

DENSO

Crafting the Core

JSAI2019

企画セッションKS-5

機械学習の安全性への取り組み －自動車業界の取り組みを中心に－

**中江 俊博
桑島 洋**

株式会社 デンソー
技術企画部 ソフトウェア先行開発室

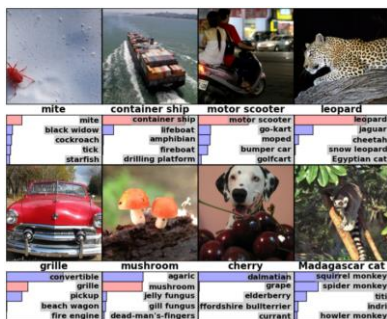
2019年6月5日



機械学習システム開発を取り巻く環境

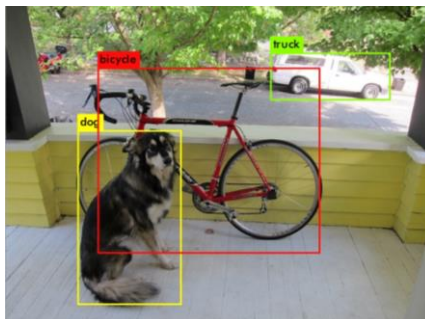
機械学習ベースの画像認識技術の進歩

物体認識



[Krizhevsky+ 2012]

物体検出



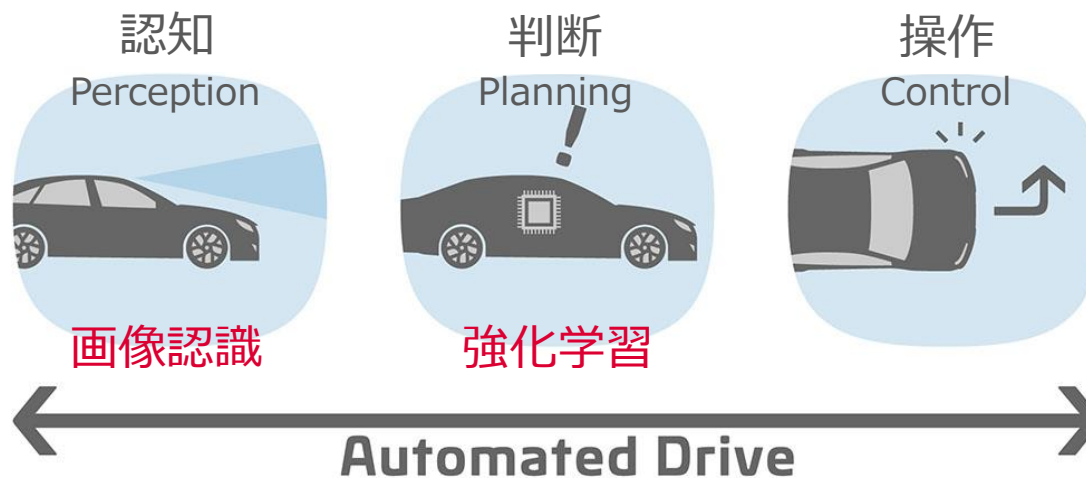
[Redmon+ 2015]

セマンティックセグメンテーション



[Badrinarayanan+ 2015]

自動運転技術への応用



機械学習システムのリスク

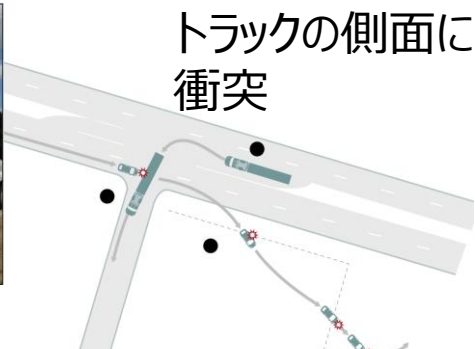
自動運転走行中の事故



2016年 Teslaの事故 (出典:NTSB)



2018年 Uberの事故 (出典:ABC 15)



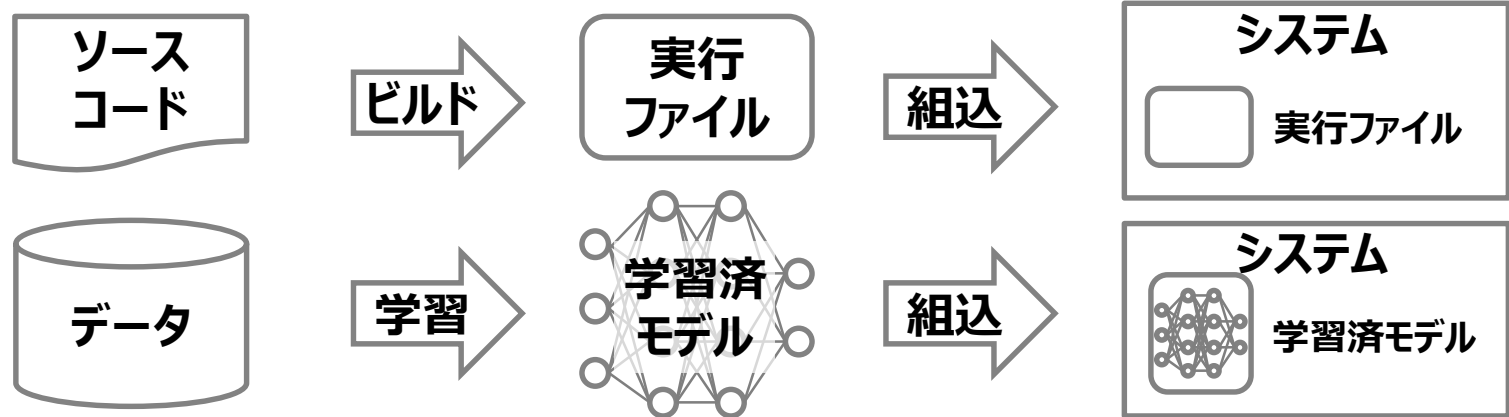
<https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF>



<https://www.nts.gov/investigations/accidentreports/pages/hwy18mh010-prelim.aspx>

オープン環境で使われる機械学習システムの安全担保は重要な課題

従来のソフトウェア開発と機械学習システム開発のギャップ



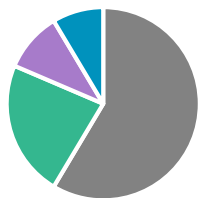
	従来のソフトウェア	機械学習
設計・実装	ソースコードを記述	データセットを作成
実行ファイル	ソースコードから 明示的に振る舞い が決定される	データから 帰納的に振る舞い が決定される
システム	演繹的に単体の 振る舞いが保証できる 前提	演繹的に単体の 振る舞いが保証できない (テストデータに依存)

データの課題

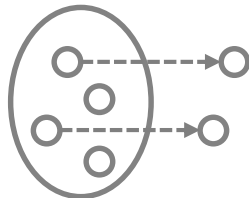
データ作成プロセス



課題



どんなデータ分布を想定すればよいのか



大量のデータからどのデータを選定するか



ラベル(クラス)をどのように定義すればよいのか



ラベル付けの品質をどう担保するか

機械学習用データの データ分布の妥当性(Validation) と データ検証プロセス(Verification) の体系化が必要

モデルの課題

説明性

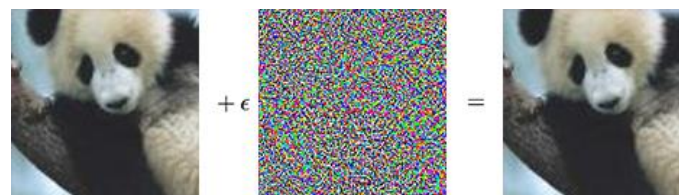


[Selvaraju+ 2016]

Explainability / Interpretability

どの開発プロセスでどう使うのか???

頑健性



パンダ

テナガザル

[Goodfellow+ 2014]

Adversarial Examples (敵対標本)

物理世界で起こり得る問題???

自動運転開発において本当に求めていること



学習済
モデル

失敗要因
を分析

要因A

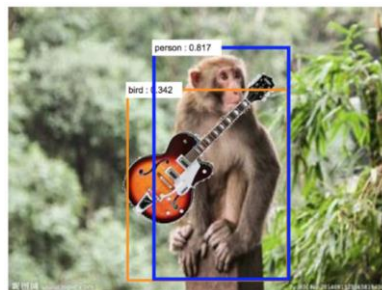
プロセス 1

B

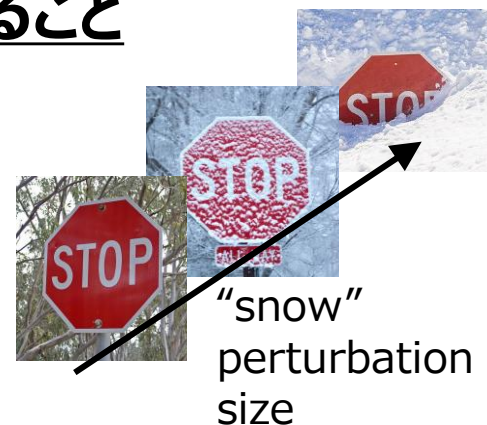
プロセス 2

C

プロセス 3



[Yuille+, Limitations of Deep Learning for Vision, and How We Might Fix Them]



"snow"
perturbation
size

要因から次のプロセスを導く説明性

レアシーンに対する頑健性 自然ノイズに対する頑健性

人でも識別が難しいシーン ～走行領域認識用データセット※を例に～

※BDD100K-Drive <https://bdd-data.berkeley.edu/>

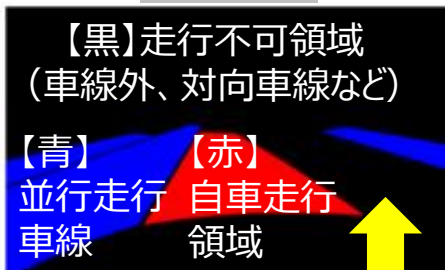
【例 1】 同一シーンに対するゆらぎ

カメラ画像



ゼブラゾーンに対し、こちらのシーンは走行不可

正解ラベル



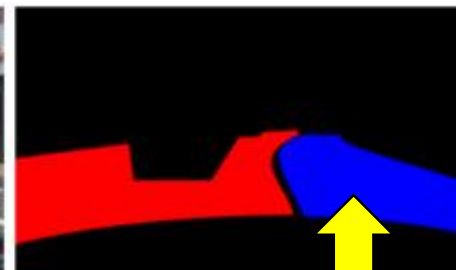
【黒】走行不可領域
(車線外、対向車線など)

【青】
並行走行
車線

【赤】
自車走行
領域



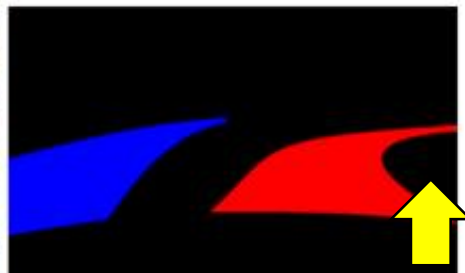
こちらのシーンは走行可能



【例 2】 シーンの見切れによるゆらぎ



対向車線かどうか識別できない



【例 3】 コンテキスト依存



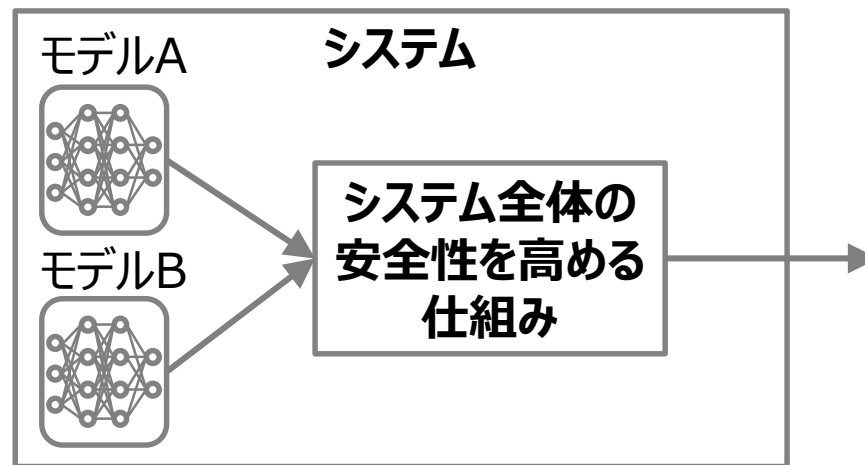
遠くの車の走行方向によって識別結果が変わる



**1 枚の画像から正解を識別するのが困難なシーンが
現実世界には存在する**

システムの課題

- トラッキング、フィルタリング # 直列型
- モデル冗長化 # 並列型
- 異常検知、Out-of-Distribution検知
- 高精度地図の活用 など



**機械学習だけに頼らず
モデルの「外側」でも安全性を高める仕組みが必要**

まとめ

- 機械学習システム開発の課題を**データ、モデル、システムの視点で俯瞰**
- 安全・品質担保の観点で、機械学習を使いこなすための**技術の体系化**が急務
 - データ作成プロセスの確立
 - 開発ニーズにマッチしたモデル技術
 - 機械学習を補完するシステム技術
- データ分布の考え方、標準プロセスなど**決め事**は一社で解決できず、**社会との合意形成**が必要

機械学習の安全性への取り組み – 各国の動向 –

■ 取り組みを北米、欧州、日本の視点で俯瞰

- 国際標準：認識系に着目しISO 21448を策定中
- 北米：従来標準に対して機械学習の影響を分析
- 欧州：AI倫理ガイドラインを発行
- 日本：AIプロダクト品質保証ガイドラインを発行
- 学会動向：欧州研究者を中心に、機械学習の安全のワークショップが発足

ISO 21448 (SOTIF) – 国際標準 –

ISO 21448 (SOTIF, Safety of the Intended Functionality) は18年11月策定開始、19年1月PAS (公開仕様書) 発表、22年発行予定

N. Becker. The Safety of the Intended Functionality, Report on ISO/TC22/SC32/WG8 activities

自動ブレーキ機能の例



認識システムの限界により、意図しないブレーキングが起こる。

- ・天候 (雨/晴れ/霧)
- ・画像認識の間違い
- ・など



トリガリング
イベント

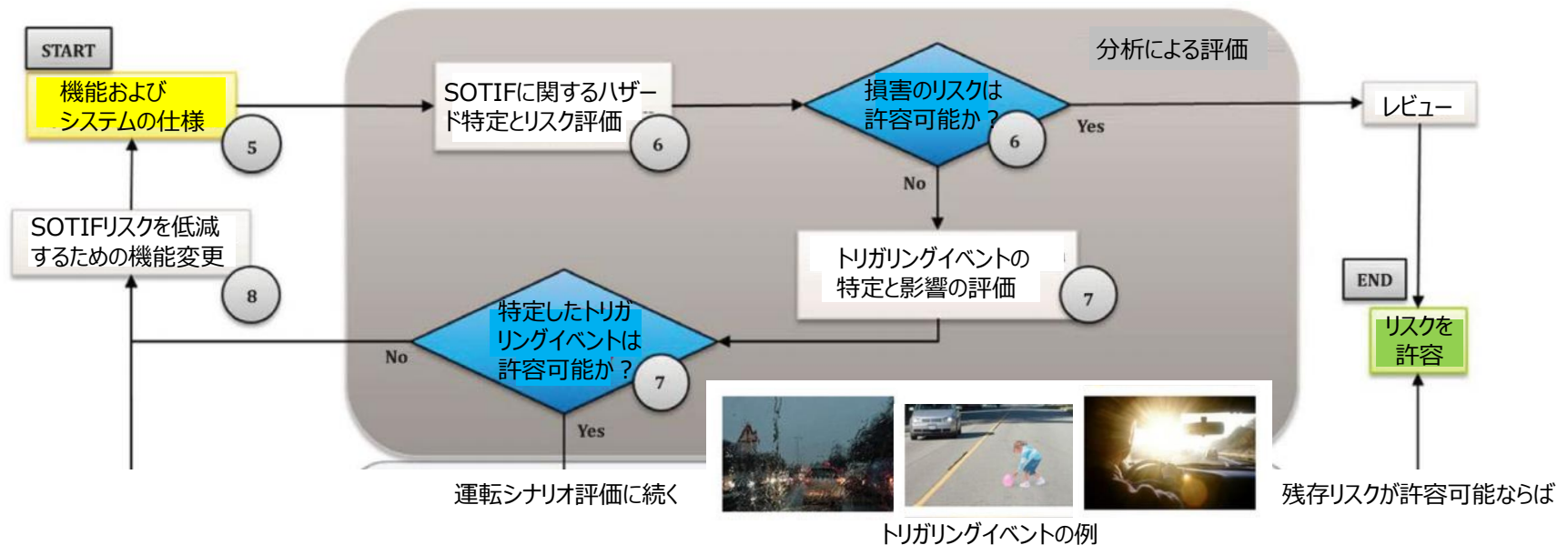
Presented by
ISO/TC22/SC32/WG8

**故障の回避、緩和を確認する機能安全に対し、
SOTIFは意図した領域内で適切に振る舞うことを確認する**

ISO 21448 (SOTIF) – 国際標準 –

トリガリングイベントの特定/影響評価とリスク低減のための機能変更を、
リスクが許容可能なレベルに低下するまで繰り返す

N. Becker. The Safety of the Intended Functionality, Report on ISO/TC22/SC32/WG8 activities

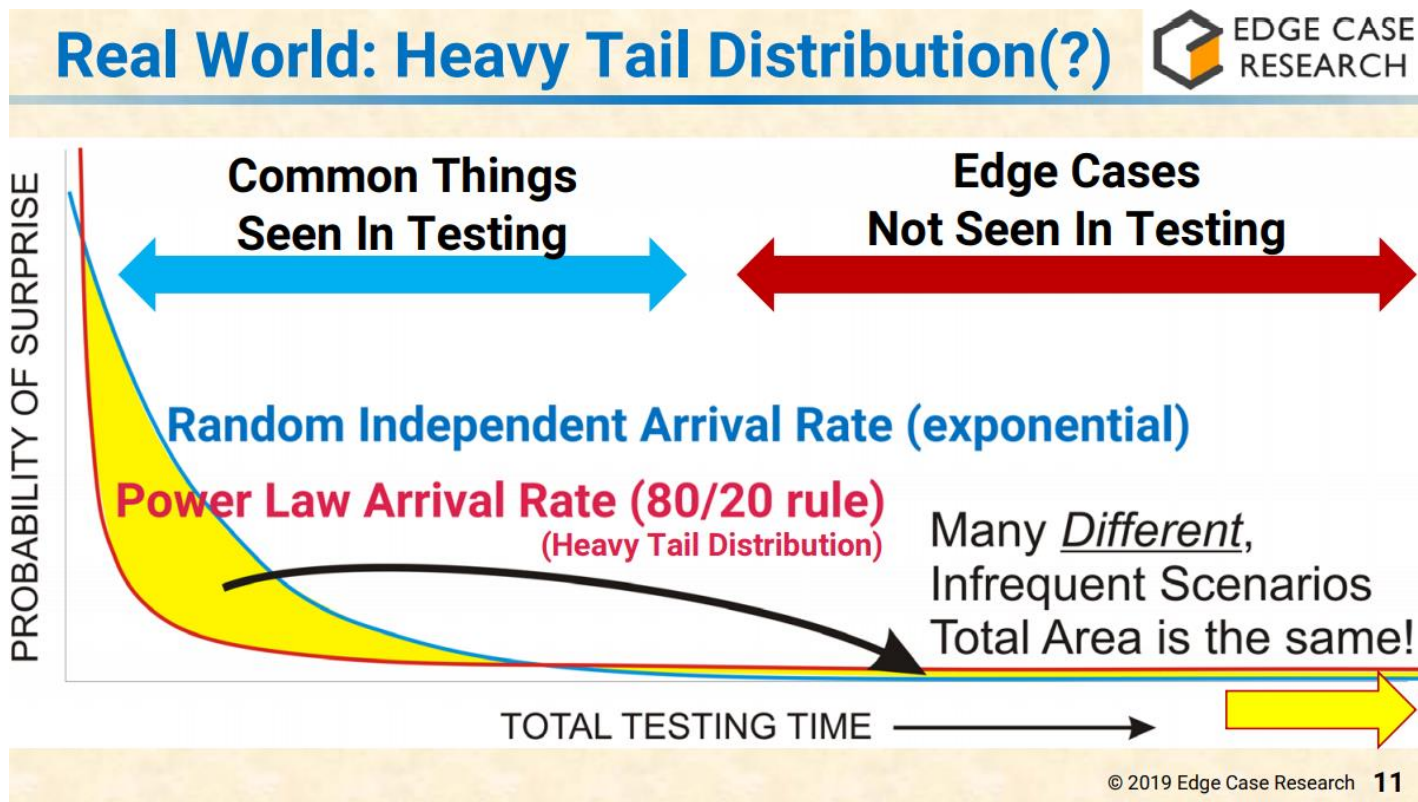


**トリガリングイベントの特定と対処を求める一方、技術はまだ無い
間違いを1件ずつ確認し対策するのは統計的機械学習には無い考え方**

現実世界のレアケース(エッジケース) – 米国 –

P. Koopman. Autonomous vehicle perception safety.

http://users.ece.cmu.edu/~koopman/talks/2104_PerceptionSafety.pdf



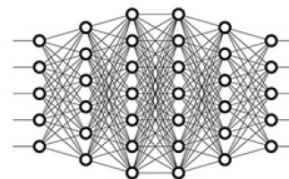
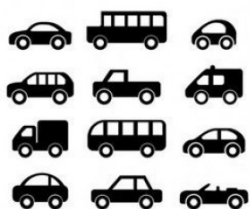
**ボリュームゾーンとロングテールをすべて集めた件数は同等と考え、
ロングテールに存在するエッジケースの重要性を主張**

ISO 26262(機能安全) の適用可能性 –カナダ–

R. Salay, R. Queiroz, and K.Czarnecki. An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software.

従来の仕様に対するプログラミングとは異なり、機械学習ベースの機能は、**完全な仕様ではないデータセットを使用した機械学習アルゴリズムによって生成される。**

Challenges of Assuring Machine Learned Components



ディープラーニングなどの高性能な機械学習のアプローチは、プログラムコードのように**人間が直接解釈できない方法で機能を表現する。**

Lack of specification

完全な仕様の欠如

Lack of inspectability

解釈可能性の欠如

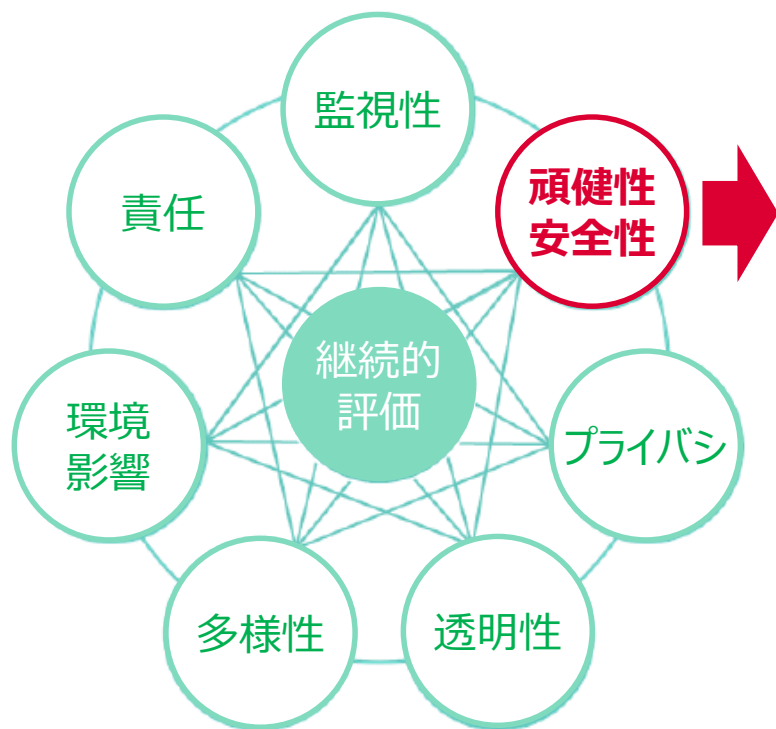
R. Salay, R. Queiroz, K. Czarnecki. An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software. SAE, 2018-01-1075, 2018; preliminary version also available at <https://arxiv.org/abs/1709.02435>

17

機械学習を使用して開発した機能に、ISO 26262で定義された保証方法を適用するには、2つの大きな障害が存在

AI倫理ガイドライン – 欧州委員会 –

High-Level Expert Group on Artificial Intelligence (AI HLEG), Ethics guidelines for trustworthy AI



7つの倫理要求

副要求	チェック項目
攻撃耐性 セキュリティ	脆弱性への対策
	攻撃に対するインテグリティと回復力
	想定外の状況や環境における振る舞い
	軍民両用の可能性
縮退運転 安全性	敵対的攻撃や想定外の状況に対する縮退運転
	各ユースケースについてのリスク評価
	ユーザや第三者に損害を与える可能性
	ユーザや第三者に与える損害の影響
精度	必要な精度の定義と必要な精度の程度
	不正確な予測をしたときに発生する損害
	許容できないほど精度低下を認識する手段
	精度を高めるための取り組み
信頼性 再現性	システムが目標を満足しているか監視する手段

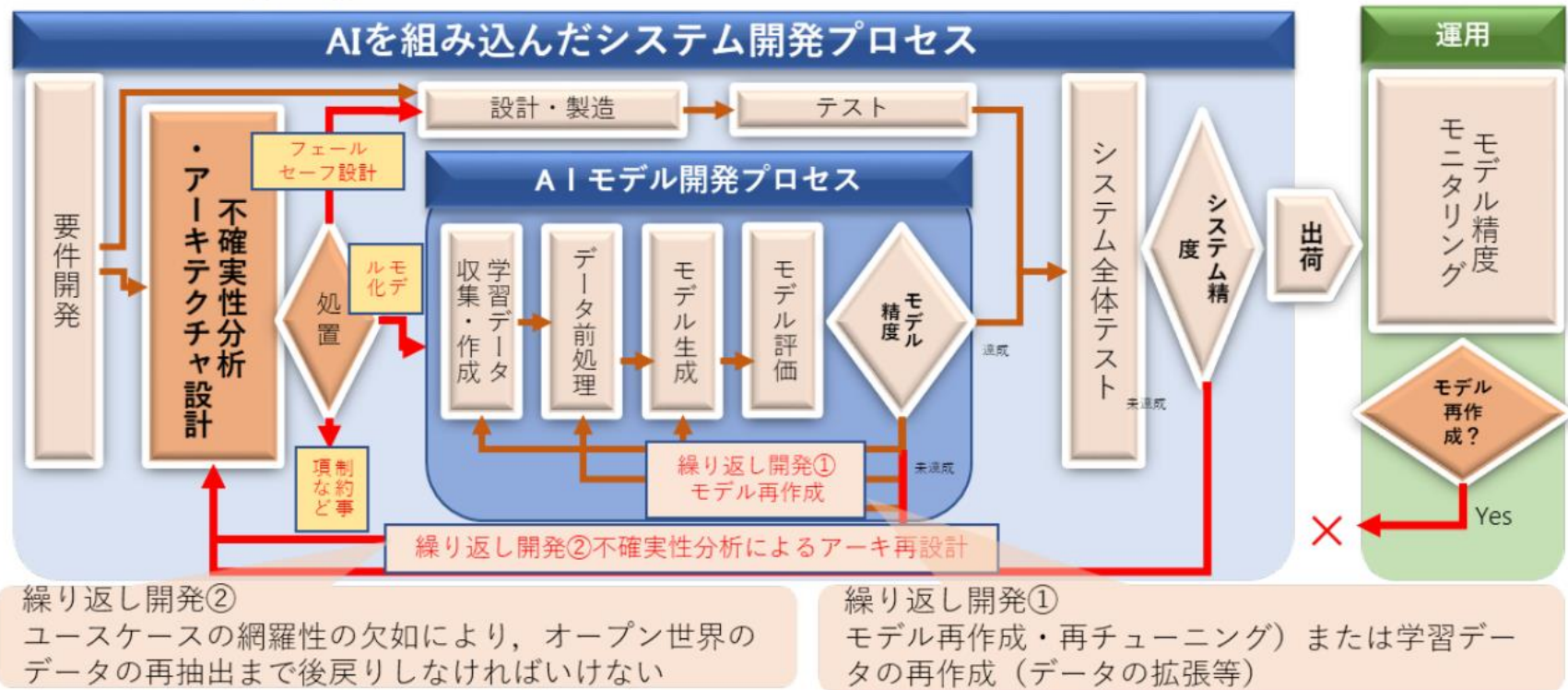
頑健性と安全性の倫理アセスメント

プロダクトだけでなく開発/運用プロセスに対する倫理要求も含まれる
19年中に試行して具体化すると共に、他の標準と整合を取る

QA4AI（自動運転グループ） – 日本 –

「AIプロダクト品質保証ガイドラインの発行について」@Open QA4AI Conference 2019 (2019/5/17)
<http://www.qa4ai.jp/QA4AI.OpenQA4AIconference.20190517.pdf>

- Process Agility向上のための繰り返し開発のプロセスを図示した



QA4AIが2019年5月にAIプロダクト品質保証ガイドラインを発行
フェールセーフ設計がAIシステム開発プロセスに組み込まれている

国際学会動向

SafeAI@AAAI (1月)

AI Safety@IJCAI (8月) WAISE@SAFECOMP (9月)

オーガナイズ



Huáscar Espinoza

CEA LIST (仏)

Han Yu

南洋理工大 (シンガポール)

Seán Ó hÉigeartaigh

ケンブリッジ大 (英)

Freddy Lecue

タレス (カナダ)

Mauricio Castillo-Effen

ロッキード・マーティン (米)

Xiaowei Huang

リバプール大 (英)

Richard Mallah

Future of Life Institute (米)

José Hernández-Orallo

バレンシア工科大 (スペイン)

Cynthia Chen

香港大 (中)

Zakaria Chihani

CEA LIST (仏)

Simos Gerasimou

ヨーク大 (英)

Andreas Theodorou

ウメオ大 (スウェーデン)

Guillaume Charpiat

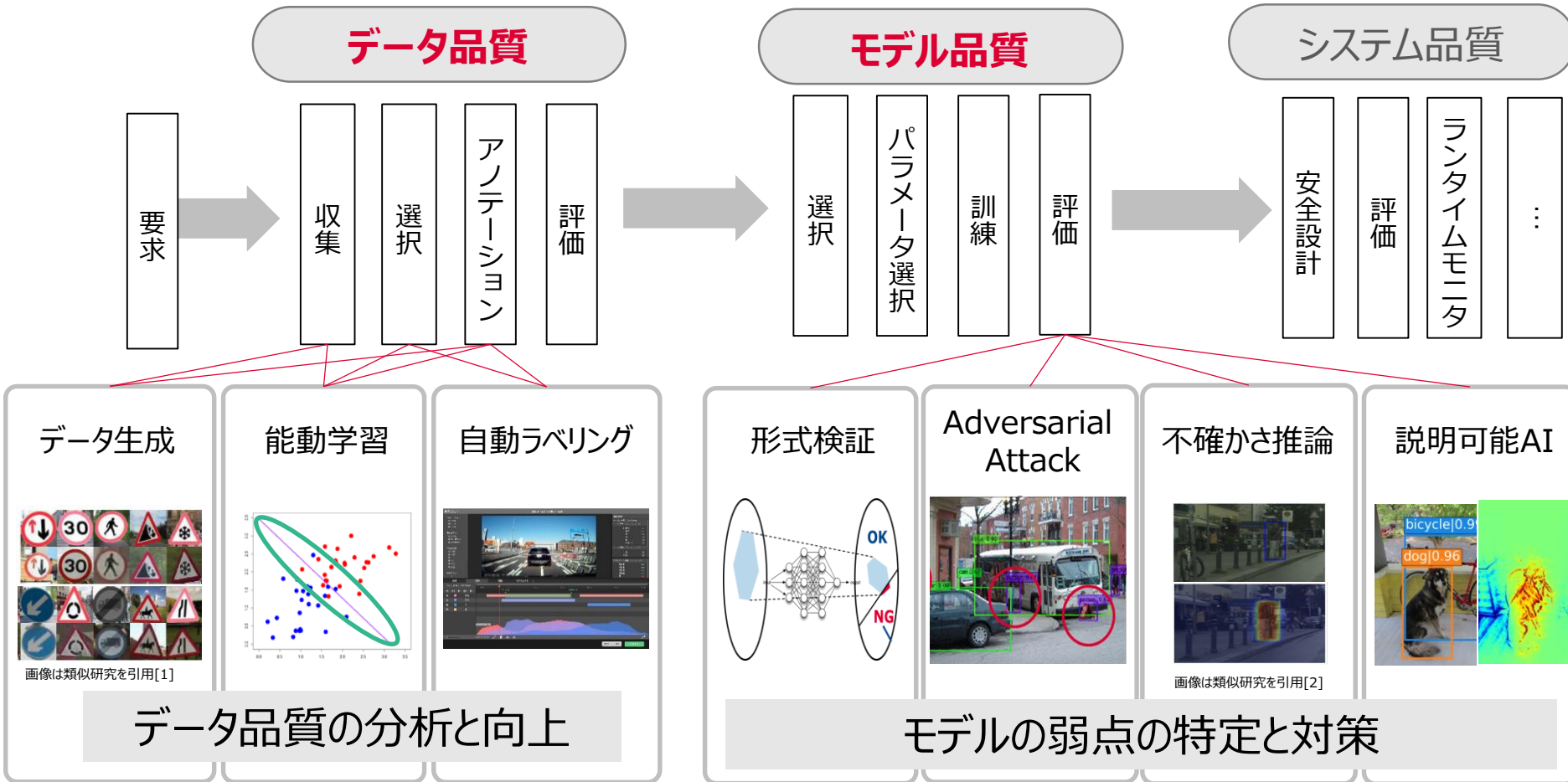
INRIA (仏)

欧州のロボット、リスク工学、検証、AI、システム/ソフトウェアなどの研究者を中心に機械学習の安全のワークショップが相次いで発足

標準化動向から考えられる技術課題

- SOTIFや自動運転システム安全の活動
 - トリガリングイベント（エッジケース）の特定
 - トリガリングイベント（エッジケース）への対処
- ISO 26262(機能安全)の適用可能性
 - 機械学習ベースの機能に対する仕様の設定
 - 機械学習モデルの解釈可能性の向上
- AI倫理ガイドライン
 - 想定外の状況や環境の検知
 - 精度低下（不確実性増大）の検知
 - ...
- AIプロダクト品質保証ガイドライン
 - 機械学習を組み込んだシステムのフェイルセーフ設計

デンソーのAI品質基盤の研究開発



データ品質、モデル品質の観点で基盤技術进行研究開発

[1] <https://www.mdpi.com/2504-2289/2/3/19/pdf>

[2] <https://arxiv.org/pdf/1711.09026.pdf>

機械学習の安全性への取り組み – 各国の動向 –

■ 取り組みを北米、欧州、日本の視点で俯瞰

- 国際標準：認識系に着目しISO 21448を策定中
- 北米：従来標準に対して機械学習の影響を分析
- 欧州：AI倫理ガイドラインを発行
- 日本：AIプロダクト品質保証ガイドラインを発行
- 学会動向：欧州研究者を中心に、機械学習の安全のワークショップが発足

- ・社会との合意形成：国際標準団体と欧州を中心に進められている
- ・技術の体系化：実証された共通技術はまだ無く探索研究のフェーズ
- ・安全工学の研究者には機械学習システム全体の安全技術を、
機械学習の研究者には機械学習モデル単体の安全技術を期待

DENSO

Crafting the Core