自動車業界における AI Safety 動向*

Trends in AI Safety in the Automotive Industry

中江 俊博 Toshihiro NAKAE
桑島 洋 Hiroshi KUWAJIMA

This paper provides an overview of the trends in the safety of automotive products equipped with AI technology, with a focus on deep learning. Specifically, it examines the evolution of AI in automotive products, the types and roles of AI in automated driving, and the challenges and technical approaches to ensuring safety. Furthermore, it also provides a detailed account of the research community on AI safety, as well as the regulatory and standardization trends in the cross-disciplinary field and the automotive industry.

Key words :
artificial intelligence, machine learning, safety, automated driving, regulation

1. はじめに

深層学習の登場による機械学習技術の進化に伴い，機械学習ベースの車載製品が開発され，社会に浸透してきた．一方で，Tesla や Uber の事故により自動運転システムの安全性に注目が集まり，一般社会の関心事になっている．自動車に関する法規，標準化において，安全性を担保するためのルール作りが進められている．

本稿では，機械学習モデルを搭載した自動運転システムの開発・運用プロセスに焦点を当て，車載製品における AI の位置付け，自動運転の安全性の考え方，AI の安全性につながる技術とプロセス，融合分野として萌芽する AI セーフティの研究コミュニティ動向，および安全性を確保するための法規・標準化の動向について概説する．

なお，本稿では人工知能（AI），機械学習（ML），深層学習（DL）の使い分けについて，データから学習する手法（学習ベース）の特性が論点となる場合は ML または DL，学習ベースかどうかは明示的に区別せず自律的な挙動を行うシステムの特性が論点となる場合は AI と呼ぶこととする．

2. 車載製品における AI の位置付け

車載製品には，エンジン，エアコン，メーターから，近年標準搭載され始めている先進運転支援システムまで様々存在する．データから周囲の環境や車・ドライバの状態を認識し，車の安全性や快適性を高めるニーズは高まっている．以下では，自動車業界における AI の安全性を議論する上での前提として，AI 技術の進化が車載製品にどのような影響を与え，現在どのような AI が車載製品に使われているかを述べる．

Fig. 1 AI applications in vehicles. The bold frames indicate AI that is equipped in market products

2.1 AI 技術の進化

車載製品では，DL の登場以前から局所特徴量と ML を組み合わせた手法が，歩行者や信号など特定の物体を認識するセンサの基盤技術として取り入れられてきた [Yurtsever 20]．従来の特徴量は，認識対象の外見的な特徴を抽出する画像処理アルゴリズムを人が設計していたため，認識対象や環境条件を予め設計に組み入れておく必要があった．しかし，DL の登場によって，高い汎化性を持つ特徴量を実際の画像から後天的に獲得できるようになり，日照変化や遮蔽などの複雑な外乱要因に対する頑健性が飛躍的に高まった．さらに，画像分類 [Krizhevsky 12]，物体検出 [Redmon 16]，セマンティックセグメンテーション [Badrinarayanan 15] など，様々な DL ベースの画像認識モデルが開発され，応用範囲が広がっていった．今や，DL の技術は車載製品に欠かせない存在となった [桑島 22]．

2.2 AI 技術の進化

では，AI の車載応用にはどのようなものがあるのだろうか．欧州自動車部品工業会（CLEPA）がまとめた一覧を Fig. 1 に示す[1][CLEPA 20]．データラベリングの効率化など開発段階で使われる機能やインフォテイメントなど人命リスクを伴わない機能を除くと，衝突被害軽減ブレーキ（AEBS）やアダプティブクルーズコントロール（ACC）で使われる認知機能，およびドライバステータスモニタ（DSM）のドライバ状態推定機能が，実際の製品に搭載されている AI の代表格である．認知機能やドライバ状態推定機能は，走る・曲がる・止まるといった自動車の走行機能の一部や，ドライバの居眠りや脇見を監視する機能の一部を担っていることから，セーフティクリティカルな車載 AI といえる．

認知機能やドライバ状態推定機能には，画像認識タスク用の教師あり学習モデルが実装されている．認知機能では，歩行者，車両，標識などの認識対象を定義したクラスとその位置を特定する DL ベースの物体検出モデル，走行可能な領域を特定するセマンティックセグメンテーション，ドライバ状態推定機能では，ドライバの視線の推定などが一例として挙げられる．以降，本稿で扱う AI 機能は，認知機能の画像認識タスクに用いられる教師あり学習モデル（以下認識 ML モデル）を想定する．

2.3 AI 技術の進化

自動運転レベルは，運転タスクの主体や運行設計領域（ODD：Operational Design Domain）などによっ

*本稿は，人工知能学会の了解を得て，「自動車業界におけるAIセーフティ動向，人工知能，Vol. 38, No. 2, pp. 210-220（2023）」より一部加筆して転載

1 [CLEPA 20] に掲載された表を著者らが和訳，一部改変．

て，レベル 0 からレベル 5 の 6 段階に分類されている．レベル 1 ～ 2 は，先進運転支援システム（ADAS：Advanced Driver Assistance System）と呼ばれ，ドライバが運転主体となる．レベル 3 以上は，自動運転システム（ADS：Automated Driving System）と呼ばれ，レベルに応じた範囲でシステムが運転主体となる[2]．本稿では，ML モデルの搭載がデファクトになっている自動運転レベル 2 以上を想定する．また，以降 ADAS と ADS を特に区別せず，自動運転システムと呼ぶ．

自動運転システムは，ML だけでなく，様々な技術の組合せによって成り立っている．例えば，認知機能においては，Visual SLAM などの自己位置推定 [Bresson 17]，物体追跡，カメラ・LiDAR・レーダなど複数のセンサから得らえた情報を統合するセンサーフュージョンを駆使し，システム全体で認識性能を高めている．

自動運転システムは，車載特有の条件下で高い性能が要求される．走行環境における様々な環境条件や認識対象のバリエーションに対して，稀にしか起こらないケース（エッジケース）を含めて対応しなければならない．また，温度変化や衝撃への耐久性，低消費電力性，リアルタイム性が求められるため，クラウド環境で画像認識を行う大規模 ML モデルをそのまま車載に適用するのは難しい．枝刈り・圧縮でリサイズした ML モデルを SoC（システムオンチップ）に統合し[3]，認識性能だけでなく上記の諸条件を評価する必要がある．

3. 自動運転における安全性

2018 年米国 Arizona 州 Tempe で発生した Uber の事故 [NTSB 18] では，自動運転システムの設計において横断歩道外の歩行者を考慮しておらず，歩行者の発見が遅れたことが事故原因の一つとされた．物体検知で，車道には歩行者はいないという想定のもと，歩行者以外のクラス（車両，自転車，その他）が頻繁に切り替わり，システムが迷い続けていた．また，トラッキング（物体追跡）の履歴も受け継がれていなかった．

Uber の事故から得られる示唆は，実世界の環境でどのようなシナリオを想定していたか，ML コンポーネントが認識に失敗したときシステムがどのような挙動を示すかが，自動運転システムの認知機能における重要な要件だということである．

従来から，自動車の安全規格には機能安全（ISO 26262）[ISO 18] が存在していた．機能安全は，ランダムハードウェア故障や，ソフトウェアのバグのような特定の状況で必ず発生するシステマティック故障に対する安全性を規定している．しかし，Uber の事故のようなケースは，意図した機能が故障によって動かなかったのではなく，そもそもそのようなシナリオを想定して機能が作られていなかった機能的不十分性によるものと考えられる[4]．そこで，機能安全を補完する規格として策定されたのが ISO 21448（SOTIF：Safety of the intented functionality）[ISO 22a] である．

3.1 ISO 21448 (SOTIF)

SOTIF は，2019 年に PAS（公開仕様書）が発行され，2022 年に IS（国際規格）として正式発効された新しい規格である．SOTIF が対象としているハザード（危害の潜在的な源）要因は，機能の不十分性とミスユースである．機能の不十分性とは，センサ，アルゴリズム，アクチュエータの仕様や性能の不十分性を指す．ミスユースとは，合理的に予見可能な，ドライバの誤った使用や過信を指す．

危険な振る舞い（センサ認識機能が目標物を見失うなど）の引き金となるような外乱要因（豪雨，逆光などの天候，カメラの汚れなどのセンサノイズ，など）をトリガ条件（triggering conditions）と呼ぶ．トリガ条件が車両レベルの危険な振る舞いまたはドライバのミスユースを引き起こし，最終的に危害に至るまでの関係を示したのが Fig. 2 である．

2　本稿では一般の読者に合わせて平易な言い回しを用いた．厳密な定義は [SAE 16]（和訳 [自動車技術会 22]）を参照いただきたい．

3　以下，訓練済み ML の重みパラメータを ML モデルと呼び，前処理，ML モデル，および後処理を SoC に統合したシステムを ML コンポーネントと呼び，それぞれ区別する．

4　[Salay 17] らは，ML モデルを実装した機能が ISO 26262 に準拠するためには，仕様の欠如と解釈可能性の欠如が大きな障害であることを示し，新たな規格の必要性を示唆した．

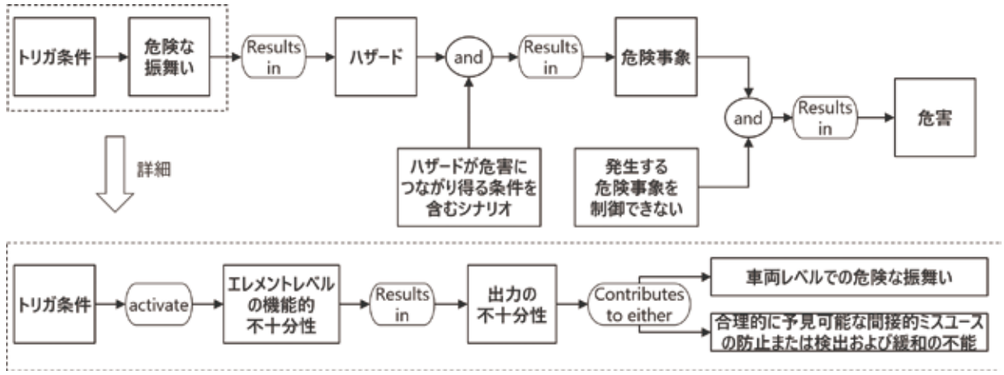

Fig. 2　SOTIF-related hazardous event model (above), SOTIF cause and effect model (below)

未知のシナリオを十分小さくし，既知の危険なシナリオを危険でないシナリオに変える SOTIF のリスク低減活動の考え方に基づき（Fig. 3），ODD や過去不具合事例から演繹的に導出できるシナリオから，車を走らせなければ発見が難しい帰納的なシナリオまで，仕様の不十分性および性能の不十分性を反復的に分析することで，危害につながるトリガ条件を洗い出す．この分析は，システムレベルだけでなく，エレメントレベル（ML コンポーネントレベル）にも適用される．

SOTIF の付録 D.2 は，ML の章になっており，本文を補完している．ML ソフトウェアの機能（e.g. 物体認識の性能）は SOTIF で扱われること（D.2.1），意図した機能を特定せずに ML 技術を利用したシステムは安全とはみなされないこと，開発者はトリガ条件を上位システムと共有し対策を講じること，ML コンポーネントのテストには，ML アルゴリズム，ML コンポーネント，車両の 3 つのレベルでテストすること（以上 D.2.3）などが述べられている．

3.2 ISO PAS 8800

ISO PAS 8800 は，2024 年の発行を目指している，AI の安全性に関する公開仕様書である．AI の安全属性，プロセス，SOTIF や機能安全と関係などが議論されている．本稿のテーマに最も関連の深い規格であるが，2023 年 12 月現在未発行のため，詳細について言及は避ける．

自動車業界では，従来の機能安全でカバーできなかった自動運転システムの機能的不十分性およびミスユースをカバーする SOTIF が作られた．AI の安全も大枠は SOTIF に含まれていくことを踏まえ，次に技術的な観点で AI の安全性について概説する．

Fig.3　SOTIF activities

4. AI の安全へのアプローチ・技術・プロセス

4.1 AI の安全上の課題

自動運転の認知機能の役割は，実世界の認識対象をセマンティックな記号（歩行者クラスなど）に変換し，後段の判断機能に伝達することである[5]．判断機能は，記号化された情報に基づいて状況を判断するため，記号化された情報の信頼性は非常に重要である．しかし，教師あり学習モデルは，視覚的特徴と記号の相関関係のパターンを学習しているだけなので，人間のように意味的な文脈を考慮することができない．ゆえに，通常は，非学習的な手法より優れた性能を発揮するが，ときに直感に反した結果を出力する場合がある．また，認識精度を高めるために ML モデルのパラメータが巨大になるほど，認識結果に至る過程を理解することが

5　直感的に認識するシステムは System1，論理的に思考するシステムは System2 と呼ばれる．

困難になるというトレードオフが生じる．

自動運転システムに使われる ML をスコープとして，安全上の工学的課題を俯瞰した研究がいくつか発表されている．[Czarnecki 18] らは，自動運転の認識 ML モデルのライフサイクルにおいて，自動運転の安全性に負の影響を与える 7 つの不確実性要因を管理するフレームワークを提案した．[Willers 20] らは，DL 手法に焦点を当てた 9 つの安全上の懸念と関連する 10 の緩和アプローチを整理した．[Mohseni 20] らは，自動運転システムにおける ML の安全性のための実用的な安全技術をレビューし整理した．[Kuwajima 20] らは，自動運転の ML システム開発における工学的課題を示した．

4.2 安全上の課題へのアプローチ

認知機能に使われる ML の信頼性への懸念および安全上の課題の先行研究を踏まえると，認識結果の信頼性をいかにして高めかつ測定できるようにするかが安全上の課題の根本といえる．認識結果として出力された物体クラス，座標位置，向きなどの情報をどこまで信用してよいか，どこまで揺らぎがあるのかが分かれば，その信頼度に基づいて認識結果を利用したり，安全方策に活用できる．

[Czarnecki 18] を基に不確実性の発生源を 4 つに集約し，上記の先行研究や開発現場を参考に，不確実性により生じる揺らぎを工学的なアプローチで抑える方法を著者らがまとめたものが Table 1 である．ML のライフサイクルの中で，不確実性による揺らぎを測り，管理し，低減するプロセスは，安全性論証[6]において重要になると考える．また，SOTIF が規定する演繹的・帰納的なトリガ条件の分析と基本的な考え方は整合している（詳細は SOTIF・AI 間の関係が PAS 8800 で規定されるのを待つ必要はあるが）．以下，観測データ，ラベル，モデル，運用ごとに，鍵となる技術・プロセスを述べる．

6 根拠資料と構造化された議論によって安全性を示すことを安全性論証と呼び，セーフティクリティカルなシステムの安全性の説明責任を果たすための活動である．

Table 1 Uncertainty and its management in the ML lifecycle

発生源	不確実性要因	不確実性による揺らぎの抑え方
観測データ	コンセプト，シーン，センサ，シナリオカバレッジ	ML の結果に影響するパターンを演繹的または機能的に洗い出す性能とリスク回避性に対するデータカバレッジ方針を示す．
ラベル	ラベリング基準，作業	ラベリング基準のばらつき，作業品質のばらつきを抑える．
モデル	アーキテクチャ，ハイパーパラメータ	上位要求を設計に反映し，デファクトの手法で最適化し評価する．
運用環境	分布シフト，レアケース	開発時からの変化を検出し最適なモデルに更新する．

4.3 観測データ

§1 均一性と被覆性

AIQM ガイドライン [AIQM 23] は，「AI パフォーマンス」「リスク回避性」の 2 つの利用時品質を定義している．AI パフォーマンスは，Accuracy など一般的に ML モデル評価に用いられる全体的な性能を表している．リスク回避性は，ML の危険な振る舞いを低減しているか，といった安全性に関係する品質特性である．問題領域（自動運転文脈でいうと ODD）に対して，データ全体として偏りなく均一にデータが含まれている均一性と，細分化した領域ごとに十分なデータが含まれている被覆性のバランスを取って両立させることを同ガイドラインは提案している．自動運転システムの ML 開発では，ODD に基づく天候などの統計的な分布が均一性に，危険なシナリオにつながるトリガ条件が被覆性に関係する．

§2 能動学習

データ選定において中核技術となるのがデータの不確実性評価による能動学習である．自動運転システムの開発では，あらゆる環境条件を網羅することは不可能であり，また収集した膨大なデータ全てに人手でラベリングをすることもまた非現実的であるため，何らかの方法で訓練データやテストデータを注意深く選定する必要がある．

不確実性が高いデータが訓練ネットワークにとって重要な情報を伝える性質を利用し，膨大なデータの中から不確実性に基づいて能動的にデータ選定する手法は能動学習という名称で知られている．DNN による予測の不確実性を MC Dropout と呼ばれるモンテカルロ（MC）法で評価する手法 [Gal 17] がある．

能動学習は，既に Tesla，Waymo など先端企業で量産開発に用いられており，自動運転開発において必要不可欠な技術になりつつある [Kargar 22]．

4.4 ラベル

§1 ラベリングガイドライン

教師あり学習の場合，通常手作業で行われるラベリングの品質は訓練結果やテスト結果に直接影響を与える．ラベリングの品質が十分でない場合，テスト結果が誤解を招き，不十分な認識性能なままフィールドで使われるおそれがある．しかし，Fig. 4 に示すように，実世界には人でも解釈に迷うシーンが存在し，ラベルの矛盾，付け間違いが発生しうる．

そこで，ラベリング基準の一貫性とラベリング作業のばらつきの低減が重要になる．ラベリングガイドライン [Willers 20] の策定によって一貫性を確保し，作業者の教育，Q&A 管理，事例の蓄積，ラベリング結果の評価等，ラベリング作業プロセスの管理によって作業のばらつきを低減する．

Fig.4 Labeling variability for the same class: examples from the BDD100K Dataset [Yu 18]

4.5 モデル

§1 ML テスト

ML テストの体系は，[Zhang 22] らによる，ML テストにおける技術，テストプロパティ，アプリケーション等のサーベイが参考になる．自動運転システムの ML 開発では，一般的に，物体認識・物体検出・セマンティックセグメンテーションなどの ML タスクに応じた指標で認識性能を評価している．また，顧客要求やトリガ条件に対する動作を確認するため，シナリオベースのテストが行われている．一方で，ニューロンカバレッジへの批判 [Abrecht 20, Pavlitskaya 22] があるように，研究が活発な ML テスト分野では研究段階の未成熟な手法も見られるため，効果・妥当性をよく確認して採用可否を検討する必要がある．

§2 頑健性

頑健性（Robustness）は ML 研究の中でも特に盛んな分野の一つであるが，[CLEPA 20] が指摘するように安全性を議論する上で定義を明確にする必要がある．自動運転の文脈においては，雨・雪によるシーンの変化やセンサノイズなどの自然摂動（Natural perturbation）と，道路標識に特殊なステッカーを張るなど攻撃者が意図的に作り出した敵対的摂動（Adversarial perturbation）に分けられる [Mohseni 20, Willers 20]．

自然摂動は，3.1 節や 4.3 節で述べた問題に帰着し，4.5 節 §1 で述べたシナリオベースのテストで評価できる．それに対し，ML モデルの開発者が意図しない出力をさせるような人工的ノイズ [Szegedy 14] を加え，攻撃的な目的で作られるのが敵対的摂動である．自動運転の文脈においては，道路標識に特殊なステッカーを貼るなど，物理世界の中でも有効な敵対的摂動により ML モデルを騙す手法が提案されている [Eykholt 18]．敵対的攻撃の再現性の問題を指摘する研究もあり [Lu 17]，攻撃の有効性の評価は定まっていない．自動車業界は，敵対的摂動は自然摂動と区別し，将来の AI セキュリティの議論に委ねている [Mohseni 20]．

§3 説明性

説明可能性（Explainability, XAI）技術は提示内容（重要特徴／重要データ）[原 20]，5 つの軸（透明性／事後性，大域性／局所性など）[亀谷 22] などの観点で体系化されているが，自動運転の文脈では，[Zablocki 22]

らが，自動運転システムの画像認識向け XAI のニーズを，エンドユーザに対するトラスト，開発者によるデバッグ，規制当局に対する説明責任の3つに分類している．このうち，デバッグ目的では，ML モデルの認識結果の背後にある原因を探るため Grad-GAM[Selvaraju 17] などの顕著性マップ（Saliency map）が使われている一方，エンドユーザ，規制当局向けには，説明結果の正確性や網羅性の面で課題があるため，自動運転文脈における XAI 技術は総じて限定的な利用に留まっている．

4.6 運用環境

§1 不確実性評価

能動学習に使われる不確実性評価手法は，運用時に分布シフトや開発時に発見できなかったエッジケースを検出するモニタリングにおいても中核技術となる．実車走行中に不確実性が高いシーンかどうか判定できれば，ML の認識結果の信頼性が低いと見なし，ML モデルの外側の安全機構につなげられる可能性がある．MC 法はサンプリングによる計算コストが高く車載用途で使えないという欠点であったが，不確実性をサンプリングではなく線形近似で評価することで計算コストを削減した手法 [Postels 19]，さらに近似精度を高めた手法 [Mae 21] など，車載に適した手法が提案されている．

5. 研究コミュニティ動向

AI セーフティは，その名の通り AI と安全が融合するときの技術課題や社会課題を扱う領域である．AI セーフティの研究開発のアプローチは，AI と安全どちらの視点を起点にするかによって，大きく 2 つに大別される．一つは AI を起点とするアプローチである．安全性に影響するとみられる ML の特性（例えば頑健性）に対し，AI/ML・データサイエンスの視点で原理を明らかにしたり解決手法を提案する．もう一つは安全を起点とするアプローチである．ML モデル・システムの開発・運用で生じる問題に対し，自動車・航空・宇宙などセーフティクリティカルなシステムを対象に培ってきた安全工学・ソフトウェア工学の手法・考え方を適用・拡張する．

Table 2　International and domestic conferences on AI safety (* denotes domestic)

研究軸	会議名
AI/ML	AISafety@IJCAI, SafeAI@AAAI, MLSafety@NeurIPS, SAIAD@ECCV, 安全性とセキュリティ研究会@JSAI*
安全工学	WAISE@SAFECOMP, AI/IoT システム安全性シンポジウム*
ソフト工学	SEMLA@Polytechnique Montréal, {WAIN, DeepTest}@ICSE, MLSE@日本ソフトウェア科学会*

5.1 国際会議・国内会議

AI セーフティ研究は，本会議に併催されるワークショップを中心に形成されている（Table 2）．母体となる本会議の研究領域は，AI/ML，安全工学，ソフトウェア工学に大別され，前述したように，AI セーフティの研究テーマは，研究母体の技術・考え方を軸にして融合分野である AI セーフティへと派生している．以下，著者らの主観により代表的なものを紹介する．

AI/ML 系では，AISafety，SafeAI において，敵対的攻撃に対する頑健性，強化学習における価値整合（Value Alignment）や報酬ハッキング（Reward Hacking），分布シフト（Distributional Shift）などの研究が提案されている．SAIAD は画像認識応用，特に自動運転システムの認知機能における研究課題がスコープである．また，人工知能学会の安全性とセキュリティ研究会は海外の AI セーフティ学会群に対応した国内初の研究会である．安全工学系では WAISE が知られ，AI システムの安全規格や安全性論証，AI システムの安全設計・リスク評価，検証・妥当性確認，AI における不確実性など，安全の考え方や検証技術を AI に拡張する研究が見られる．ソフトウェア工学系では，機械学習工学研究会（MLSE）が，2018 年設立当初から ML とソフト工学を融合した新しい研究領域を目指している点で特徴的である．MLSE は，ML の説明性，デバッグ，開発プロセスなど，幅広い領域をカバーしている．

5.2 研究プロジェクト

AI の安全性に関する主要研究プロジェクトを紹介する．VDA（ドイツ自動車工業会）が主導する AI 関連プロジェクト群 KI Familie の一つ KI Absicherung [VDA]（訳：AI セーフガード）は，自動運転の AI の安全性に関する広範な研究を行っている．英 AAIP [York] はロボットや自律システムの保証に関する研究，日 eAI [JST 未来] は自動運転や医療などの領域を対象に高信頼な AI を効率よく実現する研究を行っている．

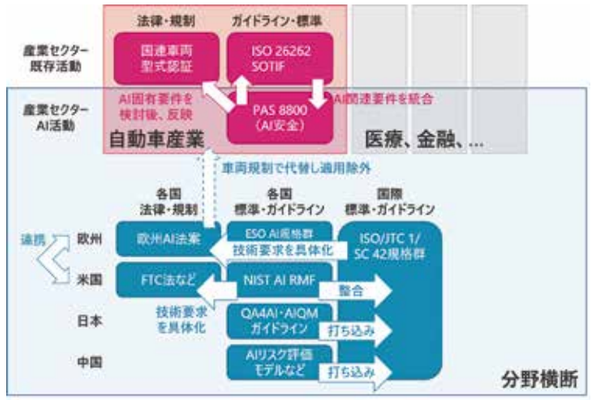

Fig. 5　The overview of regulations and international standards in intersectoral fields and the automotive industry

5.3 知識体系

研究コミュニティを中心に，AI と安全の融合分野の知識体系（BoK：Body of Knowledge）を構築する試みが行われてきた．代表例として，欧州 AI 関連プロジェクト [TAILOR] で参照されている AISafety Landscape [Espinoza 19]，安全な自律システム開発のための技術体系 [AAIP 19]，価値整合した AI を目指す姿において技術・ガバナンス要素に分解したマップ [FLI 18]，強化学習ベースの AI システムを主なスコープとして仕様・頑健性・保証の 3 領域で定義した AI セーフティ技術体系 [Ortega 18] などが挙げられる．また，従来の BoK との差分要素を分析した研究として，ML 特性と SQuaRE 品質特性とのギャップ分析 [Kuwajima 19] などがある．

5.4 注目すべき研究領域

データの実用課題に注目した近年の研究動向を紹介する．オープン環境におけるデータ分布の経時変化やロングテールに由来する課題を対象とした研究が注目されている．データセットにないデータを識別するデータ分布外（OOD：Out of Distribution）検出 [Yang 21]，未知クラスの存在を前提としたオープンセット認識（OSR：Open Set Recognition）[Scheirer 13] がある．また，NeurIPS 2021 で Andrew Ng が提唱したデータ中心の AI [Ng 21] は，データセットを固定しモデルの認識性能を向上させるモデル中心の従来の視点を変え，開発現場が取り組んでいるデータ周りの “ 泥臭い ” 作業を概念化し，その重要性に光を当てた．今後の研究の進展が望まれる．

6. 法規・標準化動向

6.1 AI ガバナンス

AI のガバナンスの枠組みは，抽象度の高い順に，(1) 最終的に保護されるべき技術中立的なゴールである AI 原則，(2) 横断的で中間的なルールである，法的拘束力のある規制（法規またはハードロー）または法的拘束力のない国際標準やガイドライン（ソフトロー），(3) 個別分野にフォーカスしたルール，に分類される [7][経産省 21].

(1) において，信頼できる AI（Trustworthy AI）は，目指すべき AI 社会を表す概念として国際的に定着している．OECD AI 原則 [OECD 19] をはじめ，各国・企業の AI 原則にこの Trustworthy というキーワードは登場し始めている．また，安全性は，様々な AI 原則のほぼ全てに含まれており [中川 20]，守るべき重要

7 [経産省 21] で定義されている (4) モニタリング／エンフォースメントは，(2) に含むと解釈できるため，本稿では除外した．

な価値の一つとみなすことができる．

国際的な議論の中で，(2) は水平的な枠組み，(3) は垂直的の枠組みとも呼ばれるが，縦横の重なり領域をどのように扱うかについて，各産業セクターの法規・標準を主導する業界団体と，分野横断的な法規・標準を主導する政府・議会や標準化団体との間で議論の争点となる．産業セクター毎に目的やリスクが異なる AI システムに対し，どこまで共通ルールを定めるのか，産業セクターの既存法規・標準と競合する要件はないのか，などを地道に調整する活動が関係者で行われる．例えば，後述する欧州 AI 法案に対して，AI の定義の広さ，適合基準の曖昧さ，技術的実現性の困難さが産業界から指摘されている [EC 20a].

Fig. 5 に分野横断および自動車業界の法規・標準の関係を示した．分野横断の枠組みで AI のガバナンスに法的拘束力を持たせるかどうかは国によって異なっている．各国に共通するルールを束ねているのは法的拘束力を持たない国際標準 ISO/IEC JTC1/SC 42（以下 SC42）[ISO 17] である．

6.2 分野横断的な法規・標準（水平的な枠組み）

§1 日本

モノづくりを強みとしてきた日本は，世界に先駆けて AI の品質や安全性に注目してきた．2018 年 AI ソフトウェア工学戦略プロポーザル [CRDS 18]，2019 年 AI 戦略 [内閣府 19] を発信した．G20AI 原則の採択と国際標準との連携により国際社会との調和を図りつつ，法的拘束力のないガイドラインによって規範と技術革新のバランスを取るソフトローを採用している．

国内の著名なガイドラインとしては，QA4AI[QA4AI 22]，AIQM[AIQM 23]，SEAMS[SEAMS 20] が挙げられる．以下，車載 AI の安全性に関連する部分を述べる [8].

QA4AI ガイドラインは，ドメイン共通の 5 つの品質保証軸の定義と自動運転などドメイン別の拡張が特徴的である．自動運転パートでは，機械学習モデルの更新と，機械学習システム全体のアーキテクチャ再設計による二重ループによる，繰り返し開発プロセスを提示している．AIQM ガイドラインは，機械学習モデルのリスク回避性など品質目標とする 3 つの外部品質と，その達成手段として問題領域分析の十分性など 9 つの内部品質を定義している．また，SC42 国内委員会が AIQM ガイドラインと AI 一般機能安全規格 TR 5469 との整合を図る国際標準化活動を行っている．SEAMS ガイドラインは，AI システムの安全設計パターンとして，AI コンポーネント自体の安全性を示すパターンと，AI コンポーネント自体は非安全系とし外部に安全メカニズムを設けるパターンを定義している．

§2 欧州

高い人権意識を有する社会である欧州は，信頼できる AI を目指した法制度作りを世界的に主導している．2019 年 AI 倫理ガイドライン [EC 19]，2020 年欧州 AI 白書 [EC 20c] および信頼できる AI の評価リスト [EC 20b] の作成を経て，2021 年欧州委員会から世界初の分野横断的 AI 関連法規となる欧州 AI 法案（EU AI Act）[EC 21] が欧州議会に提案された．同法案は，許容できない AI，ハイリスク AI，透明性義務のある AI，最小またはノーリスク AI の 4 段階で AI システムのリスクをレベル分けし，レベルに応じた要求を割り当てるリスクベースのアプローチを採用している．ただし，ハイリスクであっても厳格な既存法規が存在する応用領域は，将来の既存法規改訂時に AI 要件を検討することを条件に，現時点では適用を除外される．この除外規定により，既存法規として車両型式認証が存在する自動車業界では，後述の車両型式認証への AI 要件の追加が重要になる [9].

欧州 AI 法案の規制は抽象度が高く，標準やその他の技術仕様を用いて，具体的な技術規制を制定すると法案に明記している．それを受け，欧州委員会は，ETSI，ISO，IEC，ITU-T，および IEEE の各 AI 標準と欧州 AI 法案の要求事項との関係を分析している [Nativi 21]．このため，AI 国際標準化の重要性が増している．

§3 米国

AI 規制に対する米国のスタンスは，技術標準策定と消費者保護である．

AI 分野の世界的リーダーであることを自認している米国は [White House 19]，イノベーションを阻害しないよう規制は最低限に留め [OMB 20]，信頼できる AI の管理標準を作ることでイノベーションの促進を後押ししている．NIST（National Institute of Standards and Technology, 標準技術研究所）は，2019 年に米国 AI 標準化計画 [NIST 19] を発表した [10]．NIST が 21 年作成開始し，23 年発行された AI Risk Management Framework（AI RMF）[NIST 23] は，AI リスクを技術属性，社会技術属性，信頼原則に 3 分類し，AI リスクを管理する方法を構築する．AI RMF の策定には複数の主要 IT 企業が参画している．

一方で，FTC（Federal Trade Commission, 連邦取引委員会）は，既存法の枠組みで，社会的信用差別をもたらす AI を規制し消費者を守る立場を取っている．(本稿では安全性にフォーカスするため公平性に関する詳細は脚注を参照されたい [11]).

§4 中国

ビッグデータと応用技術に強みを持つ中国は，2030 年までに基礎研究を含めた AI 関連の全分野で世界トップ水準を目指している．技術標準を策定しイノベーションの促進を重視する米国と，公平性等に対する社会的価値観は異なるものの，戦略は類似している．

2017 年に国務院が，次世代 AI 発展計画を発表し，AI 国際標準化など 7 つの国際 AI 戦略を明確にした．これを受け，20 年に国家標準化管理委員会等の政府機関が合同で，安全・倫理，革新汎用技術などの 8 標準分類で構成する AI 標準体系を定義した国家次世代人工知能標準体系建設指南を発表した．ほぼ同時に，AI の標準研究，標準作成，国際活動をミッションとする AI 分会（TC 28/SC 42）を設立した．AI 分会は，中国 AI 戦略が定義した標準体系に基づいて国内標準を開発すると共に，ISO/IEC JTC 1/SC 42 の中国国内委員会として機能する．また，民間でも AIIA（Artificial Intelligence Industry Alliance）などが ML の機能評価等の標準を策定している [桑島 22].

一方，法規の面では，データセキュリティ法にみられるように，安全保障の観点で作られた法規の影響力が大きいのが中国の特徴である．2021 年 9 月に施行されたデータセキュリティ法は，データの分類・等級付け保護，データの国外移転に対する厳格な審査等を定めている．市場からデータを収集する自動運転開発では，同法に対応したデータ管理が必要となる．

§5 国際標準

AI 一般の国際標準の多くは ISO/IEC JTC 1/SC 42（人工知能）[ISO 17] で作られている．特に自動車業界への影響が予想されるのが，TR 5469（機能安全と AI システム）[ISO 22c] と 22989（AI コンセプトと用語）[ISO 22b] である．22989 は AI 用語を統一し，利害関係者間のコミュニケーションを円滑にするために作成されている．TR 5469 は 2023 年 12 月現在ドラフト作成中であり，内容は公開されていない．

6.3 自動車業界の法規・標準（垂直的な枠組み）

§1 国連法規

自動運転の車両型式認証を定める国連欧州経済委員会（UNECE）自動車基準調和世界フォーラム（WP.29）自動運転分科会（GRVA）では，車載 AI 利用ガイダンス案 [GRVA 23] が作成中である．2023 年 12 月現在，AI の定義が記載されている．今後，欧州 AI 法案などで扱われている AI 要件を社会からの要請として捉え，自動車業界でどう対応するか GRVA で検討されているが，AI を直接対象とした法規の策定には現時点で慎重

8 本稿のテーマと外れるが，AI システム開発の委託企業と受託企業との間の共通プロトコルとしてガイドラインを活用しているのが日本の特徴と思われる．QA4AI や AIQM は AI の品質保証，経産省の AI 契約ガイドライン [経産省 19] は AI・データの契約に関するトラブルを防ぐ用途で使われている．

9 2022 年 11 月現在の法案を基にしている．欧州 AI 法案は欧州議会で審議中のため，将来内容が変わる可能性がある．

10 また，NIST は AI RMF 策定開始とほぼ同時期に，バイアス管理草案 [NIST 22a] と説明可能 AI の 4 原則草案 [Phillips 21] を矢継ぎ早に発表している．ゼロリスク（ゼロバイアス）は達成不能であることを前提に，継続的なバイアスの特定・理解・測定・管理・削減を行う実用的なバイアス管理を目指している．

11 FTC は，2021 年 AI の社会的信用差別リスクに対応するため，3 つの従来法 – FTC 法第 5 条，公正信用報告法（Fair Credit Reporting Act, FCRA），消費者信用機会均等法（Equal Credit Opportunity Act, ECOA）– による AI の規制を示唆するとともに，AI を開発する企業に対するガイドラインを公表した [FTC 21]．これらは，肌の色・出身国等に基づく人種的なバイアスのある AI の使用や，雇用・保険・その他給付を拒否することに AI が使用される場合に関わってくる．

な姿勢である．一方，以下で述べるように，自動車業界の AI に関する国際標準化は，すでに様々な取り組みがある．

§2 国際標準

自動車業界の国際標準は，ISO/TC 22（自動車）で作られる．自動車の安全性に関連し，かつ AI とのつながりがある TC 22 配下の国際標準規格の一覧を Table 3 にまとめた．SOTIF，PAS 8800 については 3 章で述べたので，それ以外の規格について以下紹介する．

ISO/TR 4804:2020（自動運転システムの安全とサイバーセキュリティ）[ISO 20] には付録 B「深層ニューラルネットワークを用いた自動運転システムの安全関連要素の実装」が添付されている．この付録 B では，DL モデルのライフサイクルを，要求定義，仕様開発，開発と評価，展開と監視と定義し，各工程の活動をまとめている．

Table 3 International Standards on Automotive Safety (as of December 2023)

名称		会議名	AI との関係
機能安全	ISO 26262	ハードウェア故障，ソフトウェアのバグ	AI の記載なし
SOTIF	ISO 21448	機能の不十分性，合理的に予見可能なミスユース	付録で AI 開発に言及
SaFAD	ISO TR 4084	自動運転システムの安全設計（後継 TS 5083 作成中）	付録で AI 開発に言及
Safety and AI	ISO PAS 8800	AI の安全性やプロセス（作成中）	AI 自体に焦点

7. おわりに

本稿では，自動車業界において，AI は主に自動運転の認知機能とドライバ状態推定に使われ，AI に関係する機能的不十分性の安全性について規定されている SOTIF の概要を述べた．認知機能で使われる AI の安全性は，認識結果の信頼性と関連し，ML ライフサイクルの中でその信頼性を高めるために不確実性要因による揺らぎを測り，管理し，低減する技術・プロセスについて述べた．本稿の後半は，分野横断的な動きと自動車業界との関係に視点を移し，AI・安全工学・ソフト工学などの分野が融合している AI セーフティコミュニティの動向，水平的な枠組みと垂直的な枠組みから法規・標準動向を俯瞰した．

SOTIF，機能安全と整合した AI の安全規格 PAS 8800 が作られ，それが自動車業界の AI セーフティの基本となるであろう．今後，自動車業界全体で構築した標準的な考え方に基づき，各社が製品の安全性と共に AI の安全性について説明責任を果たすことが求められる．

参考文献

[JSAI 01] 人工知能学会編集委員会：人工知能学会誌原稿執筆案内，Vol.16, No.1，pp.**-***（2001）.

[AAIP 19] AAIP: Body of Knowledge, https://www.york.ac.uk/assuring-autonomy/guidance/body-ofknowledge/（2019）

[Abrecht 20] Abrecht, S., Akila, M., Gannamaneni, S. S., Groh,K., Heinzemann, C., Houben, S. and Woehrle, M.: Revisitingneuron coverage and its application to test generation, incomputer safety, reliability, and security, SAFECOMP 2020 Workshops（2020）

[AIQM 23] AIQM 機械学習品質マネジメント検討委員会：機械学習品質マネジメントガイドライン（AIQM ガイドライン）第 3 版（2023）

[Badrinarayanan 17] Badrinarayanan, V., Kendall, A. and Cipolla, R.: SegNet: A deep convolutional encoder-decoderarchitecture for image segmentation, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 39, No. 12, pp. 2481-2495（2017）

[Bresson 17] Bresson, G., Alsayed, Z., Yu, L. and Glaser, S.: Simultaneous localization and mapping: a survey of currenttrends in autonomous driving, IEEE Trans. on Intelligent Vehicles, Vol. 2, No. 3, pp. 194-220（2017）

[CLEPA 20] CLEPA: POSITION PAPER Artificial Intelligence for a coherent regulatory framework that ensures safety and trust（2020）

[CRDS 18] CRDS: AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立，科学技術振興機構戦略提案・報告書（2018）

[Czarnecki 18] Czarnecki, K. and Salay, R.: Towards a frameworkto manage perceptual uncertainty for safe automated driving, Computer Safety, Reliability, and Security, pp. 439-445（2018）

[EC 19] EC: Ethics guidelines for trustworthy AI（2019）

[EC 20a] EC: Artificial intelligence-ethical and legal requirements（2020）

[EC 20b] EC: Assessment List for Trustworthy Artificial Intelligence（ALTAI）for self-assessment（2020）

[EC 20c] EC: White Paper on Artificial Intelligence: A European approach to excellence and trust（2020）

[EC 21] EC: EU AI Act（2021）

[Espinoza 19] Espinoza, H., Yu, H., Huang, X., Lecue, F., Hernández-Orallo, J., hÉigeartaigh, S. c. and Mallah, R.: Towards an AI safety landscape: An Overview, https://www.aisafetyw.org/ai-safetylandscape（2019）

[Eykholt 18] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D.: Robust physical-world attacks on deep learning visual classification, Proc. IEEE Conf. on Computer Vision and Pattern Recognition（CVPR）（2018）

[FLI 18] FLI: Value Alignment Research Landscape, https://futureoflife.org/valuealignmentmap/（2018）

[FTC 21] FTC: Aiming for truth, fairness, and equity in your company's use of AI（2021）

[Gal 17] Gal, Y., Islam, R. and Ghahramani, Z.: Deep Bayesian active learning with image data, Proc. 34th Int. Conf. on Machine Learning, Vol. 70, ICML'17, p. 1183-1192, JMLR（2017）

[GRVA 23] GRVA: ECE/TRANS/WP.29/GRVA/2023/17: Proposal for a draft resolution with guidance on Artificial Intelligence in the context of road vehicles (2023)

[原 20] 原 聡：機械学習モデルの判断根拠の説明～ Explainable AI 研究の近年の展開～，画像センシングシンポジウム（2020）

[ISO 17] ISO: ISO/IEC JTC 1/SC 42（2017）

[ISO 18] ISO: ISO 26262-1:2018 Road vehicles — Functional safety（2018）

[ISO 20] ISO: ISO/TR 4804:2020 Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation（2020）

[ISO 22a] ISO: ISO 21448:2022 Road vehicles — Safety of the intended functionality（2022）

[ISO 22b] ISO: ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology（2022）

[ISO 22c] ISO: ISO/IEC CD TR 5469 Artificial intelligence — Functional safety and AI systems（2022）

[自動車技術会 22] 自動車技術会：JASO テクニカルペーパ TP18004 自動車用運転自動化システムのレベル分類及び定義（2022）

[JST 未来] JST 未来社会創造事業：Engineerable AI プロジェクト，https://engineerable.ai/

[亀谷 22] 亀谷由隆：説明可能 AI 技術のこれまでとこれから，信学会基礎・境界ソサイエティ Fundamentals Review, Vol. 16, No. 2, pp. 83-92（2022）

[Kargar 22] Kargar, I.: Blog post: Active learning, data selection, data auto-labeling, and simulation in autonomous driving，https://medium.com/aiguys/active-learningand-data-autolabeling-in-autonomous-driving-5d6bec956a38（2022）

[経産省 19] 経済産業省：AI・データの利用に関する契約ガイドライン（2019）

[経産省 21] 経済産業省：AI 原則の実践の在り方に関する検討会我が国の AI ガバナンスの在り方，ver1.1（2021）

[Krizhevsky 12] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, Vol. 25, Curran Associates, Inc.（2012）

[Kuwajima 19] Kuwajima, H. and Ishikawa, F.: Adapting SQuaRE for quality assessment of artificial intelligence systems, 2019 IEEE Int. Symp. on Software Reliability Engineering Workshops（ISSREW）（2019）

[Kuwajima 20] Kuwajima, H., Yasuoka, H. and Nakae, T.: Engineering problems in machine learning systems, Machine Learning, Vol. 109, No. 5, p. 1103-1126（2020）

[桑島 22] 桑島 洋，平田雄一，中江俊博：自動車業界における機械学習システムの品質確保の事例，システム／制御／情報，Vol. 66, No. 5, pp. 187-194（2022）

[Lu 17] Lu, J., Sibai, H., Fabry, E. and Forsyth, D.: NO need to worry about adversarial examples in object detection in autonomous vehicles, arXiv:1707.03501（2017）

[Mae 21] Mae, Y., Kumagai, W. and Kanamori, T.: Uncertainty propagation for dropout-based Bayesian neural networks, Neural Networks, Vol. 144, pp. 394-406（2021）

[Mohseni 20] Mohseni, S., Pitale, M., Singh, V. and Wang, Z.: Practical solutions for machine learning safety in autonomous vehicles, SafeAI（2020）

[内閣府 19] 内閣府：AI 戦略 2019（2019）

[中川 20] 中川裕志：人工知能学会チュートリアル 3A4-TS-2, AI 倫理とガバナンス：世界動向と産業界の取り組み（2020）

[Nativi 21] Nativi, S. and De Nigris, S.: AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework, Technical Report, European Commission（2021）

[Ng 21] Ng, A.: Data-centric AI Resource Hub, https://datacentricai.org/（2021）

[NIST 19] NIST: A Plan for Federal Engagement in Developing AI Technical Standards and Related Tools in response to Executive Order（EO 13859）（2019）

[NIST 22a] NIST: AI Fundamental Research-Managing AI Bias（2022）

[NIST 23] NIST: AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework（2023）

[NTSB 18] NTSB: Collision between vehicle controlled by developmental automated driving system and pedestrian tempe, Arizona, 2018（2018）

[OECD 19] OECD: OECD AI Principles（2019）

[OMB 20] OMB: Guidance for Regulation of Artificial Intelligence Applications（2020）
[Ortega 18] Ortega, P. A., Maini, V. and the DeepMind safety team: Bui lding safe arti f icial intell igence: specification, robustness, and assurance, https://deepmindsafetyresearch.medium.com/building-safeartificial-intelligence-52f5f75058f1（2018）
[Pavlitskaya 22] Pavlitskaya, S., Yikmis, S. and Zollner, J.: Is Neuron coverage needed to make person detection more robust ?, CVPR 2022 TCV workshop（2022）
[Phillips 21] Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A. and Przybocki, M. A.: Four principles of explainable artificial intelligence, Technical report（2021）
[Postels 19] Postels, J., Ferroni, F., Coskun, H., Navab, N. and Tombari, F.: Sampling-free epistemic uncertainty estimation using approximated variance propagation, Proc. IEEE Int. Conf. on Computer Vision（ICCV）（2019）
[QA4AI 22] QA4AI コンソーシアム：AI プロダクト品質保証ガイドライン（QA4AI ガイドライン）（2022）
[Redmon 16] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection, Proc. IEEE Conf. on Computer Vision and Pattern Recognition（CVPR）（2016）
[SAE 16] SAE: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles（2016）
[Salay 17] Salay, R., Queiroz, R. and Czarnecki, K.: An Analysis of ISO 26262: Using machine learning safely in automotive software, arXiv:1709.02435（2017）
[Scheirer 13] Scheirer, W. J., Rezende Rocha, de A., Sapkota, A. and Boult, T. E.: Toward open set recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 35, No. 7, pp. 1757- 1772（2013）
[SEAMS 20] SEAMS プロジェクト：人工知能搭載システムの安全設計ガイドライン（SEAMS ガイドライン）（2020）
[Selvaraju 17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization, Proc. IEEE Int. Conf. on Computer Vision（ICCV）（2017）
[Szegedy 14] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: Intriguing properties of neural networks, Proc. Int. Conf. on Learning Representations（ICLR）（2014）
[TAILOR] TAILOR: Foundations of Trustworthy AI-Integrating Reasoning, Learning and Optimization（EU Horizon 2020 Project）
[VDA] VDA: KI Absicherung, https://www.kiabsicherungprojekt.de/en/
[White House 19] White House: Maintaining American Leadership in Artificial Intelligence（2019）
[Willers 20] Willers, O., Sudholt, S., Raafatnia, S. and Abrecht, S.: Safety concerns and mitigation approaches regarding the use of deep learning in safety-critical perception tasks, SAFECOMP（2020）
[Yang 21] Yang, J., Zhou, K., Li, Y. and Liu, Z.: Generalized out-ofdistribution detection: A survey, arXiv:2110.11334（2021）
[York] York, of U.: Assuring Autonomy International Programme, https://www.york.ac.uk/assuring-autonomy/
[Yu 18] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V. and Darrell, T.: BDD100K: A diverse driving dataset for heterogeneous multitask learning, arXiv:1805.04687（2018）
[Yurtsever 20] Yurtsever, E., Lambert, J., Carballo, A. and Takeda, K.: A survey of autonomous driving: common practices and emerging technologies, IEEE Access, Vol. 8, pp. 58443- 58469（2020）
[Zablocki 22] Zablocki, c., Ben-Younes, H., Pérez, P. and Cord, M.: Explainability of deep vision-based autonomous driving systems: Review and challenges, J. of Computer Vision, Vol. 130, pp. 2425-2452（2022）
[Zhang 22] Zhang, J. M., Harman, M., Ma, L. and Liu, Y.: Machine learning testing: Survey, landscapes and horizons, IEEE Trans. on Software Engineering, Vol. 48, No. 1, pp. 1-36（2022）

著者

中江 俊博
なかえ としひろ
ソフト生産革新部
製品に搭載される AI の品質保証，ソフトウェア開発効率化のための AI 活用に従事

桑島 洋
くわじま ひろし
ソフト生産革新部　博士（工学）
AI 品質・AI 安全の研究開発，社内プロセス構築，標準化活動に従事