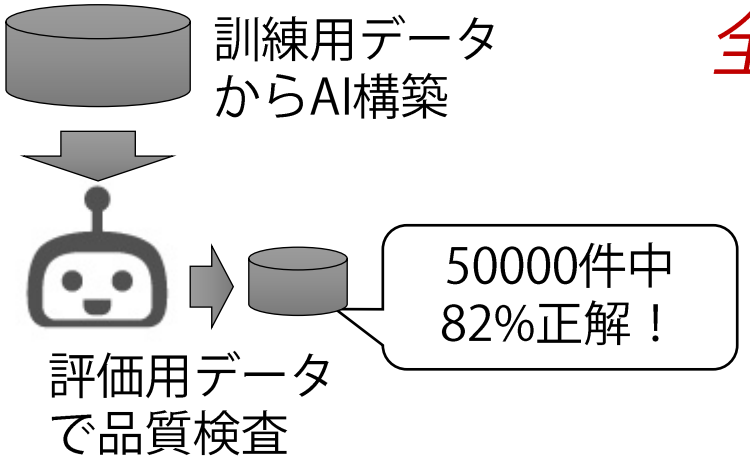
2. 高々数十個程度に絞って
パラメーターを探索・調整

→

- 致命的な誤りに対して
高い効果の修正が可能
- 成功していた部分に影響を
与えない修正を追及可能
- 複数の修正案を
効率よく提示可能

改めて：AIと安全性のギャップ

古典的なAI品質



総体構築
全体平均

改善反復
個別評価

従来の安全性論証

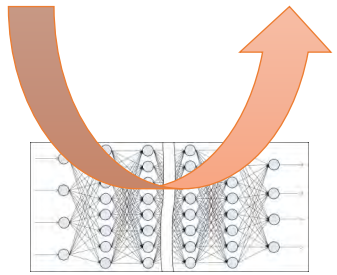
故障	ハザード	許容可否・対応
ブレーキランプへの 配線が切れたら？	ブレーキを踏んだときに 後ろの車が気づかず衝突！	リスク高・許容できない → 線を二重化しよう
…	…	…

“Engineerable AI” プロジェクトでの研究：実証側面

番号	AIの誤り種別	シーン	ハザード	リスクレベル	AI評価	許容可否
001	誤分類： 歩行者 → バイク搭乗者	自車の前の 歩行者	ブレーキせず 歩行者に衝突	5	誤り 4%	○
002	誤分類： バイク搭乗者 → 歩行者	近距離で追随する 後方車両	不要なブレーキで 後方から衝突	3	誤り 35%	×
…	…	…	…	…	…	…

1. 安全性の専門家による分析
→ 数百件のリスク要因

2. リスク要因を踏まえた細粒度のAI性能評価

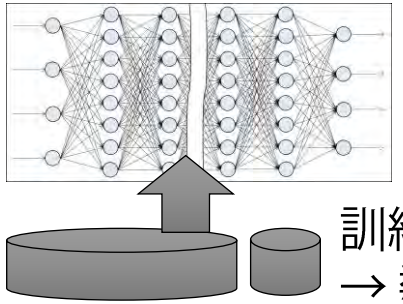


3. リスク要因を踏まえたAI性能修正の実証

※ 上記は簡易化したイメージ
実際は ISO 21448 (SOTIF) 的な分析プロセス
→ 20個超の安全性指標を評価

“Engineerable AI” プロジェクトでの研究：技術側面

従来



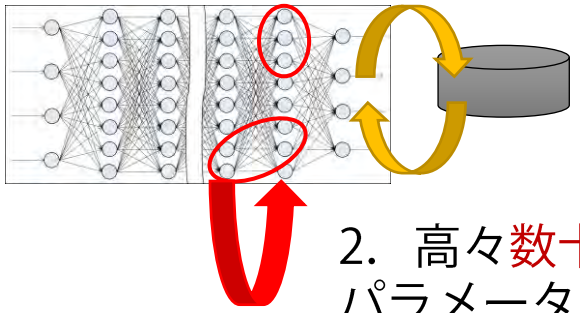
“Changing anything changes everything”

訓練データ追加による再訓練
→ 数百万個以上の
パラメーターを「シャッフル」

更新結果が不確実・制御困難！

意図せぬ性能劣化

提案：深層学習デバッグ技術



1. 訓練データや運用データから
致命的な誤りの要因を分析

2. 高々数十個程度に絞って
パラメーターを調整

調整対象を絞るから

- 致命的な誤りに対して
高い効果の修正が可能
- 成功していた部分に影響を
与えない修正を追及可能
- 複数の修正案を
効率よく提示可能

従来プログラムの欠陥に対する技術を発展！

[https://www.nii.ac.jp/news/release/2023/0317.html]

技術例：DistrRep（動機）

[with Calsi+, Distributed Repair of Deep Neural Networks, ICST'23]

- 「誤りによるリスク」は複数種類ありレベルが異なる
  - 例：eAIプロジェクトでの安全性ベンチマーク
    - 「歩行者をバイク搭乗者と間違える」
      → 誤った「ブレーキなし」 → 致命的な事故
    - 「バイクを自転車と間違える」
      → 誤った「ブレーキ」 → 遅延・後ろからの衝突（可能性・影響やや低）
  - 混同行列のようにラベルの組だけ考えても多数の可能性
  - 実際には「緊急シーン」「渋滞時」など属性の考慮も
  - 標準での記載例：“characterize performance on safety related subsets of input space” (ANSI/UL 4600)

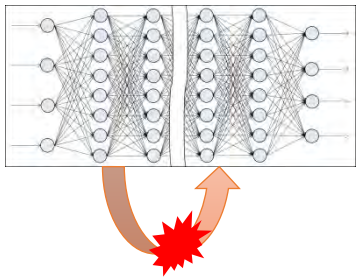
技術例：DistrRep（概要）

[with Calsi+, Distributed Repair of Deep Neural Networks, ICST'23]

[https://www.nii.ac.jp/news/release/2023/0317.html]

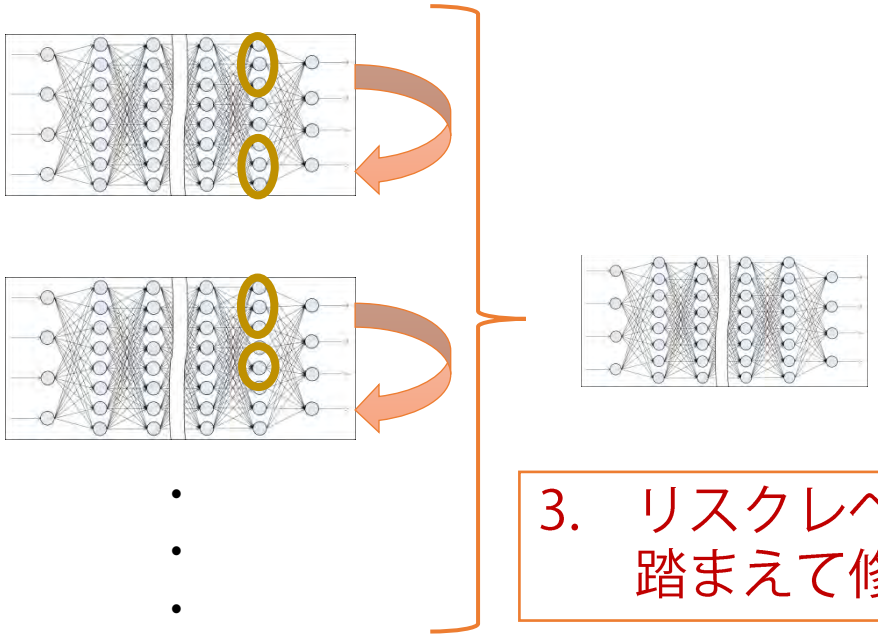
ケースバイケースの
系統的なリスク分析

画像識別AIの誤り種類	影響の分析	リスクレベル	現AIでの発生率評価
歩行者を別の物体と誤識別	歩行者が目の前にいるとき…	5	受入れ不可
別の物体を歩行者と誤識別	…	3	受入れ可
…	…	…	…



すべてを考慮したAI修正のため
DNNの数百万超のパラメーターを
調整することは従来技術では困難！

DistrRep技術による
DNN修正



3. リスクレベルによる優先度を踏まえて修正案を統合

それぞれの種類の誤りごとに、
1. 要因となっているパラメーター群を絞り込み
2. その誤りを修正するためのパラメーター変更を探索
→ 認識誤りの種類ごとの修正案をまず作成

目次

- 機械学習型AIの品質・トラスト
- 探索アプローチによる技術例
- 大規模言語モデル・生成AIによる変化

大規模言語モデル・生成AIの影響

- 方向１：自動運転の一部になる＝テスト・評価の対象に
  - オープンな知識・エッジケースを扱える（か評価する）
  - 例：「ビニール袋だから止まらなくてよい」，「カフェの椅子が路上に転がっているので止まる」など訓練を経ず判断
- 方向２：テスト・評価の道具として使う
  - 詳細な属性ラベル付けの道具になる
  - 例：「路上にいる歩行者」，「レアな形状の車」など，幅広い属性ラベルを付けたり検索したりできる

方向１：大規模言語モデル・生成AIの評価

- 現在はテキスト入出力を中心にかなりの取り組み・議論
  - 想定すべき入出力が広い
  - 評価基準があいまい，広い，評価自動化の実装が困難
  - …
  - 自動運転だと，画像・動画を扱う想定なので，評価のコストがさらに高くなる？
    （例：評価データの準備）

本章での用語	SQuaRE for AI	[Guo+, arXiv23]
QC01：回答性能	Functional Correctness	
QC01-1：自然言語処理における回答性能		Question Answering, Knowledge Completion, Reasoning
QC01-2：ツール活用に関する回答性能		Tool Learning
QC01-3：創造性・多様性に関する回答性能		–
QC01-4：制御可能性	User Controllability	–
QC02：事実性・誠実性	Functional Correctness	Question Answering, Knowledge Completion, Truthfulness
QC02-1：一般的な知識に対する事実性・誠実性		
QC02-2：与えた知識に対する事実性・誠実性		
QC02-3：根拠の説明性・妥当性		–
QC03：倫理性・アラインメント	Societal and Ethical Risk Mitigation	Ethics and Morality
QC03-1：公平性		Bias
QC03-2：安全性		Toxicity, Risk Evaluation
QC03-3：データガバナンス		Risk Evaluation
QC04：頑健性	Robustness	Robustness Evaluation
QC05：AI セキュリティ	Security	Robustness Evaluation

[QA4AIガイドライン2024.01版における品質特性の整理]

方向２：大規模言語モデル・生成AIの道具としての活用

- 一通りのタスクが何でもできる
  - が，LLMだけにEnd-to-Endでやらせるのがベストとは限らない
- 不定型・不完全なものでも「常識」でさばいてくれる

➡必要な準備・適用の前提が低く，今までと違うことができる

- データのフォーマット・スキーマなどの統一・形式化や，
  大量データを用いた事前訓練などが（従来よりは）なくてもよい
- 日本語における言葉・言い回しのぶれを苦にしない
- 明文化していないことも，一般的な「常識」なら補完してくれる
- 構造化・形式化されたデータを扱う従来技術にもつなげられる

取り組み例：画像認識AIの探索的テスト

- 画像認識AIの「系統的な弱点」を見つけたい
  - 訓練データ不足などに起因する，特定状況の不十分な認識性能
  - AIQM・QA4AIガイドライン，ANSI/UL 4600，ISO PAS 8800など
  - これまでは人間が定めた固定のデータラベル内でのテスト・評価
    - 例：「都市部で歩行者が近くにいる画像」
  - コストが高すぎるとともに，オープンな世界の一部しか扱えない

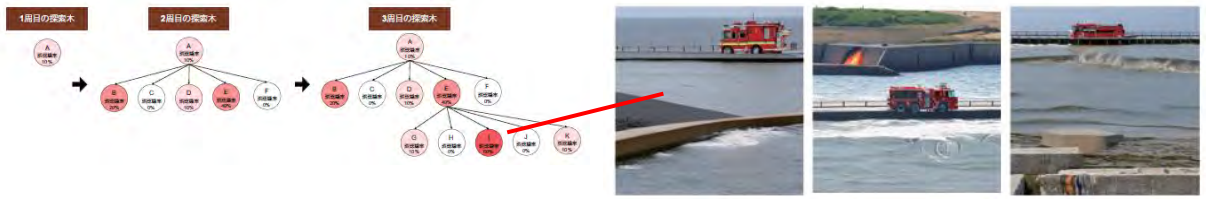
➡ 対話型生成AIの「常識」による多様な概念のカバー
＋ 画像生成AIによる低コストなテスト・評価

[Torikoshi+, AdaSniper: An Adaptive Automated Approach for Systematic Error Detection in Image Recognition Models, FOSE'24]

取り組み例：画像認識AIの探索的テスト

- 対話型生成AI (GPT)：アイディア出しを繰り返し
  - 今まで見つけた「系統的な弱点」を踏まえ，次に探る状況を列挙
  - 誤り検出重視・多様性重視は調整可能
- 画像認識AI (StableDiffusion)：認識性能評価用の画像生成
  - 補足：自然画像と生成画像の弱点検出能力も別途比較済み

探索の過程を表現した木



例：「防波堤近くの消防車」が弱点

GPTが似たような弱点と概念をまとめて
「水が多い状況での消防車」と報告することも可能

[Torikoshi+, AdaSniper: An Adaptive Automated Approach for Systematic Error Detection in Image Recognition Models, FOSE'24]
[Yokoyama+, Investigating the Applicability of Image Generative Models to Weakness Detection Tasks, SES'24]

全体まとめ：AIと品質

- 高い不完全さ・不確実性が原則に
  - 要求・環境・傾向の膨大さ ＋ 実装した挙動の理解困難性
  - 実世界や社会に大きく踏み込むことが増加
  - 不確かなものに対する試行錯誤の反復継続
  - めまぐるしく変わる技術と世界

ステークホルダーの議論を通し不完全・不確かなものを受け入れ
自動化された探索・テスト・測定・監視による継続的進化を
&
新しい技術・時代の変化を楽しみましょう！

2024年度 宇宙航空安全・ミッション保証シンポジウム
「自律化技術・AI」×「Assurance」

宇宙機へのAI搭載に向けた ソフトウェア品質保証の取り組み

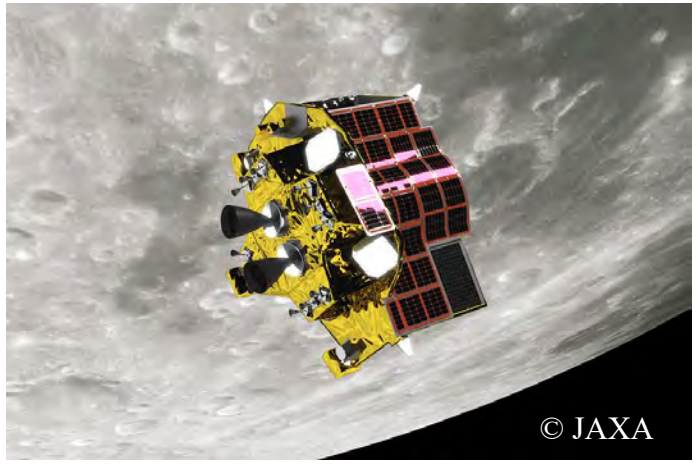
2025年1月15日

宇宙航空研究開発機構
安全・信頼性推進部
神戸 大輔

アジェンダ

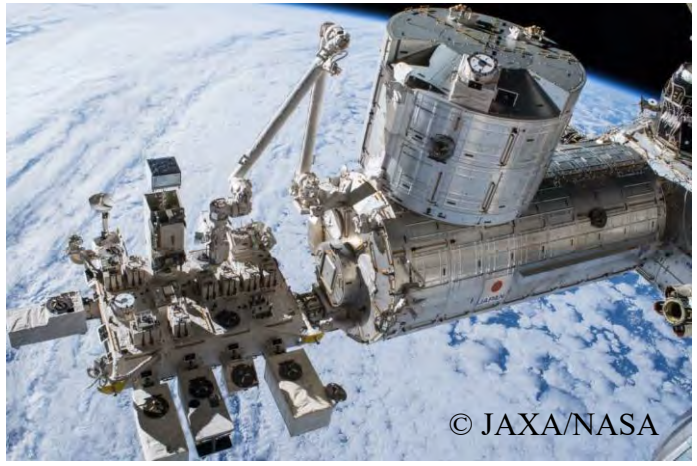
- 宇宙機ソフトウェアに対する従来の品質保証
- 宇宙機へのAI搭載に向けたソフトウェア品質保証の取組み

宇宙機ソフトウェアの特徴



- 衛星/探査機
  - Attitude Control, Data Handling, Mission, Sensor
  - Fail Safe → Fail Operative
  - FDIR: Fault Detection Isolation and Recovery
  - Rewritable
  - Assembly / C language
  - Reuse
  - Bit upset by radiation

- 有人システム
  - Central Control system, Robot Arm (manipulator), payload … so many software
  - Safety Requirement (2-Fail Operative)
  - Complex and large amount of FDIR
  - Include the crew (human) in system control loop
  - Ada / C language



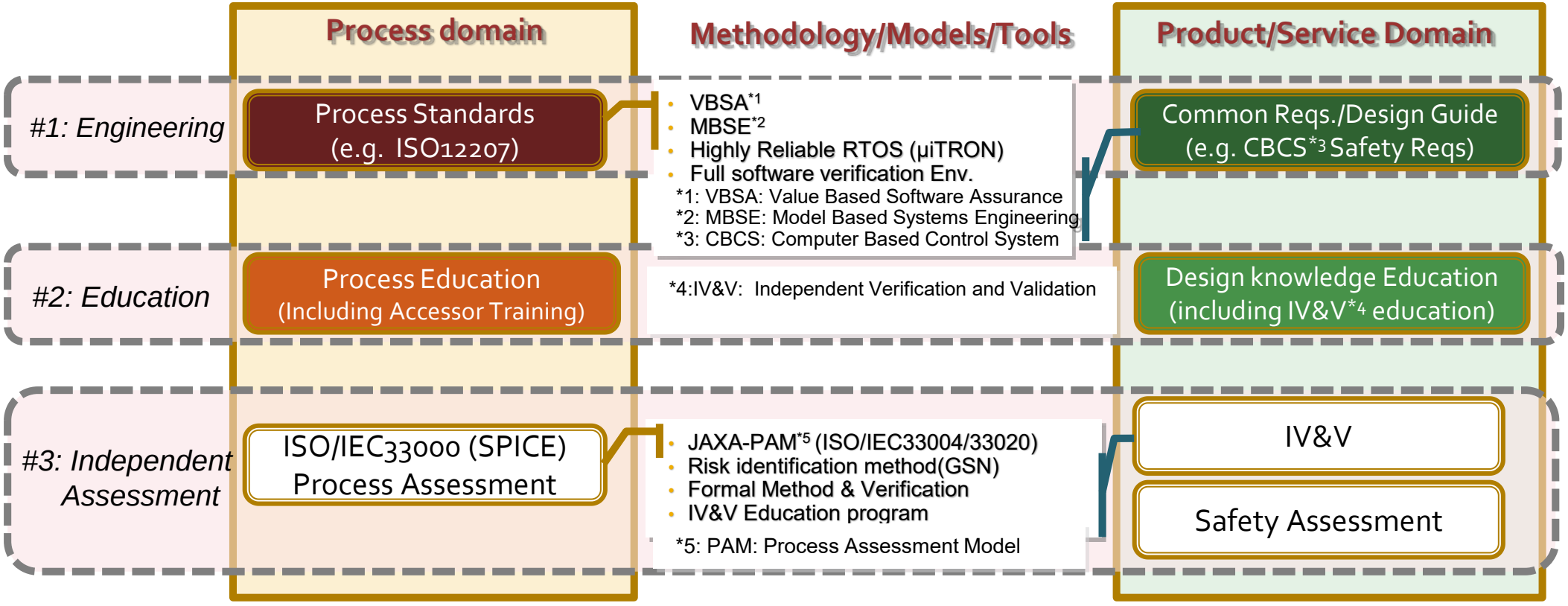
- ロケット
  - Guidance Navigation Control etc.
  - High Reliability / Voting
  - Assembly / C language
  - Hard real-time

宇宙機ソフトウェアの品質保証アプローチ

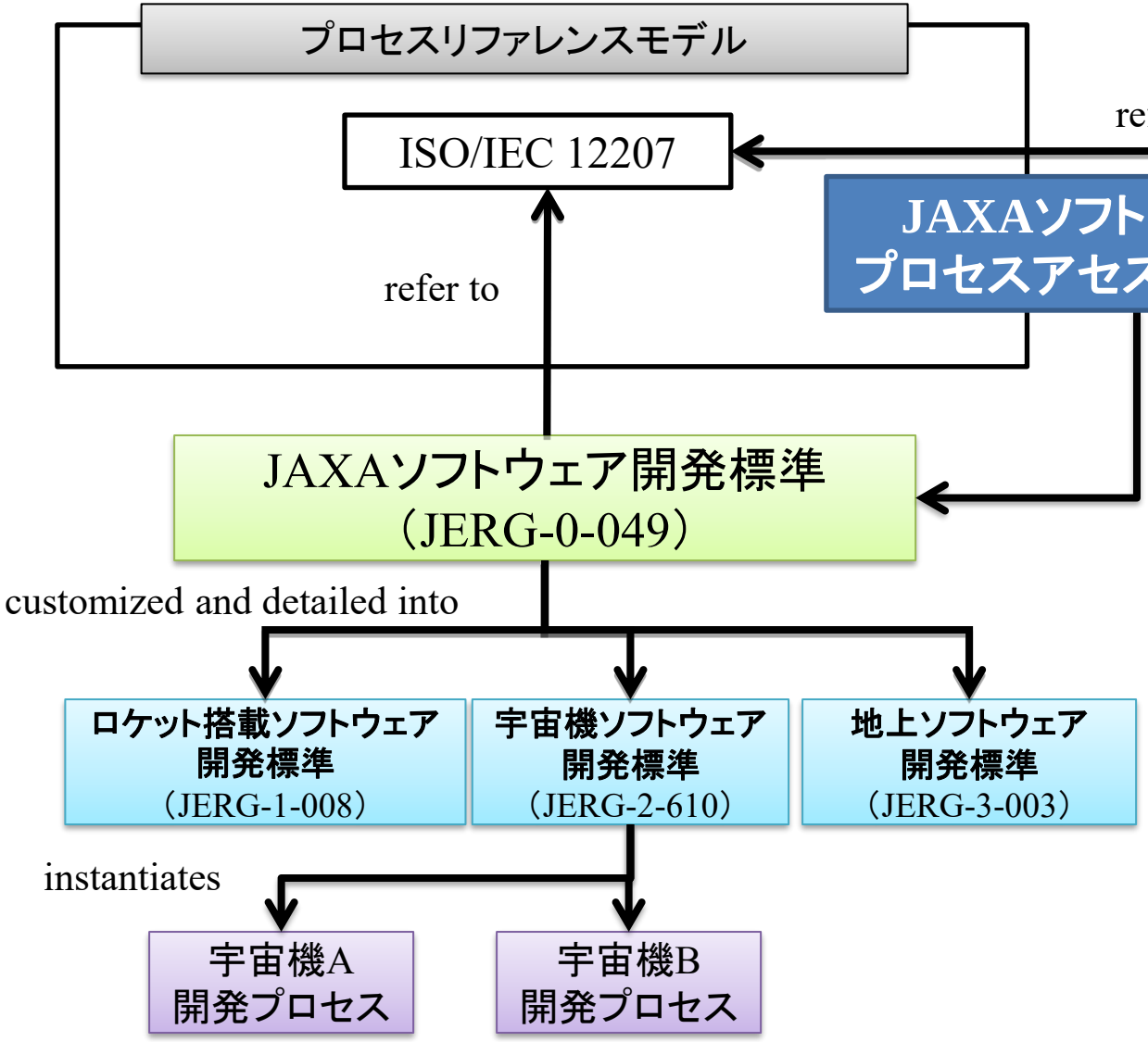
First Priority *"Mission Success of space systems"*

No matter how high the reliability or process maturity, a positive success is more important.

Approaches to achieve the goal



ソフトウェア開発標準とプロセスアセスメント



- 対象システムに応じたソフトウェア開発標準（プロセス要求）の策定と適用
- プロセスアセスメントモデル（ベストプラクティス）の開発と活用

ソフトウェア開発標準（プロセス要求）

4　一般要求事項
4.1　テーラリング

5　主ライフサイクルプロセス

5.3　開発プロセス
5.3.1 プロセス開始の準備
5.3.2 全開発プロセス適用事項
5.3.3 コンピュータシステム要求分析
5.3.4 コンピュータシステム方式設計
5.3.5 ソフトウェア要求分析
5.3.6 ソフトウェア設計
5.3.8 ソフトウェア製作
5.3.10 ソフトウェア統合
5.3.11 ソフトウェア統合試験
5.3.12 目標プラットフォームへのインストール(組み込み)
5.3.13 コンピュータシステム統合およびコンピュータシステム総合試験
5.3.14 ソフトウェア製品の供給と導入
5.3.15 ソフトウェア製品の受け入れ

5.4　運用プロセス
5.4.1 プロセス開始の準備
5.4.2 運用試験
5.4.3 ソフトウェアを含むコンピュータシステムの運用
5.4.4 運用結果の管理
5.4.5 顧客およびユーザサポート

5.5　保守プロセス
5.5.1 プロセス開始の準備
5.5.2 問題把握および修正分析
5.5.3 修正の実施
5.5.4 ソフトウェアの書き換え
5.5.5 ロジスティクス支援の実施
5.5.6 保守およびロジスティクス結果の管理
5.5.7 移行
5.5.8 ソフトウェア廃棄

6　支援ライフサイクルプロセス

6.1　文書化プロセス
6.1.1 プロセス開始の準備
6.1.2 作成
6.1.3 発行
6.1.4 維持・改訂・廃棄

6.2　構成管理プロセス
6.2.1 プロセス開始の準備
6.2.2 構成識別
6.2.3 構成変更管理
6.2.4 構成変更状況の記録
6.2.5 構成変更状況の評価
6.2.6 リリース管理と出荷
6.2.7 構成監査の実施

6.3　品質保証プロセス
6.3.1 プロセス開始の準備
6.3.2 製品およびサービス品質の保証
6.3.3 プロセスの保証
6.3.4 品質システムの保証
6.3.5 品質保証記録の管理

6.4　検証プロセス
6.4.1 プロセス開始の準備
6.4.2 検証
6.4.3 検証結果の管理

6.5　妥当性確認プロセス
6.5.1 プロセス開始の準備
6.5.2 妥当性確認
6.5.3 妥当性確認結果の管理

6.6　共同レビュープロセス
6.6.1 プロセス開始の準備
6.6.2 プロジェクト管理レビュー
6.6.3 技術レビュー

6.7　アセスメントプロセス
6.7.1 プロセス開始の準備
6.7.2 アセスメントの実施

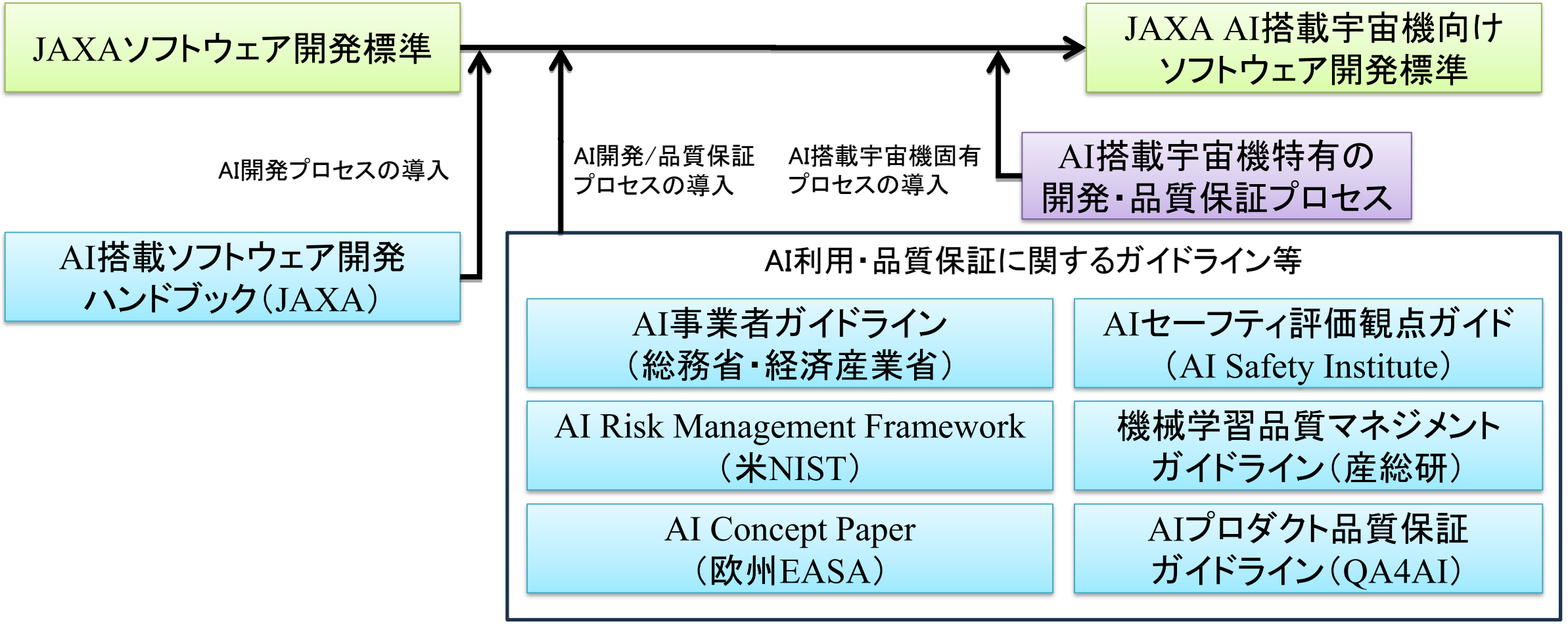
6.8　問題解決プロセス
6.8.1 プロセス開始の準備
6.8.2 問題の解決
6.8.3 予防
6.8.4 傾向分析

ソフトウェアライフサイクルにおける開発/運用/保守/支援プロセスの要求

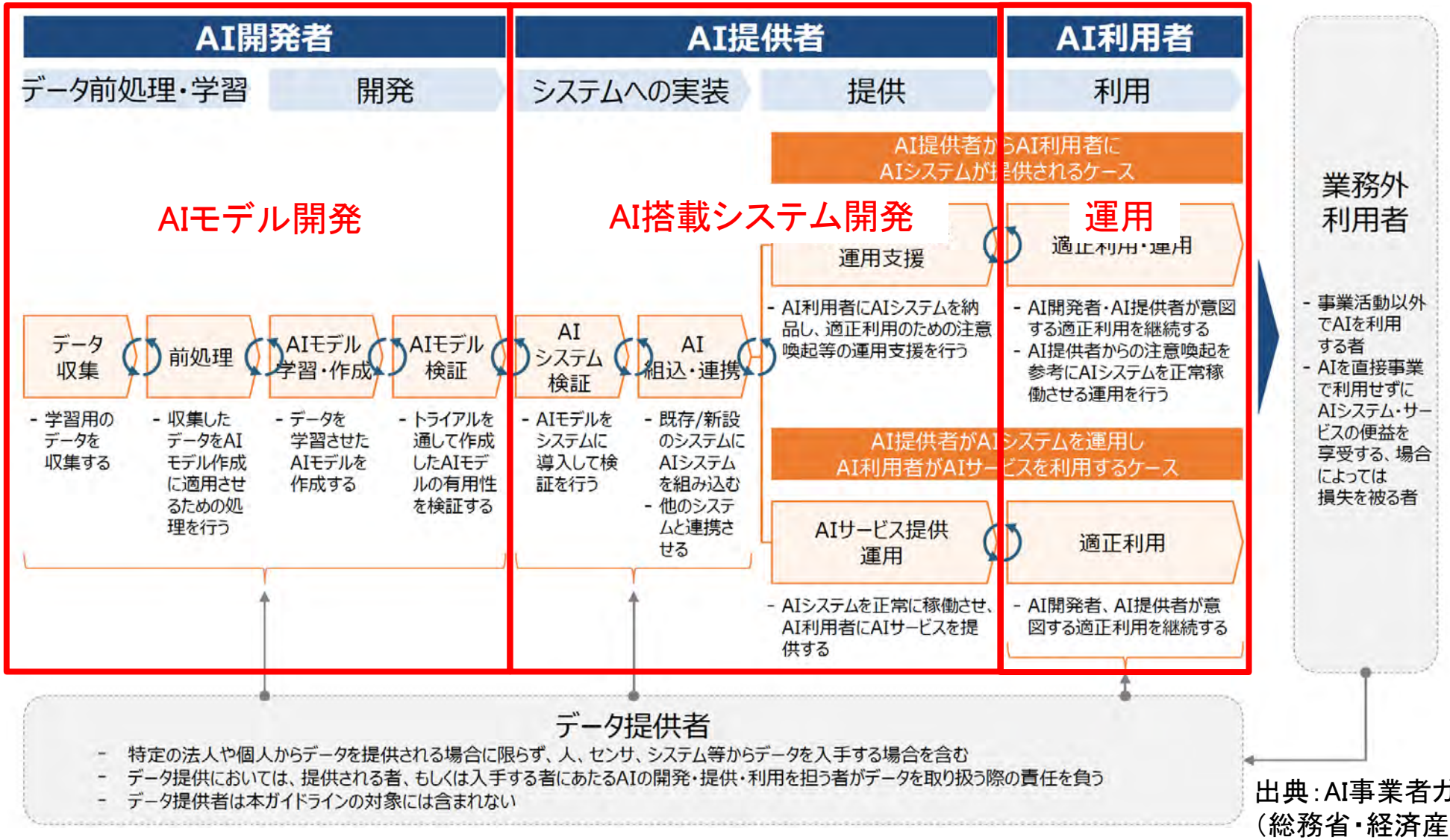
AI搭載により、必要となるプロセス・アクティビティを拡張

AI搭載に向けたソフトウェア開発標準（プロセス要求）の整備

- ソフトウェア開発標準へのAI固有の開発/品質保証プロセスの拡張



AIの開発から利用までのバリューチェーン



出典：AI事業者ガイドライン 第1.0版（総務省・経済産業省）

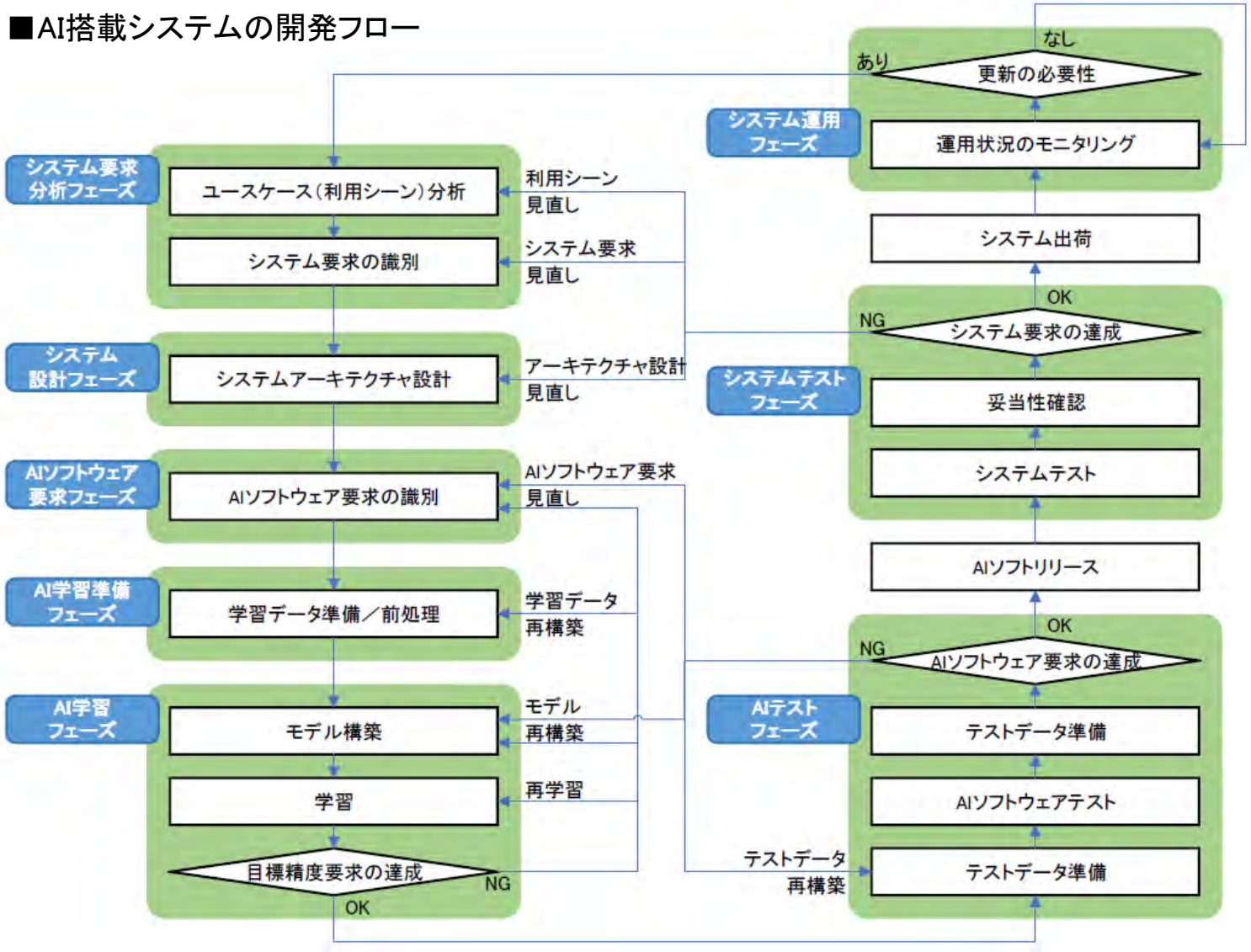
AI搭載システム開発プロセス（例）

■従来型とAI搭載システムの開発プロセスの比較

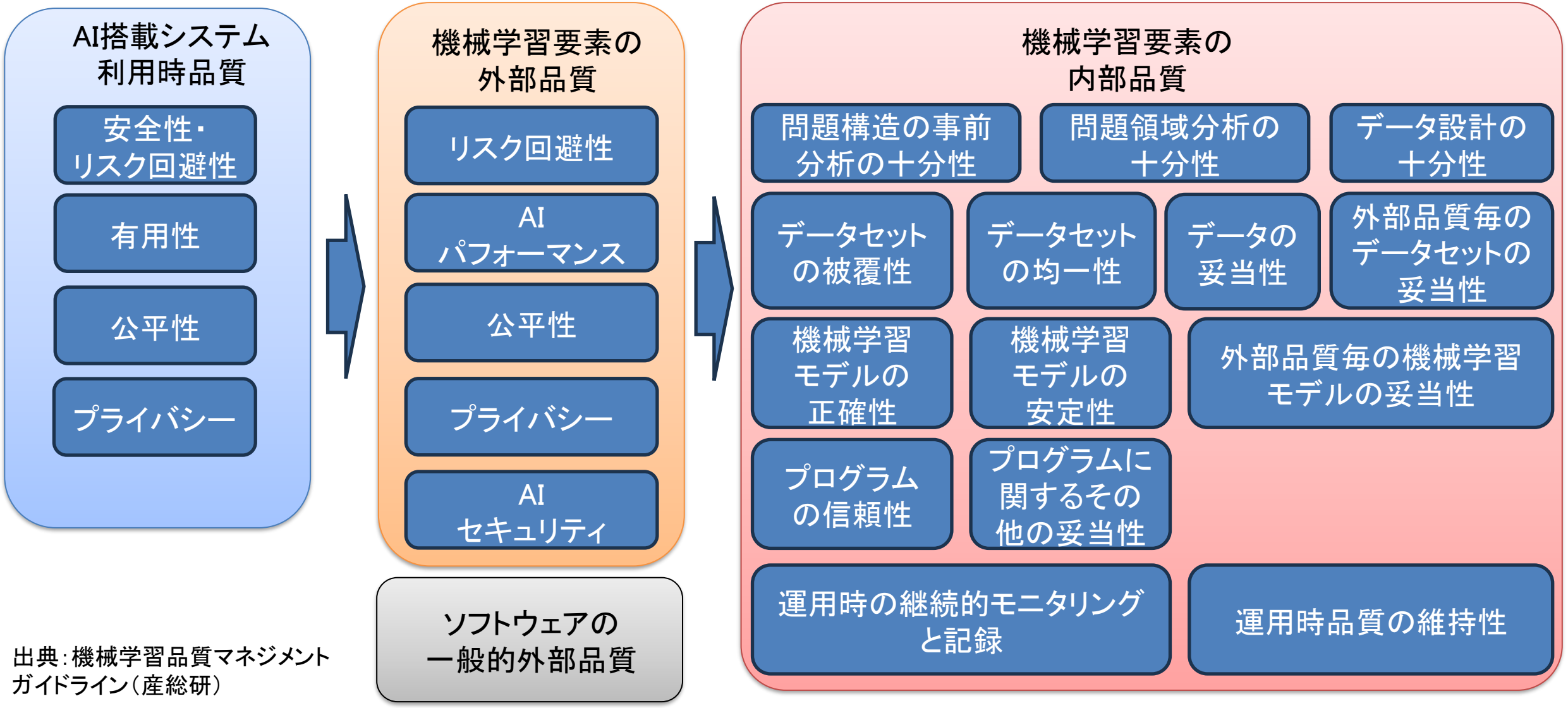
レガシーシステム開発プロセス	AI 搭載システム開発プロセス
システム要求分析フェーズ	システム要求分析フェーズ
システム設計フェーズ	システム設計フェーズ
ソフトウェア要求フェーズ	AI ソフトウェア要求フェーズ
ソフトウェア設計フェーズ	AI 学習準備フェーズ
ソフトウェア製造フェーズ	AI 学習フェーズ
ソフトウェアテストフェーズ	AI テストフェーズ
ソフトリリース	AI ソフトリリース
システムテストフェーズ	システムテストフェーズ
システム出荷	システム出荷
システム運用フェーズ	システム運用フェーズ

出典：AI搭載ソフトウェア開発ハンドブック
（JAXA 第三研究ユニット）

■AI搭載システムの開発フロー



AI搭載システム品質実現の構造



出典：機械学習品質マネジメント
ガイドライン（産総研）

AI搭載宇宙機の品質保証における課題

- 利用環境の多様性/未知性
  - ミッション毎に利用環境が異なり、AIによる推論精度確保が困難
- データセット準備の困難さ
  - AIモデル開発に必要なデータセットの量・質（十分性/被覆性/均一性/妥当性）の確保が困難
- 安全要求の高さ
  - AIモデルの推論精度が低い場合にも、安全確保が必要

まとめ

- 本発表では、宇宙機へのAI搭載に向けた、ソフトウェア品質保証の取組み状況をご紹介した。

- 今後は、現行のJAXAソフトウェア開発標準を拡張する形で、AI固有の開発・品質保証プロセスを整備するとともに、AIが搭載された宇宙機固有の品質保証の課題の対策を検討予定。

着陸後、月面で航法カメラ(CAM-PX)
により撮像された月面画像

着陸後、マルチバンド分光カメラによる
スキャン撮像により得られた月面画像
(JAXA、立命館大学、会津大学)

宇宙航空安全・ミッション保証シンポジウム
2025.1.15

小型月着陸実証機 SLIMの成果:

自律的な航法誘導制御系の事前検証と運用結果を中心に

宇宙航空研究開発機構
宇宙科学研究所　宇宙機応用工学研究系
福田 盛介
研究開発部門　第一研究ユニット
植田 聡史

▶ SLIMミッションの目的

SLIM(Smart Lander for investigating Moon)は、以下の2つの目的を達成することで、将来の月惑星探査に貢献することを目指したJAXAプロジェクト(2016年4月～)。

【目的A】月への高精度着陸技術の実証を目指す

- 従来の月着陸精度である数km～10数kmに対して100mオーダーを目指す
- キーとなる技術は、「画像照合航法」「自律的な航法誘導制御」および「細かく推力調整可能な推進系」

【目的B】軽量な月惑星探査機システムを実現し、月惑星探査の高頻度化に貢献する

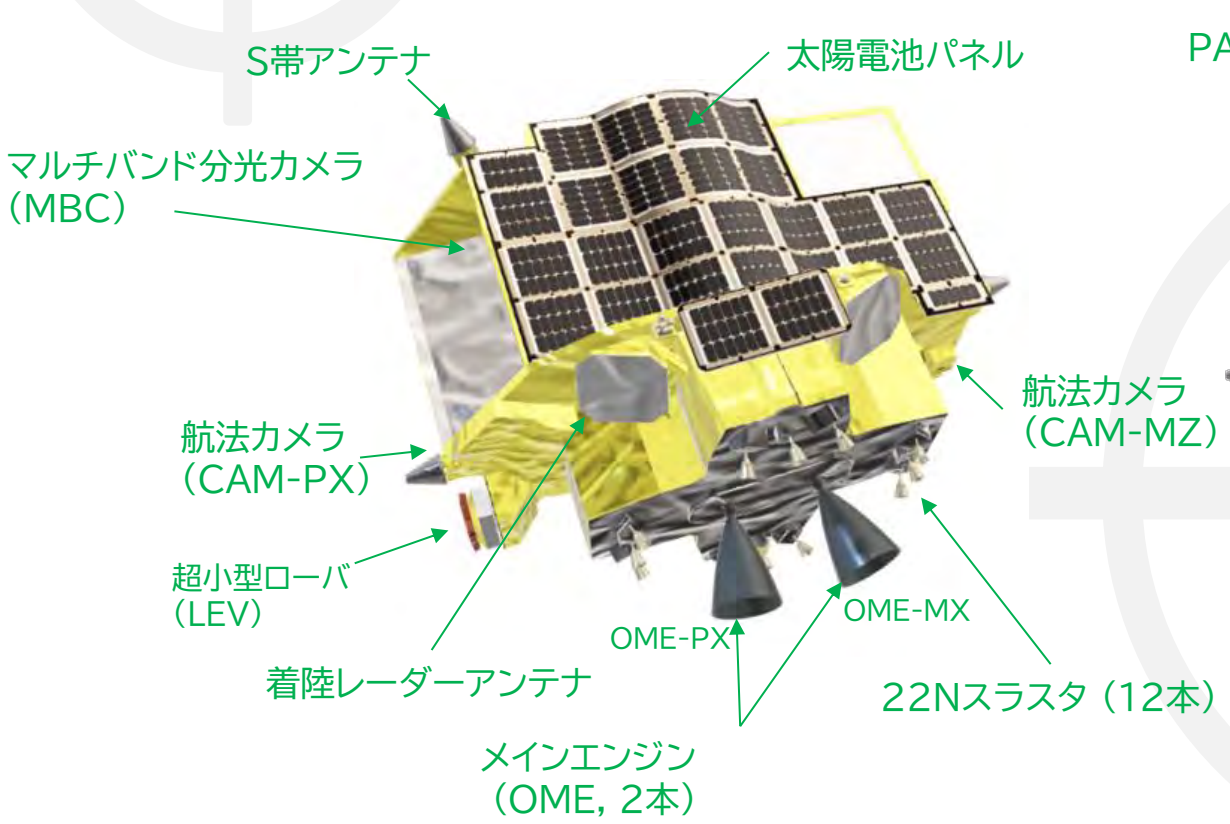
- 小型・軽量で高性能な化学推進システムの実現
- 宇宙機一般で中核をなす計算機や電源システムの軽量化

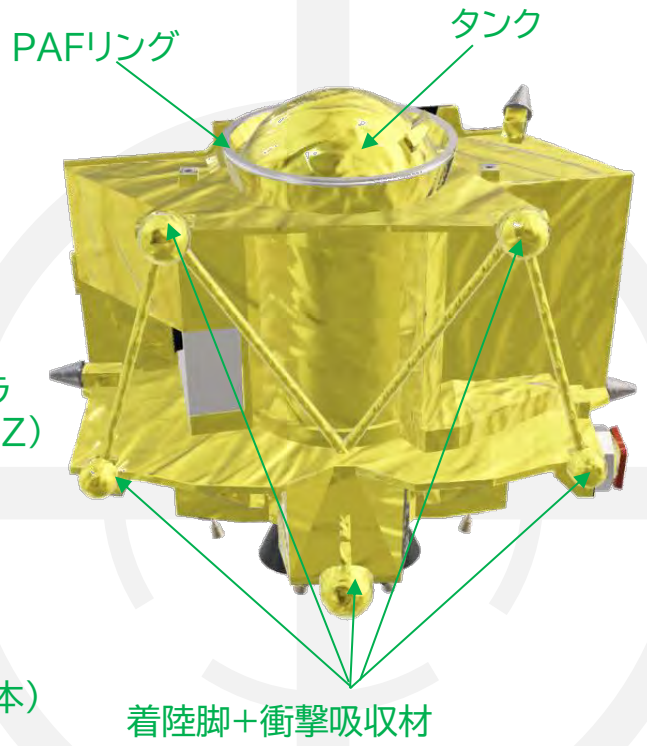
SLIM探査機外観

打上前の最終準備中のSLIMフライトモデル
（2023年 @種子島宇宙センター）

SLIM探査機外観

- 質量:200kg(推薬なし) / 約700-730kg(打ち上げ時)
- 高さ:約2.4m、縦:約1.7m、横:約2.7m





軽量化のため燃料・酸化剤一体型タンクを採用しており、これが探査機主構造を兼ねています。
電気コンポーネントはデジタル技術を積極的に採用し、統合化・小型軽量化を図りました。
(統合化計算機、電力制御分配器、Sバンド送受信機など)

▶【参考】最近の月着陸機との比較

- 参考として、近年打ち上げられた月着陸機の着陸精度及び質量を以下に示す。
- いずれも公開情報によるもので、詳細不明なところもあるが、SLIMの達成した着陸精度及び小型軽量化の世界における位置づけを理解する一助にはなる。

【各国の着陸機比較】※公開・報道情報を踏まえ整理

	SLIM	HAKUTO-R (M-1:1号機)	Chandrayaan -3	Luna-25	Peregrine Lander (M1:1号機)	Nova-C (IM-1:初号機)
機関	**JAXA (日)**	Ispace社 (日)	ISRO (印)	Roscosmos社 (露)	Astrobotic社 (米)	Intuitive Machines社(米)
打上げ時期 着陸結果	**2023年9月 成功**	2022年12月 失敗	2023年7月 成功	2023年8月 失敗(着陸せず)	2024年1月 失敗(着陸せず)	2024年2月 成功
着陸機等質量 ※打上げ時(燃料込み)	**約715kg**	約1,000kg	約3,900kg	約1,750kg	約1,480kg	約1900kg
※(燃料除く)	**約200kg**	約340kg	(不明)	約800kg	約480kg	約620kg
画像照合による高精度航法	**搭載**	非搭載	非搭載	非搭載	試験搭載 (着陸には不使用)	搭載
目標着陸精度 (km)	**0.1km**	数km ※同社記者会見に関する報道情報による	4km×2.4km	30km×15km	24km×6km	(詳細不明だが結果は数km)
主要ミッション	**高精度着陸 技術実証**	民間月面着陸	月面着陸、 科学ミッション	月面着陸、 科学ミッション	民間月面着陸	民間月面着陸

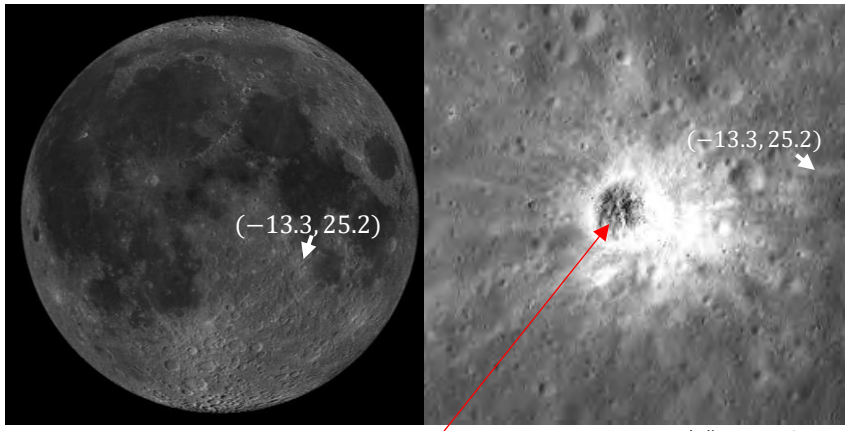
なお上記以外に、中国の嫦娥5号が2020年12月1日に成功しているが、これを含めても、**2020年代に入ってからの世界における月面軟着陸の成功率は5割程度。**

SLIMの着陸目標地点

SLIMの着陸目標地点は、以下の通り*1（平均地球/極軸系（ME））

経度： 25.24889 [deg] / 緯度：-13.31549 [deg]

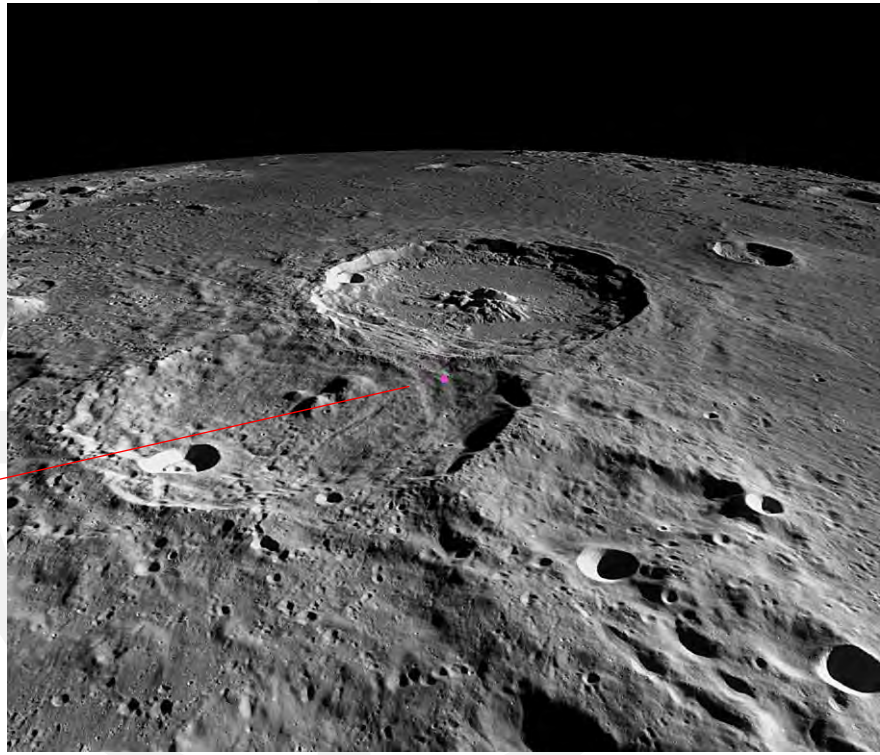
- なお従来、着陸目標地点としては、NASA、ISRO等の協力先を除き緯度/経度で小数点第一位までを公表していた。
- これは、科学的意義と着陸安全性を両立する地点の精位置情報自体に意義があると考えていたためである。
- 実際、SLIMでも、長い議論の末に着陸目標地点を選定した経緯がある。



出典:NASA/LRO

SHIOLIクレータ

*1… より厳密には、着陸目標地点は月面画像上で定義されており、その緯度経度がこの値となっている。

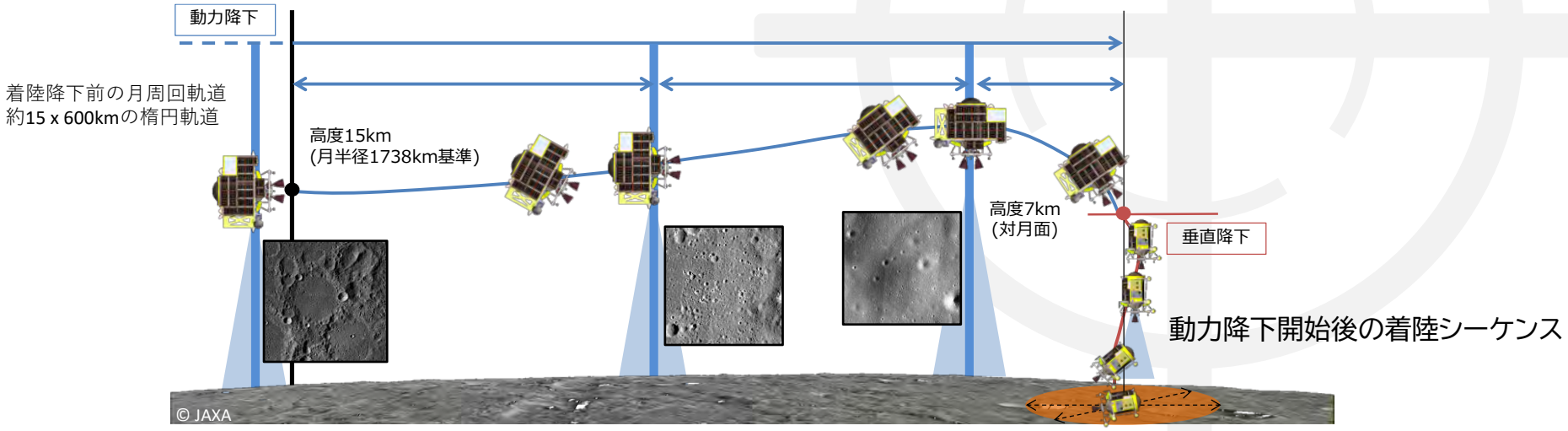


©藤井大地, LRO

出典：X, @dfuji1
https://x.com/i/status/1748103951336227113

SLIMの着陸シーケンスの概要

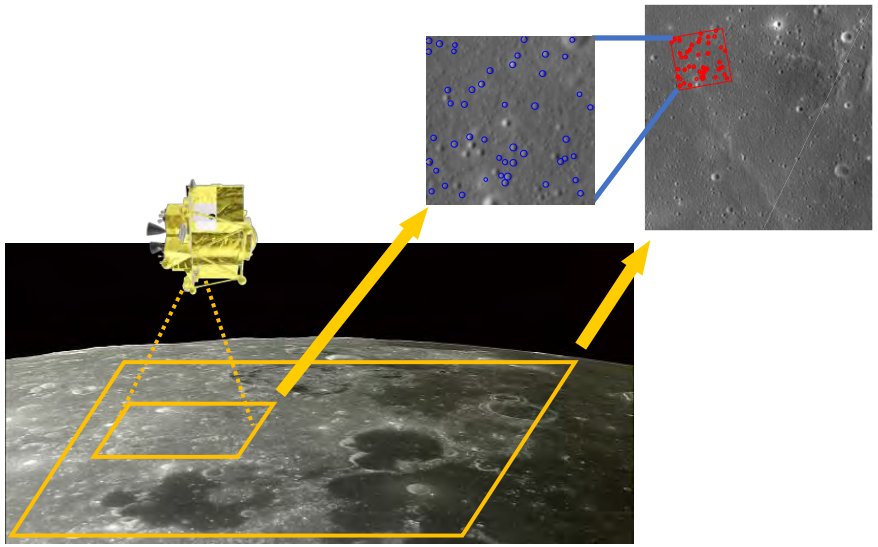
1. 月周回軌道から着陸降下を開始、**航法カメラによる画像航法**を行って高精度に自身の位置を推定しながら、**自律的な航法誘導制御**により、月面上の目標地点に接近。
2. 目標地点上空からは、着陸レーダによる高度・地面相対速度の精密な計測も開始し、航法誘導に反映。
3. 着陸地点上空約50mで画像ベースの障害物検出を行い、危険な岩などを自律的に避けて着陸する。すなわち、上空50mまでは着陸のピンポイント性を追求するが、それ以降はむしろ着陸の安全性を優先して障害物回避を行う。



動力降下開始後の着陸シーケンス

▶ 地形照合航法の必要性

- 月では地球周辺と異なり、GPSなどのGNSS測位を利用できない。地上局を用いた軌道決定では通常、視線直交方向の精度が悪く、リアルタイムでの位置推定も困難である。機上でのIMUデータの伝搬には誤差が蓄積するため、従来の月着陸機の着陸精度は数km～数10km程度にとどまっていた。
- また、着陸地点の決定などに用いる過去の周回機の月面撮像画像などと、航法誘導に用いる座標系での位置情報（緯度・経度など）の間に誤差やずれがある場合は、最終的な着陸精度に直接影響する。
- このため、100m級の精度で月面に着陸するためには、何らかの方法で月面各処の位置にひもづけられた「地形」情報を参照・照合した航法が必須となる。 SLIMでは大学等の有識者と協働し、ロバスト性や搭載性に優れたクレータベースの光学画像航法を開発した。



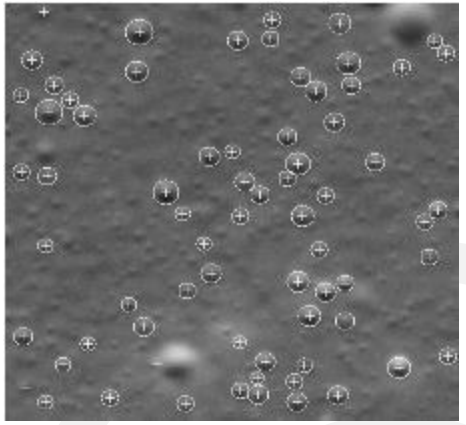
クレータ情報を用いた光学画像航法のイメージ

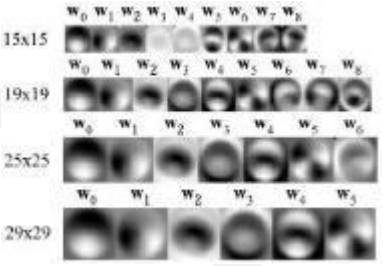
光学画像航法の研究室でのデモの様子

SLIM画像照合航法のアルゴリズム

主成分分析ベースのクレータ抽出アルゴリズム

- 探査機上の処理は主成分との積和演算と閾値処理のみで構成可能(計算量の削減)
- 比較的サイズの小さな30～50個程度/画像のクレータを使用

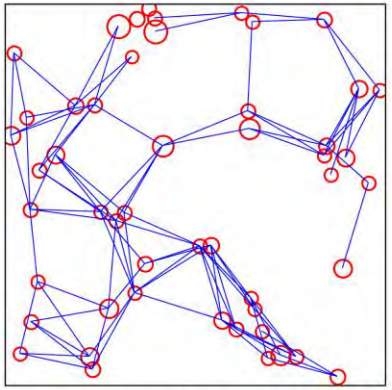




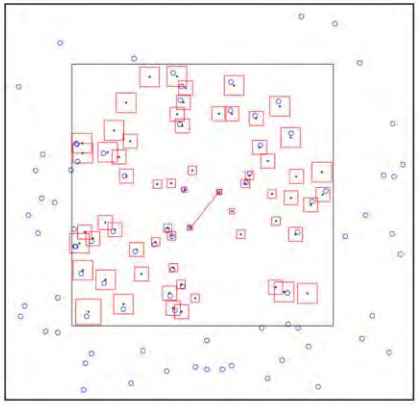
クレータマッチングアルゴリズム

- クレータ抽出結果をクレータ地図(位置情報)と点群レベルでマッチング
- クレータ地図は「かぐや」や米LROなどの月周回機が取得した画像を用いて準備し、SLIMに搭載した

Detected craters



Line segment features of craters



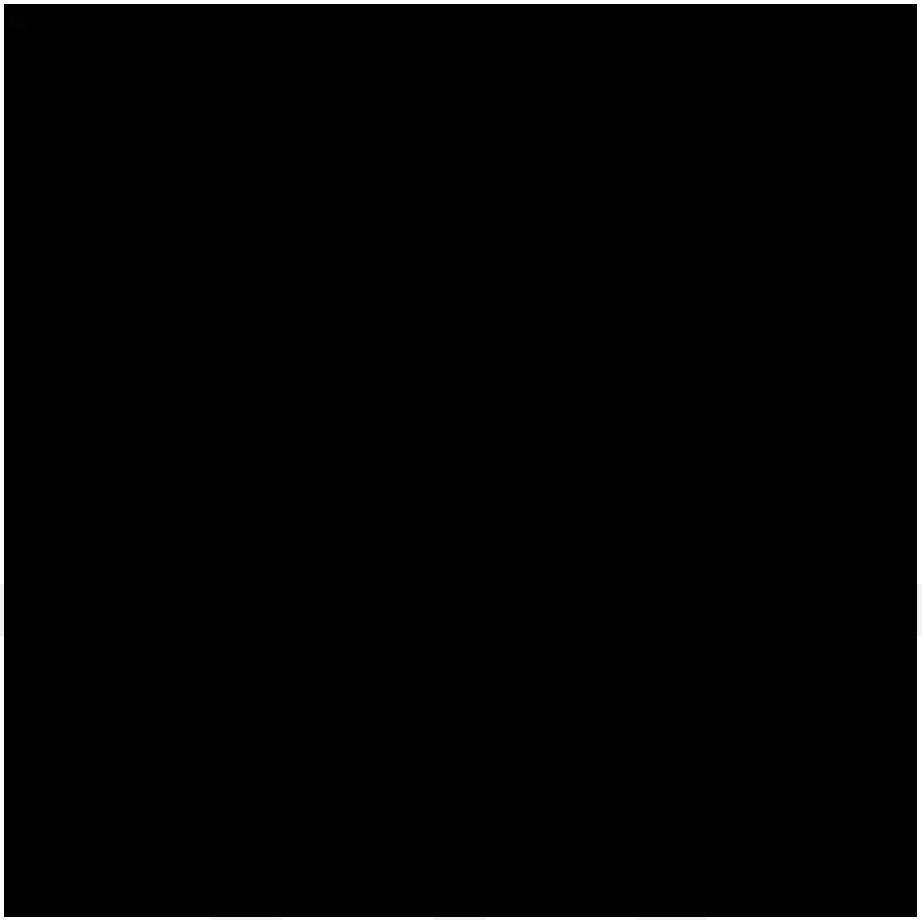
Line segment matching

関係の大学・研究室と共同で研究開発したこれらの処理を宇宙用FPGAに実装した

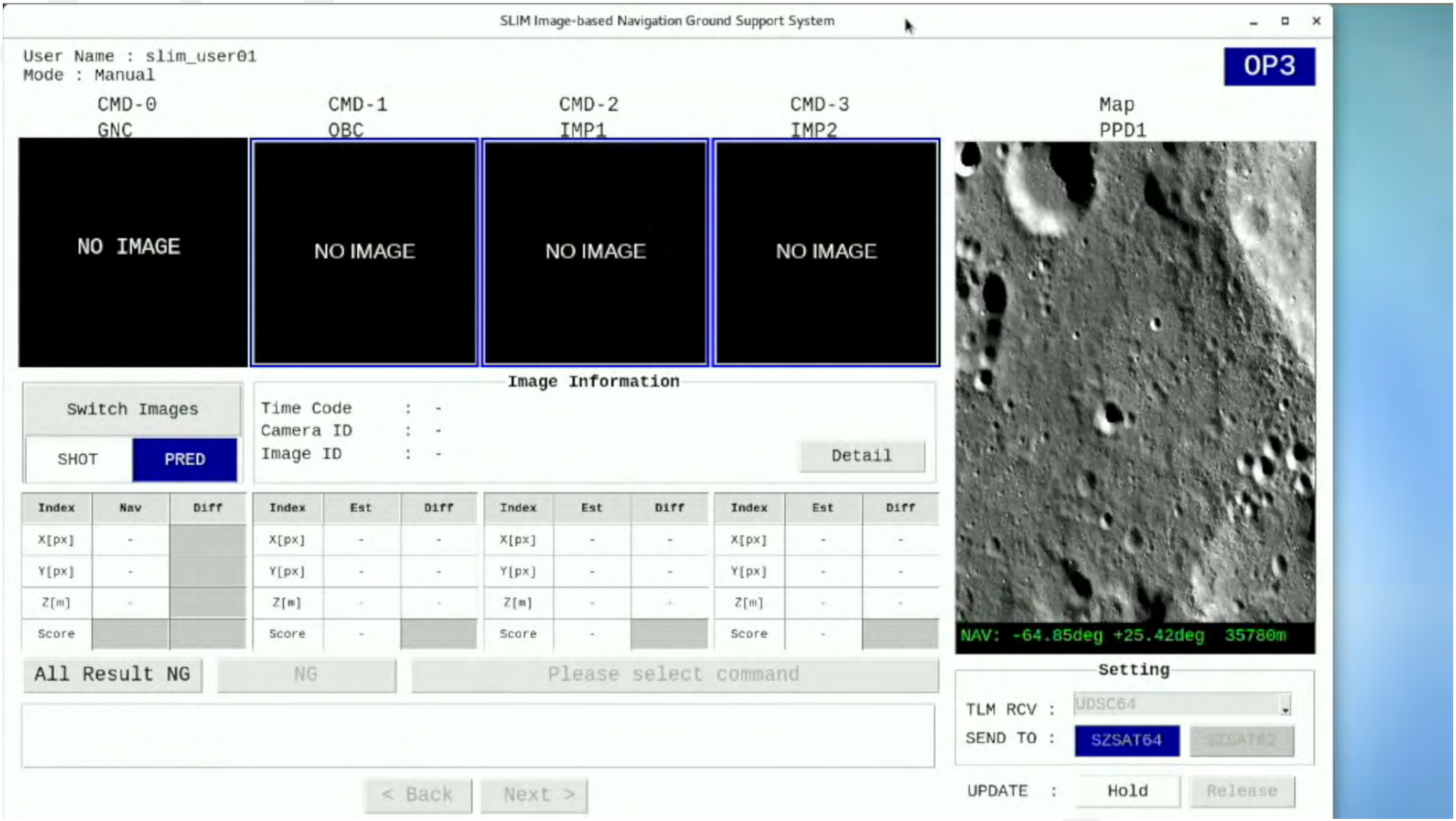
協力: 明治大、電通大、東京都立大

障害物検知

- 着陸点周辺の岩(ボルダー)などの障害物の存在状況は、事前に入手可能な最高分解能の周回機による画像により確認しているが、SLIMの着陸に支障をもたらすボルダーの全ては打ち上げ前には検出できない。
- したがって、月面上空50mにおいて、画像ベースの障害物検知を行う。
- 画像中の局所領域の輝度の分散から障害物を識別し、視野内で最も安全な領域を誘導系に出力する。
- 高度50m以下では着陸目標点ではなく最も安全な領域を目指して降下する。

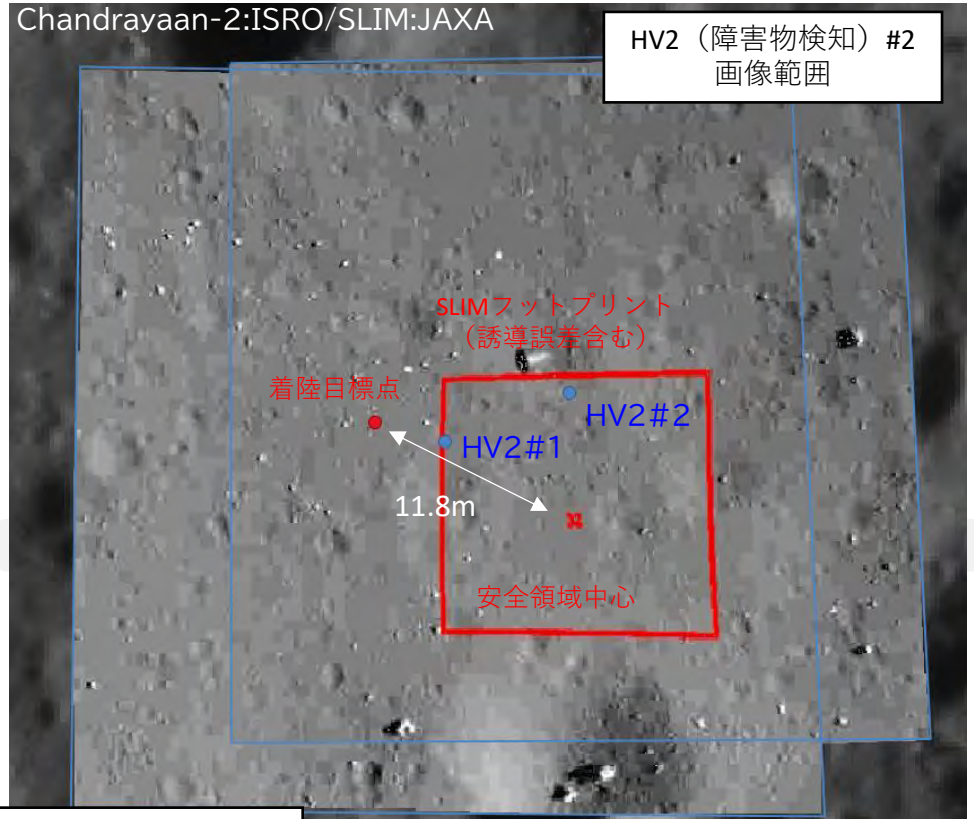


着陸運用での画像照合航法の結果



▶ ピンポイント着陸精度評価 / 高度50m付近で評価

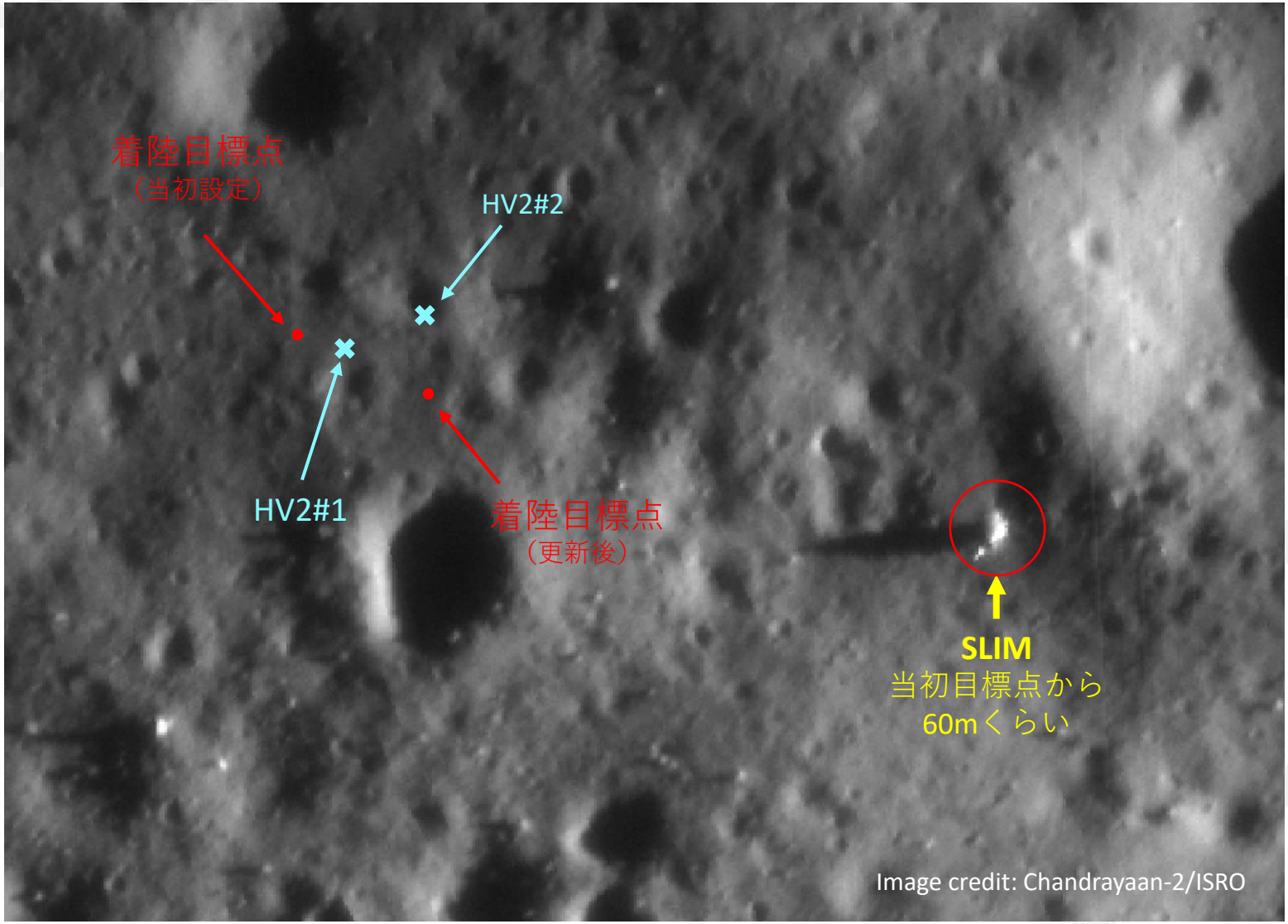
- 右下図は、障害物回避時の画像付近を拡大して示したもの。青い丸点が、1回目（#1）、2回目（#2）の撮像画像（青枠）の中心位置を示しており、この位置がSLIMの水平位置に相当し、その着陸目標地点からの距離を右下表に示している。
- この図から、障害物検知実施中の位置精度は、#1で3～4m程度、#2で10m程度だったと推定される。
- 前述の通り、この高度付近以降は着陸安全性を優先して着陸目標が設定される（障害物回避）。その意味で、<u>ピンポイント着陸精度としては、概ね10m程度以下</u>と評価している。
- さらに、#2については既に後述する異常事象により東へ流されている可能性が高く、<u>実際のピンポイント着陸精度は3-4m程度だった可能性が高い</u>。



シーケンス	想定高度	着陸目標点との水平距離
HV2#1	50m	3.4m
HV2#2	50m	10.2m

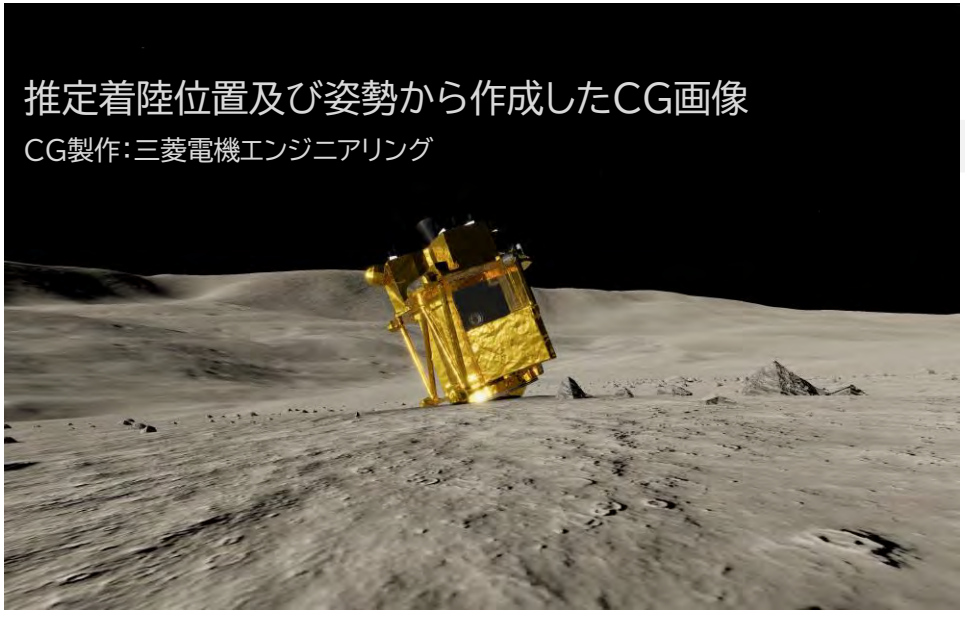
障害物回避以降は、図中の“安全領域中心”を目指した着陸へと移行する。着陸目標地点への正確な着陸より、障害物回避による安全性を優先するため（今回は着陸目標地点から11.8m離れた地点を目指すことになる）。従って、ピンポイント着陸精度については、障害物回避前の精度で評価することが、実力を評価することになる。

▶ 着陸後のSLIMを撮像したインドの周回機の画像



▶ 推定されている着陸姿勢

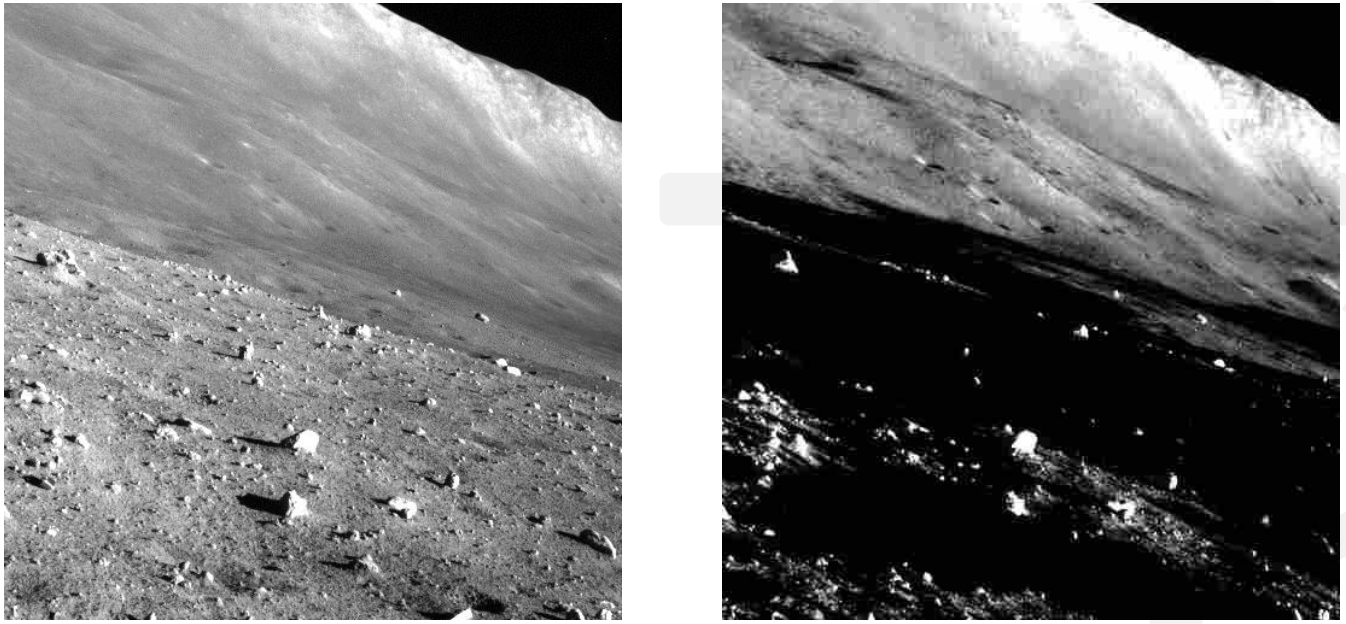
- 着陸後、最終的な探査機の姿勢は、探査機の各種データや着陸後の運用状況から、下図のようにメインエンジンが上を向いたほぼ鉛直の姿勢で、太陽電池パネルが西を向いた姿勢と考えている（当初計画的は太陽電池パネルが上向きとなる姿勢を予定）。
- 着陸は月面上の“午前中”であり太陽は東に位置していたため、着陸時点では太陽電池からの電力発生が失われた状態となった。そのため、バッテリ残量をモニタしながら、所定の手順に従って着陸降下中のデータをダウンロードし、科学観測機器の運用を一部実施した後、コマンドによりバッテリを電気的に切り離す措置を行った（短絡故障による探査機永久故障を避けるため）。結果、探査機は一旦電源オフとなった（1/20 2:57頃）
- 着陸直前にSLIMから分離した2基の小型ローバ（LEV-1、LEV-2）が連携して取得した月着陸後のSLIMの画像により、ほぼ推定通りの姿勢で静定していることが確認された。



LEV-2が月面のSLIMを撮像した実画像

越夜運用

- SLIMの着陸地点は1月31日に月の夜を迎え、SLIMは休眠状態に入ったが、2月25日に再度探査機との通信を確立し、越夜後の動作を確認した。その後、3月、4月と計3回の越夜に成功した。
- ラジオアイソトープ等を用いずに越夜を成功した探査機の例は少なく、SLIMの越夜運用で得られたデータから後続のミッションに資する知見が得られる可能性がある。
- 電子部品についても、先進的な部品実装の耐久性や、集積化による部品点数の削減の影響など、興味深い話題が多い。



1回目の越夜後に航法カメラにより撮像した着陸地点周辺の風景（左2/25, 右2/29）

着陸後、月面で航法カメラ(CAM-PX)
により撮像された月面画像

着陸後、マルチバンド分光カメラによる
スキャン撮像により得られた月面画像
(JAXA、立命館大学、会津大学)

宇宙航空安全・ミッション保証シンポジウム
2025.1.15

小型月着陸実証機 SLIMの成果:

自律的な航法誘導制御系の事前検証と運用結果を中心に

宇宙航空研究開発機構
宇宙科学研究所　宇宙機応用工学研究系
福田 盛介
研究開発部門　第一研究ユニット
植田 聡史

SLIMの自律的な航法誘導制御系

SLIM航法誘導制御系は「着陸降下」「軌道制御」で主要な役割を果たす。搭載系に組み込まれた自律機能に対して地上からコマンドによりパラメータ設定を行いミッションを実行する。

【着陸降下】 月周回軌道から減速し100m以内の精度で着陸

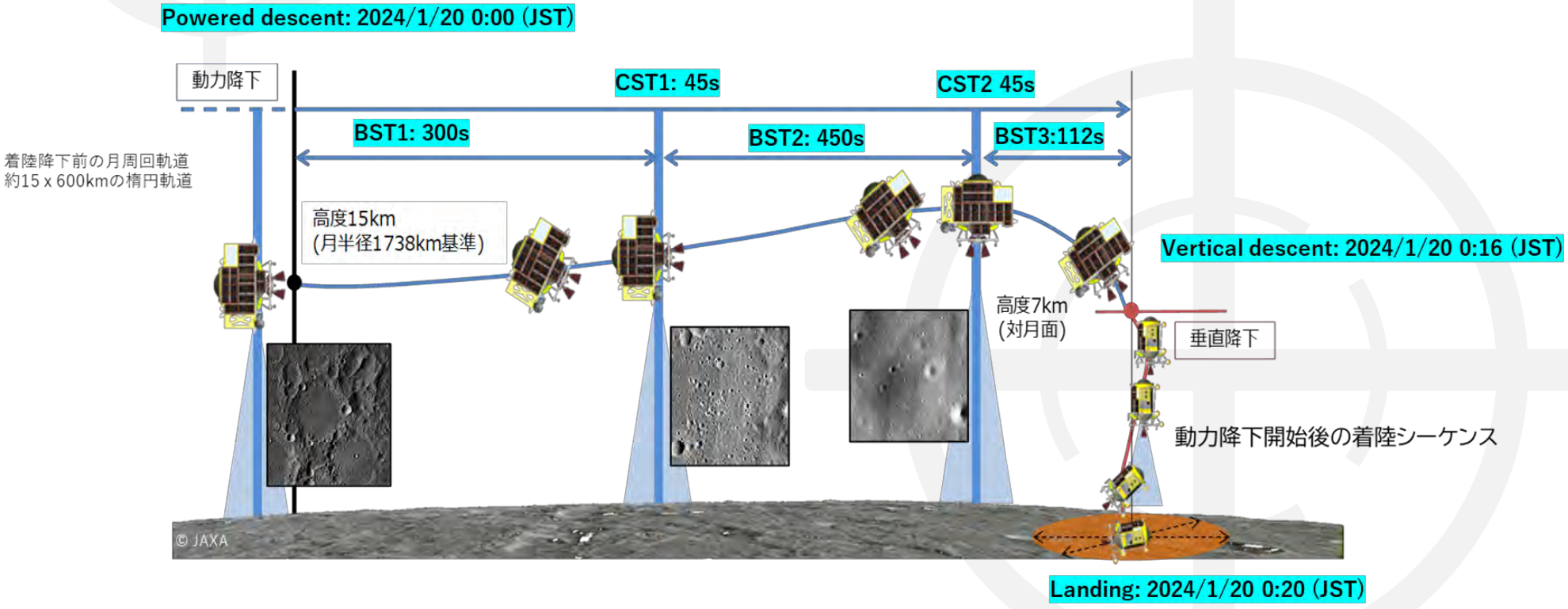
- 画像照合航法により自律的に高精度航法値(位置・速度)を得る
- 目標地点に向けた飛行経路および制御目標値を自律的に生成する
- 上記のためのパラメータ設定を着陸降下開始前にアップロード
- 加えて、目標軌道から外れるなどのオフノミナル対応のための自律FDIR[*1]を実装

【軌道制御】 ロケット分離から着陸降下開始までの軌道遷移

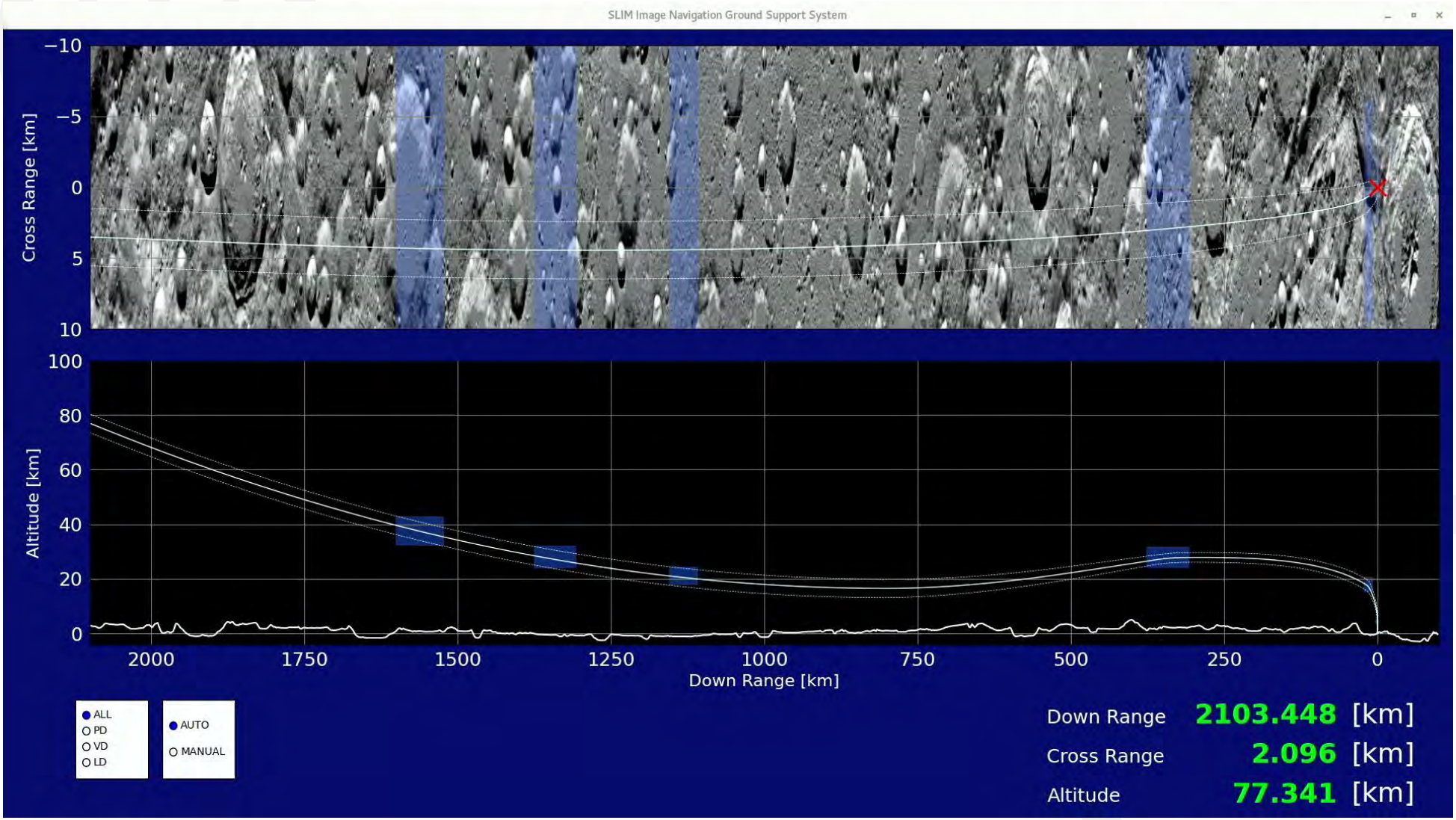
- 軌道変更のためのエンジン噴射・姿勢変更のシーケンスを自律的に実施
- 地上システムで取得した軌道決定値に基づき軌道制御計画を立案
- 軌道制御のためのパラメータ設定を各軌道制御の直前にアップロード
- 加えて、意図しないエンジン停止などのオフノミナル対応のための自律FDIRを実装

[*1] FDIR:Fault Detection, Isolation and Recovery

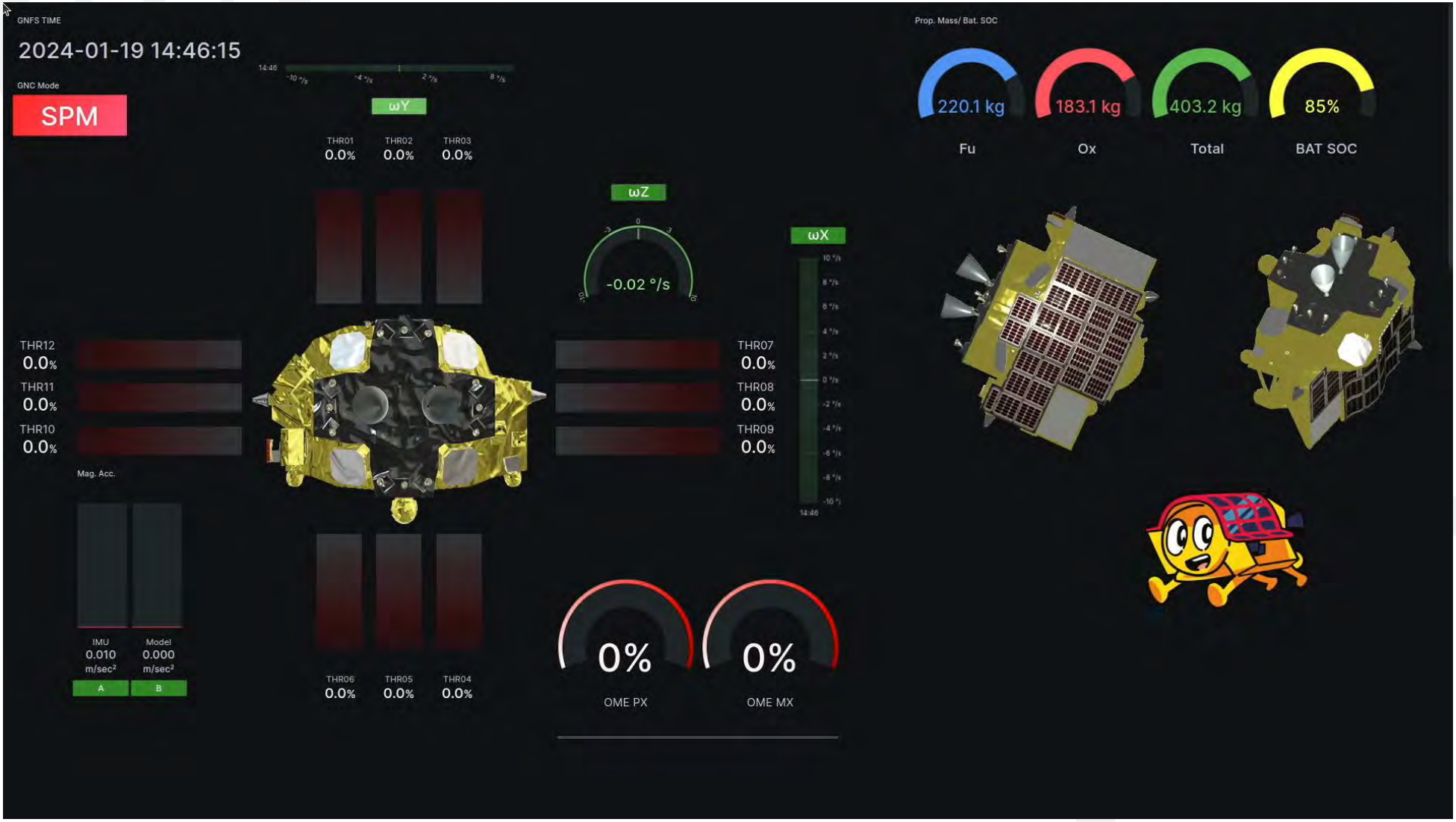
月面へのピンポイント着陸降下シーケンス



着陸降下時の高精度航法値

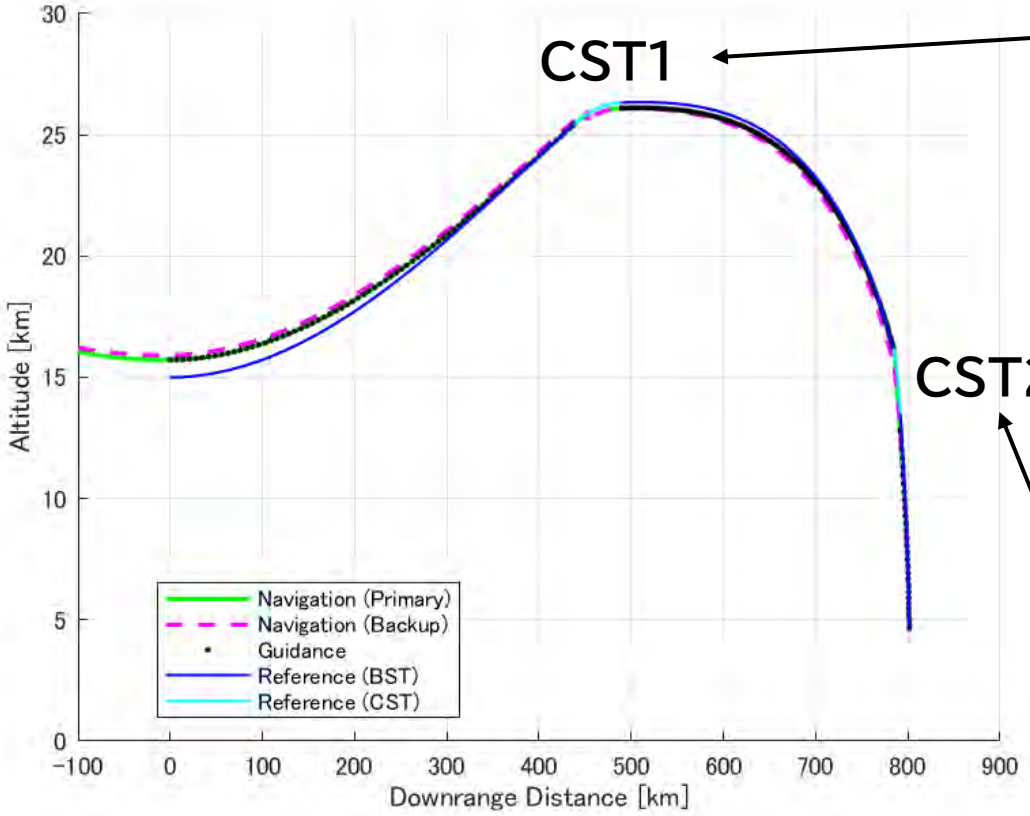


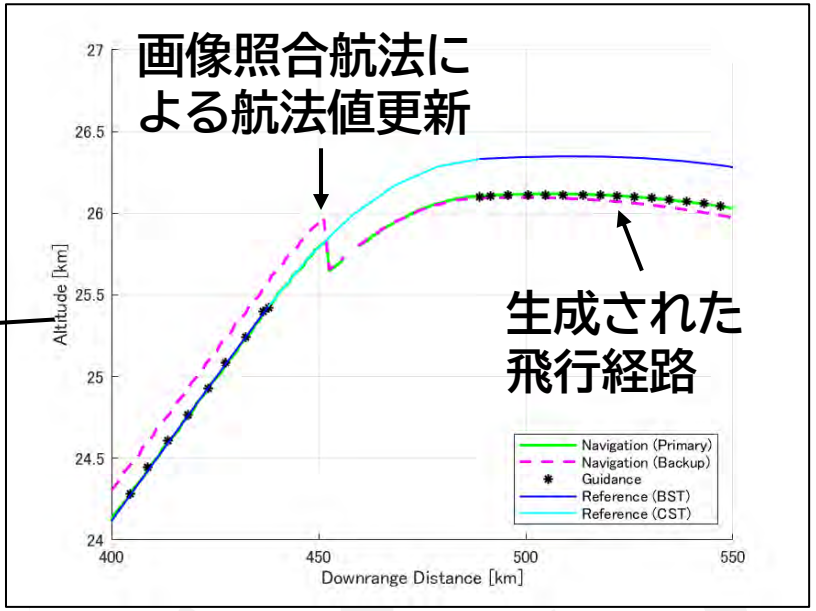
着陸降下時の姿勢制御・エンジン噴射

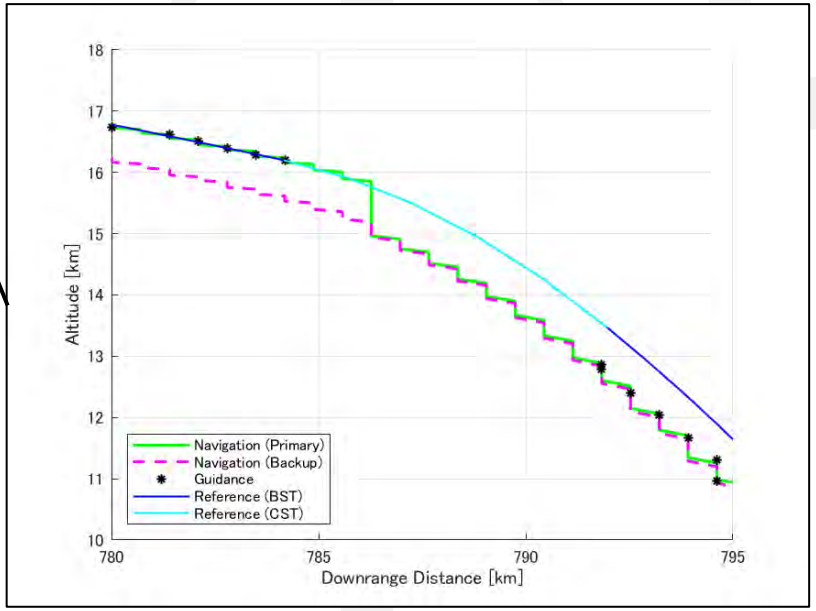


目標地点に向けた飛行経路の自律生成(ダウンレンジ・高度)

飛行結果:
ダウンレンジ距離と高度

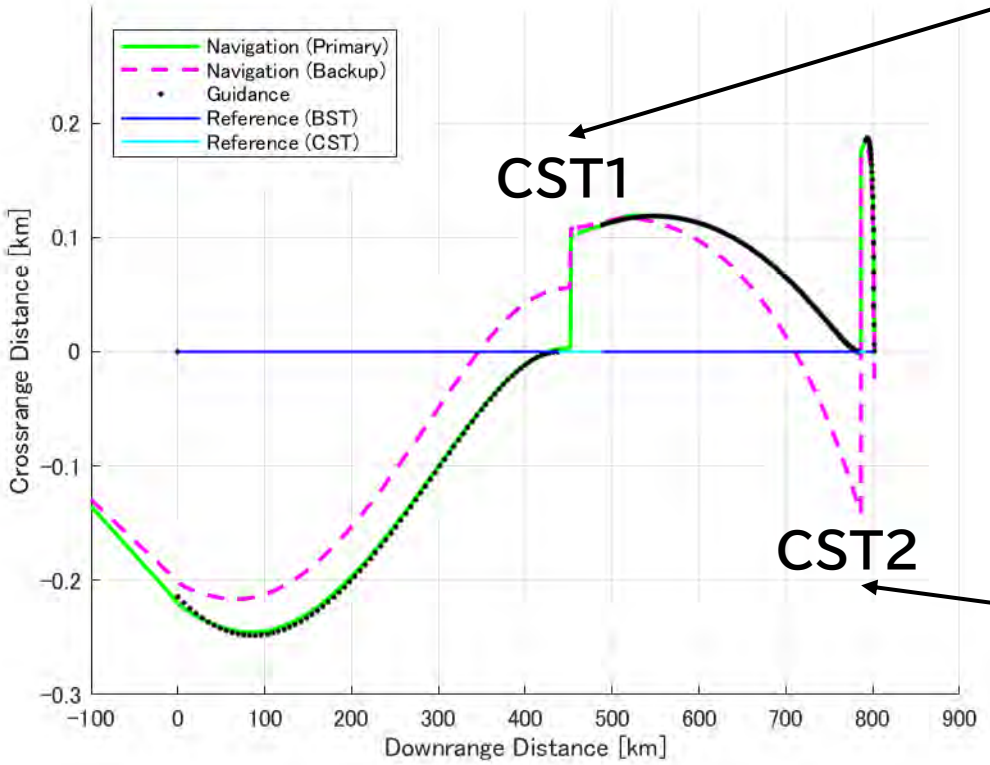


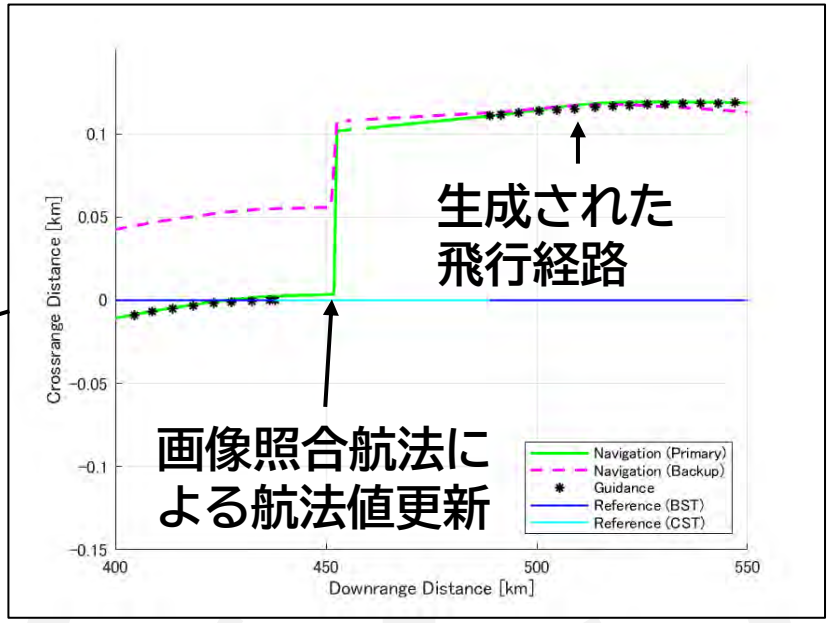


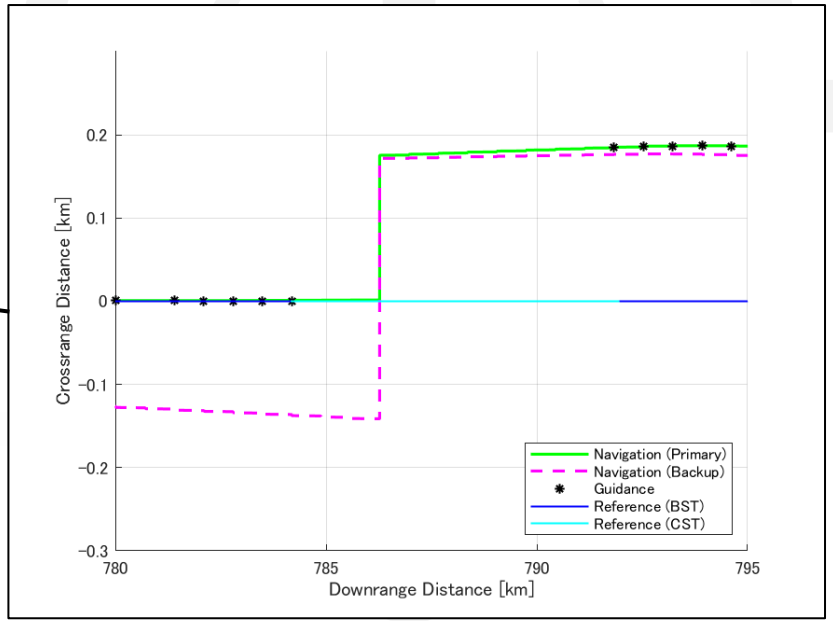


目標地点に向けた飛行経路の自律生成（クロスレンジ）

飛行結果：
ダウンレンジ距離とクロスレンジ距離

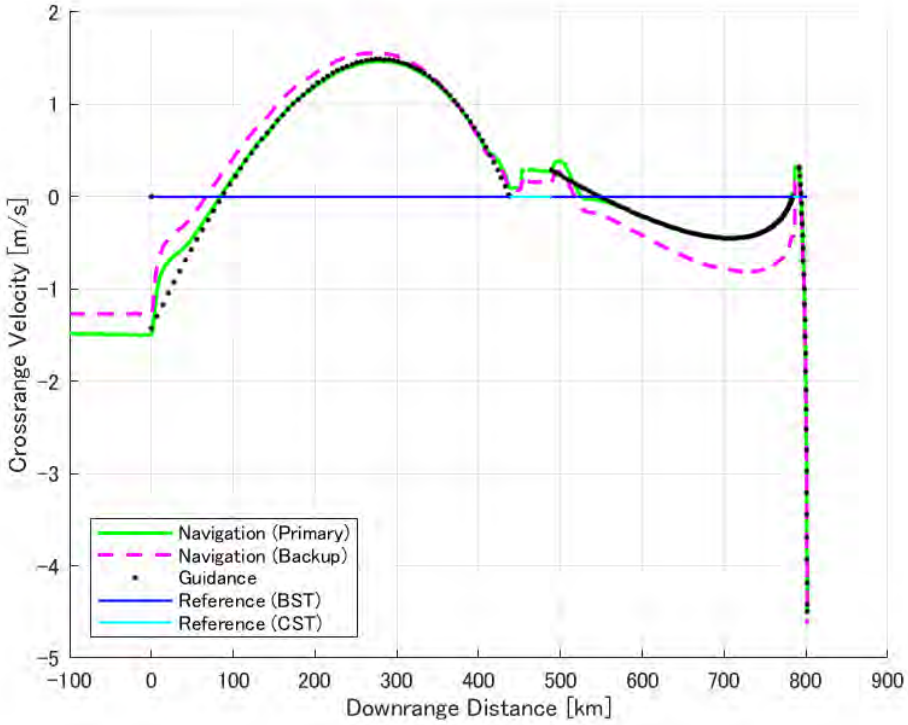




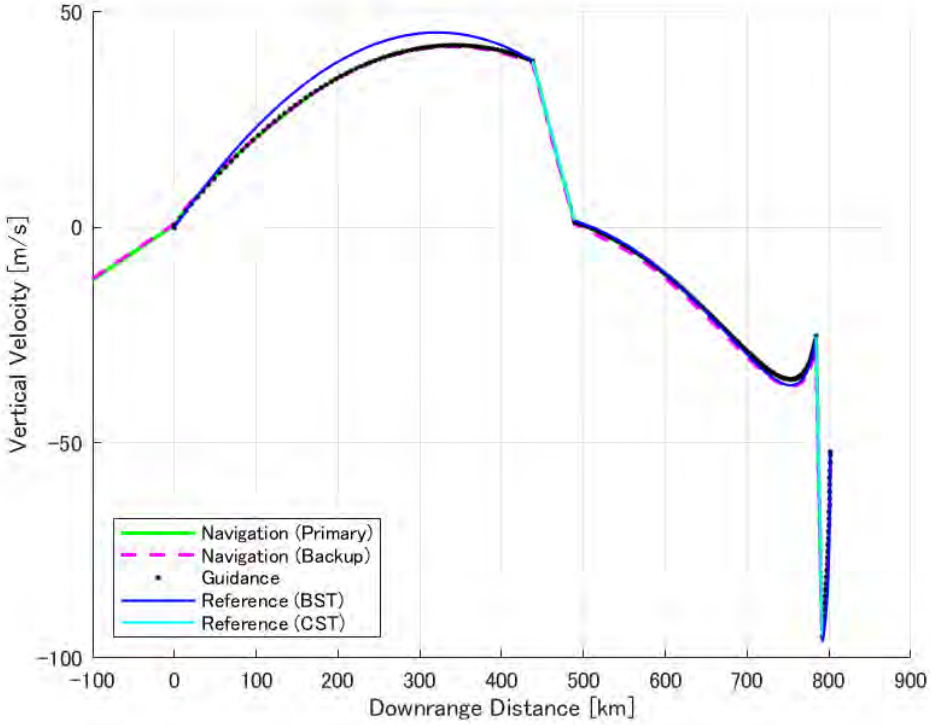


目標地点に向けた飛行経路の自律生成(速度)

飛行結果:
ダウンレンジ距離とクロスレンジ速度



飛行結果:
ダウンレンジ距離と垂直速度



SLIM航法誘導制御での機械学習技術の活用

SLIM航法誘導制御では様々な形で機械学習技術を活用している。活用形態は搭載ソフトウェアへの実装、地上システムでの実装、機能検討での活用など多岐にわたる。

【搭載ソフトウェアへの実装】 回帰モデルをアルゴリズムとして実装

- 動力降下フェーズの誘導精度高精度化のためのエンジン噴射開始時刻補正値算出
- 動力降下フェーズのオフノミナル対応として推力加速度値を回帰モデルにより補正

【地上システムへの実装】 ガウス過程回帰によるコマンド値生成

- 動力降下フェーズのオフノミナル対応として軌道速度の減速を優先し軟着陸を実現する誘導に移行した場合の姿勢角指示コマンド値の生成に使用

【機能検討での活用】 過大推力対応機能の設計に強化学習を活用

- 過大推力に起因する並進制御誤差拡大時(オフノミナル)の対応として軌道面外方向に正弦波で姿勢を振ることで過剰推力を逃がす機能を搭載ソフトに実装した
  - ➢ 過大推力に対応する補助制御則を事前知識無しで強化学習により生成した結果を踏まえ搭載系の機能を検討した

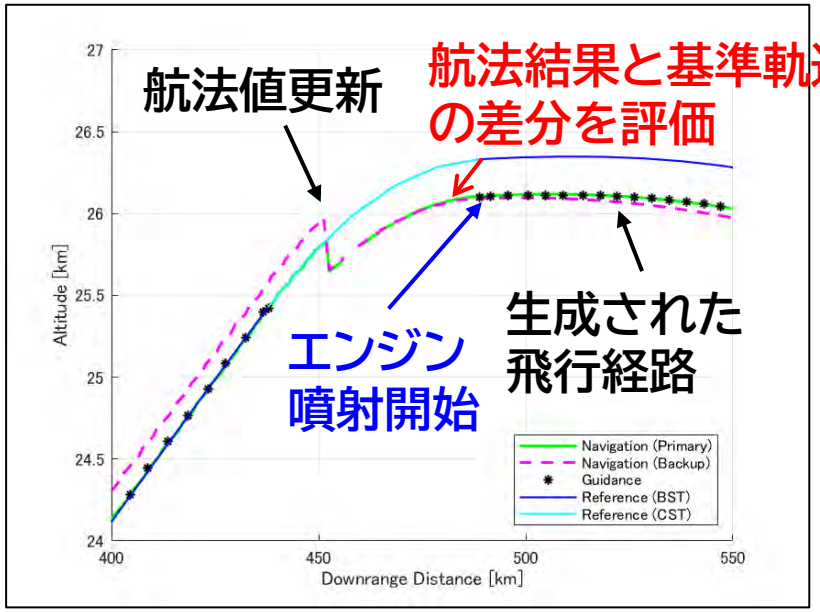
上記に示した各機能のうちノミナルシーケンスで必ず動作するのは「動力降下フェーズの誘導精度高精度化のためのエンジン噴射開始時刻補正値算出」のみである

SLIMでの機械学習の活用：エンジン噴射開始時刻補正

SLIM動力降下フェーズの飛行経路自律生成では、終端での位置・速度をともに高精度に実現する必要がある。エンジン噴射開始時刻を補正することで高精度化が可能という特性があり、エンジン噴射開始時刻補正値算出のために機械学習に基づく回帰モデルを適用した。

回帰モデル

$$Y = \gamma_0 + \sum_{i=1}^{8} \gamma_i X_i + \sum_{i=1}^{8} \gamma_{ii} {X_i}^2 + \sum_{i=1}^{8} \gamma_{iii} {X_i}^3 + \sum_{i<j} \gamma_{ij} X_i X_j + \sum_{i<j<k} \gamma_{ijk} X_i X_j X_k + \sum_{i \neq j} \gamma_{iij} {X_i}^2 X_j$$



目的変数：エンジン噴射開始時刻補正値

Variable	Unit	Example 1	Example 2
ΔT	s	+2.67	-3.68
ΔA	n.d.	-0.01	0.00
Δm	kg	-0.41	+0.38
ΔP_β	m	-3641.98	+6720.62
ΔP_ε	m	-301.96	-266.67
ΔP_r	m	-1102.63	+542.30
ΔV_β	m/s	+0.91	-0.95
ΔV_ε	m/s	-0.90	+1.13
ΔV_r	m/s	-3.83	+5.03

説明変数：航法結果と基準軌道の差分

オフノミナル対応自律FDIR(動力降下)と機械学習技術の活用

SLIMの目的の一つである「軽量な月惑星探査機システム」の実現のため、機器の冗長構成を最小限とした。アルゴリズムや運用上の工夫でオフノミナルに対応する設計思想とした。

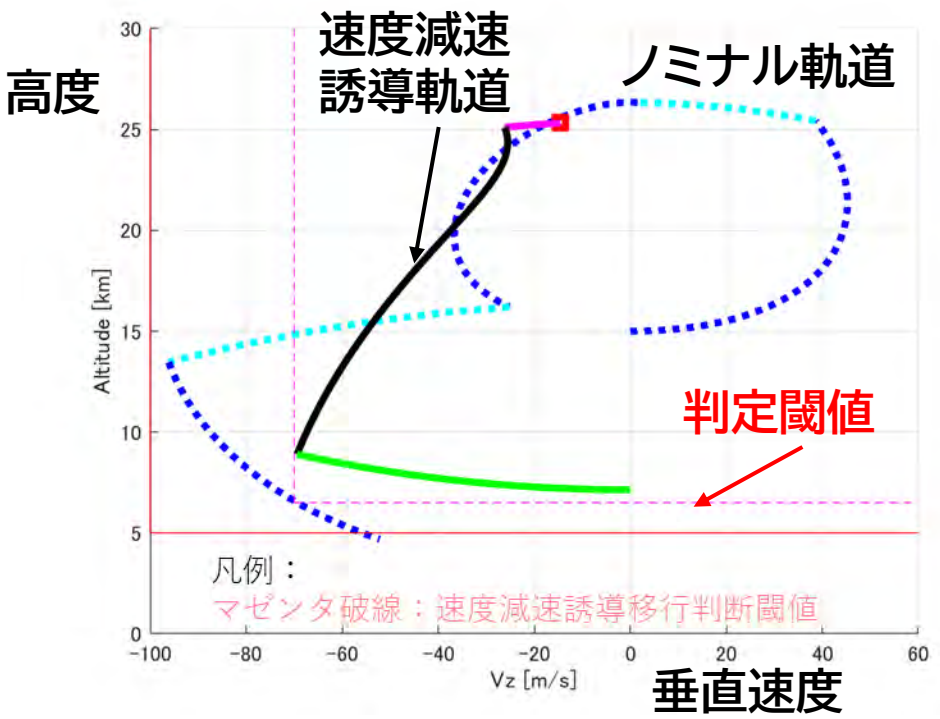
異常の状態	観測される事象	事象発生時の対応	機械学習技術の活用
推力過大	並進制御誤差拡大 (進行方向逆方向)	面外方向に正弦波で姿勢を振ることで過剰推力を逃がす	機能検討に活用: 推力過大時の補助制御則を強化学習により生成した結果を活用
推力不足	並進制御誤差拡大 (進行方向順方向)	次区間の誘導計算ロバスト性向上で対応	搭載ソフト実装: 次区間の推力加速度値を回帰モデルにより補正
姿勢制御異常	姿勢制御誤差拡大	エンジン噴射を一時的に停止し、外乱トルクを低減させる	なし
誘導計算異常	エンジン噴射開始時刻補正値が許容下限値を下回る	エンジン噴射開始時刻補正値の超過分に応じて誘導計算に適用する推力加速度を補正する	搭載ソフト実装: 推力加速度値を回帰モデルにより補正
	誘導計算が正常に実行されたことを示すフラグがFailを示す	着陸降下基準軌道通りの誘導を実施する	なし
高度の異常低下	高度航法値が月面への衝突が懸念されるレベルまで低下	地上からのコマンドによりオフノミナル誘導シーケンスへ移行し月面衝突を回避しながら水平速度を低下させる	地上システムに実装: 軌道速度の減速を優先し軟着陸を実現する誘導アルゴリズム(管制設備側に実装)をガウス過程回帰モデルにより実現

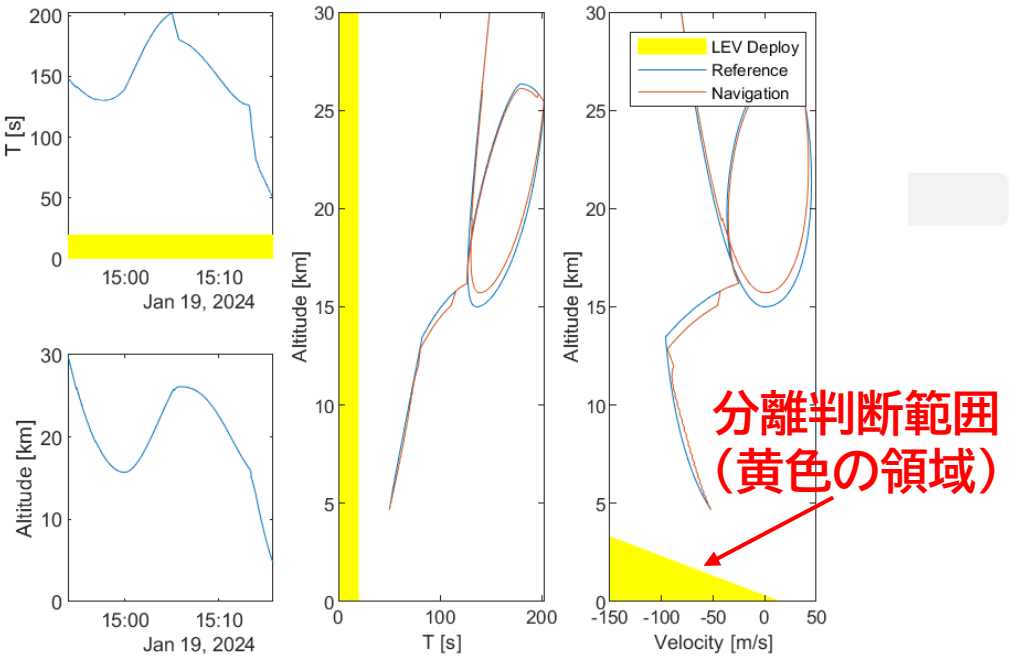
自律システムに対する運用者による緊急判断とコマンド運用

SLIM動力降下フェーズの自律FDIR機能は、オフノミナルな条件においても当初目的を達成するために、搭載ソフトによる自律的な判断・移行を実行することとしている。ただし、ピンポイント着陸をあきらめ軟着陸のみ実現する、または、超小型ローバのみを月面に着陸させるという判断に対応した機能は運用者による地上からのコマンド送信により実行される。

速度減速誘導移行判断
高度・垂直速度が所定の範囲を超えた場合に移行

超小型ローバ分離判断
墜落までの予想時間に基づき分離コマンド送信を判断





SLIM航法誘導制御系の検証

SLIM航法誘導制御系は過去の宇宙機開発で蓄積された航法誘導制御技術とピンポイント着陸のための先進的な高精度航法誘導技術の組み合わせで構成されている。自律機能、機械学習、地上からのコマンド運用を含むシステムとなっており、ソフトウェア検証では従来的な検証活動に加えてこれらの事情を加味したケース設定および評価を実施した。

【ノミナル検証】 モンテカルロシミュレーションによる要求達成確認

- 搭載ソフトウェアおよびシミュレータを用いた着陸降下シーケンスに関する1000ケースのシミュレーションによりピンポイント着陸精度(100m)を達成することを評価
  - ➢ 前提となる基準軌道やパラメータを更新するごとに実施

【End-to-End検証】 仕様範囲外やオフノミナル時の挙動確認

- 従来的なFDIR機能確認を全て実施したうえで、航法センサや推進系の性能が仕様範囲外を示した場合、地上からのオフノミナル対応コマンドが発行された場合、搭載系に設定するコマンド値を変更した場合などを想定したソフトウェアシミュレーションを実施
  - ➢ ケース設定ではオフノミナル対象となる機器を網羅しながら各設計者の経験・知見も反映

【コマンド検証】 軌道制御・着陸降下開始前に実施するコマンド計画の最終検証

- 軌道制御や着陸降下開始のための搭載ソフトウェア設定値を送信するコマンド計画ファイル(手順書)に設定したパラメータを適用した最終検証をソフトウェアシミュレーションにより実施
  - ➢ 運用直前のため正確なシミュレーション実施に加え評価および判断を限られた時間内に完了する

まとめ：SLIMの自律的な航法誘導制御系

- SLIM航法誘導制御系の役割・運用結果
  - 高精度航法と目標地点に向けた飛行経路生成を自律的に実行する
  - ピンポイント着陸精度達成に貢献
- SLIM航法誘導制御での機械学習技術の活用
  - 活用形態は搭載ソフトウェアへの実装、地上システムでの実装、機能検討での活用
  - 回帰モデルに基づくエンジン噴射開始時刻補正値算出について軌道上での動作実績あり
- 運用者による緊急判断とコマンド運用
  - 当初目的達成をあきらめることになる緊急運用は地上から運用者が判断してコマンド送信
- SLIM航法誘導制御系の検証
  - 過去の宇宙機開発で蓄積された航法誘導制御検証がベース
  - その上で先進的な高精度航法誘導技術が搭載されていること、自律機能・機械学習・地上からのコマンド運用を含むシステムとなっていることを考慮し、ノミナル検証、End-to-End検証、コマンド検証の3段階で実施

2024年度 宇宙航空安全・ミッション保証シンポジウム
「自律化技術・AI」x「Assurance（アシュアランス）」

自動車の自動運転・運転支援システムにおける AI導入とアシュアランス

2025年 1月15日

株式会社本田技術研究所 先進技術研究所
フェロー
杉本 洋一

目次

- 交通事故ゼロ社会と 自由な移動の喜び
- 運転支援と自動運転の概要
- 実用化された自動運転レベル3における 安全論証活動
- 自動運転・運転支援システムに対するAI導入と Safety Assurance
- 最後に

目次

創業者の安全への想い

「人命尊重」「積極安全」

交通機関というものは
人命を尊ぶものである

Hondaの安全目標（2021年発表）

HONDA

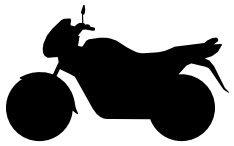
交通事故ゼロのモビリティ社会

Honda 安全の大義

リアルな世界を
感性・五感豊かに感じて楽しむ

Honda 安全理念

共存安全
すべての交通参加者に安全・安心を提供



2050年 全世界に於いて
Hondaの二輪・四輪が関与する交通事故死者ゼロを目指す（保有）

新車だけではなく、市場に現存するすべてのHonda車および相手歩行者、自転車をも対象とするチャレンジングな目標

Hondaが目指す方向性

HONDA

人々に自由な移動の喜びを提供し続けていくこと

CREATE

自由な移動＝普遍的・本質的価値

TRANSCEND

時間と空間の制約から人を解放

AUGMENT

人のあらゆる可能性の拡張

目次

SAE 自動車用運転自動化システムのレベル定義

HONDA

操縦の主体 ＝ ドライバー

運転支援

レベル0

ドライバーの運転

レベル1

前後 or 横方向の運転支援

レベル2

前後&左右方向の運転支援
ドライバーの監視は必要

レベル3

システム監視による
自動運転
ドライバーによる
バックアップ
（限定領域）

レベル4

システム監視による
自動運転
（限定領域）

レベル5

完全自動運転

自動運転

操縦の主体 ＝ システム

操縦：認知・予測・判断・操作の行為を行うこと

レベル1, 2は運転支援　レベル3以上が自動運転

自動運転　二つのアプローチ

HONDA

SAE
運転自動化
レベル

5
4
3
2
1

自動運転

運転支援

完全自動運転

新移動サービス
Mobility
as a Service

Source 引用：
経済産業省ニュースリリース

中山間地
限定コース
低速

Source 引用：NEDO

高速道
隊列走行

市街地
限定エリア

レベルの高さと
技術難易度は一致しない

全国高速道から
一般道全域へ

パーソナルカー

限定

走行環境条件（地理的、道路種別、環境条件、交通状況、速度、等）

何時でも
どこでも

限定領域　Operational Design Domain（ODD）

安全運転支援／自動運転の カテゴリー

HONDA

パーソナル自動運転（Dual Mode）

Level 3

ドライバー ⇄ システム

運転支援（運転負荷軽減）

ACC, LKASなど

Level 1 - 2

ドライバー

ACC: Adaptive Cruise Control
LKAS: Lane Keep Assist System

ドライバレス・モビリティサービス

Level 4

Source 引用：
経済産業省ニュースリリース

乗員　システム

事故回避支援

衝突被害軽減ブレーキ、ESCなど

SAEレベルとは関係なし

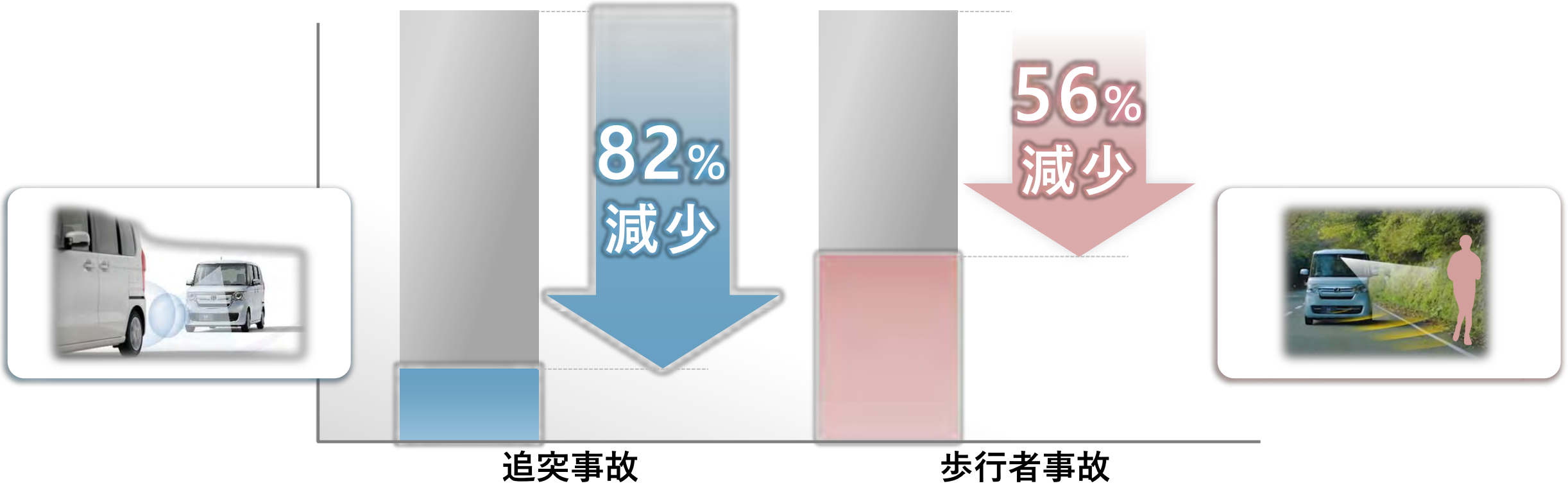
ESC: Electronic Stability Control

安全運転支援搭載車の事故削減効果

HONDA

追突事故発生率は82%
歩行者事故発生率は56%減少※

Honda SENSING搭載車の事故削減効果（N-BOX）



出典：公益財団法人交通事故総合分析センターのデータを基にHondaが独自に算出

※N-BOX（2011年11月～2017年8月）AEB非搭載車に対する現行N-BOX（2017年9月～2020年12月）の登録台数当たり交通事故死傷者数調査結果の差分。
公益財団法人交通事故総合分析センターのデータを基に、2020年内にN-BOXが1当の人身事故による死傷者数について調査。

全方位安全運転支援システム

HONDA

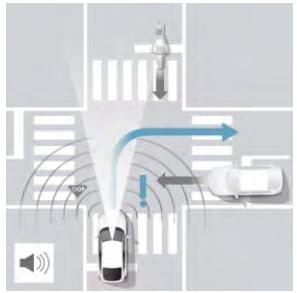
自動運転開発で培った技術を活かし、様々な事故シーンへの対応を拡大

HondaSENSING搭載機能＋5機能

衝突軽減ブレーキ（CMBS）

機能拡大　交差点：出会いがしら
歩行者：車両 側方/対向対応

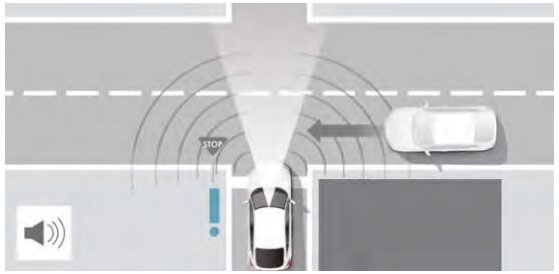
二輪四輪交差車両対応



右左折時の横断歩行者対応

前方交差車両警報

低速走行または発進する際の交差車両情報

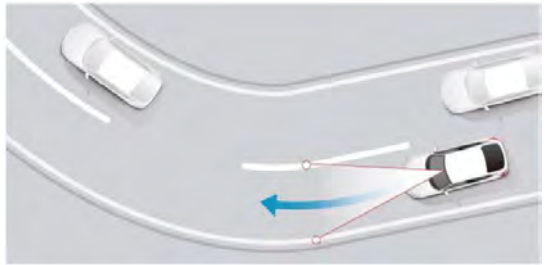


車線変更時衝突抑制機能

衝突回避の為ハンドル操作を支援

アダプティブクルーズコントロール（ACC）カーブ車速調整機能

カーブ手前で車線の曲率を読み取り、車速調整



車線変更支援機能

システムが車線変更に伴うハンドル操作を支援

世界初のレベル3自動運転　Honda SENSING Elite

HONDA

Honda SENSING Elite の提供価値

HONDA

高速道路本線・自動車専用本線上での
車線維持支援・車線変更支援・渋滞時運転機能を提供

走行シーン

ドライバー価値

目次

- 交通事故ゼロ社会と 自由な移動の喜び
- 運転支援と自動運転の概要
- **実用化された自動運転レベル3における 安全論証活動**
- 自動運転・運転支援システムに対するAI導入と Safety Assurance
- 最後に

レベル3の壁とは?

HONDA

自動運行装置

カメラ・レーダー・ライダー・360度状況検知
安全論証・実車/シミュレーション検証　など



適用

レベル4/5
自動運転（限定領域）/ 完全自動運転車

レベル3
条件付自動運転車
（限定領域）

レベル2
●運転支援車

レベル1
●運転支援車

要求

政府の求める安全技術ガイドライン

合理的に予見される
防止可能な人身事故が
生じない※

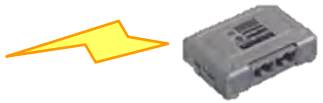
※ 国土交通省‘自動運転車の安全技術ガイドライン

Honda SENSING Eliteのシステム構成

HONDA

自車位置認識

バックエンドサーバー



TCU（通信ユニット）

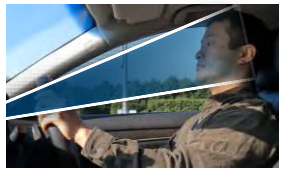
マルチGNSSアンテナ

地図ECU
＋高精度地図

自車位置

ドライバー状態検知

ドライバー
モニタリングカメラ

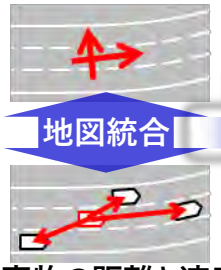


把持センサー
操舵トルクセンサー

ドライバ状態

行動計画

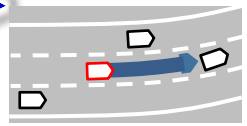
座標照合と白線補正



地図統合

ローカル
マップ
処理

Main-ECU



障害物の距離と速度

最適な行動計画

周辺車両

車両制御

システム状態

HMI

センター・ディスプレー
（NAVI）

ヘッドアップ
ディスプレー

ステアリング
インジケーター

全面液晶メーター

外界認識

Radar Fusion

Camera

Radar

Sub-ECU1

LiDAR Fusion

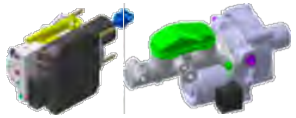
Camera

LiDAR

Sub-ECU2

車両制御・機能冗長

＜止まる＞
ブレーキ冗長

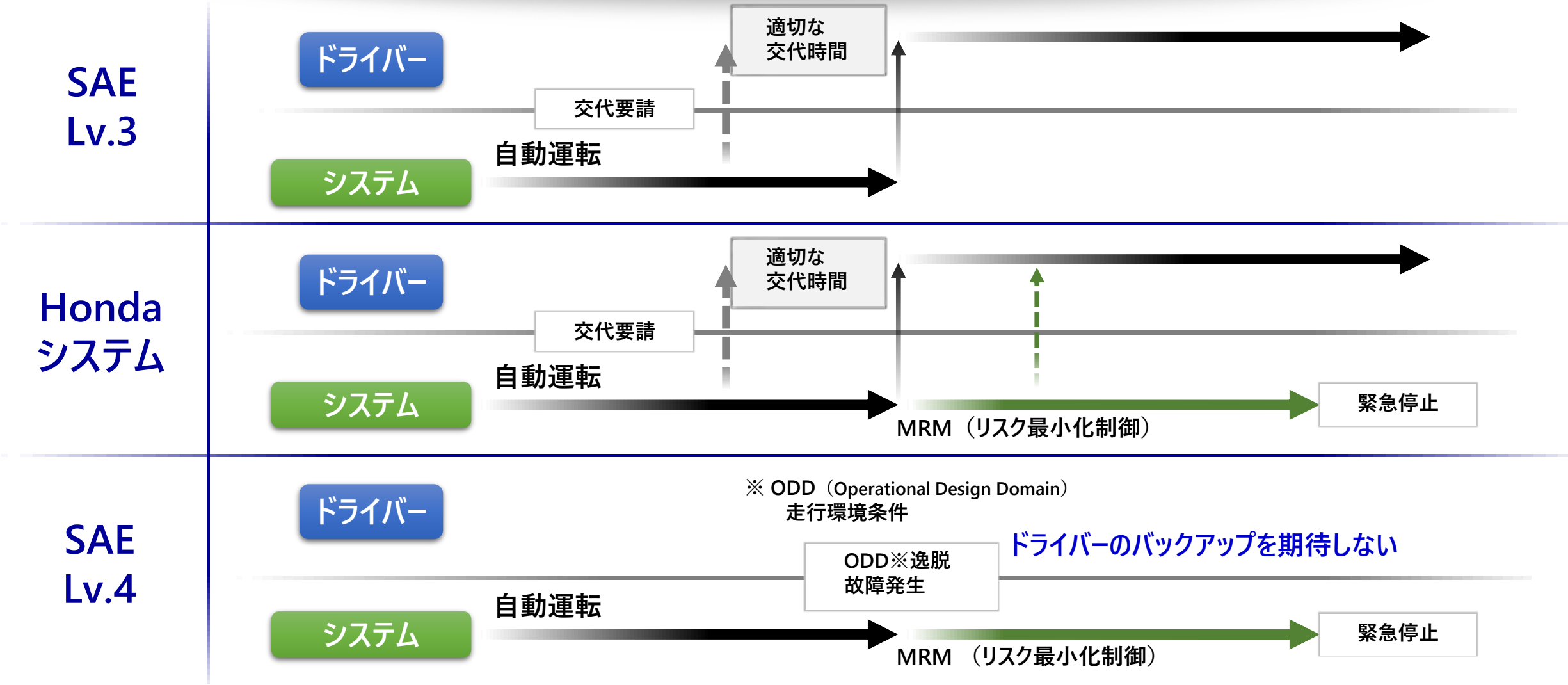


＜曲がる＞
EPS冗長化

＜電源＞
2nd バッテリー

安全な運転交代の考え方　SAEレベル定義と Hondaのシステム

HONDA



運転交代時の安全性を考慮し リスク最小化制御の導入を基準に先駆けて決定（2017年発表）

安全な運転交代の考え方

HONDA

ドライバーへの運転操作要求は 視覚・聴覚・触覚で確実に伝達
万が一要求に応じない場合はリスク極小となるように停車

トラフィックジャムパイロット

← 運転操作要求 →

緊急時停車支援機能

視覚（ディスプレー/ メーター / インジケーター）による警告

聴覚による警報（弱）

聴覚による警報（強）

触覚（ベルト振動）による警報

ハザード/ホーンによる
外部通報

停車

自動運転の開発プロセス

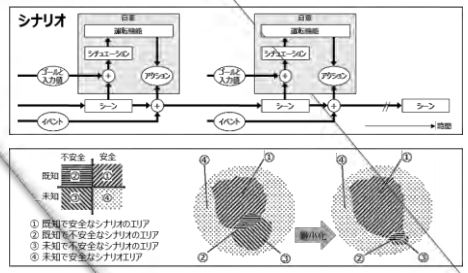
HONDA

安全性・信頼性を 最重視した開発

自動運転のシステム安全性を達成する為に 国際的 基準・標準を組み込んだ開発プロセスを構築
安全目標の設定から、システム仕様の設計検証 ならびに 実装の妥当性検証 まで網羅的に対応

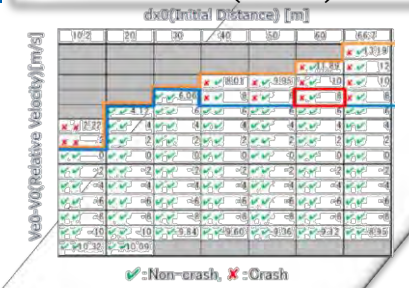
自動運転 安全の基準	ACSF/FRAV/VMAD （UN-ECE）	道路運送車両法 （国交省）	道路交通法 （警察庁）

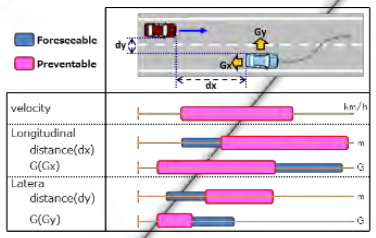
ISO 21448 SOTIF（draft）



ISO 26262 Functional Safety

ISO 34502 Scenario-based safety evaluation framework（draft）



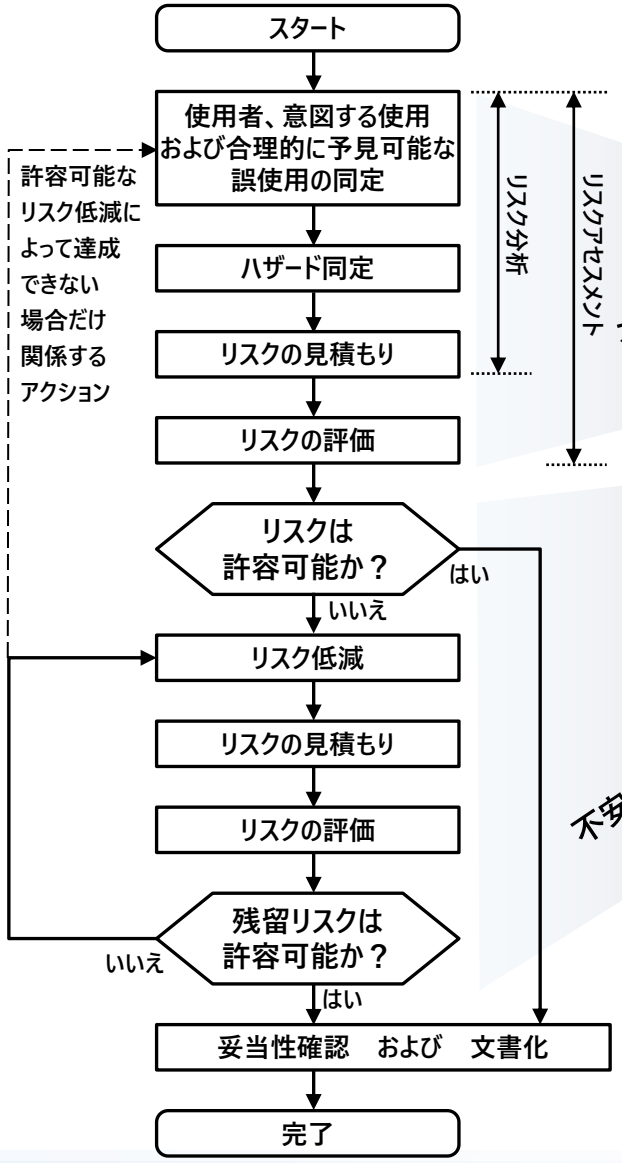


ISO Guide51 に沿った論証Vプロセス

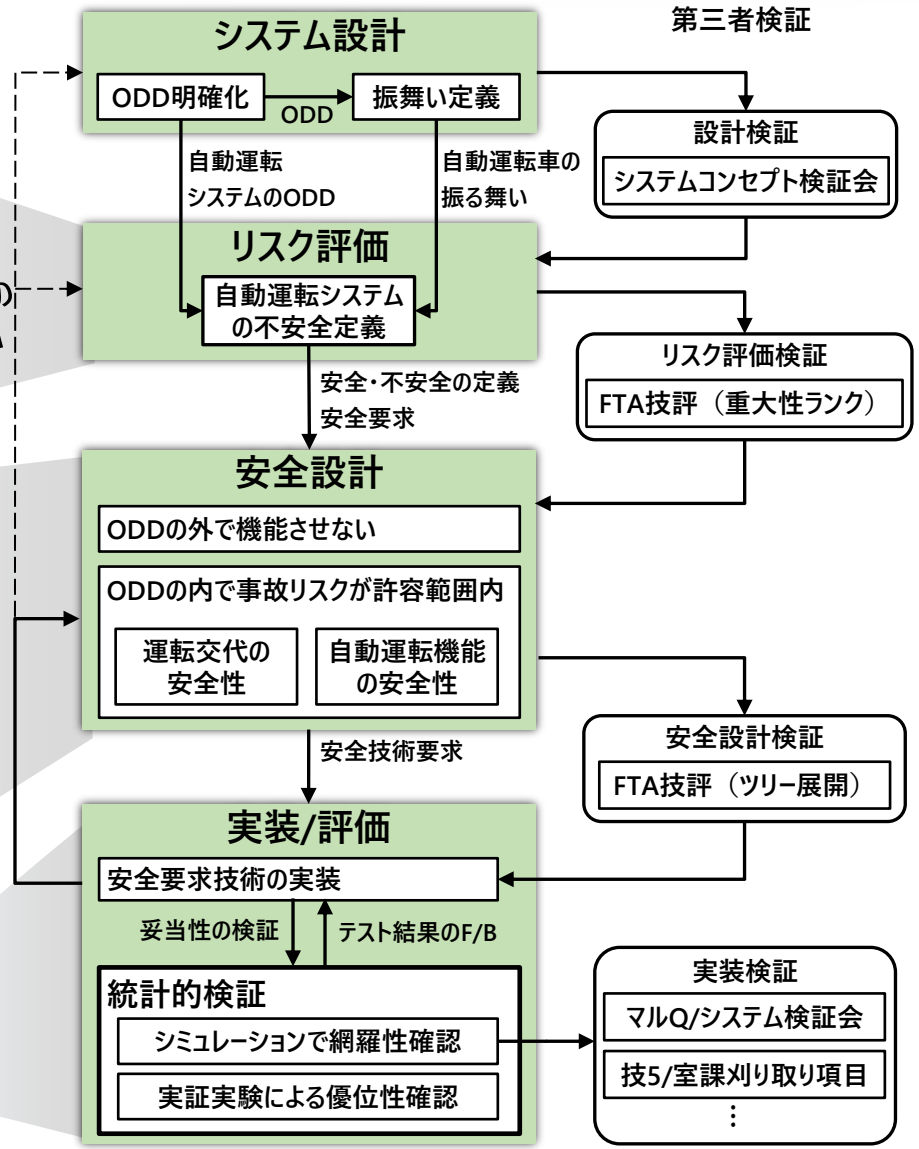
Hondaの安全設計プロセス

HONDA

【ISO Guide51 のリスク低減プロセス】



【Hondaの安全設計プロセス と 体制】



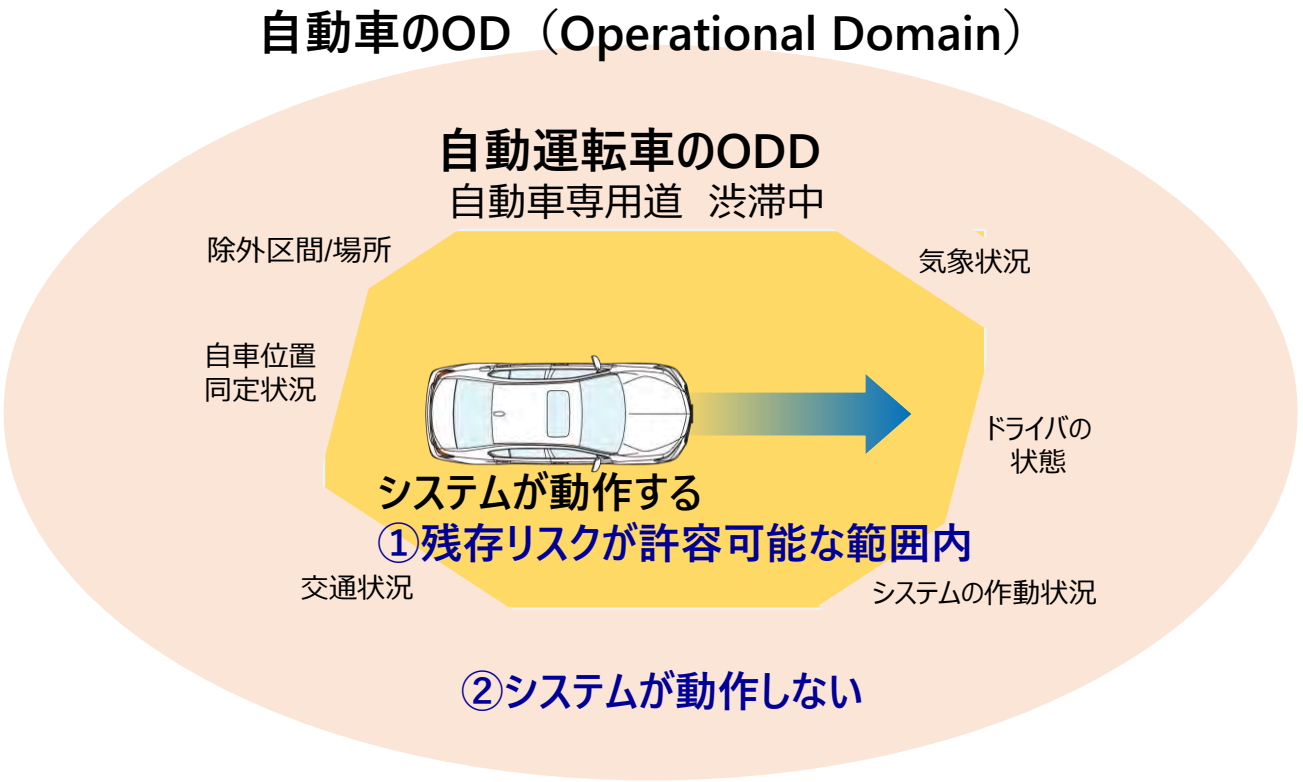
Honda	SOTIF	FUSA	Scenario Validation	SaFAD	UL4600
◎	◎	○	○	○	◎
○	◎	○	○	○	◎
◎	◎	○		◎	◎
○	○	○			○
◎	○			◎	○
◎	○	○		◎	○
○	○	◎		○	○
○	○	◎	◎	○	
◎	○		◎	○	
◎	Hondaは実証実験による優位性の確認含め網羅的に対応できている				

自動運転における 安全論証の考え方

HONDA

ODD（Operational Design Domain）とは

特定の「運転自動化システム」または その機能が動作する様に専用設計された運行条件



①と②を証明することによって
自動運転の安全性を論証することが可能

ODDを適切に設計し 自動運転システムを動作させる領域を限定することによって、
発生しうる危害も限定する事が可能となる

自動運転車の安全性

HONDA

How safe is safe enough?

ISO Guide 51「安全側面 −規格への導入指針」

安全：許容できないリスクが無いこと

リスクはゼロではない　絶対安全はない⇒ リスクは残る

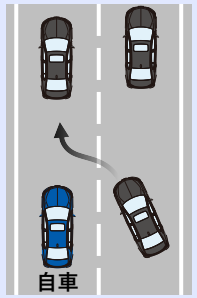
許容できるリスク

現在の社会価値に基づく一定の状況下において、受け容れられるリスクのレベル

ITARDA 事故データベース
（統計データ、ミクロ調査データ）
ヒヤリハット データ

自動運転車は
合理的に 予見可能 × 防止可能 な
人身事故を引き起こさない

有能で注意深いドライバー
同等以上の回避性能



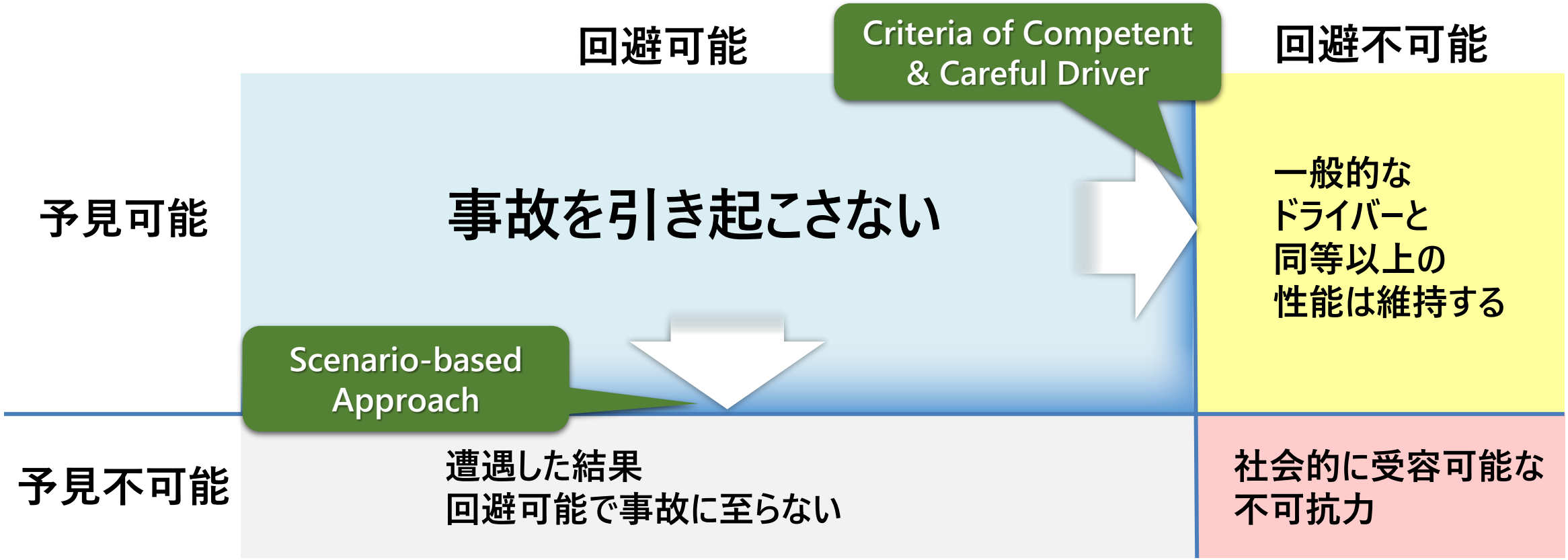
例：
割込み車両等

安全設計と評価

HONDA

システムは 合理的に予見され 回避可能な事故を引き起こさない

国交省「自動運転車の安全技術ガイドライン」/ UN R157（2020）



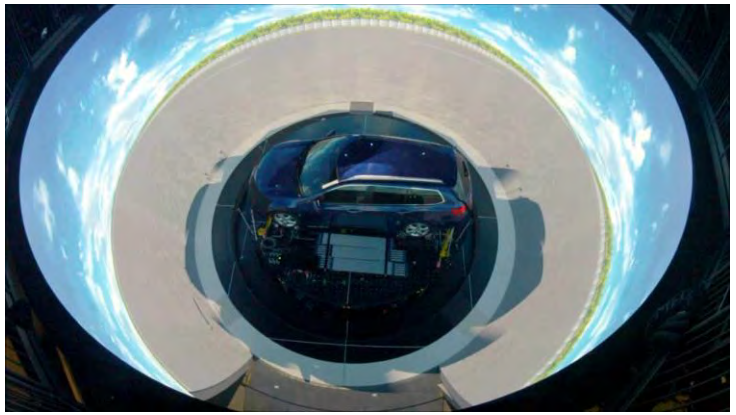
ISO 21448 SOTIF（Safety of the Intended Functionality）を先取り

Competent and Careful（C&C）Driverの基準

HONDA

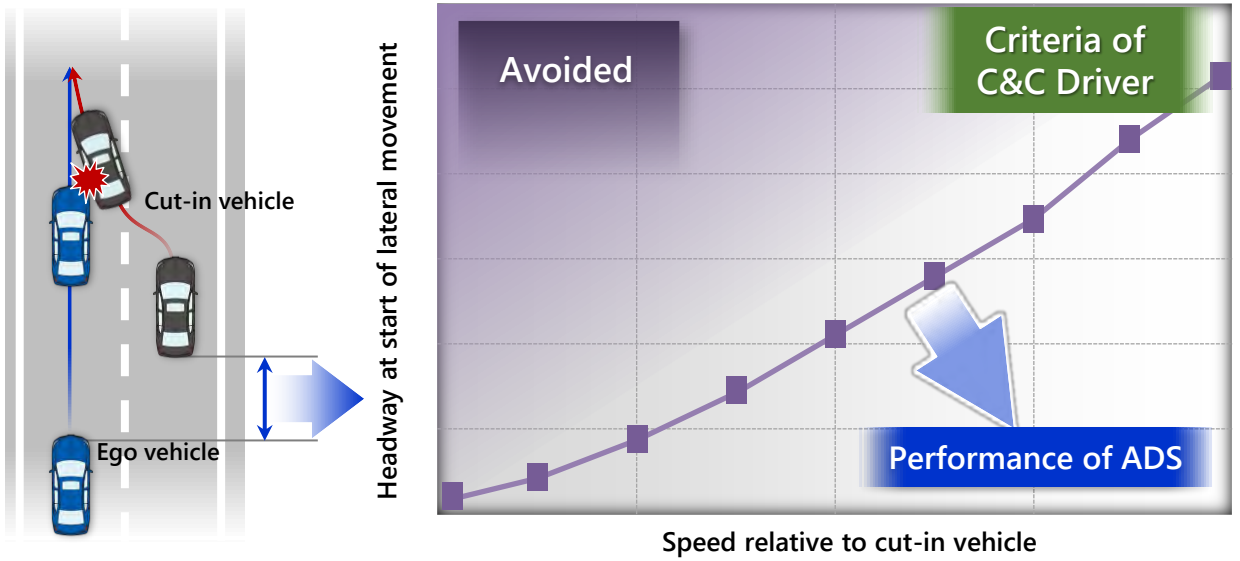
- ドライビングシミュレータ

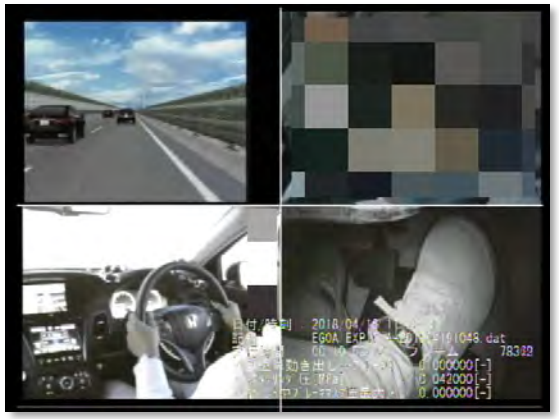
Vision-focused Type



Motion-focused Type

- 例：割り込みシナリオ



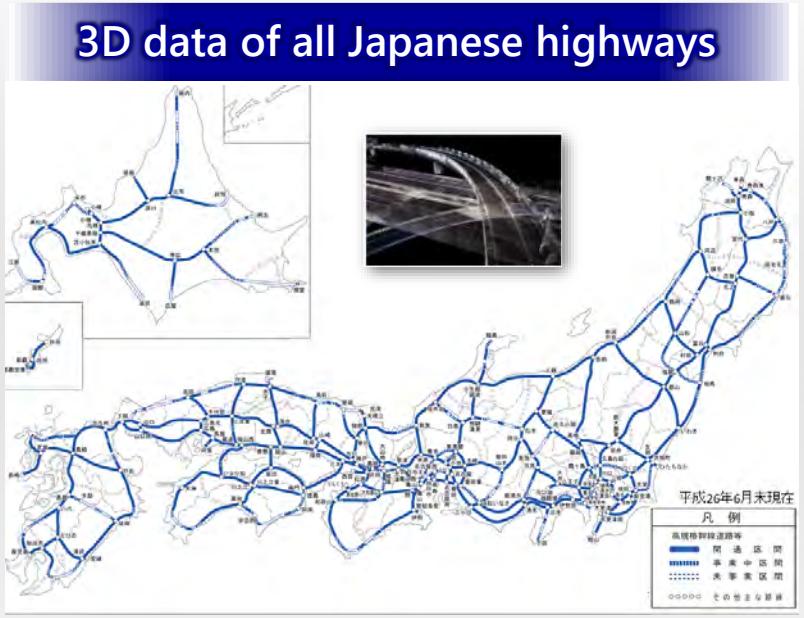


一般被験者のテストによって、C&C driver の基準を導出

シナリオベースの検証パターン生成

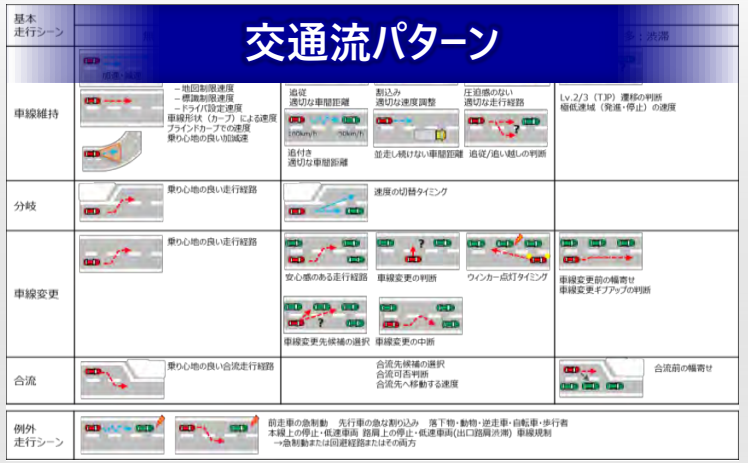
HONDA

国内全高速道路のモデル化

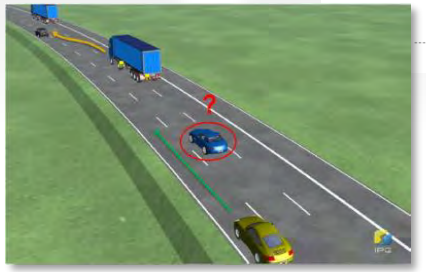


3D highway archive data（Mobile Mapping System data）is converted into a simulation model

交通流パターンの設計



国内全高速道路モデル



網羅的交通流パターン

モデル環境上で あらゆる交通流を再現するための 検証パターンを作成

Scenario Based Safety Evaluation Framework（ISO 34502）を先取り

安全性の検証

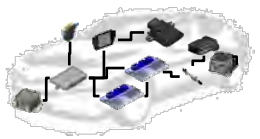
HONDA

シミュレーション

先進のコンピューターシステムを用いた
約1,000万通りのシミュレーション

ドライビングシミュレーター

モデル・イン・ザ・ループ・
シミュレーション
（MIL）



ハードウェア・イン・ザ・ループ・
シミュレーション
（HIL）

相関

公道検証

全国約130万kmの実証実験

実際の走行環境を網羅する実証実験

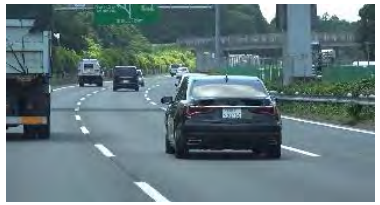
立体交差

トンネル

実証実験の様子

高速道路本線

渋滞



自動運転車（レベル3）の型式指定を取得

HONDA

国土交通省
Ministry of Land, Infrastructure, Transport and Tourism

令 和 ２ 年 １１ 月 １１ 日
自動車局審査・リコール課

世界初！　自動運転車（レベル3）の型式指定を行いました

国土交通省は、本田技研工業株式会社から申請のあった車両（通称名：レジェンド）に対し、自動運行装置を備えた車両としては世界初の型式指定を行いました。

１．概要

本田技研工業（株）から申請のあった自動運行装置を搭載した自動運転車（レベル３）について、（独）自動車技術総合機構交通安全環境研究所における保安基準適合性の審査を踏まえ、本日、世界で初めて型式指定を行いました。

自動運転車については、交通事故の削減、高齢者等の移動手段の確保、物流分野における生産性向上等、我が国が抱える様々な社会課題の解決に大きな役割を果たすことが期待されています。そのため、自動運転に係る政府全体の戦略である「官民ＩＴＳ構想・ロードマップ」（ＩＴＳ総合戦略本部決定）において、市場化・サービス化に係るシナリオと目標を掲げ、国土交通省を含め官民一体となって早期実現に向け取り組んでおります。（別紙１）

同ロードマップにおいて、高速道路の自動運転車（レベル３）の市場化目標時期が 2020 年目途とされていることを踏まえ、国土交通省では、昨年５月の道路運送車両法の一部改正に基づき、本年３月、世界に先駆けて自動運転車の保安基準を策定するなど、早期導入に向け制度整備を進めてきました。（別紙２）

国土交通省としては、引き続き、自動走行分野において世界をリードし、様々な車社会の課題解決に大きく寄与する自動運転の一層の実用化、普及に取り組んでまいります。

国土交通省
報道発表資料より引用

2020年11月11日　世界初となる 自動運転車（レベル3）の型式指定

目次

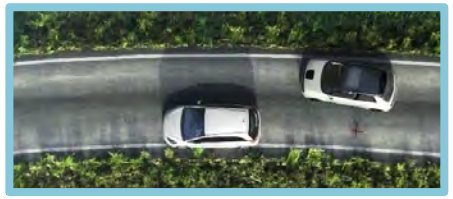
- 交通事故ゼロ社会と 自由な移動の喜び
- 運転支援と自動運転の概要
- 実用化された自動運転レベル3における 安全論証活動
- **自動運転・運転支援システムに対するAI導入と Safety Assurance**
- 最後に

知能化モビリティが 実現する未来社会

HONDA

いつでも、どこでも、どこへでも、人とモノの移動を「交通事故ゼロ」「ストレスフリー」で可能とし
「自由な移動の喜び」を一人ひとりが実感できる社会

苦手シーンを自動通過/運転支援
ストレスフリーで自由に移動

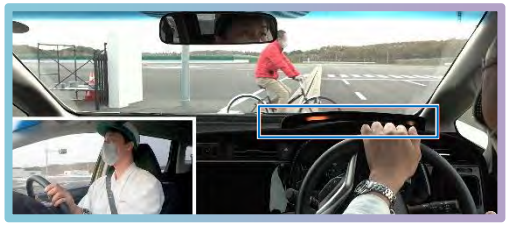


日常の短距離移動を
安全・快適に

人とモノのお気軽・安全
“短距離”移動

長距離を楽々移動

あらかじめリスクを知って
自分で回避



“もしも”のときは
自動で回避

自動運転/
運転支援車

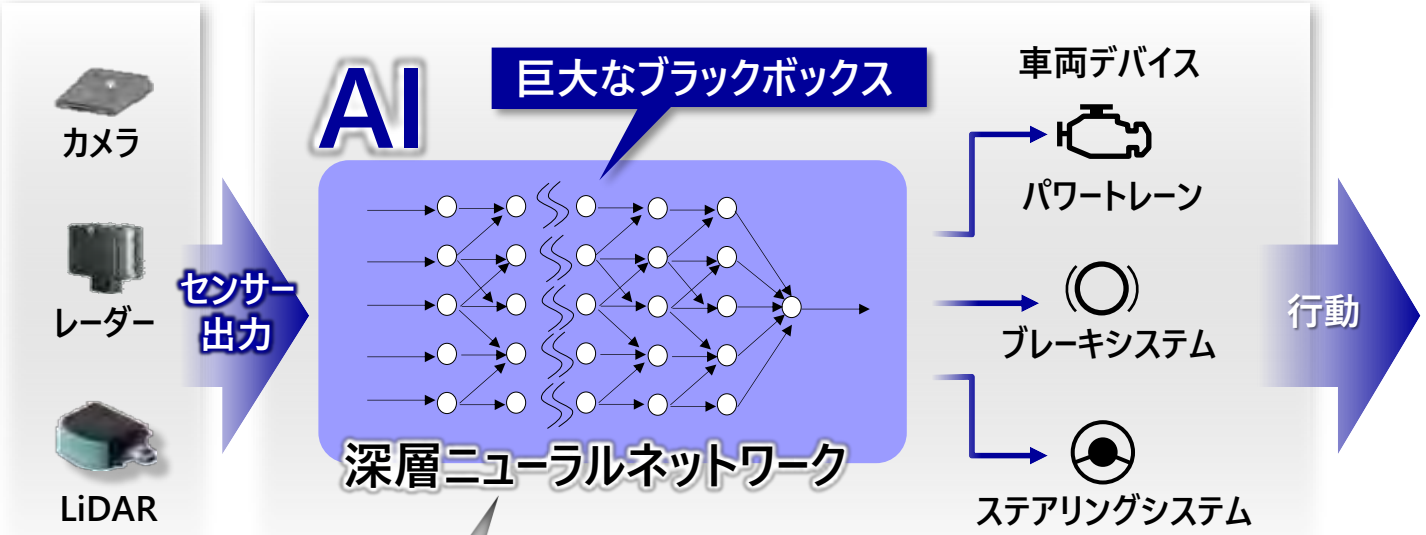
知能化
モビリティ

知能化
マイクロモビリティ

自動運転・運転支援システムへのAI適用の課題

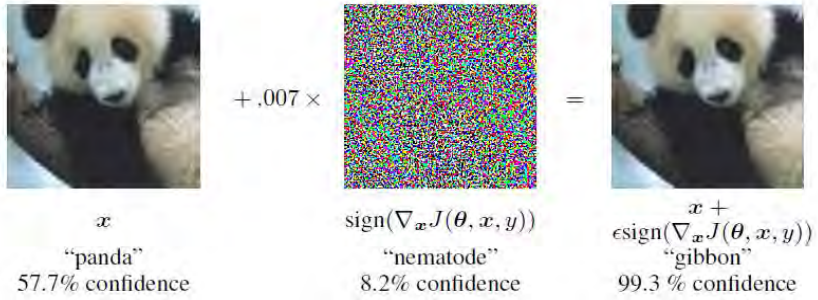
HONDA

End-to-End型自動運転システム



AIが自動運転へ適用され，
十分なロバスト性を示さなかったとき，
何が起きる？

AIのロバスト性課題（例）



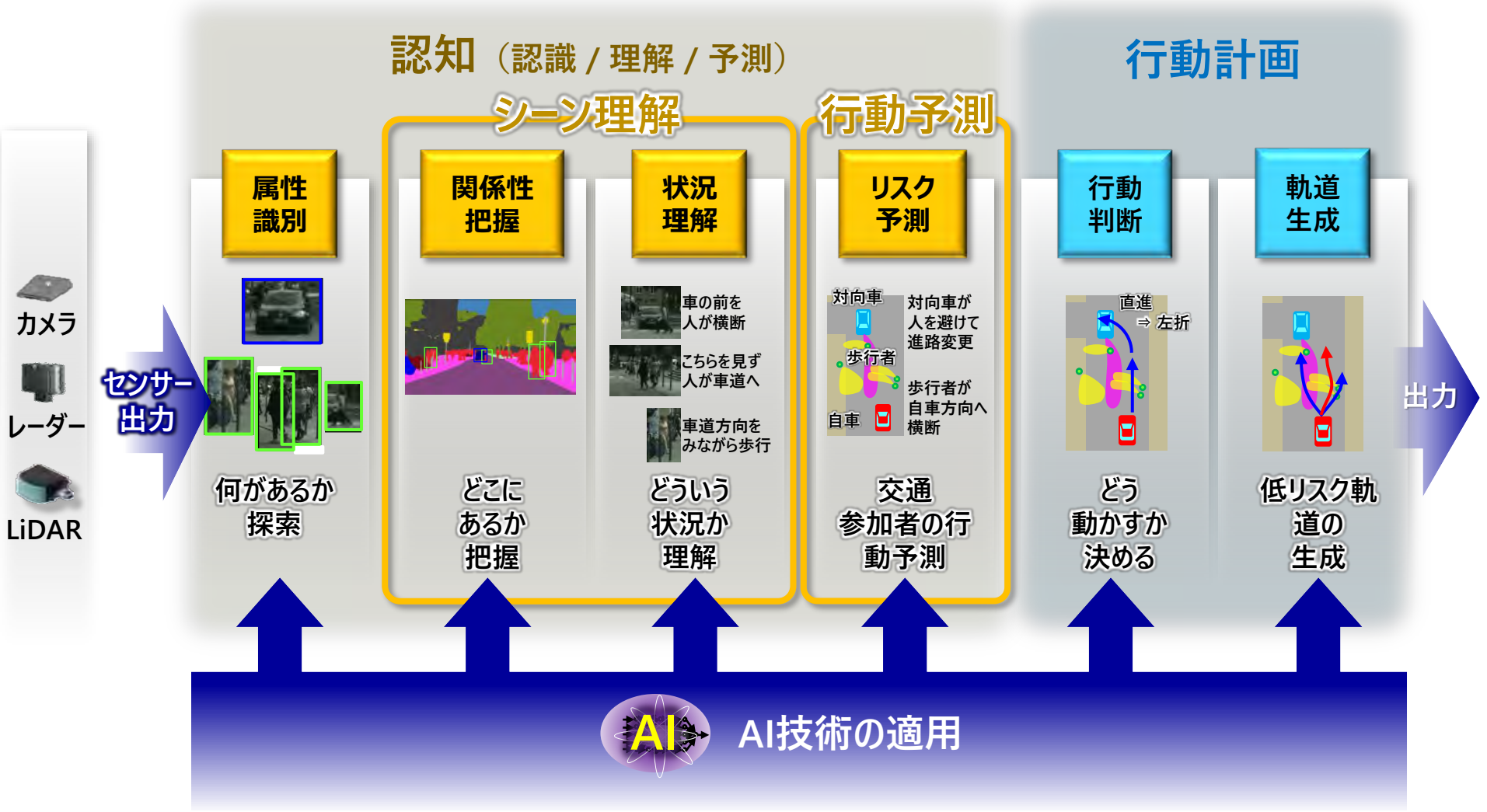
人間は"Panda"として認識

AI は "Gibbon"として認識

Source：Ian J.Goodfellow, et al., Explaining and Harnessing Adversarial Examples, ICLR（2015） 8

AIを用いた自動運転・運転支援システムの構成

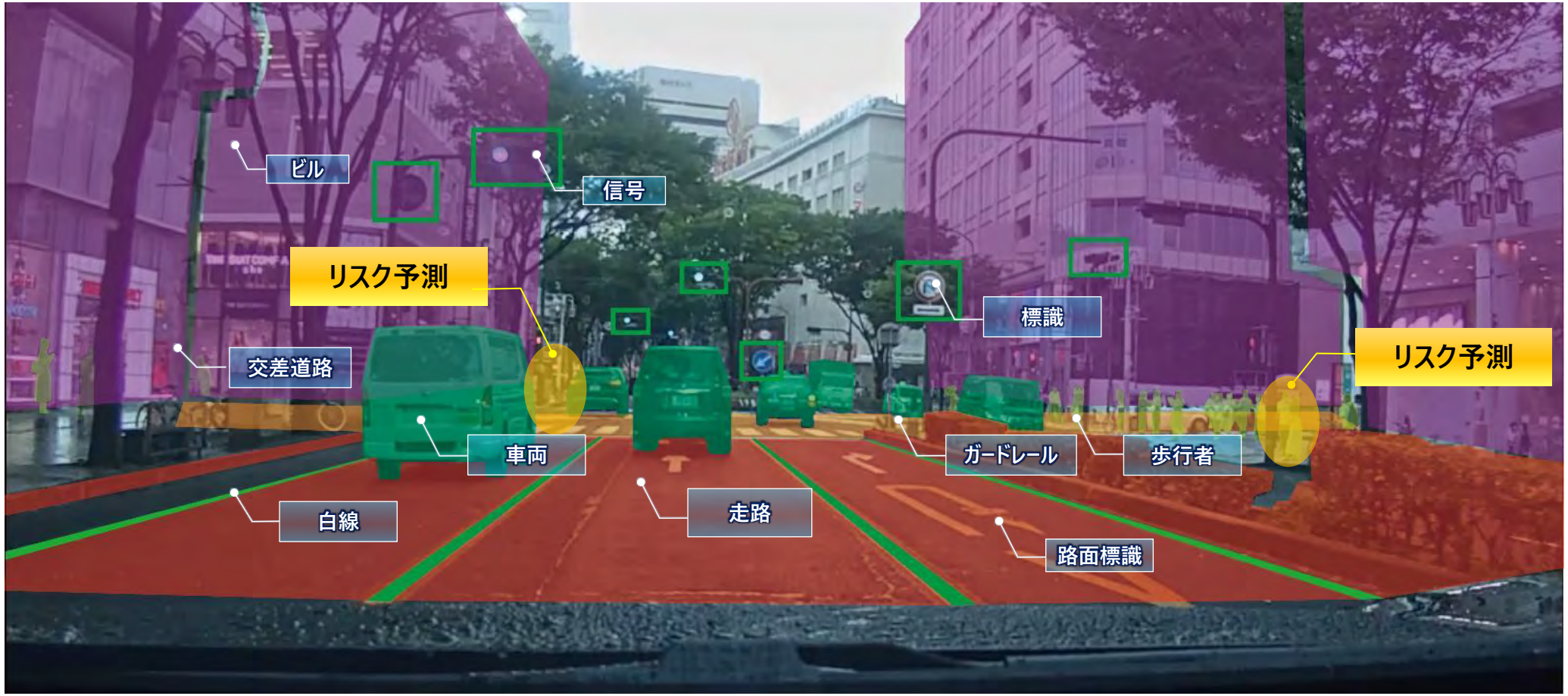
HONDA



認知領域に対するAI技術活用

HONDA

ビル、ガードレール、交差道路など複雑なシーンの認識が可能



見通しの悪い交差点・路上にある“危険リスクの予測”により一般道での運転支援を実現を目指す

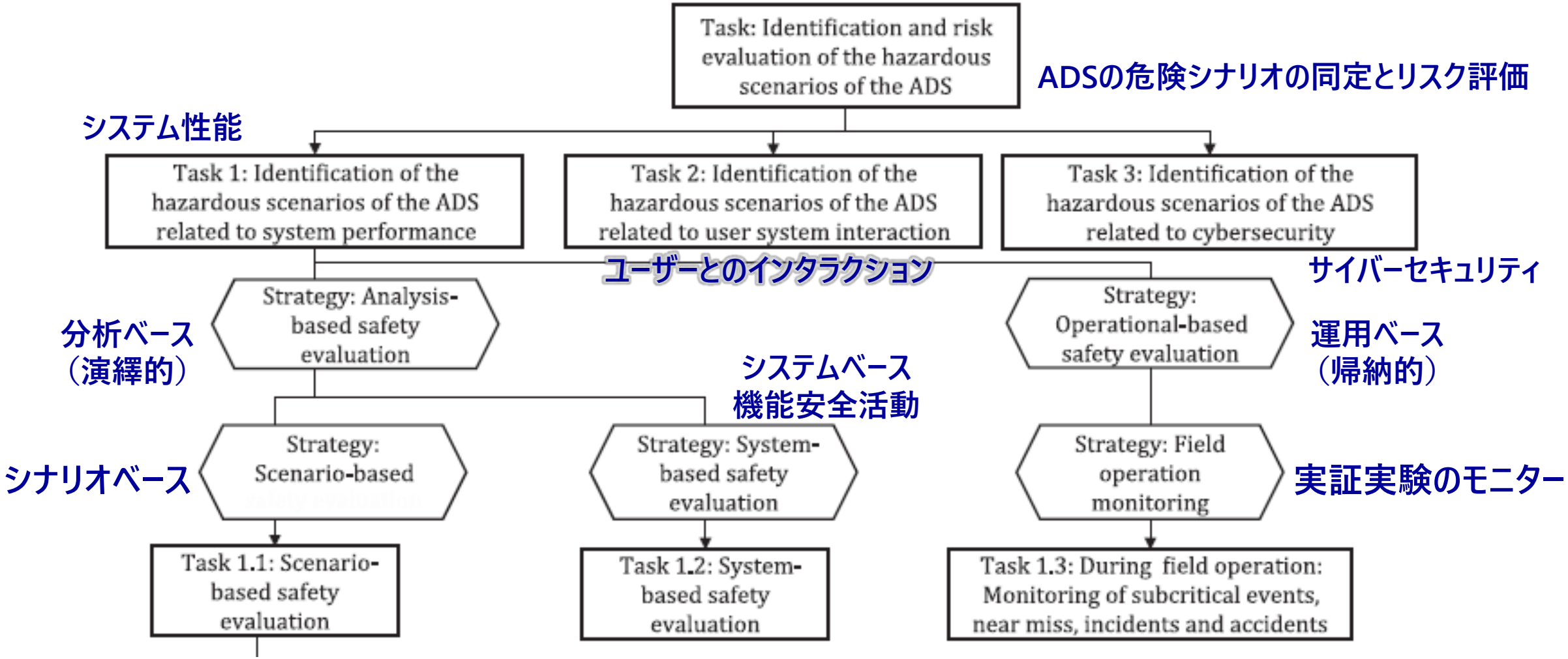
自動運転・運転支援システムに対するEnd-to-Endアプローチ

HONDA

課題：explainable, interpretable？　hallucinationは？

自動運転のSafety Assurance -シナリオベースアプローチと実証的アプローチ-

HONDA



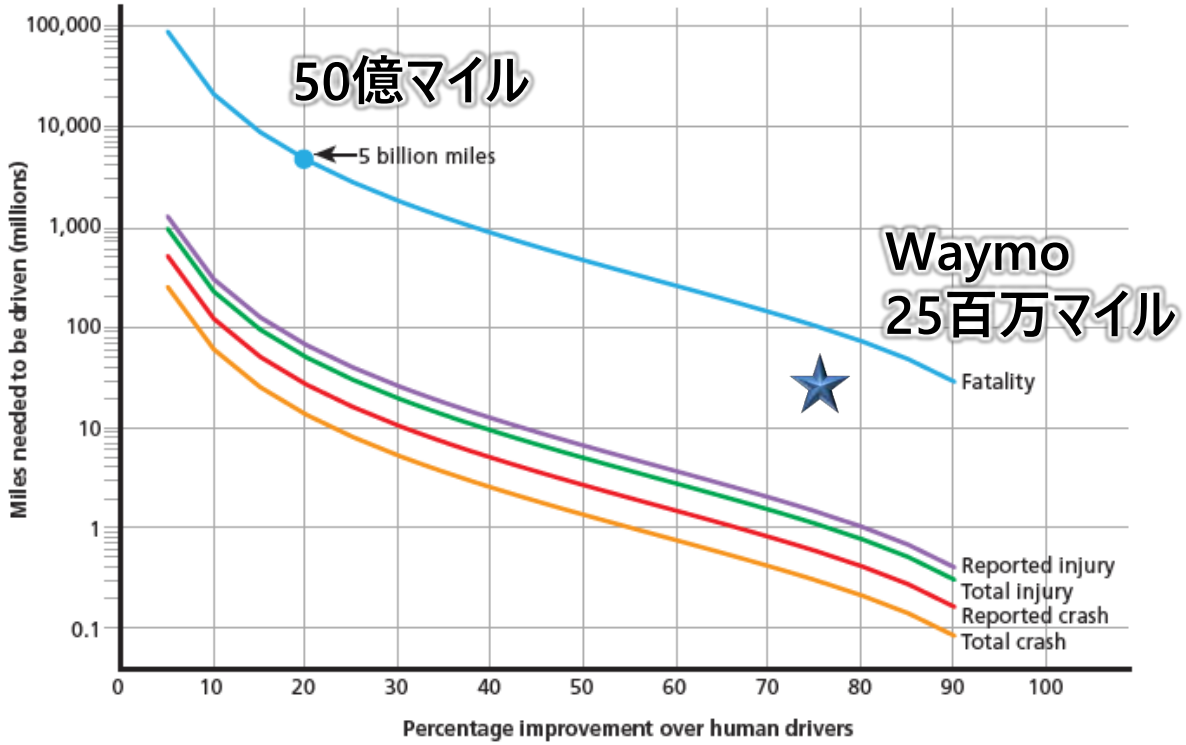
Road vehicles — Test scenarios for automated driving systems — Scenario based safety evaluation framework

出典：ISO 34502:2022

実証的アプローチ

HONDA

Figure 3. Miles Needed to Demonstrate with 95% Confidence that the Autonomous Vehicle Failure Rate Is Lower than the Human Driver Failure Rate



SOURCE: Authors' analysis.
NOTE: The results depend upon the estimated failure rate of autonomous vehicles. This is shown on the horizontal axis and defined as a percent improvement over the human driver failure rate. The comparison can be made to the human driver fatality rate (blue line), reported injury rate (purple line), estimated total injury rate (green line), reported crash rate (red line), and estimated total crash rate (orange line).
RAND *RR1478-3*

出典：Rando report [Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability? | RAND]

Waymo 走行実績 ～2024/7

Locations	RO Miles through July 2024
Los Angeles	1.097M
San Francisco	7.134M
Phoenix	17.049M
Austin	28K

出典： Waymo HP [Safety Impact]

シナリオベースアプローチ　ISO34502

HONDA

自動運転の安全性の考え方

合理的に予見可能で防止可能な人身事故は起こさない

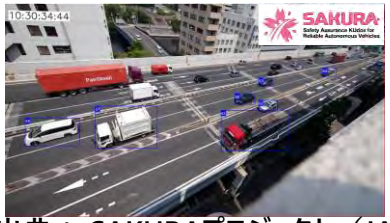
予見可能
実環境データ等から対象範囲を抽出

防止可能
事故回避が可能な範囲

事故は起こさない
人とシステムを比較

【ISO34502 の 安全性検証プロセス】

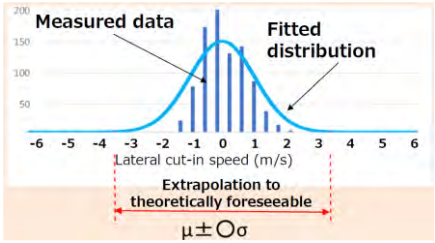
実環境の交通データ



出典： SAKURAプロジェクト（JARI）

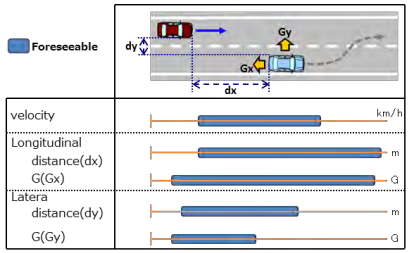
網羅的な機能シナリオ

予見可能性

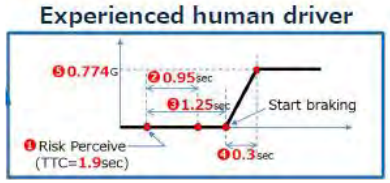


論理的な機能シナリオ

※事故シナリオ含む

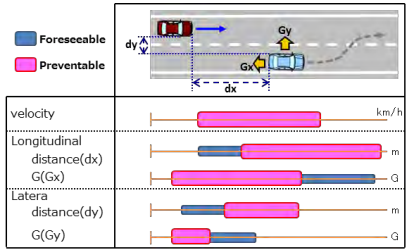


回避可能性

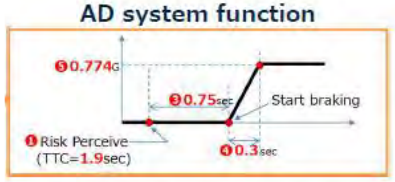


具体的な試験シナリオ

※回避すべきシナリオ

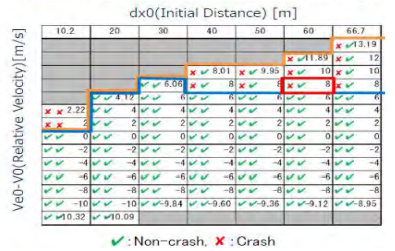


システム機能モデル



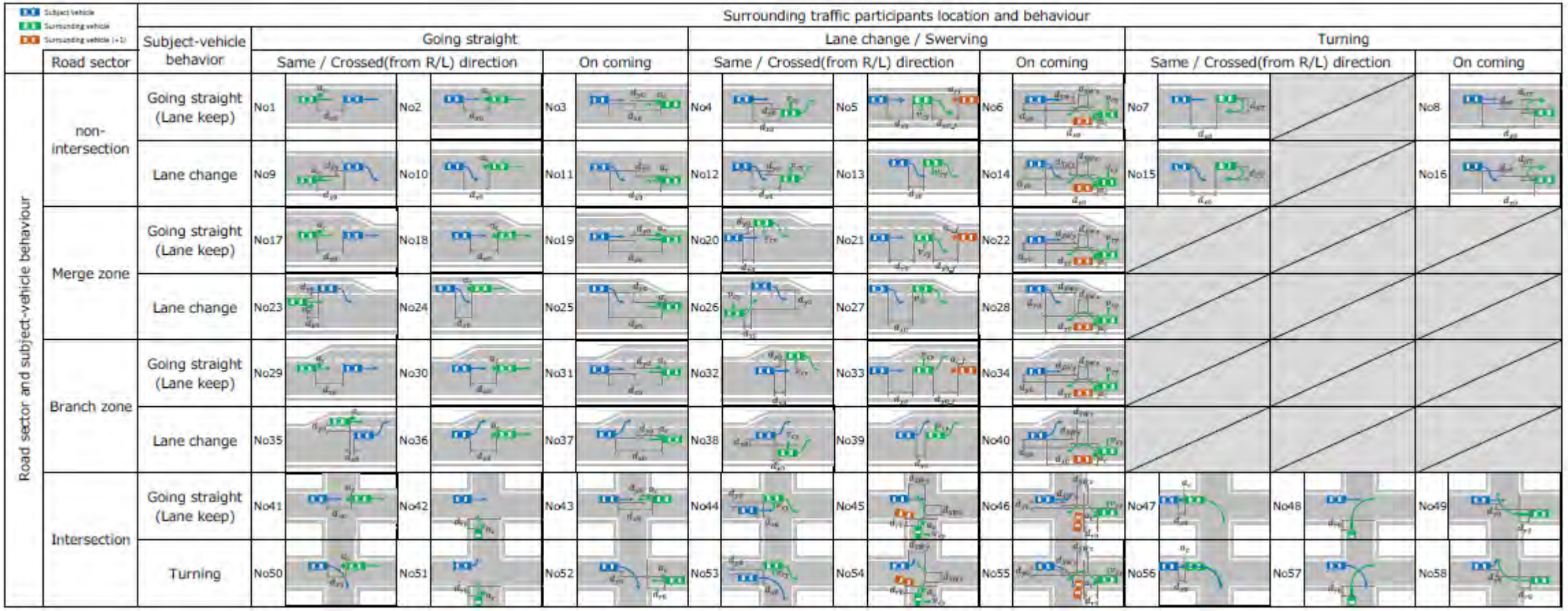
安全性の評価

※人とシステムの比較



一般道　交通外乱シナリオ

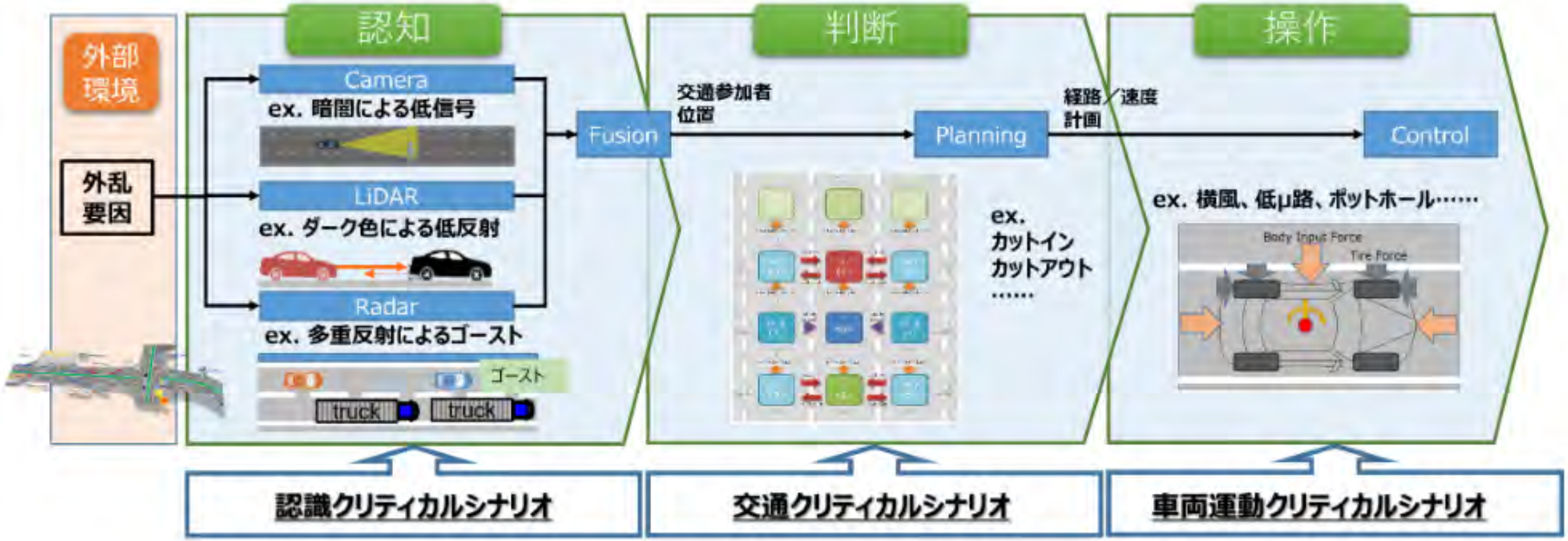
HONDA



一般道は 58のシナリオに集約

出典：自動運転の安全性評価フレームワーク Ver 3.0 日本自動車工業会

シナリオの体系化



三要素に分解して、体系的に整理

出典：経済産業省HP　日本発の自動運転システムの「シナリオに基づく安全性評価フレームワーク」に関する国際標準が発行されました　（METI/経済産業省）

安全性検証手法

HONDA

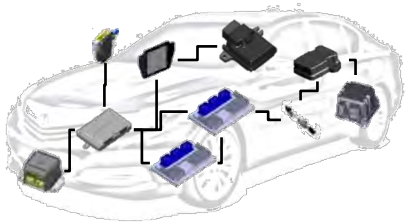
Driving Simulator

ドライバーの感性 応答性などを評価

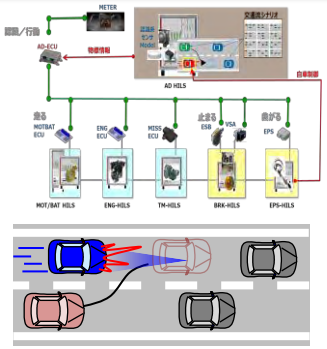
対象　ドライバー

実ECU 検証

対象　ECU実装ソフトウェア



HIL

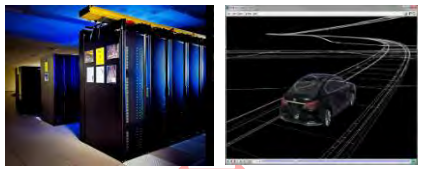


危険なシーンをシミュレーション

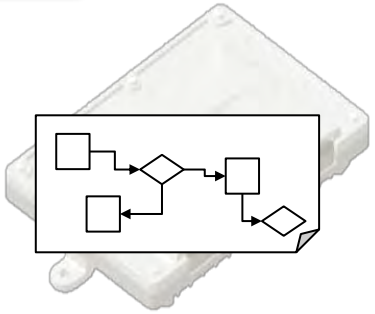
Hardware In the Loop Simulation

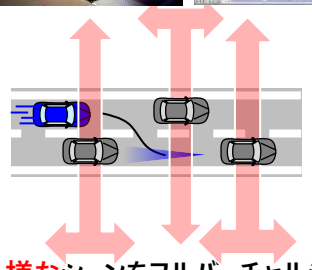
MIL

アルゴリズム検証



対象　ソフトウェア





多様なシーンをフルバーチャルシミュレーション

Model In the Loop Simulation

実交通流・実環境で総合的に評価

対象　全システム/ハード

実車検証

立体交差

S.A / I.C

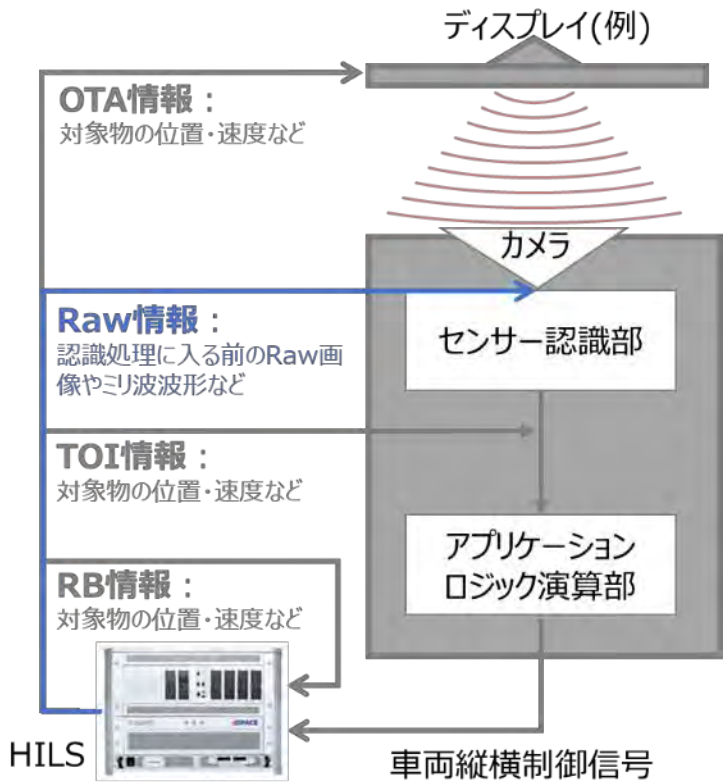
ジャンクション

Raw Data インジェクションを用いたカメラ認識処理の検証

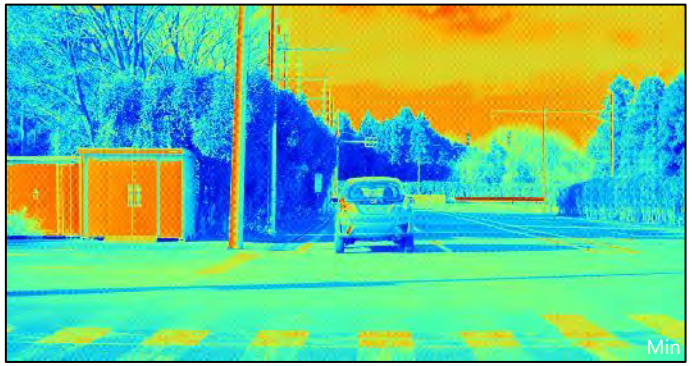
HONDA

CGを元にカメラのRawデータを作成し カメラの認識処理へ直接入力

インジェクションの考え方

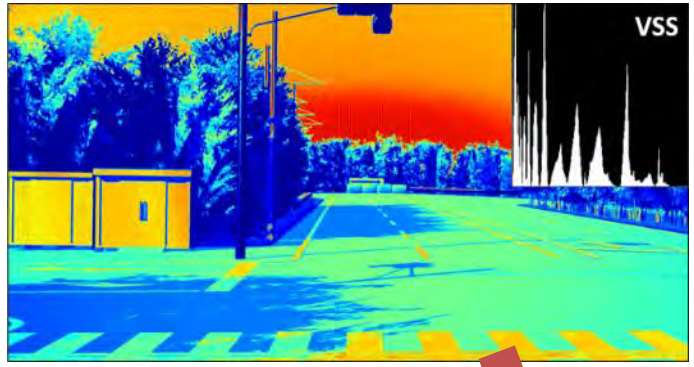


実車



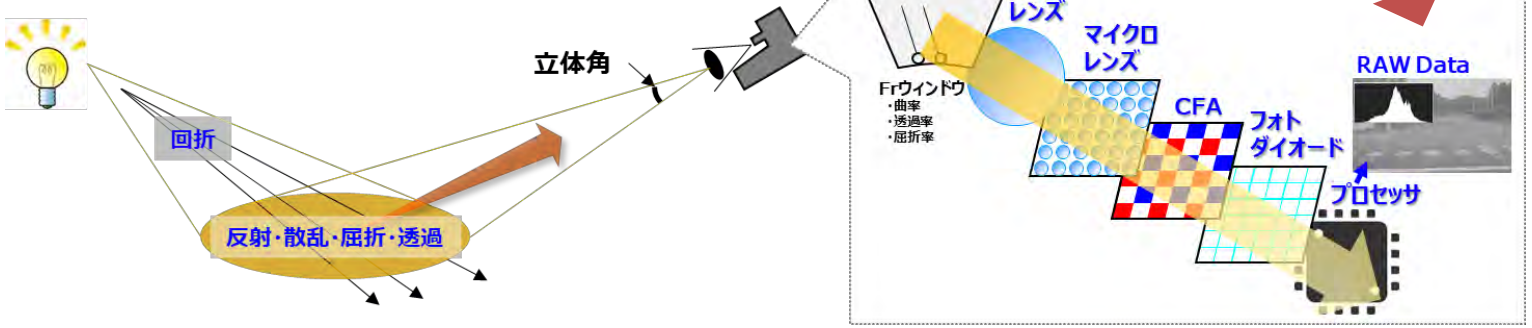
イメージャー電圧値[V]

CG



CGから
インジェクション用
Rawデータを作成

光源からカメラ認識のイメージ図



デジタルツインを核にした開発検証環境

HONDA

実車計測と仮想検証を融合させることで AIなどの開発を高速進化

デジタルツイン　World Modelの開発

HONDA

デジタルツインを核に
高速でソフトウェアを進化可能な開発環境を構築

目次

- 交通事故ゼロ社会と 自由な移動の喜び
- 運転支援と自動運転の概要
- 実用化された自動運転レベル3における 安全論証活動
- 自動運転・運転支援システムに対するAI導入と Safety Assurance
- **最後に**

最後に

- AIを適用した自動運転・運転支援システムの安全性を論証するには
  従来からの機能安全活動に加え SOTIFやシナリオベースなどの
  分析的アプローチによる検証が一層重要に
- 実走行車両からの data-drivenによるデジタルツインの活用も
  加速してゆく
- AIの適用には 開発プロセス管理、トレーサビリティ、透明性が重要
  ⇒ 規格、基準の重要性が高まってゆく

HONDA

The Power of Dreams

How we move you.

CREATE ▸ TRANSCEND, AUGMENT

DENSO
Crafting the Core

JAXA宇宙航空安全ミッション保証シンポジウム

デンソーにおけるAI品質保証の仕組み

2025/1/15

中神 徹也

株式会社デンソー
ソフト生産革新部

Agenda

1. 世の中のAI動向とデンソーの取り組み
2. デンソー社内におけるAI品質保証の仕組みづくり
    1. AIプロセス開発
    2. AIツール・AI技法開発
    3. AI品質保証支援・AI人材育成
3. まとめ

1

世の中のAI動向とデンソーの取り組み

世の中のAI動向

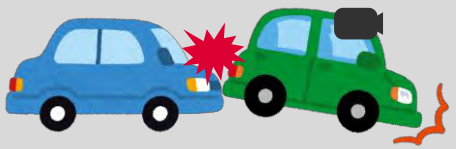
深層学習 ⇨ 画像　音声　テキスト

日本ソフトウェア科学会
機械学習工学研究会

[市場]
深層学習の登場によって
AIが様々な領域に浸透

[アカデミア] [1]
AIを工学的に扱うための
研究が活発化

深層学習の登場でAIの産業応用が加速



[Tesla, Uber] [4]
Autopilotモードの
利用による事故

[MS, Google] [5]
人種差別的なAIに
対する世論の反発

AIのリスクに対する社会の不安

[GAFA,Bosch,Conti] [2]
AI人材獲得・育成を
加速

[Tesla,Waymo,ME] [3]
MLOps技術による
AI開発の高速化

AIの積極的な活用が競争のカギ

[法規・標準] [6]
各国政府、ISOが
AIリスク対策に着手

[Bosch,NEC] [7]
自社のAI倫理ガイド・
品質ガイドを発信

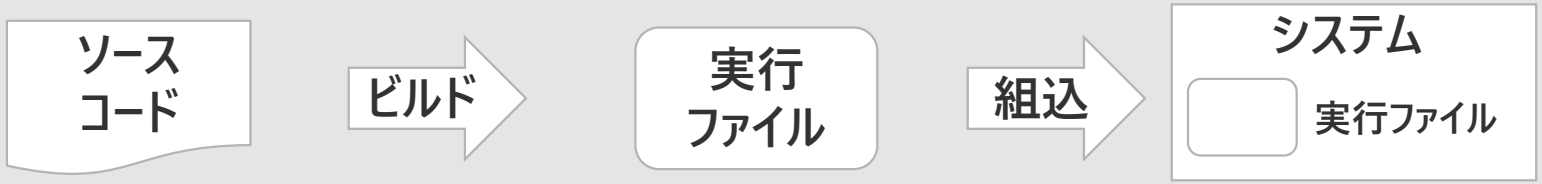
AIの社会受容性に関するルール化の動き

世の中がAIの研究から応用フェーズに入り、AIの社会受容性に関する取り組みを開始

[1] https://sites.google.com/view/sig-mlse
[2] https://www.bosch.co.jp/press/group-2001-02/
[3] https://www.youtube.com/watch?v=Q0nGo2-y0xY
[4] https://www.youtube.com/watch?v=w2VWAoLrzE0
[5] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist
[6] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[7] https://www.bosch.com/stories/ethical-guidelines-for-artificial-intelligence/

AI動向に関する技術的な背景

AIの振る舞いは訓練用データセットに基づいて決まるため、**AIの学習結果はブラックボックス**になる。

	設計・実装	実行ファイル	システム
従来のソフトウェア			
	ソースコードを記述	ソースコードから **明示的**に**振る舞い**が決まる	演繹的に単体の **振る舞いを保証できる**前提
AI（機械学習）			
	データセットを作成	データから**帰納的**に **振る舞い**が決まる	演繹的に単体の **振る舞いを保証できない** （テストデータに依存）

デンソーでは、どのようAIを開発したのかを説明できるようにするために、AI品質保証に取り組んでいる

デンソーの取り組み

事業部主体でAI搭載製品の品質を確保

デンソーのAI搭載製品群

安全・安心

利便・快適

新事業

社会に受容される仕組み

[1] ISO [2] [3]

世の中のAIルール（各国・国際の法規・標準）

整合

規程類

既存規程類

AI品質管理規程

AI設計基準（品質）

体制

AI品質分科会

デンソーの社内標準（基準・仕組み）

世の中のAIルールに対応できるようにするため、デンソーの社内標準づくりを開始

[1] https://pixabay.com/ja/vectors/%E5%9B%BD%E9%9A%9B%E9%80%A3%E5%90%88-%E9%9D%92-%E3%83%AD%E3%82%B4-uno-303670/
[2] https://pixabay.com/ja/vectors/%E6%AC%A7%E5%B7%9E%E9%80%A3%E5%90%88-%E3%83%A8%E3%83%BC%E3%83%AD%E3%83%83%E3%83%91-%E5%9B%BD%E6%97%97-155207/
[3] https://pixabay.com/ja/vectors/%E3%83%A6%E3%83%8A%E3%82%A4%E3%83%86%E3%83%83%E3%83%89-%E5%9B%BD%E6%97%97-%E5%B7%9E-26177/

2

デンソー社内におけるAI品質保証の仕組みづくり

AI品質保証における役割分担

【方針】AI法規・標準の要求事項対応時のバラツキを抑えるため、AI全社組織・事業部の役割分担を行う

AI法規,基準,業界標準 等

調査

AI全社組織

①AIプロセス開発
・世界・業界の法規標準動向をAI品質規程類に継続的取り込み
・事業部の使い勝手を踏まえたAI品質規程類の継続的改善

②AIツール・AI技法開発
・AI品質適用のための標準開発
（AI品質確保を効率化するツールなど）

③AI品質保証支援・AI人材育成
・事業部のAI品質保証担当やAI品質開発担当の育成・支援
・実プロジェクトへのAI品質適用に関する技術相談

支援

事業部

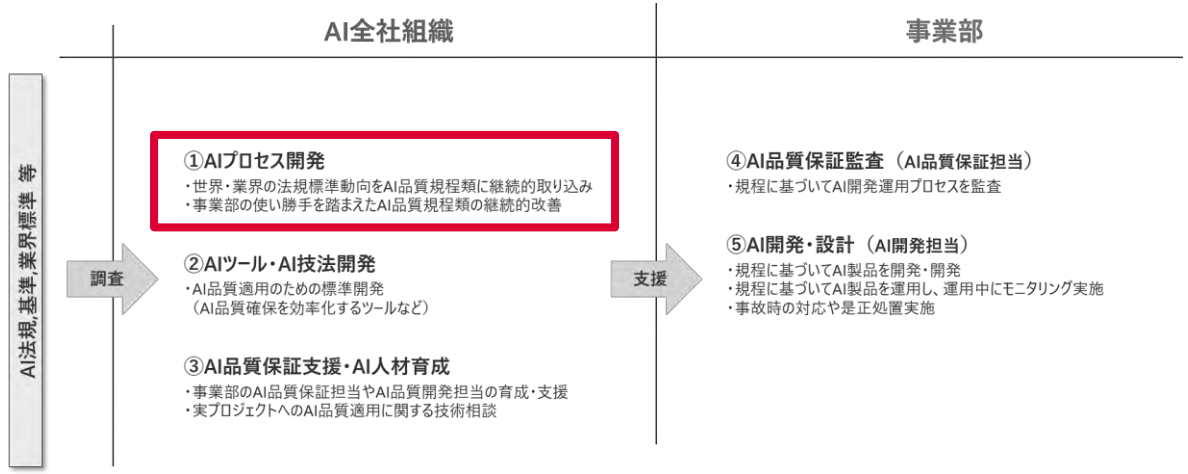
④AI品質保証監査（AI品質保証担当）
・規程に基づいてAI開発運用プロセスを監査

⑤AI開発・設計（AI開発担当）
・規程に基づいてAI製品を開発・開発
・規程に基づいてAI製品を運用し、運用中にモニタリング実施
・事故時の対応や是正処置実施

専門知見・経験を持つAI全社組織が、技術・社会変化を捉えてAI品質保証の仕組みを作成・更新・普及

AI搭載製品・サービスを持つ各々の事業部が、AIプロセスに基づく活動として主体的にAI品質保証を実施

①AIプロセス開発



AI品質保証の仕組みづくりの方針

目的①（マネジメント領域）：　法規・標準のAI要求事項を満たす
目的②（エンジニアリング領域）：　AI要求事項に含まれない技術的な内容を補完するため、デンソーが必要と考える内容を世間相場に沿って定義

デンソーでは、以下の枠組みでAI品質保証の仕組みを構築している。

方針①（マネジメント領域）

AI要求事項に基づいて管理規程を作成

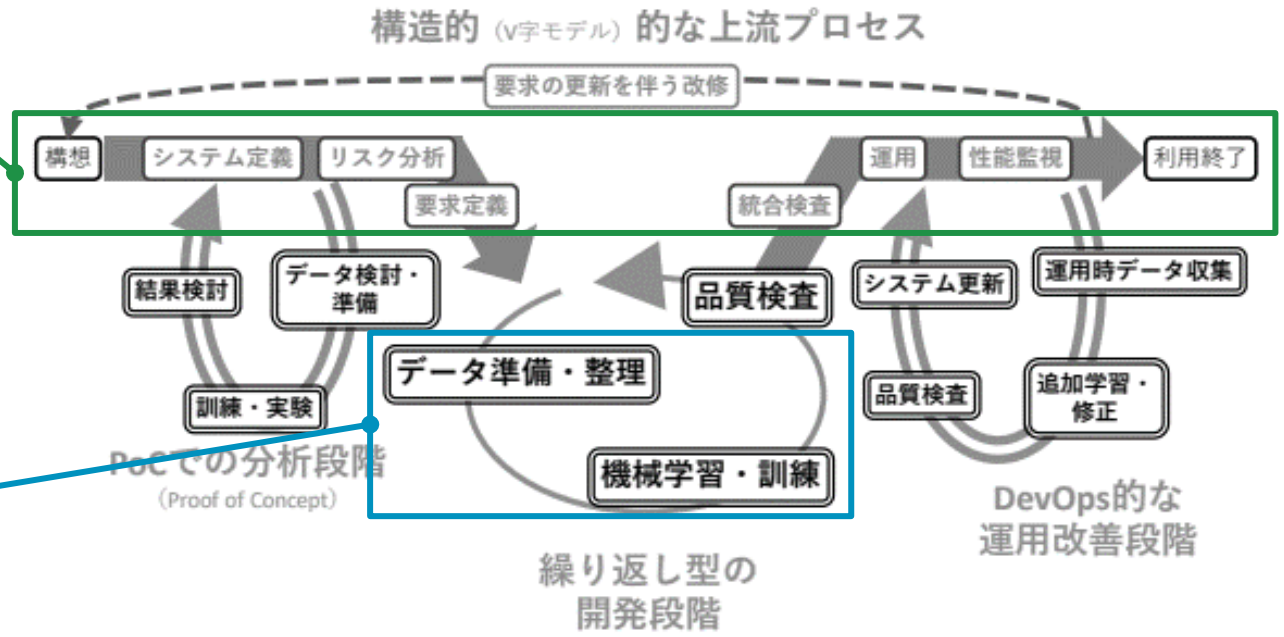
アウトプット：AI品質管理規程

方針②（エンジニアリング領域）

論文・ガイドライン※1に基づいて設計基準を作成

アウトプット：AI設計基準（品質）

※1：Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving、機械学習品質マネジメントガイドライン（AIQM）など

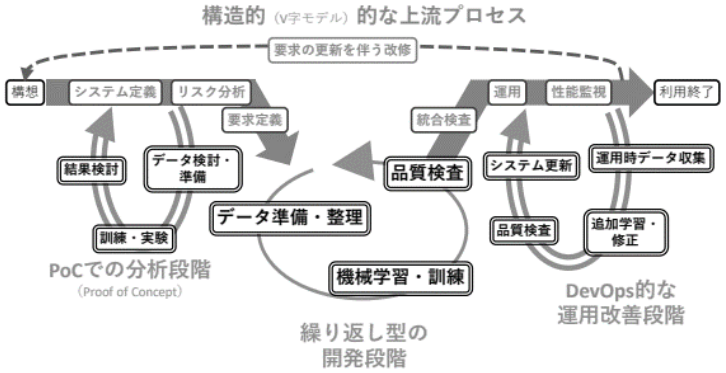


出典：AIQM第3版　混合型機械学習ライフサイクルプロセスの概念図

方針①（マネジメント領域）に対する構築アプローチ

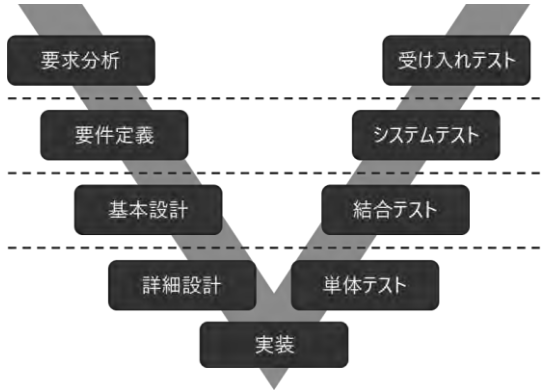
方針①（マネジメント領域）
AI要求事項に基づいて管理規程を作成
アウトプット：AI品質管理規程

開発者は、AI開発において、繰り返し型の開発プロセスとV字型の開発プロセスの**両方に対応**する必要あり。



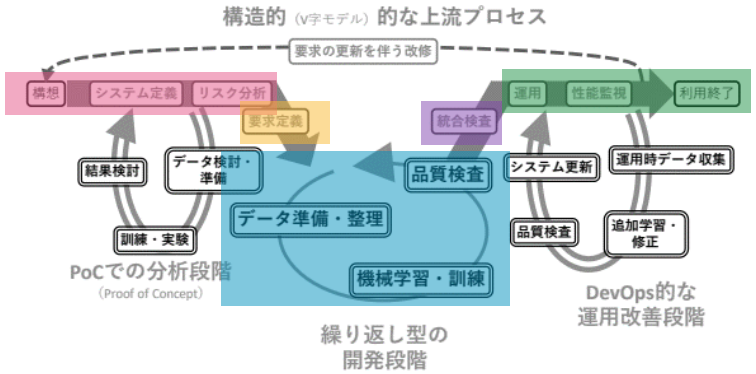
繰り返し型のAI開発プロセス

どちらのプロセスにも準拠していることを示す必要あり



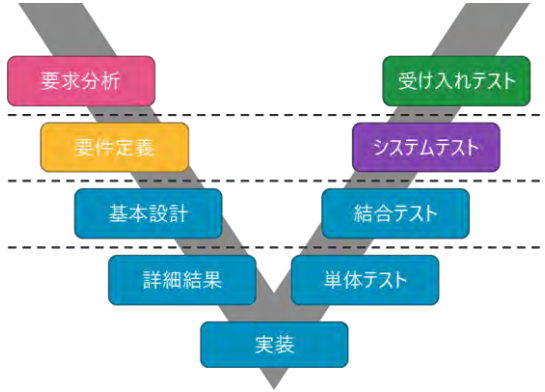
V字型の開発プロセス

AI開発プロセスの各活動を分解して従来プロセスに反映することで、開発者は**1つだけの開発プロセスに対応**すればよい。



繰り返し型のAI開発プロセス

従来プロセスと整合するように反映



V字型の開発プロセス

従来の製品開発からスムーズに移行するため、繰り返し型のAI開発プロセスをV字型の開発プロセスに反映

AI品質管理規程の概要

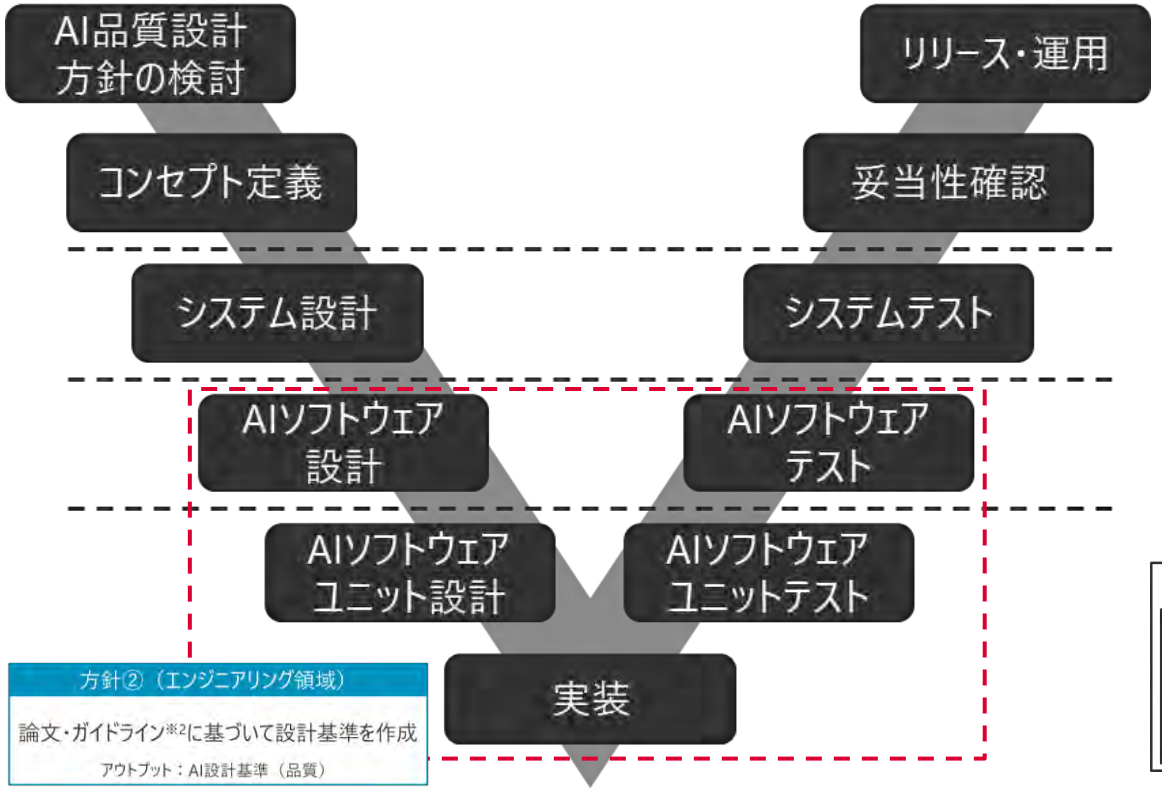
方針①（マネジメント領域）

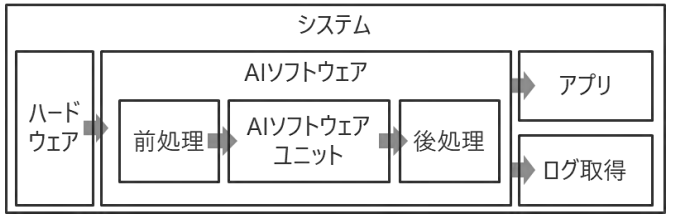
AI要求事項に基づいて管理規程を作成

アウトプット：AI品質管理規程

AI品質管理規程では、**AI品質を担保する上で必要な活動**をプロセスとして定義している。

AI品質管理規程





参考：開発対象のシステム構成の例

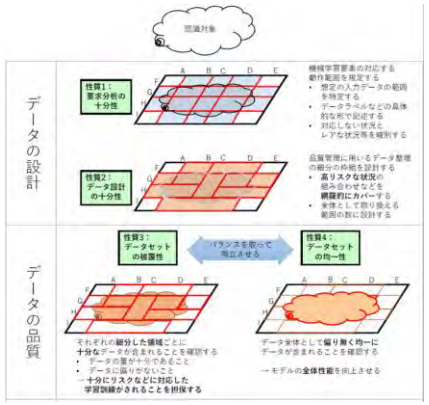
品質管理プロセスに関連するAI要求事項を取り込んだ規程を作成

方針②（エンジニアリング領域）に対する構築アプローチ

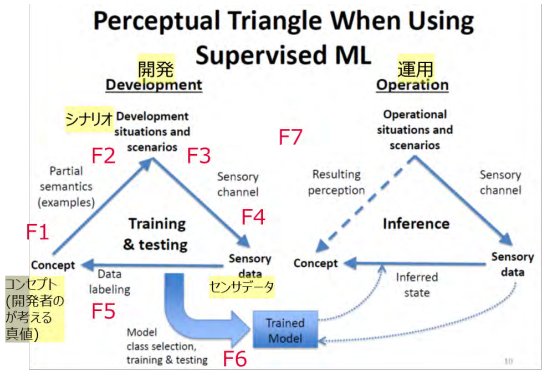
方針②（エンジニアリング領域）
論文・ガイドライン※1に基づいて設計基準を作成
アウトプット：AI設計基準（品質）

デンソーでは、車載製品／非車載製品・サービスの両方を扱っており、どちらの製品に対しても対応できるようにするため、
IT業界と自動車業界の両方の考えを取り入れる方針とした。

IT業界



自動車業界



F1:コンセプトの不確実性
(例)「歩行者」の定義がぶれる
F2:シナリオのカバレッジの不確実性
(例) 未知の「歩行者」のパターンが存在
F3:シーンの不確実性
(例) 遮蔽、霧、夜間、人の重なりなど
F4:センサ特性の不確実性
(例) ダイナミックレンジ、ノイズ、搭載位置など
F5:ラベルの不確実性
(例) クラスの間違い、BBOXのズレ・バラツキ
F6:モデルの不確実性
(例) パラメータの初期値、選定する最適化手法
F7:運用時の不確実性
(例) データ分布の変化、レアケース

K. Czarnecki and R. Salay. Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving. WAISE2018.

AI品質を担保する際の要点を抽出

AI品質を担保するための
基本的な考え方

各業界で言及されている内容に基づいて考えづくりを実施

AI設計基準（品質）の概要：4つのゆらぎ

方針②（エンジニアリング領域）

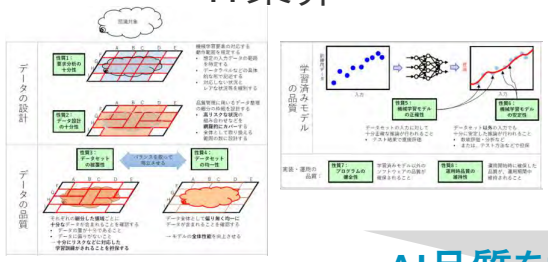
論文・ガイドライン※1に基づいて設計基準を作成

アウトプット：AI設計基準（品質）

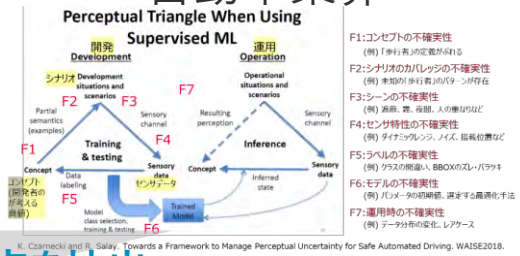
AI品質を担保する基本方針

AI開発プロセスの構成要素に対し、**品質低下や不確実性要因によってゆらぎが発生することを防止**する。

IT業界



自動車業界



AI品質を担保する際の要点を抽出

AI品質を担保するための基本的な考え方

	ゆらぎの種類	AI品質を担保するための基本的な考え方
データ収集時の考え方を抽出	観測のゆらぎ	AIの結果に影響するパターンを演繹的かつ帰納的に洗い出す
		パフォーマンスとリスク回避性各々に対するデータカバレッジ方針を示す
ラベリング時の考え方を抽出	ラベルのゆらぎ	ラベリング基準のばらつき、ラベリング作業品質のばらつきを抑える
モデル構築時の考え方を抽出	モデルのゆらぎ	適切なモデルを選定し、デファクト手法によりチューニングと評価を行う
運用時の考え方を抽出	運用のゆらぎ	開発時からの変化を検出・最適なモデルに更新する
	4つのゆらぎ	

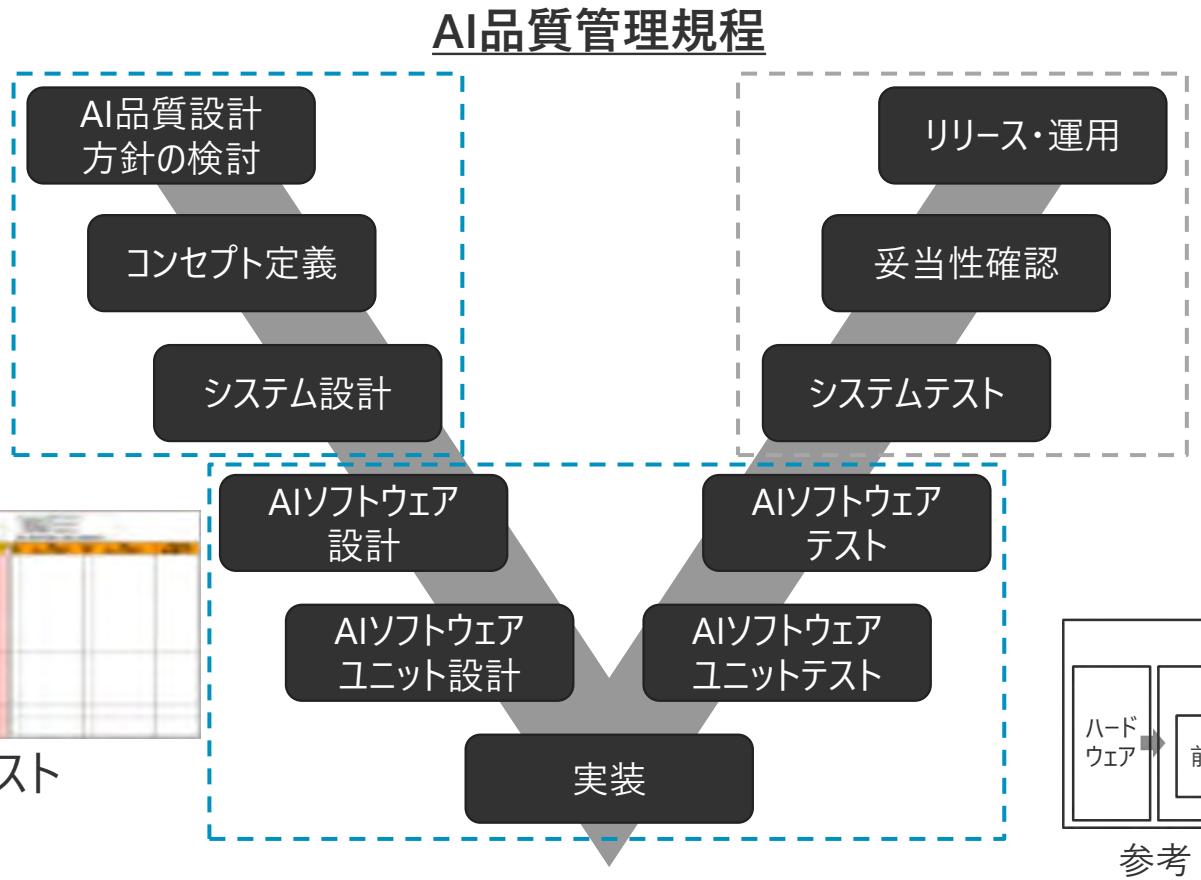
AIに含まれる"ゆらぎ"を抑えるための基本的な考え方に基づいてAI設計基準を定義

AI品質管理の実装

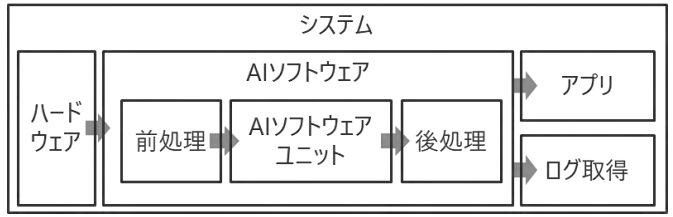
AI開発プロセスを評価するためのチェックシートも作成している。

システム設計
チェックシート

AIソフトウェア設計・テスト
チェックシート



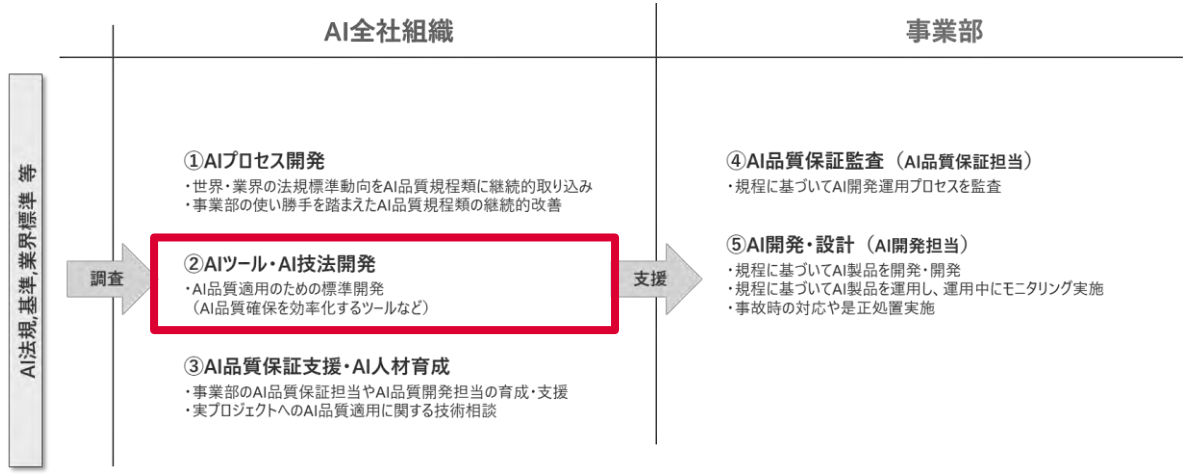
通常時の開発と同様に
テスト・運用を実施



参考：開発対象のシステム構成の例

開発の各フェーズにおいて、AI品質保証の実施状況を確認できる仕組みを構築

②AIツール・AI技法開発



AIリスクアセスメント手法　今回のスコープ

24年10月に品質月間テキスト「AIリスクアセスメント ガイドブック」[1]が出版された。

今回は、本ガイドブックのうち、以下の**赤枠部分**について紹介する。

章構成

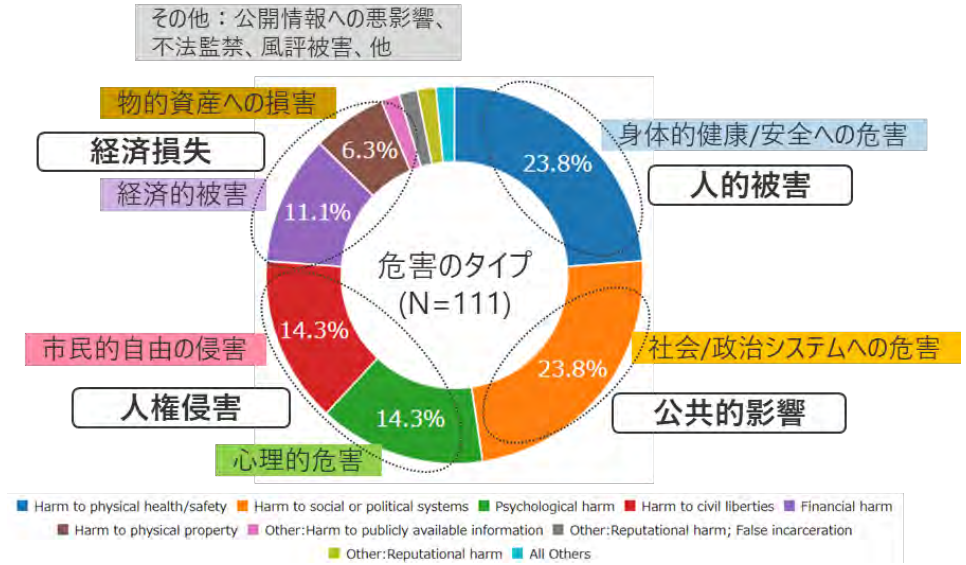
AIリスクアセスメント手法と適用事例が掲載されており、デンソーでも活用している

[1] https://webdesk.jsa.or.jp/books/W11M0100/index/?syohin_cd=6924476

AI品質適用のための標準開発　AIリスクの定義

AI起因のリスクの種類

Partnership on AI[1]が運営する"AI事故データベース(AIID[2])"には、2014年以降に世界で報道されたAI関連の事故情報が登録されている。この情報に基づいて、AIリスクが定義されている。



AI関連の事故情報（内訳）

AIリスク

リスク観点	定義
人的被害	• 死亡、障害、異音等の人に対する健康被害
経済損失	• 誤課金、誤請求等ユーザ資産への影響 • 機能やサービス停止、契約違反で生じる顧客事業への影響 • 契約違反、知的財産侵害などの損失
公共的影響	• 機能やサービス停止で生じる社会的影響、環境への影響
人権侵害	• 社会的な権利を享受する機会を奪ったり、社会的な評価を貶めたりするなど、個人の信用差別

AI関連の事故情報の調査結果に基づくと、AIリスクは4つのリスク観点に分類可能

[1] https://partnershiponai.org/
[2] https://incidentdatabase.ai/taxonomies/

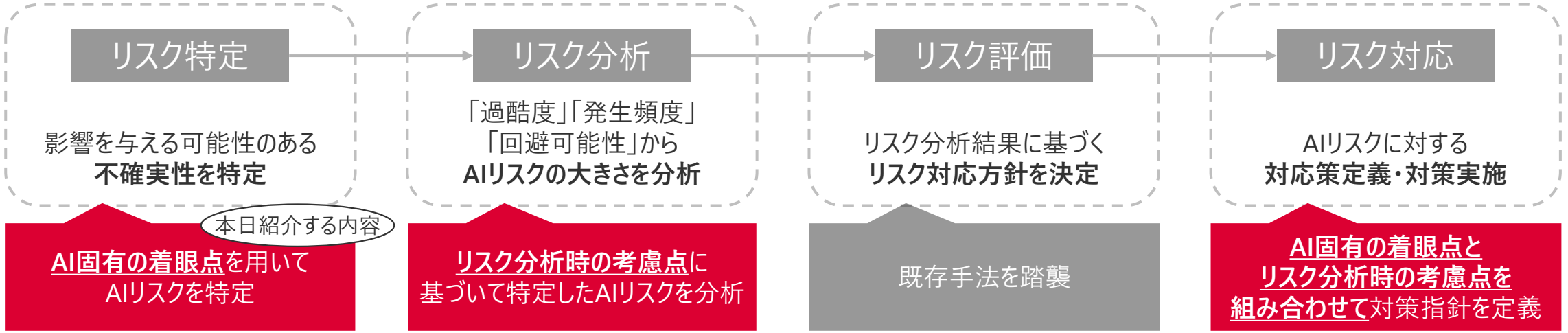
AI品質適用のための標準開発　AIリスクアセスメント手法

AIリスクアセスメント手法の構築方針

他のリスクアセスメント手法に**アドオンして利用**できるようにするため、ドメイン固有の内容が含まれていないかつ、他の多くのリスクアセスメント手法が参照しているISO31000に対し、**AI固有の観点を取り入れる**ことにより既存手法を拡張する。

AIリスクアセスメントの手順概要

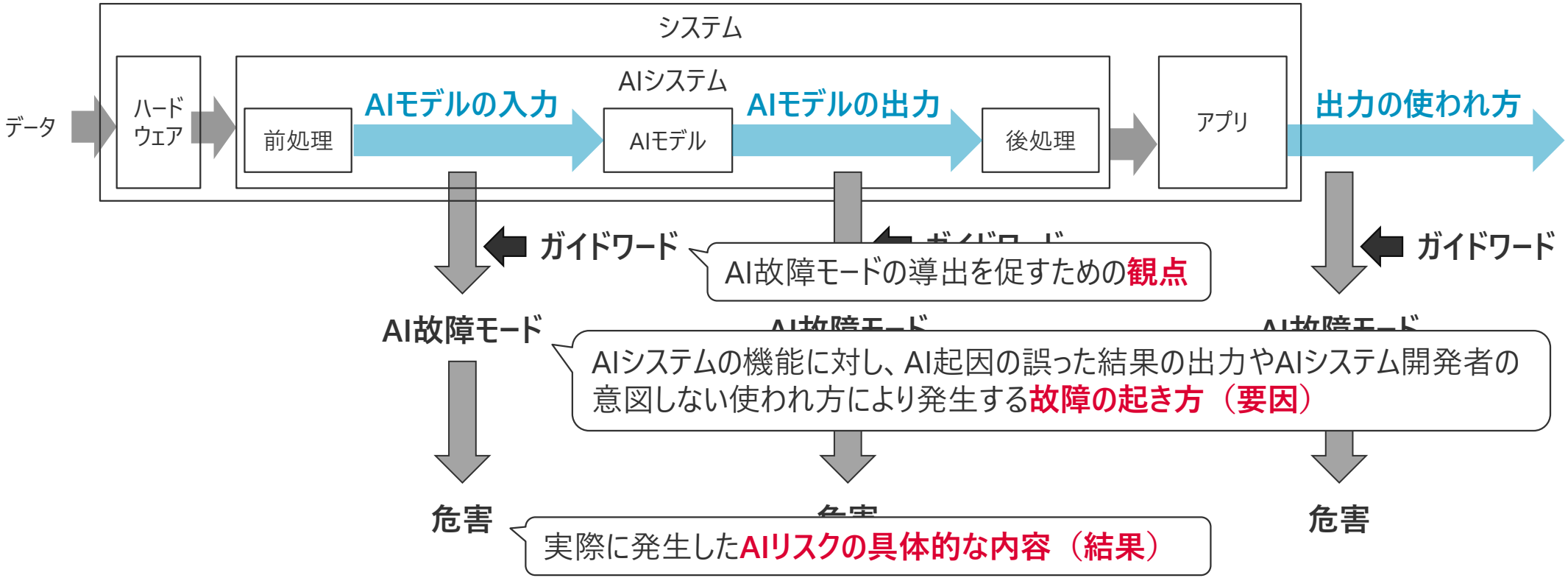
参考：ISO31000:2018(en)



AI固有の観点を取り入れたAIリスクアセスメント手法を利用

AIリスクアセスメント手法　リスク特定

AIリスクは「観測のゆらぎ」「ラベルのゆらぎ」「モデルのゆらぎ」および、EU AI Actの「ハイリスク用途」が影響して発生すると考え、本手法ではAIリスクの発生源を**AI固有の着眼点として定義**[1]し、着眼点に基づいてAIリスクを特定する。



AI固有の着眼点に基づいてリスクを特定

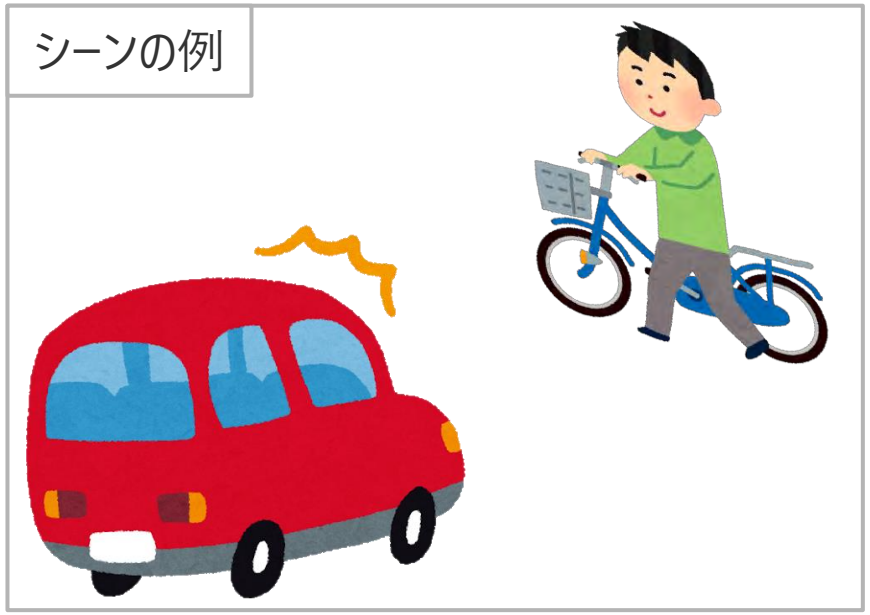
[1] https://jsqc.org/178sympo/

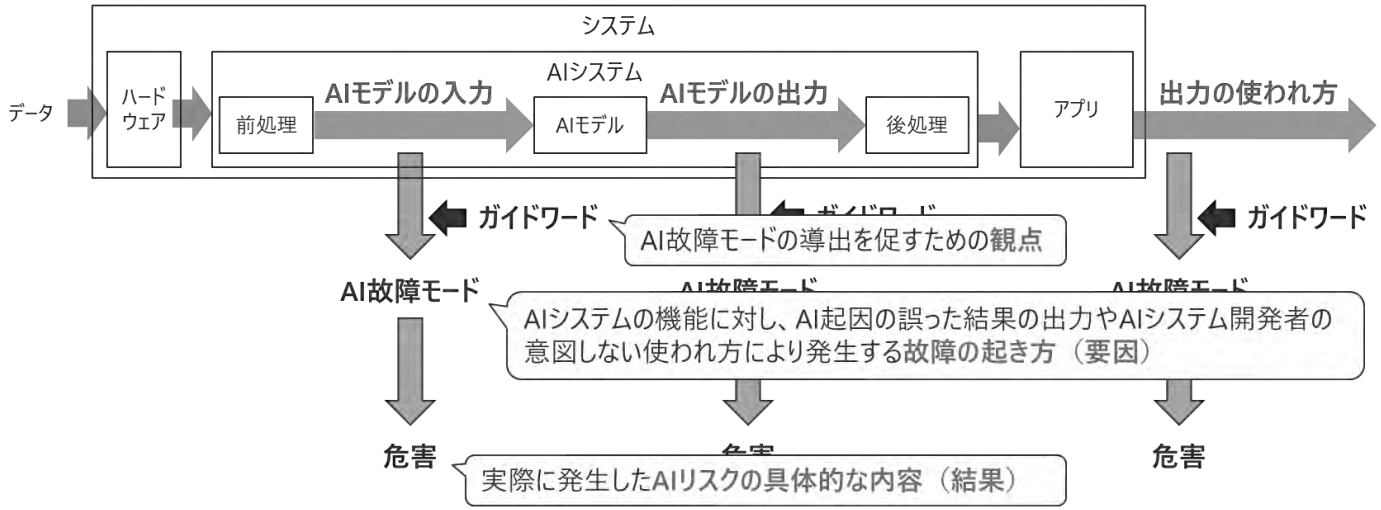
AIリスクアセスメント手法　リスク特定例

「AIを用いた歩行者認識機能が搭載された自動運転システム」におけるリスク特定の例

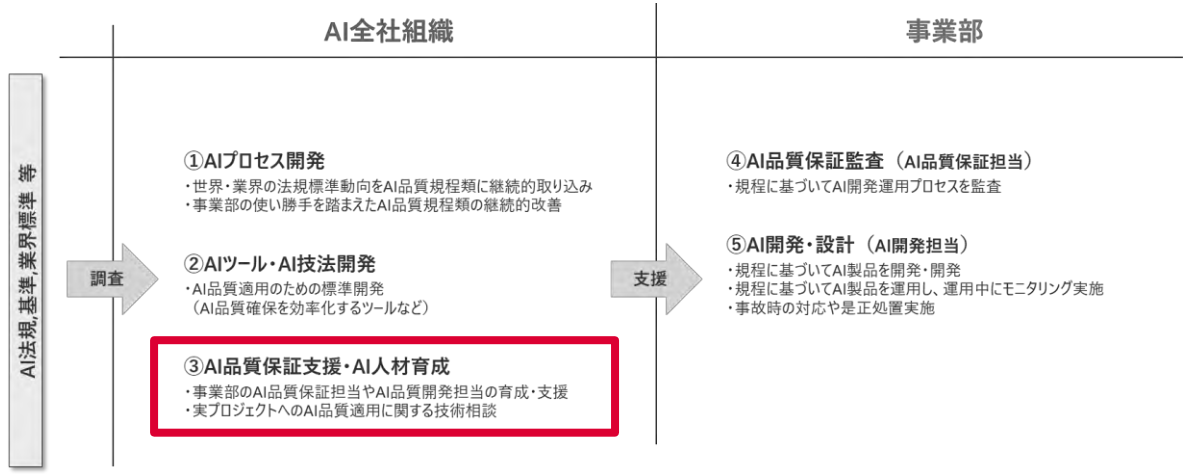
AI固有の着眼点	ガイドワード	AI故障モードの例	危害の例
AIモデルの入力	網羅性	**夜間の歩行者**を見落とす	【人的被害】歩行者を見落として接触することで、歩行者が負傷する。
	センシティブ情報	**センシティブ情報**を学習時に利用する	【人権侵害】AIによる判定にバイアスが生じ、特定の性別・人種の歩行者を見落として接触することで、歩行者が負傷する。
AIモデルの出力	例外処理	**自転車を押している歩行者**の分類が決まっていない	【人的被害】歩行者の認識／未認識の繰り返しにより、適切な制御ができず歩行者と接触し、歩行者が負傷する。
出力の使われ方	誤使用	歩行者を見落とした場合はドライバーが衝突回避する必要があることを、**ドライバーに認識させていない**。	【人的被害】歩行者の見落としに対してドライバーが衝突回避をしなかったことにより、歩行者と接触することで、歩行者が負傷する。

シーンの例





③AI品質保証支援・AI人材育成

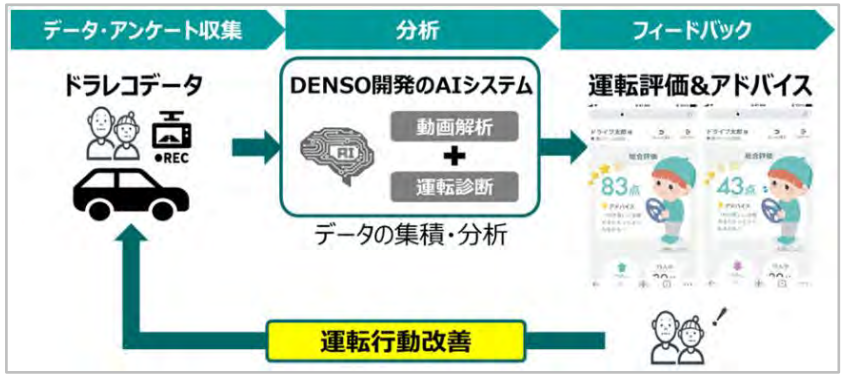


AI品質保証支援の実績

AI品質保証支援を実施している製品・サービスの例を以下に示す。

先進安全／自動運転[1]

農業用自動収穫機[2]



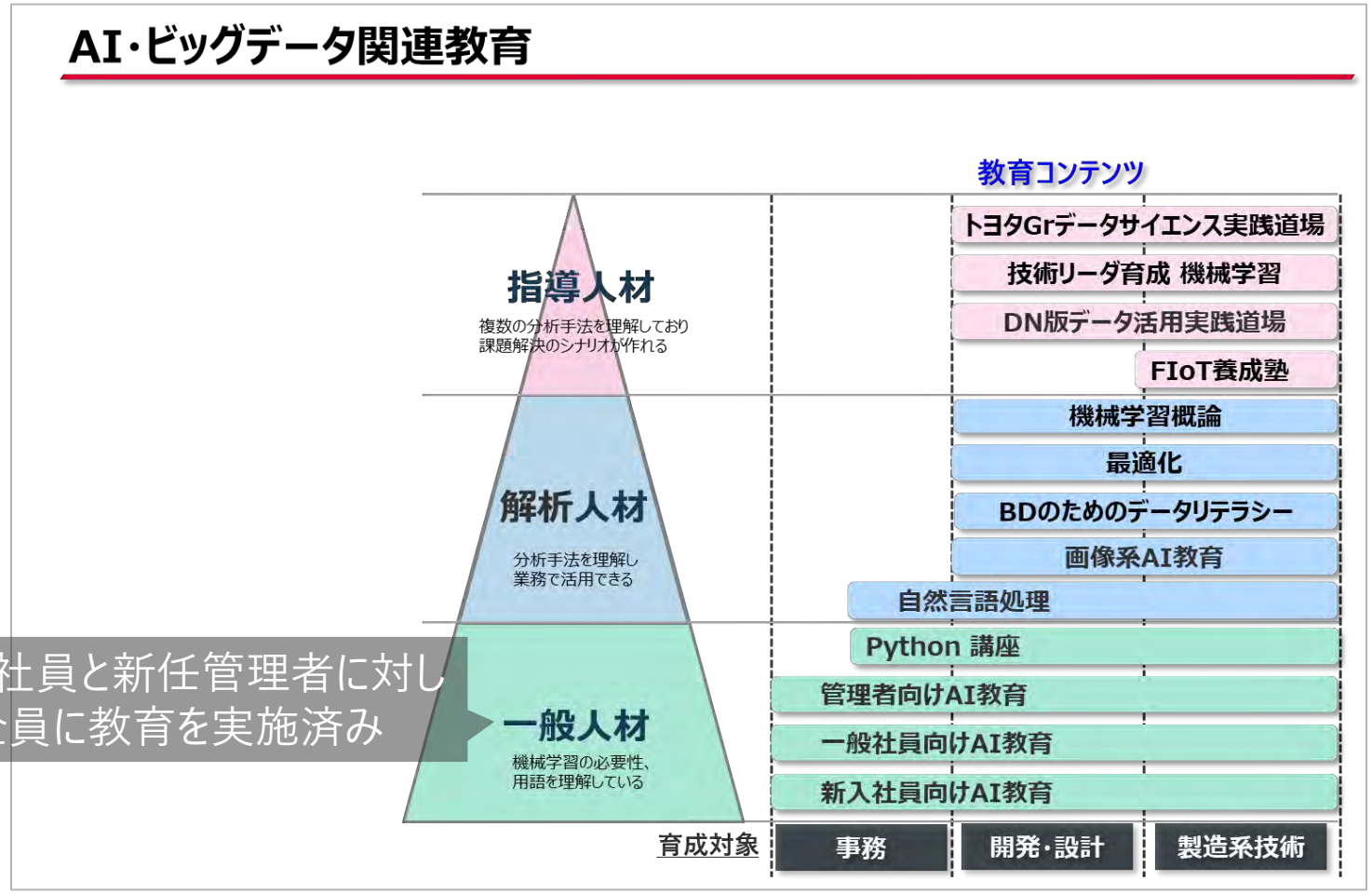
AI運転診断システム[3]

デンソーで取り扱っている製品・サービスに対してAI品質保証支援を実施

[1] https://www.denso.com/jp/ja/driven-base/tech-design/techlinks_ai_1/
[2] https://www.denso-agri.com/technology/
[3] https://global.toyota/jp/newsroom/corporate/37860847.html

AI人材育成に向けた取り組み

デンソーでは、AI人材育成に向けた教育体制の整備も進めている。



新入社員と新任管理者に対し
全員に教育を実施済み

AI関連の教育コンテンツをカリキュラムに盛り込み、関係者全員の基礎教育完了を目指す

3

まとめ

まとめ

世の中のAI動向とデンソーの取り組み

世の中全体でルール化の動きが高まっているため、デンソーでは社内標準づくりを進めてきた。

デンソー社内におけるAI品質保証の仕組みづくり

下記の3つを軸として、AI品質保証の仕組みづくりを進めている。

① AIプロセス開発
- 品質管理プロセスに関連するAI要求事項を取り込んだ規程を作成
- AIに含まれる"ゆらぎ"を抑えるための基本的な考え方に基づいてAI設計基準を定義
- 開発の各フェーズにおいて、AI品質保証の実施状況を確認できる仕組みを構築

② AIツール・AI技法開発
- AI固有の観点を取り入れたAIリスクアセスメント手法を利用

③ AI品質保証支援・AI人材育成
- デンソーで取り扱っている製品・サービスに対してAI品質保証支援を実施
- AI関連の教育コンテンツをカリキュラムに盛り込み、関係者全員の基礎教育完了を目指す

AIによるイノベーションの促進とともに、AI品質保証に対するデンソーとしての説明責任を果たす

Attached documents

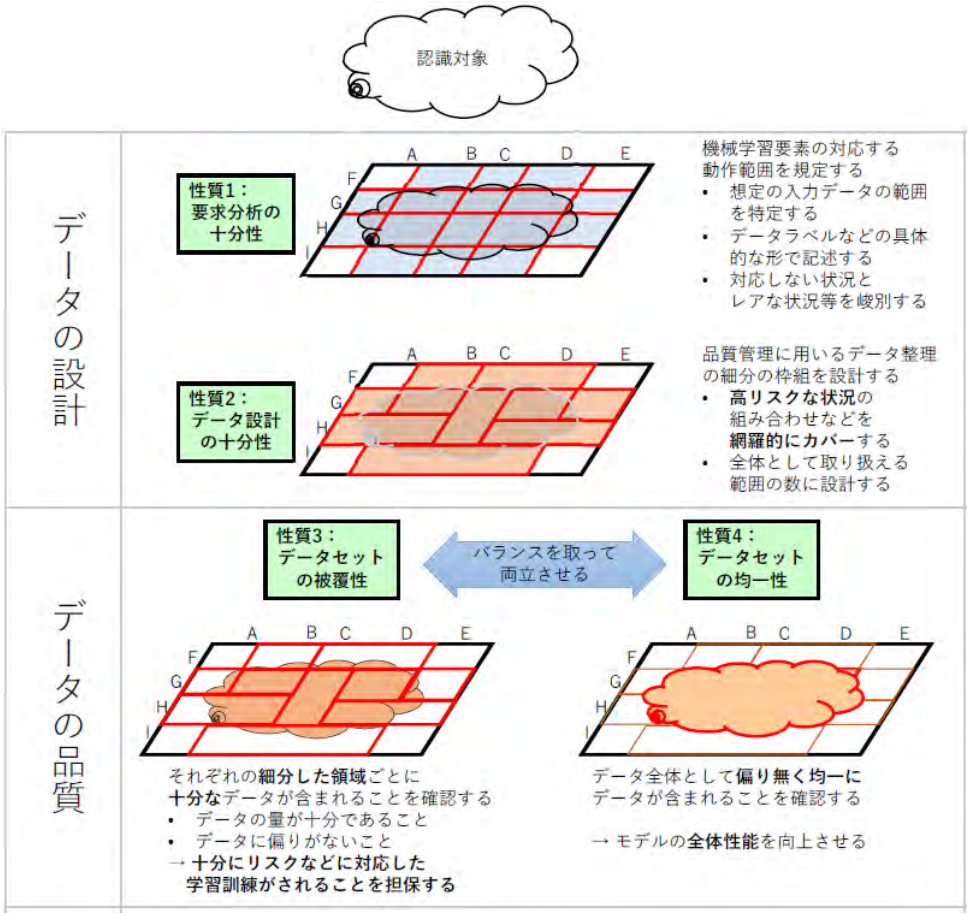
参照した論文・ガイドライン：IT業界

方針②（エンジニアリング領域）

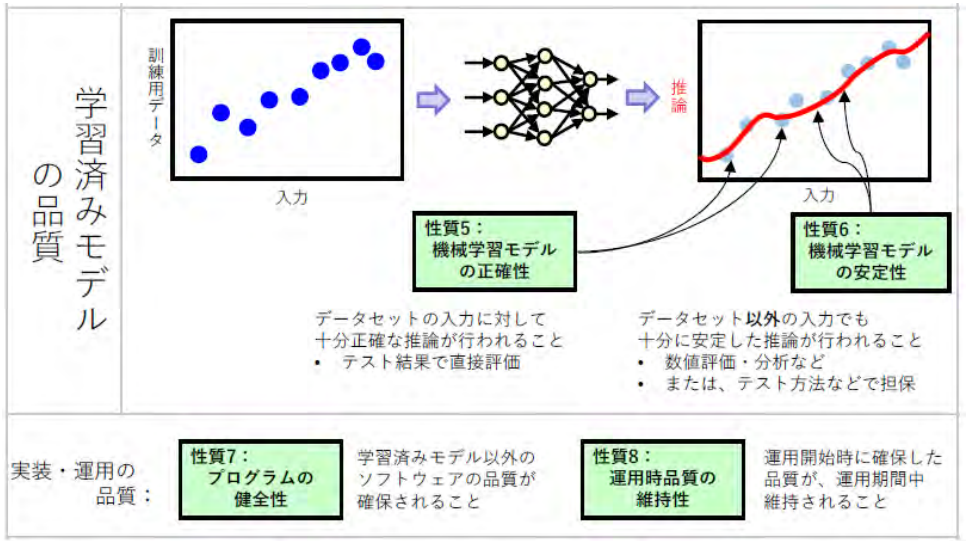
論文・ガイドライン※1に基づいて設計基準を作成

アウトプット：AI設計基準（品質）

機械学習で実装されたソフトウェアコンポーネント（機械学習要素）の内部品質を管理する際の特性軸について言及



出典：AIQM第1版　着目する内部品質特性



出典：AIQM第1版　着目する内部品質特性

AI開発における品質管理指針を定めたガイドラインを参照

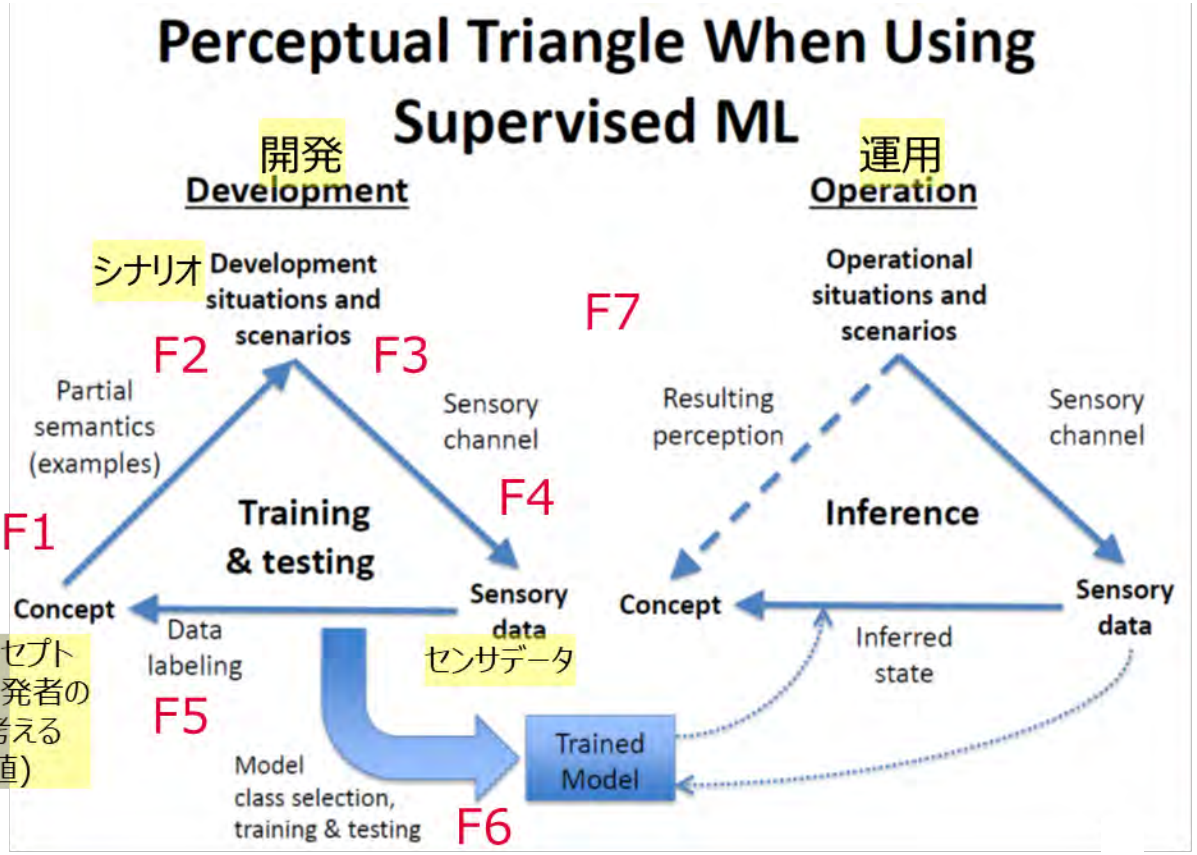
参照した論文・ガイドライン：自動車業界

方針②（エンジニアリング領域）

論文・ガイドライン※1に基づいて設計基準を作成

アウトプット：AI設計基準（品質）

教師あり学習を用いた機械学習モデルに入り込む7つの認識不確実性について言及



K. Czarnecki and R. Salay. Towards a Framework to Manage Perceptual Uncertainty for Safe Automated Driving. WAISE2018.

F1:コンセプトの不確実性

(例)「歩行者」の定義がぶれる

F2:シナリオのカバレッジの不確実性

(例) 未知の「歩行者」のパターンが存在

F3:シーンの不確実性

(例) 遮蔽、霧、夜間、人の重なりなど

F4:センサ特性の不確実性

(例) ダイナミックレンジ、ノイズ、搭載位置など

F5:ラベルの不確実性

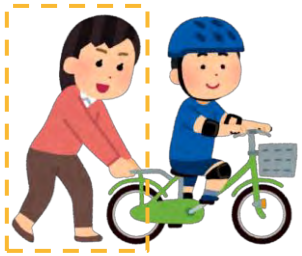
(例) クラスの間違い、BBOXのズレ・バラツキ

F6:モデルの不確実性

(例) パラメータの初期値、選定する最適化手法

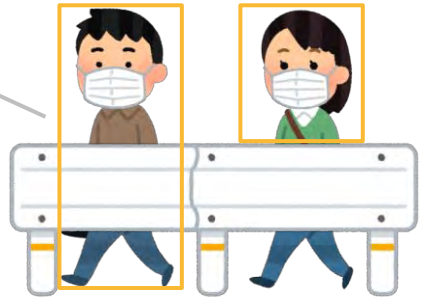
F7:運用時の不確実性

(例) データ分布の変化、レアケース



自転車を押している人

仮装している人



全身をラベリング　　上半身のみラベリング

自動運転プロセスの文脈でおさえるべき観点を網羅した論文を参照

AI設計基準（品質）の概要：調査結果の整理

方針②（エンジニアリング領域）

論文・ガイドライン※1に基づいて設計基準を作成

アウトプット：AI設計基準（品質）

AI品質に対して、**不確実性要因はネガティブ要因、品質特性はポジティブ要因の位置づけ**として捉えると、
不確実性要因と品質特性はそれぞれ以下のように整理できる。

不確実性要因と品質特性

類似する項目をまとめた上で考え方を構築

- F1：コンセプトの不確実性
- F2：シナリオカバレッジの不確実性
- F3：シーンの不確実性
- F4：センサ特性の不確実性
- 性質1：要求分析の十分性
- 性質2：データ設計の十分性
- 性質3：データセットの被覆性
- 性質4：データセットの均一性

→ データ収集時の考え方を抽出

- F5：ラベルの不確実性

→ ラベリング時の考え方を抽出

- F6：モデルの不確実性
- 性質5：機械学習モデルの正確性
- 性質6：機械学習モデルの安定性

→ モデル構築時の考え方を抽出

- F7：運用時の不確実性
- 性質8：運用時品質の維持性

→ 運用時の考え方を抽出

- 性質7：プログラムの健全性 → 通常のソフトウェアの品質保証活動にて対応

AI品質を担保するための基本的な考え方

ゆらぎの種類	基本的な考え方
観測のゆらぎ	AIの結果に影響するパターンを演繹的かつ帰納的に洗い出す
	パフォーマンスとリスク回避性各々に対するデータカバレッジ方針を示す
ラベルのゆらぎ	ラベリング基準のばらつき、ラベリング作業品質のばらつきを抑える
モデルのゆらぎ	適切なモデルを選定し、デファクト手法によりチューニングと評価を行う
運用のゆらぎ	開発時からの変化を検出・最適なモデルに更新する

4つのゆらぎ

DENSO

Crafting the Core

加速しよう、未来を。
VERISERVE

変化し続けるLLMモデルをプロダクトに組み込む際のテストの考え方

株式会社ベリサーブ
研究開発部
須原秀敏

2025.01.15

自己紹介

- 経歴
  - 2010年ベリサーブ入社
  - ～2014年：車載系のテストエンジニア
  - ～2020年：車載ドメインのテストの研究開発
  - ～現在：テスト・品質保証の研究開発＆ベトナムでの開発拠点立ち上げ～運営
- 現在のテーマ
  - モデルベースドテスト
  - 説明可能なテスト（リスクベースドテスト）
  - AIを使ったテスト・品質保証
  - AIシステムのテスト・品質保証
- 社外活動
  - JTC1/SC7/WG26エキスパート、JSTQB技術委員、QA4AIコンソーシアムメンバー

本日のテーマ

1. 変化するコンポーネント（LLM）を含むシステムをどう品質保証するか
2. 曖昧さを含む出力をするシステムをどう品質保証するか
3. ふるまいが確定的でないシステムをどうテストで網羅するか
4. 品質保証できるための説明可能性を持ったシステムをどう設計するか

本日のテーマをAIQMの内部品質特性にマッピング

1. 変化するコンポーネント（LLM）を含むシステムをどう品質保証するか
   E-1. 運用時品質の維持性
2. 曖昧さを含む出力をするシステムをどう品質保証するか
   C-1. 機械学習モデルの正確性
3. ふるまいが確定的でないシステムをどうテストで網羅するか
   C-2. 機械学習モデルの安定性
4. 品質保証できるための説明可能性を持ったシステムをどう設計するか
   説明性、透明性

各品質特性は「機械学習品質マネジメントガイドライン Rev. 4.2.0.0113」（AIQM）より引用

お伝えしたいこと

- **自動リグレッションテスト**の重要性は
  ますます高まってきている
- メタモルフィックテスティングは有用だが、
  銀の弾丸ではない
- 作り手やサービス提供者が安心できる
  システムの**作り方**が重要

本日のテーマ

1. 変化するコンポーネント（LLM）を含むシステムをどう品質保証するか
2. 曖昧さを含む出力をするシステムをどう品質保証するか
3. ふるまいが確定的でないシステムをどうテストで網羅するか
4. 品質保証できるための説明可能性を持ったシステムをどう設計するか

LLMを含むシステムの課題

A. LLMが日々進化しており、LLMの差し替えなどにより**想定しない変化**が起きる可能性がある

B. LLMは自然言語を扱っており、出力が**機械的に判断しにくい**

課題Aの具体例

モデル名	リリース日
gpt-4o (2024-05-13)	2024/5/13
gpt-4o-mini (2024-07-18)	2024/7/18
gpt-4o (2024-08-06)	2024/8/6
gpt-4o (2024-11-20)	2024/11/20

同じモデル（名）でもどんどん
異なるモデルがリリースされる
「運用時品質の維持」が課題

課題Bの具体例

全く同じ質問をしても、異なる回答が
返ってくることがある
「機械学習モデルの正確さ」が課題

導入している解決策

A. LLMが日々進化しており、LLMの差し替えなどにより**想定しない変化**が起きる可能性がある

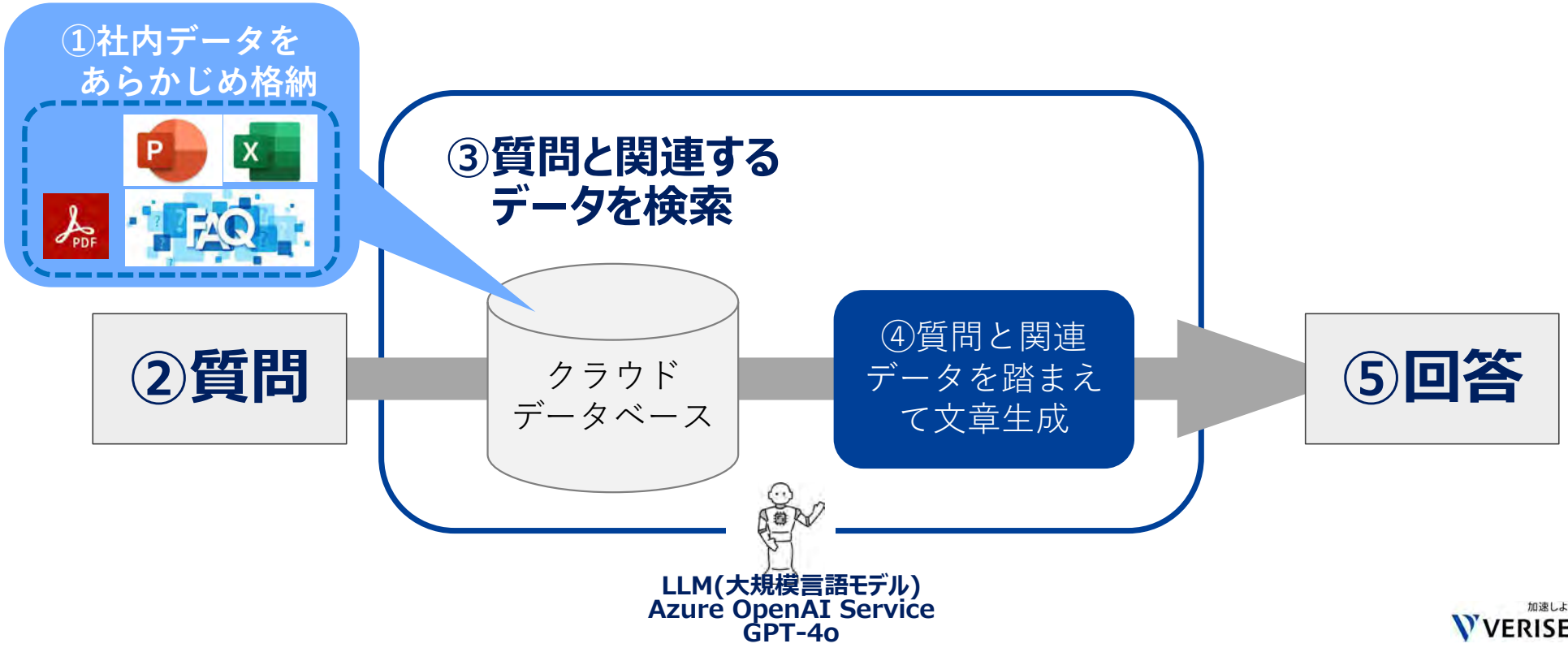
継続的リグレッションテスト

B. LLMは自然言語を扱っており、出力が**機械的に判断しにくい**

確率的な判定

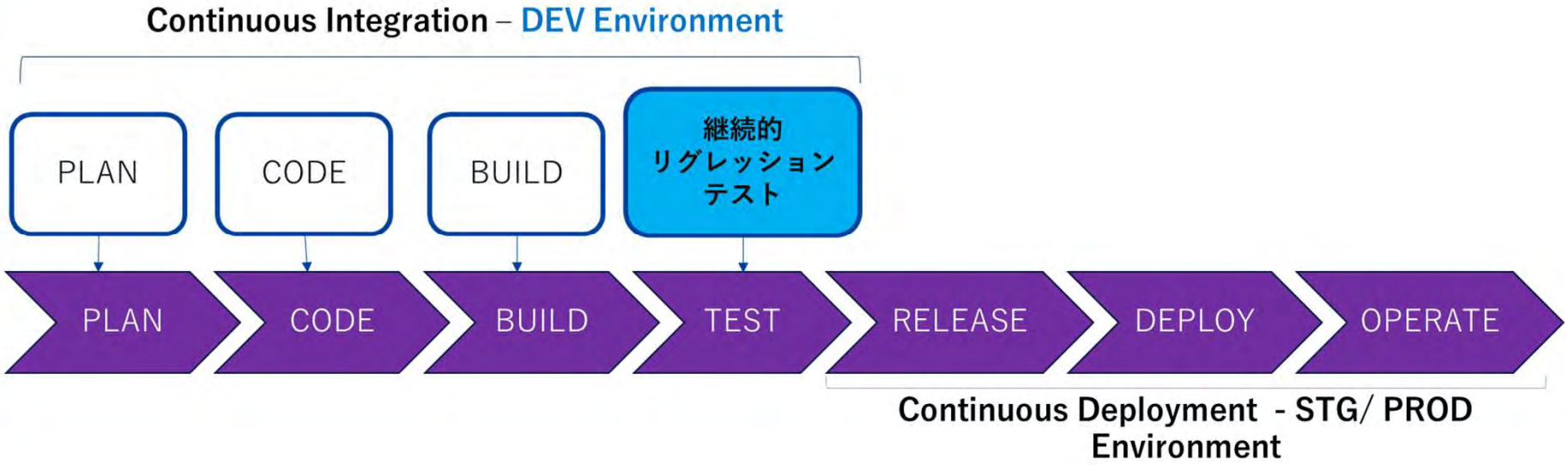
対象プロダクト

- RAG（Retrieval-augmented generation）を用いた社内手続きや規程について回答するチャットボット



解決策A: 継続的リグレッションテスト

- 想定しない変化に対して、継続的にチェックをする仕組み
  - 意図的なモデル変更時
    - 例えば、GPT-4oからGPT-4o-miniにモデルを差し替えた場合に期待と大きく外れた動きをするようになってしまっていないかを確認する
  - 意図しない定常時



解決策B: 確率的な判定

- 幅を持った判定をするために、自然言語用の判定を導入
  - Ragas：OSSのRAG評価フレームワーク
  - LangCheck：Citadel AIがリリースしたLLM評価ツール
  - オリジナルのテストケース
- 例えば表のような結果が出るため、実験結果から閾値を設定し、判定

		True Answer	Wrong Answer
RAGAS	Answer Semantic Similarity	0.96	0.82
	Answer Correctness	0.90	0.25
Langcheck	Reference_based_text_quality	0.72	0.16
	Reference_free_text_quality	0.6	0.3

つまり…

- 日々進化するコンポーネントをプロダクトに組み込むのは、リスクが低くない
- それに対し、これぐらいのテストに通っていればユーザーに継続的に価値を提供できるだろう、というラインを設定し、継続的リグレッションテストという形で具現化した
- 自然言語というハードルを確率的な閾値設定で除去した

本日のテーマ

1. 変化するコンポーネント（LLM）を含むシステムをどう品質保証するか
2. 曖昧さを含む出力をするシステムをどう品質保証するか
3. ふるまいが確定的でないシステムをどうテストで網羅するか
4. 品質保証できるための説明可能性を持ったシステムをどう設計するか

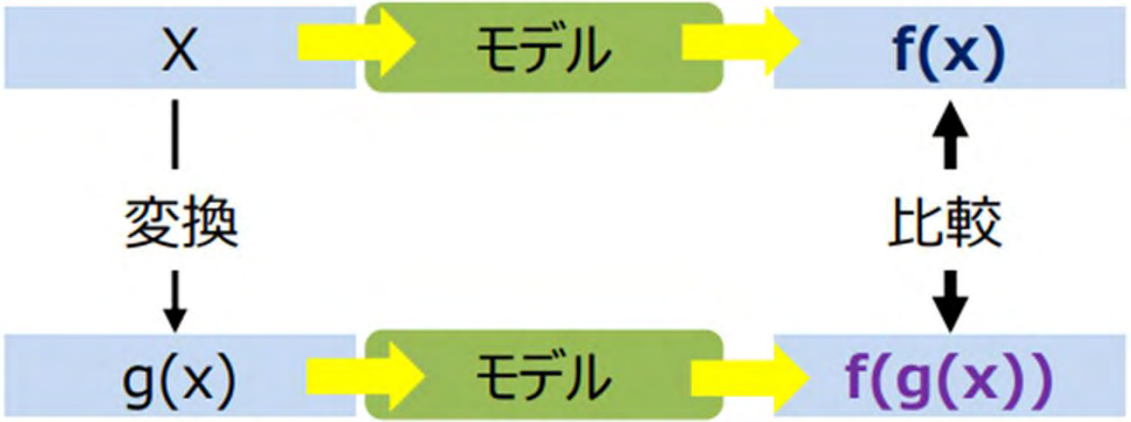
ふるまいが確定的でないシステムの課題

- プログラムでふるまいが決まるわけではないため、通常のプログラムに対するテスト技法（i.e. 同値分割）が使えない
  - 同値分割：正常範囲の入力を与えれば、同様のふるまいをするだろうという仮定に基づいたテスト技法

同値と扱うことなく、網羅的なテストが必要になる

メタモルフィックテスティング

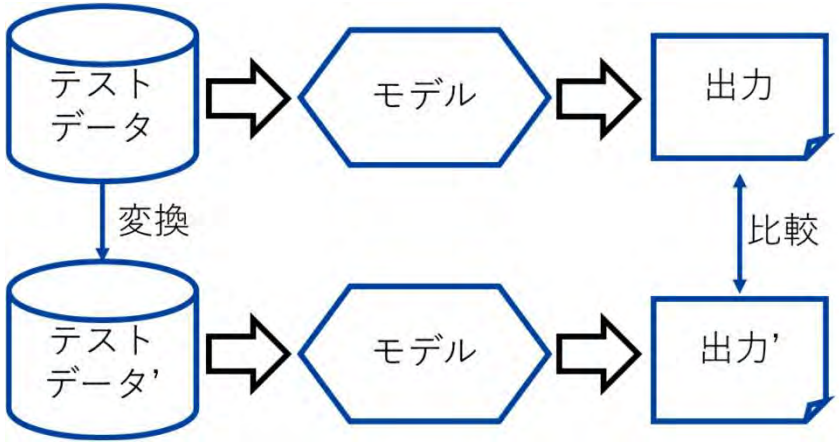
- オラクル問題の解決策といわれる
  - 「部分オラクルの代表的な方法にメタモルフィック・テスティング（MT）がある」（AIQM第4版より）
  - テストケースの増幅の一手法ととらえられる
  - 変換のロジックのことをメタモルフィック関係と呼ぶ



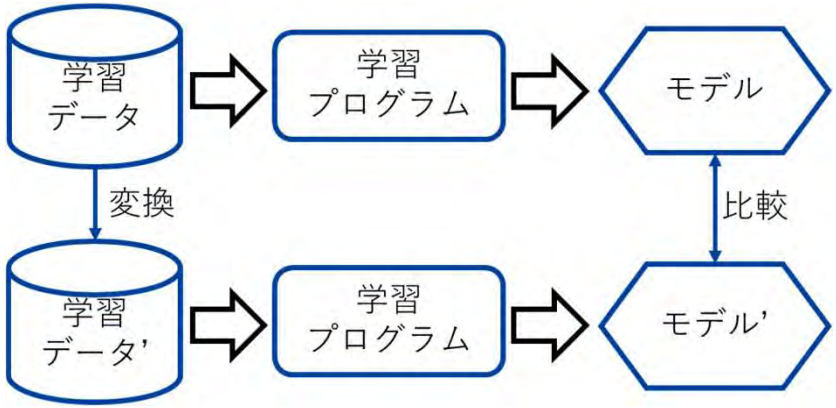
「データ品質を利用したメタモルフィックテスティングによる 機械学習・深層学習モデルの評価」より

メタモルフィックテスティングの使いどころ

A. モデルに対するテスト



B. 学習プログラムに対するテスト



つまり…

- メタモルフィックテスティング等の技法により、オラクル問題を解決しつつ比較的網羅的なテストを作ることは可能
- 一方で、作れるテストは「部分」であるため、メタモルフィックテスティングのみですべてが解決するわけではないことに注意
- メタモルフィックテスティングの使いどころは、モデルそのもののテストと学習プログラムのテストの両面がある

本日のテーマ

1. 変化するコンポーネント（LLM）を含むシステムをどう品質保証するか
2. 曖昧さを含む出力をするシステムをどう品質保証するか
3. ふるまいが確定的でないシステムをどうテストで網羅するか
4. 品質保証できるための説明可能性を持ったシステムをどう設計するか

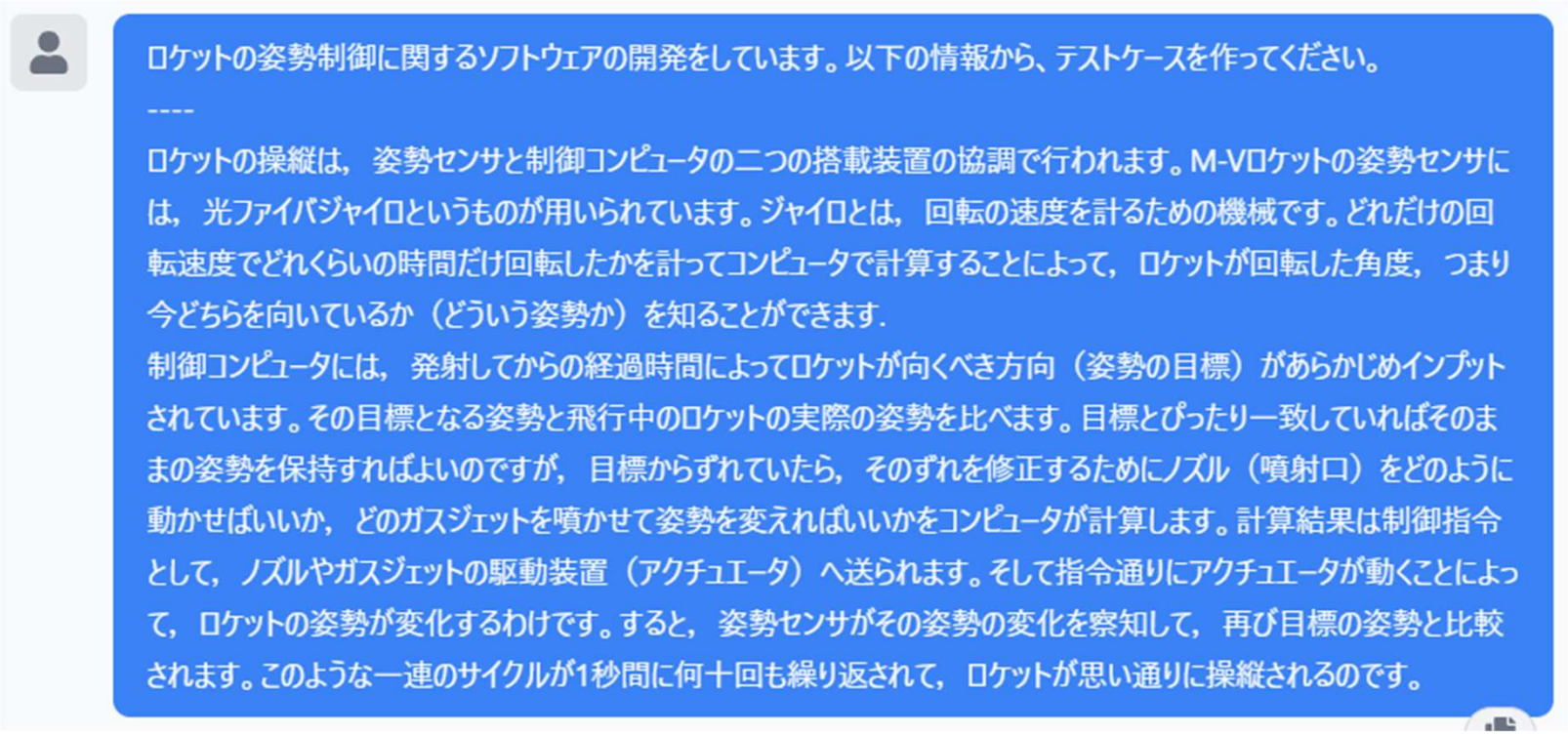
ブラックボックスなAIを組み込んだシステムの課題

- 動作の内容が説明可能である、という説明性の欠如
- **いくつかの処理段階が必要なタスク**をAIに任せることにより、なぜそういう結論になったかが**説明できない**

当社のAI活用の本丸

「良いテスト（ケース）をAIに作らせたい」

→よし、仕様書をAIに食わせてテストを作らせよう！



仕様情報は https://www.isas.jaxa.jp/j/column/electric/05.shtml より引用

当社のAI活用の本丸（つづき）

何かしらは作ってくれるが、
なぜそのテストケースでよいのかの説明ができない

→説明するための技術として、旧来から研究しているモデルベースドテストを活用

テストケース 1: 静止状態からの初期テスト

- 目的: ロケットが静止状態から目標姿勢に向けて適切に動作を開始できるかを確認します。
- 入力:
  - 初期姿勢データ: すべての軸で角速度が0
  - 目標姿勢データ: 一定の角度 (例: ピッチ20度)
- 期待結果:
  - 制御コンピュータが適切なノズルとガスジェット指令を生成し、最小時間で目標姿勢に達する。

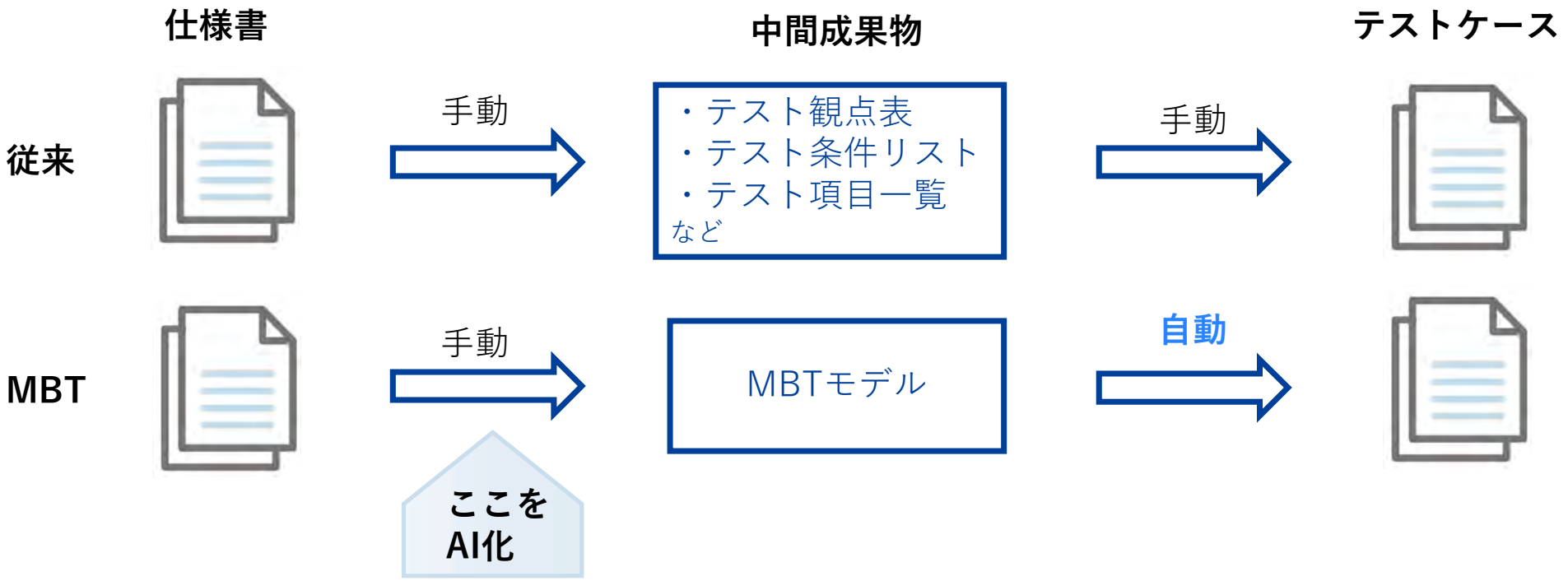
テストケース 2: 継続的偏差の補正

- 目的: 飛行中に生じる姿勢の偏差が継続的に補正されるかを確認します。
- 入力:
  - 姿勢データ: ゆっくりとしたランダムノイズ付き変動
  - 目標姿勢データ: 固定 (例: ヨー10度)
- 期待結果:
  - 制御システムがランダムノイズに対してロケットの姿勢を安定的に保つ。

テストケース 3: センサーデータ誤差の影響確認

- 目的: センサーデータの誤差が制御にどのような影響を与えるかを確認します。
- 入力:

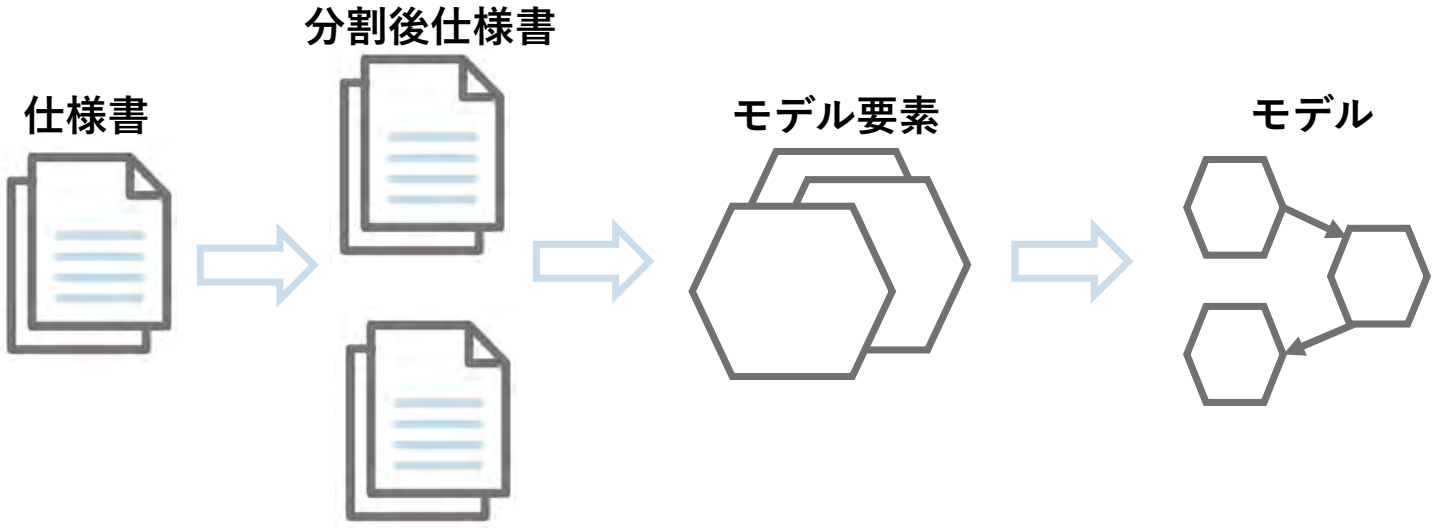
モデルベースドテスト（MBT）とは



MBTのモデルをAIに作らせることで、説明性の維持と、AIの制御可能性の維持を試みている

さらに細分化

- 対象の仕様の選定、モデルの要素の抽出などのプロセスに分割



分割後のプロセスの一部をAI（＋人）に任せることで、
システムとしての説明性、制御可能性を上げている

つまり…

- 全部AIに任せると、何が起きているかわからないが、プロセスを区切って途中で人（もしくはプログラム）の目を入れると、説明性、制御可能性が上がる

まとめ

- **自動リグレッションテスト**の重要性は
  ますます高まってきている
- メタモルフィックテスティングは有用だが、
  銀の弾丸ではない
- 作り手やサービス提供者が安心できる
  システムの**作り方**が重要

参考文献

- Ragas：https://docs.ragas.io/en/latest/index.html
- LangCheck：https://docs.ragas.io/en/latest/index.html
- 機械学習品質マネジメントガイドライン Rev. 4.2.0.0113

加速しよう、未来を。

VERISERVE

E.O.F.

24-007

HQW!
Hello, Quality World!

イノベーションを加速させる
知恵と品質技術にアクセスする
テクノロジーライフメディア

www.veriserve.co.jp/helloqualityworld/