AI品質管理　グループ討議　結果報告

フィジカルAIの品質管理

「どこまでの範囲なら安全に作れるか」を決めて，その範囲で作る

（1）
議論する範囲と課題

（2）
商品化の条件と評価軸

（3）
今後のリスクと低減方策

「フィジカルAI」担当グループ　2026年8月７日～8日

観点（1）

議論する範囲と，議論したい課題

今回，議論する範囲

- **人と同じ空間で物理的に作業するAI**（家事，厨房，介護など）を対象とする
- 「安全に作ると受容がないもの」になりやすい．**どこまでの範囲なら安全に作れるかを見極め，その範囲で作る**，という順序で考える
- **人や動物との相互作用も視野に入れる**．既存の家事ベンチマークはここを対象外としている
- 対象は製品としての成立可否．PoC/B2B/B2Cで満たすべきレベルに幅がある

本日の問い

- **どこまでの範囲なら，安全**に作れるか
- その範囲を，**どの軸で表現**するか
- 範囲外に残るリスクを，**誰がどう引き受ける**か

① 「どこまでしてよいか」の基準がない

- 何をしたら不安全か，という**危険源のリスト**が未整備
- 火も使わず，尖ってもいないが，**やってはいけないこと**がある
- 例：皿が割れると鋭い縁ができる，といった**二次的な危険源も**抜けている

② 評価がタスク単位に留まっている

- ベンチマークはタスクを1個ずつしかチェックできていない
- **マルチタスク，ロングホライズンの評価が不足**している
- タスクとしては成功，でも**不安全**な場合（SafeVLA-Bench の問題提起）

③ 暗黙知をどう獲得し，どう検証するか

- 人にはできるが，**やり方を説明できない領域が残る**
- 深層学習はゴールだけを与え，**やり方は教えずに学習させることは可能**
- ルールベースで表現できるのか（自動運転車では？　強化学習では？）

参照：Stanford BEHAVIOR（家庭内活動を詳細に分類したベンチマーク．task と failure mode の整理に利用可能）／SafeVLA-Bench

観点（2）

商品化できる条件と，品質評価の軸

残存リスクが「社会受容性」の範囲に収まれば商品化できる．安全側に振り過ぎると，受容される製品（需要）にならない

商品化の条件
（メーカーとしての免責範囲／条件）

- **何が起きるか分からないものは製品として成立しにくい**.
  医療なら成立し得るが，家事向けは難しい.
- **B2Cでは使用法に制限を設けにくい**.
  「子供を近づけないでください」は成立しない.
- **判断根拠を明文化**していれば，精度75%でも使える場面がある．無ければ使えない.
- **フェイルオペレーション／フォールバック設計**.
  AI抜きでも従来から議論されている論点.

品質評価の軸の候補

「どこまで」の軸

- **自律性**（SAEの自動運転レベル相当）と人の介入範囲
- **対応幅**（versatility）＝タスク対応の広さ
- **安全要求レベル**（AISL 等）
- **社会受容の段階**：PoC → B2B（専門家）→ B2C

「対策方法」の軸

- **本質安全 /Simplexアーキテクチャ**（従来方式とAI の切替）
- 人側の管理（運用でカバー：**Human Oversight/HITL**）
- **説明可能性**（なぜそうしたのか，何をしたいのか）
- モード分け（人が近づいたら止まる 等）

人が接しない → **資格のある人しか接しない** → **一般人が接する（B2C）**

右へ行くほど使用条件を限定できず，満たすべきレベルが上がる．キッチンロボットも介護ロボットも，業務用か家庭用かで要求が変わる

「対応幅」の参照系として，Stanford BEHAVIOR の家事タスク分類（1,000活動，タスクと危険源の対応づけ）が使える.

観点（3）

今後のフィジカルAIのリスクと，リスク低減方策

1 暗黙知の獲得

- 介護：人をベッドから車いすへ移す動作はマルチタスク.
  被介護者の姿勢制御，体の向きの変更，負荷低減，時間効率.
- 強化学習の前に，ベテランの振舞から報酬関数を学ばせる(逆強化学習).
  模倣ではなく背景の戦略を学習．生成AIでの獲得にも期待.
- 正例と負例，事故対応のシナリオも用意する（機転が効くように）.
  （レジリンスの強化）

2 評価項目の漏れと汎用性

- 本当に必要な項目がリストに入っているか，を先に問う
- ロバスト性：予定の行動がうまくいかない時に，気づけるか／対処を考案できるか
- 同時並行の作業（ソースとパスタを同時にゆでる）も評価の対象に含める

3 unknown-unknown と残存リスク

- テストは特定のケースしかできない．その近傍/隙間のケースが大丈夫か？
- 場合分け → 各場合の割合の見積り → その上でテスト．モニタユーザで多様性を確保.
- 「何億km走行しても起きない」という主張には uk-uk が含まれる.
  リスクの大きさで絞り込む.

4 人共存の安全と，時間の概念

- 人や動物との相互作用そのものを安全の対象にする
  （既存のベンチマークは対象外としている部分）
- 時間の概念：シーケンス(いつどの順でどれだけ待つか)．長時間の連続評価でも残るリスク．人との時間の捉え方の違いによるリスク
  →時間制約パラメータ制御または学習が必要
- モード分け(人が近づいたら止まる)と，AIの役割設計で議論が変わる

国際標準化/ガイドラインの役割・位置づけ

- 危険源リスト，評価軸，自律性レベルといった「どこまで」の共通言語を用意する場として位置づける
- 医療安全の国際標準（JCI 認定），院内の医療の質/安全管理部門でのインシデント＆対処情報の蓄積は，第三者ガバナンスの参照モデルになる
- コツが明示知化されると，うまくいく．一方でベテランは必ずしも手順どおりに動かない（レジリエンス・エンジニアリング）.

フィジカルAIに対する期待は，いい方向での創発性につながる．人との相互作用で互いに高め合う共進化の効果も.