細やかな要求に応じて安全なAIを仕立て上げる

eAIプロジェクト
自動運転の安全性に向けたAI修正技術

石川 冬樹, Paolo ARCAINI, 吉岡 信和, eAI プロジェクト

どんな研究？

機械学習・深層学習技術を用いたAIソフトウェアに対して，細やかな要求やリスクを考慮し対応するための工学技術に取り組んでいます．

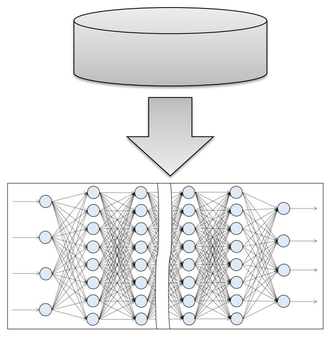
何ができる？

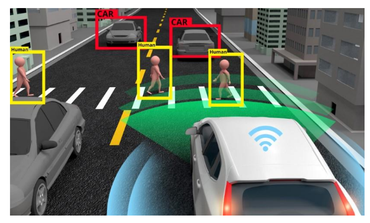
プログラムに対する欠陥箇所推定や自動修正の技術を適合させ，AIソフトウェアに対しても，要求・リスクに応じた「狙った修正」を行えます．

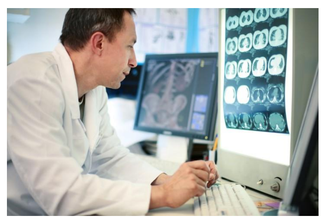
eAI プロジェクトの概要

機械学習・深層学習技術を用いたAIソフトウェア：
データを用いた訓練によるパラメータ調整を通して機能を構築・修正
（深層学習では数百万以上のパラメータ）

1．大量データが必要となる
2．避けたい誤りの考慮など細かい調整ができない

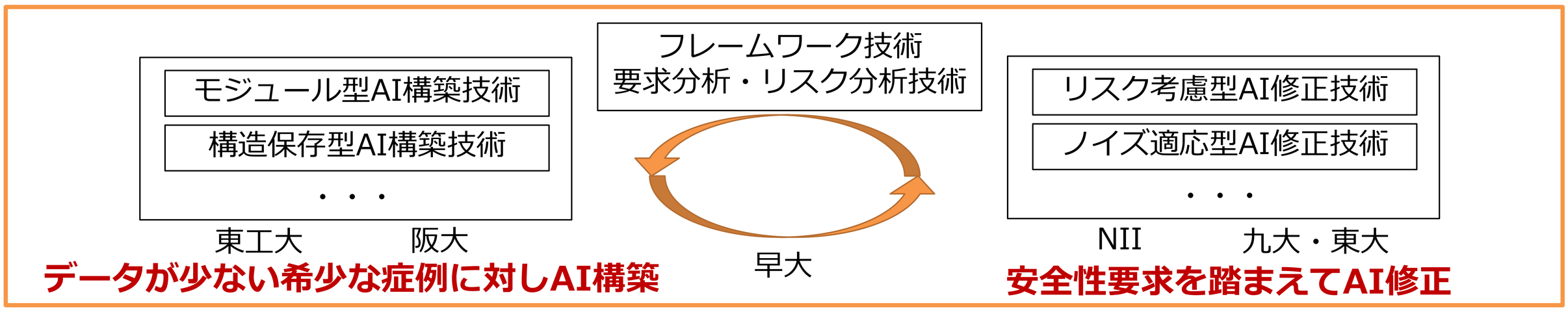






安全・信頼が重要な分野におけるAI活用の障害

要求やリスクに応じAIを仕立て上げるフレームワーク → 医療・交通の二分野で実証



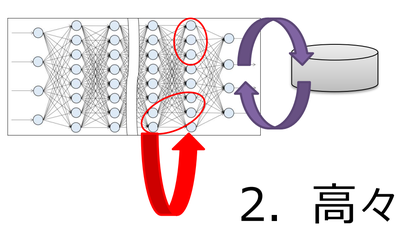
AI修正技術へのアプローチ

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	近距離の後方車両	不要なブレーキで後方から衝突	3	誤り 35%	×
…	…	…	…	…	…	…

自動車企業や安全性の専門家とともにAI誤りの影響を分析してベンチマークを設定

データ追加や訓練ハイパーパラメータの調整では特定の誤りを減らすような修正になかなかたどり着かず，意図せず他の誤りが増えてしまうことも

従来ソフトウェアプログラム向けの欠陥箇所推定・自動修正技術の考え方を活用！



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を調整

技術例： DistrRep [ICST'23]

- 複数種別の誤り率を優先度を踏まえた調整を重視
- 異なる誤り種別ごとに対して修正案を作成し，優先度を踏まえてそれらを統合

技術例： NeuRecover [SANER'22]

- 更新字の重要な対象の見落とし増加防止を重視
- 過去のバージョンと比較することで，修正すべきパラメーターを絞り込み

連絡先：国立情報学研究所　アーキテクチャ科学研究系　石川　冬樹
URL：https://research.nii.ac.jp/~f-ishikawa/ Email：f-ishikawa@nii.ac.jp

ツールキット公開中
https://github.com/jst-qaml/eAI-Repair-Toolkit/