Rev. 1.0 (2018-12-17)

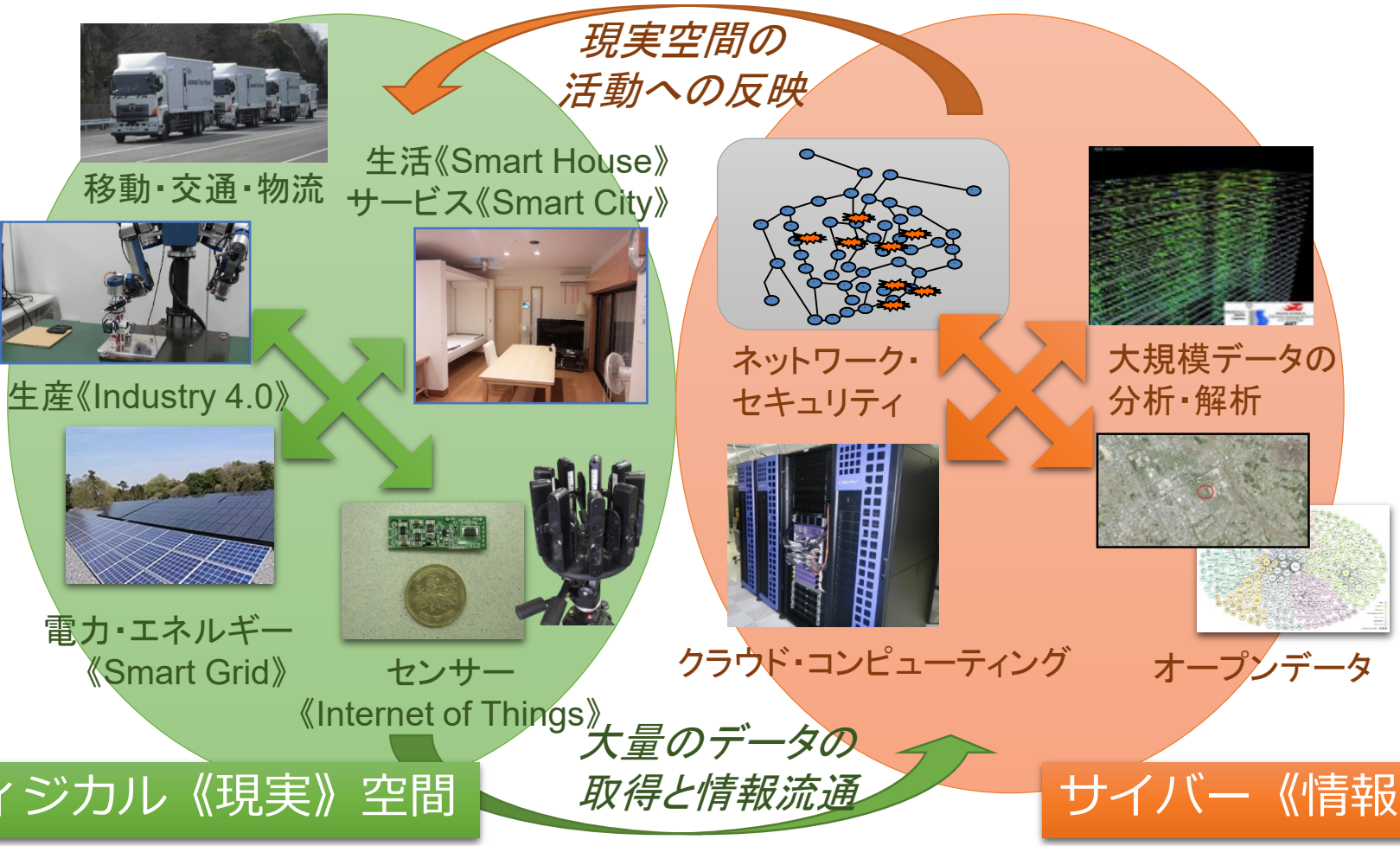
CPS & AI 時代のソフトウェア セキュリティ・セーフティ

2018年12月17日

産業技術総合研究所
大岩 寛

Cyber Physical System (CPS)

コンピュータ（サイバー空間） と 現実社会 の密連携・一体化



CPS とソフトウェア

- コンピュータシステムの「動作意図」を定義するのはソフトウェア
- ソフトウェアが人の命を預かる時代
  - 航空機・鉄道車両
  - 自動車
  - 消費者機械
  - インフラ
  - 医療・ヘルスケア

CPS とソフトウェア

- CPS の本質的な難しさ
  - ソフトウェアの本質 = 入力と出力の関係の記述
  - 「実世界」が入力

⇒ 起こりうる入力の全貌を列挙しきれない

  - 例えば、車の自動運転

    - » 天気・気温・湿度・路面状況・時間帯・明るさ
    - » 障害物の距離、大きさ、位置、数、速度、種類
    - » 自分の速度、タイヤの状態、前後車の速度
    - » 運転手の状態（起きてる・寝てる）etc... etc...
  - どんなに書き起こしても、全てを網羅したと言えない

⇒ むしろ、書き切れないことこそが実世界の本質

機械学習AIへの期待

- 機械学習によるシステム構築への期待
  - ★「**新しいプログラム作成技術**」
  - 人間が分析しきれない事象に対して、
  - とにかく大量のデータを与えることで、
  - 何らかの判断基準を自動的に生み出してくれる

- CPS的なソフトウェアの開発費削減
- 本質的に分析しきれない「実社会」への対応
- 「実社会」の絶え間ない変化への追従
- 人間の思いつかない解法・ロジックの発見

機械学習AIへの期待

- 特に、画像認識などへの大きな期待
  - 人間が直感的に認識していて「理由を説明できない」＝人間の知識をプログラム化できない分野
  - 近年のディープラーニングなどの技術進展で著しく認識精度・構築効率が向上
  - CPSにおける広い応用の範囲
    - 実世界でモノを動かすには、空間の認識が必須

機械学習AIへの期待

- 特に、画像認識などへの大きな期待
  - 人間が直感的に認識していて「理由を説明できない」＝人間の知識をプログラム化できない分野
  - 近年のディープラーニングなどの技術進展で著しく認識精度・構築効率が向上
  - CPSにおける広い応用の範囲
    - 実世界でモノを動かすには、空間の認識が必須

機械学習AIへの期待と不安

- 一方で、実世界の応用を阻む「不安」
  - あまりにも「ブラックボックス」なAIの動作
    - 事前に判断が予測できない
      - どのようなときに判断を間違えるのか、わからない
      - ＝ どのようなときに信用して良いのか、わからない
    - 判断の理由が説明できない
      - AIが判断を間違えたときに、間違えた理由が分からない
      - 理由が分からないので確実な修正もできない

⇒ 品質を説明できない

- 学習時のデータに対しては、正答率などを評価している
- しかし、それが本番動作時の正答率を説明できない

機械学習AIへの期待と不安

- 実世界の応用を阻む「品質の問題点」
  1. 顧客に安心して買ってもらえない
     - 一定品質を保証できないので、PoCから先に進めない
  2. 誤動作時に責任を取れない/逃れられない
     - 「無過失」を証明できないので、
       想定外の結果に全責任を負う羽目になりかねない
  3. 売買契約において不利になる
     - 「想定内の瑕疵」と「当初想定を超えた機能拡張」が
       区別できない ⇒ 延々とメンテナンスする羽目になる
     - 安いデータでいい加減に作った手抜きのAIと、
       きちんと慎重に作ったAIが区別できない ⇒ 競争に負ける

AI & CPS 時代の セーフティ・セキュリティ

- 実世界を安全・セキュアに保ち続けるため

  ⇒ AI & CPS 時代のソフトウェアのための
  セーフティ・セキュリティの技術開発

1. 機械学習など「新たな形態のソフトウェア」の品質保証の枠組みの構築
2. CPS の本質である「実世界」の無限の多様性への対応

従来のソフトウェア安全性確保の考え方

- リスク対策の透明性・説明性を重要視
  - 想定されるリスクを分析し全て列挙する
  - それぞれのリスクに対して、対策を検討
  - 各リスクに対応する安全性機能の「実現方針」を確認する
    - 「人が作ったものならば、実装したときの考え方・『方針』を説明・確認できる」
      - 方針が確認できれば、一通り「だいたい」正しく実装されているだろう、と推定
      - 誤りが無いことを、個別にテスト等で確認
  - 製作工程全体のプロセスを管理することで、システム全体の説明性を確保

→翻って「リスク対策が説明・確認できるように、作る」

従来のソフトウェア安全性確保の考え方

- そのままでは機械学習AIには適用できない
  - あるのは大量の学習データ
    - 想定されるリスクケースを
      全て網羅しているか、わからない
  - 学習AIプログラムに学習データを与えると、
    何らかの結果が返ってくる
    - 学習データに対して、うまく動くことは確認できる
    - でも、なぜうまく動くのか、理由は説明できない

    ⇒ 本番の別のデータに対してうまく動くことは、
    期待であって説明できない

    ⇒ リスク対応の説明性・透明性が確保できない

    新しい方法論と枠組みが必要

AI品質保証プロジェクト

- 機械学習AIの品質保証の枠組みの開発

  CPS 向け AI ソフトウェアのための品質の基準 + 開発ガイドライン + 品質保証技術

  – AI & CPS 実装の品質要件・レベルの設定
  – 標準的な開発実装プロセスのガイドライン策定
  – 品質を担保する具体的な管理目標の設定
  – 具体的な品質確保のための技術・ツール開発

AI品質保証プロジェクト

アウトプット①:
フォーラム・デファクト標準としての
安全性基準・確認ガイドライン
→ 将来のJIS/ISO化を想定

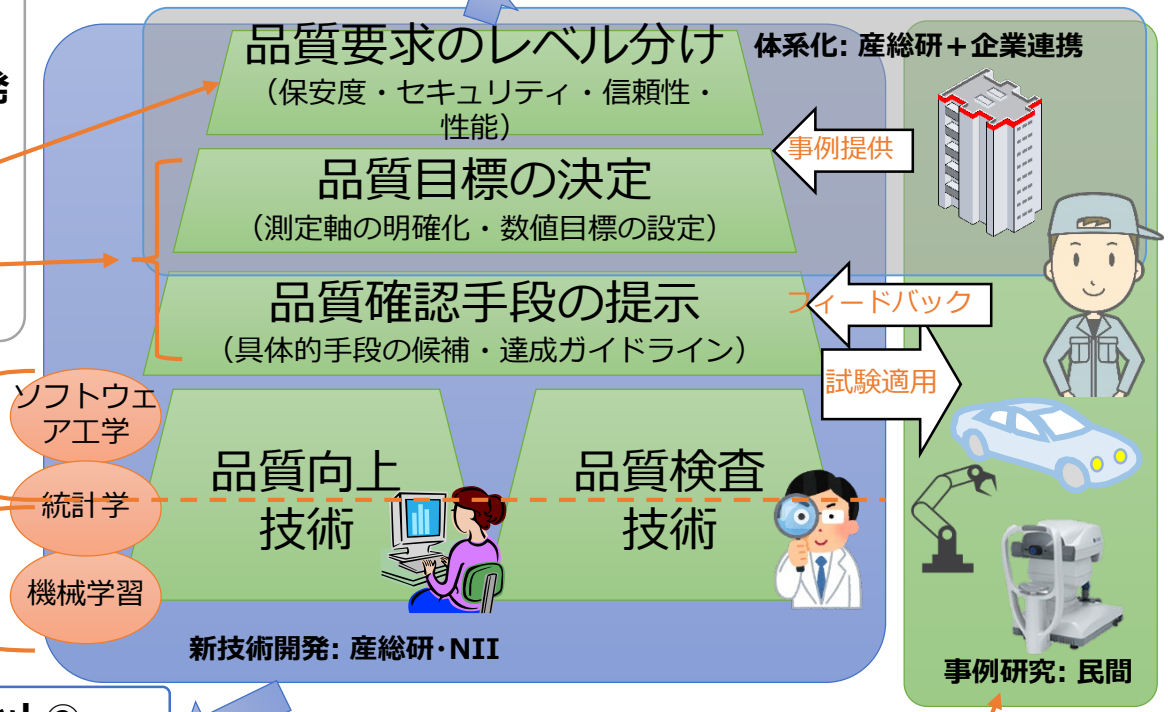
実施項目1:
品質要件の明確化に関する研究と
品質保証エコシステム（保証プロセス）の開発

1-1: AI利用製品の品質要件の明確化と
レベル分けに関する研究
1-2: AI利用製品の品質保証のための
実装プロセスに関する研究

AIソフトウェアに特化した
具体的な品質保証技術

実施項目2: AIの品質を実装時・検査時・
実用時それぞれで管理し担保する
技術の研究開発（短期）

実施項目3: 高品質AIシステムのための
AI基礎技術の研究開発(中長期)



アウトプット②:
具体的技術・ツール

実施項目4: 製品レベルでの品質保証実証研究
グランドチャレンジを設定し、実応用を研究として実施

アウトカム③:
産業界の技術力・
国際競争力強化

- 社会基盤としての基準作りと、それを実現する技術をセットで並行開発
- 産業界との密接な連携
  - 〔先導研究から〕企業との連携による実事例分析・基準作りへの反映
  - 〔本格研究から〕実事例での実際の技術適用とフィードバック

AI品質保証プロジェクト

- 研究体制 （NEDOプロジェクト）
  - 産総研
    - サイバーフィジカルセキュリティ研究センター
    - 人工知能研究センター・人工知能研究戦略部
    - ロボットイノベーション研究センター
  - 国立情報学研究所
  - 民間企業
    - AI品質マネジメント検討委員会
    - 具体的なグランドチャレンジ分野の設定と実応用の製品開発を通じた技術開発

予定されるアウトプット

- AI/CPS 品質保証ガイドライン・基準
  - 品質要件ごとのレベル分けの標準
  - 品質要求レベルごとの開発プロセスガイド
    - 具体的な品質測定軸と数値目標
    - 品質管理プロセスへの要求事項
  - 具体的な分析方法の候補リスト・実用ガイド

＋

- 品質改善・リスク低減のための具体的技術

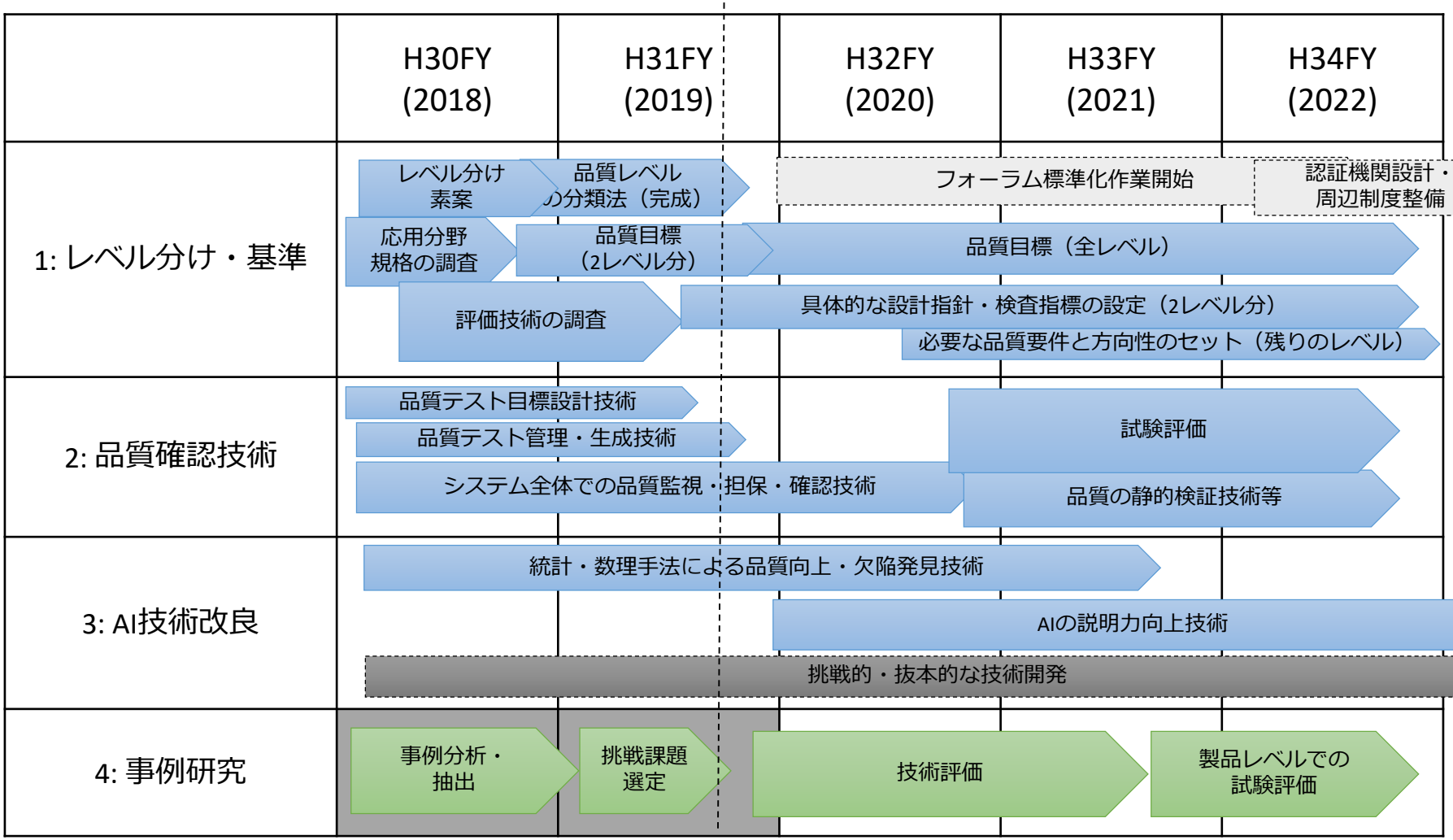
開発する技術例

- リスク分析の精度向上
  - CPS応用でのリスク分析の標準手順・テンプレート
    - CPSリスクのバリエーションの洗い出し手法
    - CPSリスク分析の品質評価（手順 or 技術）
- 学習元データの品質確保の基準と技術
  - データの集め方・データ量
  - データのクレンジング・教示の準備方法
  - データの品質・網羅性のチェック技術

開発する技術例

- AI & CPS 向けの品質検査・管理技術
  - ソフトウェア工学的なアプローチでの技術開発
  - 学習による実装に適したソフトウェアテスト技法
  - 学習パラメータと実装をセットにした静的検査
    - ニューロネットワークのモデル検査
    - 出力パラメータの安定性・整合性の検査
- AI技術そのものの改良
  - 安定性の高い (＝セキュリティリスクの少ない) 学習モデル
  - 学習結果の品質に結びつく「習熟度」の指標化
  - 学習結果からの「説明」の導出 (Explainable AI)

研究開発スケジュール



来年中にガイドライン草案、4年程度で実用化 ＋ 長期的・根本的なAI技術改良もセット
認証・標準化を視野に入れた活動を展開

まとめ

- CPS & AI の発展には
  ソフトウェアの安全性等への不安を
  取り除くことが必要である。
- 新しいソフトウェア応用には、
  新しい安全性・セキュリティなどの
  品質向上のための技術が必要となる。

- CPS & AI のための品質保証に関する
  研究開発に熱く取り組んでいきます。