資料7

規制の精緻化に向けた
「ソフトウエアの類型化」の枠組み

横断分野:　上野山　勝也

議論の背景にある論点

（論点１）機械学習を用いたAIについては、ルールベースで記述されるシステムとは異なり、与えられたデータに基づく統計的な処理を行うものであることから、

- 「予測可能性や安全性の確保を達成することが従来のシステムに比して、必ずしも容易ではない」
  - AIは、データの入力によって学習を積み重ねることによって継続的に品質向上を図るものであり、随時アルゴリズムの内容が変わっていく。
  - ディープラーニングを用いたAIについては、どのような要素が影響して一定の判断が下されたかを事後的に検証することが、事実上困難（アルゴリズムのブラックボックス化）

（論点２）これらの課題を踏まえた上で社会受容性を確保しつつ、AIを社会実装していくためには、次のような対応が必要となること

- （未来投資会議における実証事業を含め、ユースケースをもとに横断的施策を整備することが必要）。
  - AIの信頼性評価に関する基準・ガイドライン・ガイドブック等の策定（例）
    - AIが利用される場面のシリアスさに応じた「リスク回避性」「AIパフォーマンス」「公平性」等の要求レベルの整理
  - （ベンチャー企業等が）AIの信頼性を実証できるテストベッド/技術検証基盤の開発（注：規制省庁のキャパビルを含む）
  - アップデートごとの認証ではなく、アップデートのプロセスに着目した認証制度の在り方

横断分野:
規制の精緻化に向けた 「ソフトウエアの類型」と「ソフトウエア評価法」

特に、ソフトウエアの審査における対応方法は？

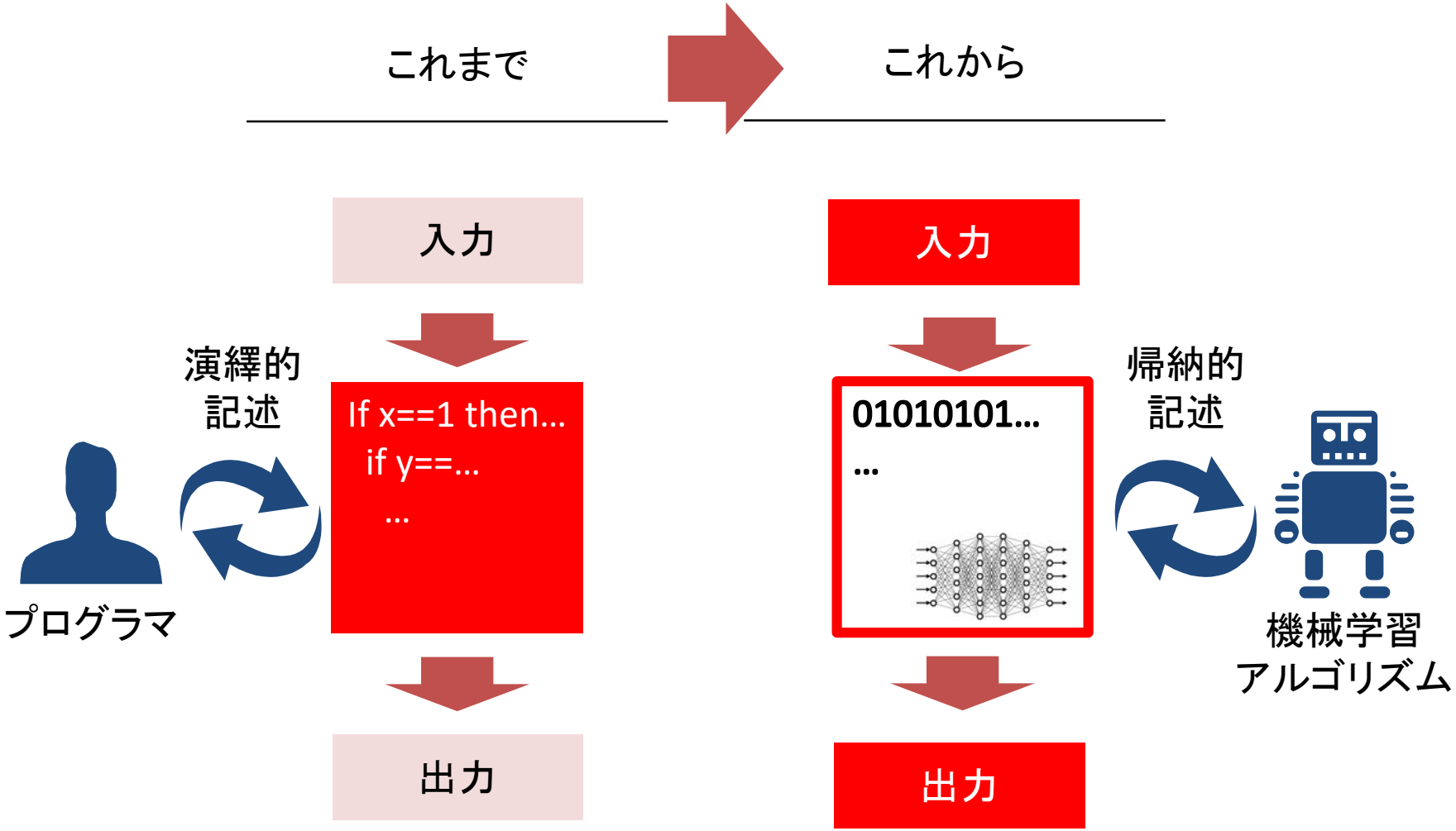
- ①信頼性評価/②公平性/③解釈性/④有事の責任の所在 をどのように、
  認可方法/ガイドライン に反映させるべきか？

本日

– **論点A： 対応方法を、何で、どのように、変えるべきか？（本日）**
  - X： どのような挙動のソフトウエアを？（これまでのソフトウエア, 機械学習, AI etc）
  - Y： どのような運用方法で？（Static or Dynamic etc）
  - Z： どのような領域に適応するのか？

– **論点B： ソフトウエアの評価の実現可能性（技術観点から）**
  - 品質/解釈性/公平性 etc

旧来型のソフトウエアと
新たなソフトウエア(データ循環型ソフトウエア) の違い



X： 規制の精緻化に向けたソフトウエアの類型(案)

- ゼロベースで考えるのではなく、これまでのソフトウエアと知能化周辺のソフトウエアの違いは何か？ への着目を推奨

タイプ０：これまでのソフトウエア：

- INPUTとOUTPUTが**１：１** 対応

タイプ１：機械学習型ソフトウエア ：

- INPUTとOUTPUTが**１：１** 対応
- 解釈性/公平性のみが論点

タイプ２：機械学習型ソフトウエア – 確率的出力

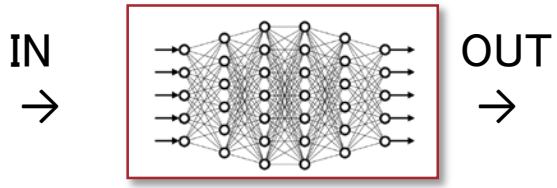
- INPUTとOUTPUTが**１：N** 対応
- 上記に加え、品質評価も論点

タイプ３：機械学習型ソフトウエア–自律運動あり

- INPUTとOUTPUTが **N：M** 対応
- 上記に加え、責任の所在も論点

Y：ソフトウエアのアップデートの方法

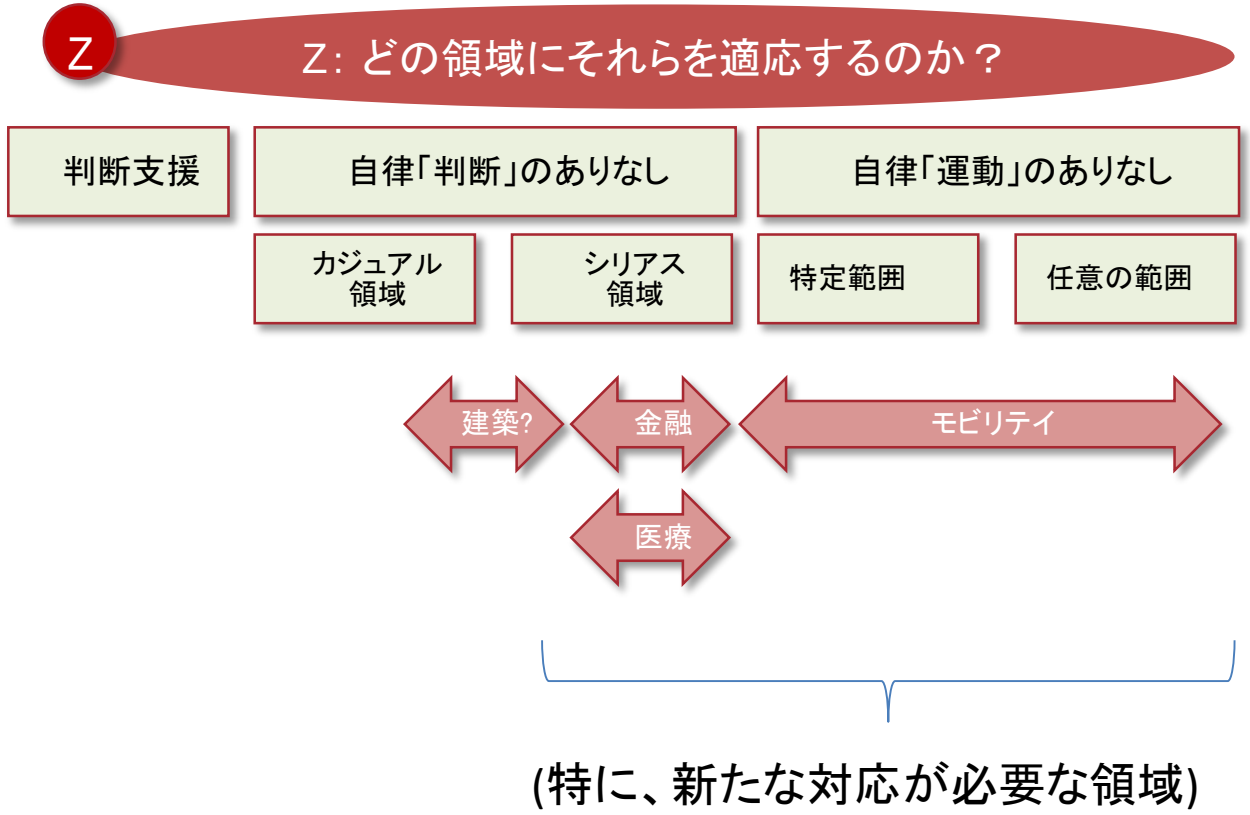
- 下記のどのタイプかにより、認可の方法や、必要なガイドラインが異なりそう



1. アップデートなし：
   - ローンチ後、ソフトウエアは不変
2. アップデートあり：特定のタイミングにおいてアップデートが走る
   - 一斉にアップデートさせる (サーバーサイドから or 他の方法)
3. アップデートあり： 自律的にアップデートが走る
   - 書き換わる対象
     - 3-1: パラメーターのみが書き換わる
     - 3-2: モデル全体が書き換わる

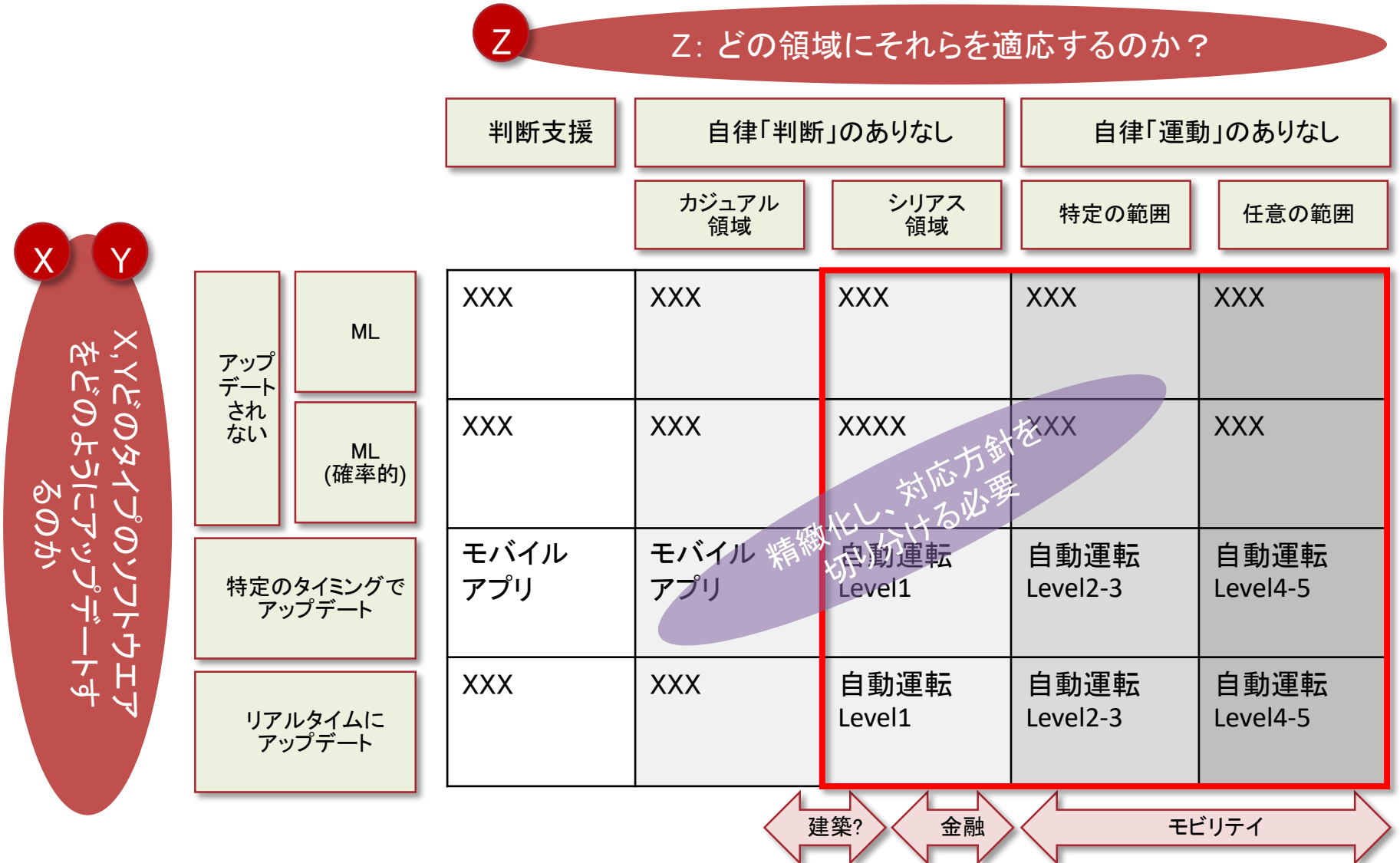
Z： どのような領域に適応するのか？

- これまでのソフトウエアのほとんどは、「自律判断」と「自律運動」を行わない
- 行う場合も、シリアスな判断や、特定領域での自律運動であった



(特に、新たな対応が必要な領域)

X:Y:Zの枠組みで見たときに対応方法はどう変わるか？
：①信頼性評価/②公平性/③解釈性/④有事の責任の所在 をどのように、認可方法/ガイドライン に反映させるべきか？



まとめ

- 前提として、既存のソフトウエアと新たなソフトウエアの差分に着目した検討を推奨
- 規制の精緻化に向けた「ソフトウエアの類型を切り分ける枠組み(X,Y,Z)」を提示
  - X: ソフトウエアの仕組み：演繹的記述 OR 機能的技術、1:1の出力 OR 確率的出力
  - Y: アップデートの方法：なし OR 特定タイミングにアップデート OR リアルタイム
  - Z: 適応領域：　自律判断の有無と自律運動の有無、カジュアル or シリアス etc
- 上記枠組みを一つのフレームワークとして、①信頼性評価/②公平性/③解釈性/④有事の責任の所在 をどのように、認可方法/ガイドライン に反映させるべきか？を精緻化することは有用そう
- 実際にどのようなソフトウエア評価 (信頼性評価 等) が可能か、は別途詳細化が必要