

JAI-Trust

2026-09-08
Satoshi Sekine

AI安全性の評価のためにはベンチマークが必要

- ◆ 本プロジェクトはAI開発者・研究者を中心としたボトムアップな活動
 - NII-LLMC主催の安全性シンポジウム(2025夏)での議論
 - AI開発者・研究者がコミュニティーベースのベンチマークの必要性を共鳴
 - 2025秋に開始（約60名が発足ミーティングに参加）
 - 2025年度は主にNII-LLMCの予算で推進
- ◆ AISIの責務であるAI安全性の評価とマッチ
 - 2026年度からはAISIの旗の下で実施
 - 本会議は半年間の活動の披露の場であり、これからの方向性を形作っていく場

なぜボトムアップなAI安全性ベンチマークが必要か、どう作るか

- ◆ 特定の組織が独自の基準を定めるだけでは不十分
 - 自らのモデルを自らが評価することが困難、または、不誠実
 - 多様な応用分野で十分な信頼性基準を確保しにくい
- ◆ 様々な参加者が納得できる安全性の定義・評価・実践を構築する
 - コミュニティーでの構築が必要
- ◆ 産官学の知見を持ち寄りながら、LLMの安全性ベンチマークを協働で構築
 - ベンチマークに対する様々な視点

開発者の立場から

(All Japan / One Teamで)

1 団体が規定した安全性ではなく、
コミュニティとしての基準

抽象的なガイドラインではなく
具体的な基準・データ・ツール

具体的なLLMの安全性の評価基準を構築し

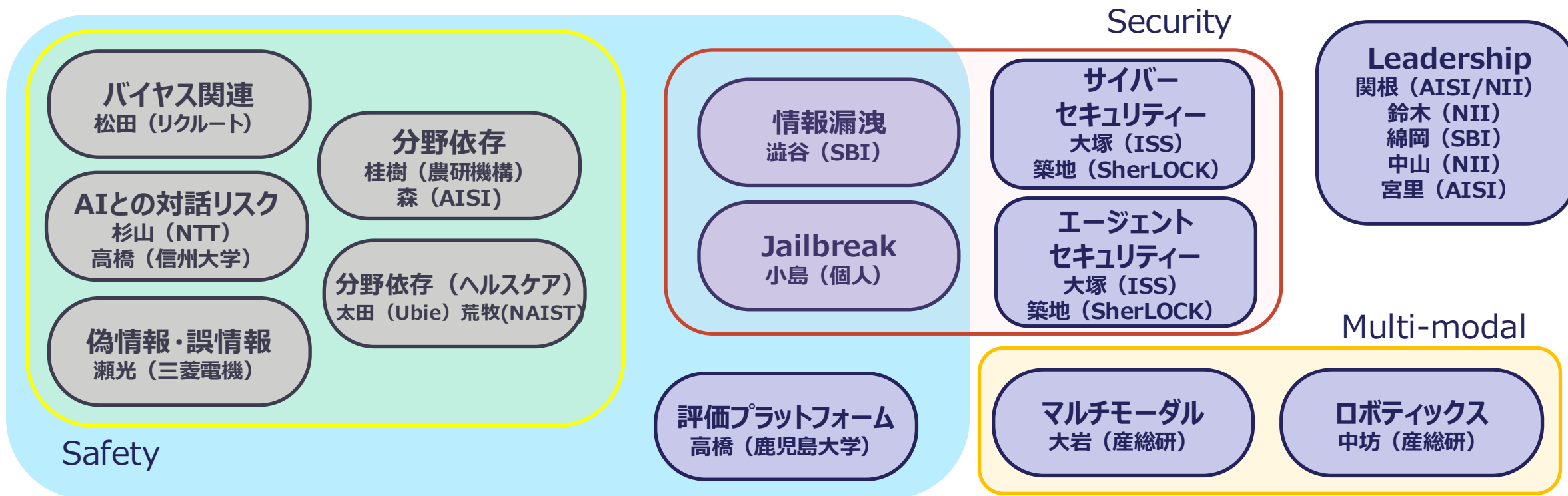
(ベンチマーク / 評価基準 / 評価ツールを構築 & 提供し)

世に問う

安全性の定義は押し付けられない
最終的に世間の合意が必要

- ◆ ユーザーが生成AIを安心安全に利用できるようにする
- ◆ 「Safety & Security」の全般を対象
- ◆ 利用場面、内容、対策の個別化により分類し、分科会形式で運営
- ◆ 参加者は基本ボランティア
 - 企業、アカデミア、官庁など様々な方が参加中
 - 活発に活動する方のみのslackに154名が参加（2026/9/2現在）
 - 予算は、計算資源、データ構築、ツール作成に活用

JAI-Trustは、AIの安全性評価のベンチマークを構築し評価を行う
「物差しを作り」「自ら評価できる能力を持つ」



産官学の協力の下、One Team/All Japanのコミュニティー。100名以上の参加者。12の分科会。

2025年10月にNII安全性WGで誕生。2026年4月にAISIIに移管
コミュニティベースで、実際に手を動かす人だけがslackに参加

- ◆ 月1回の運営会議
- ◆ 各分科会は1～3週に1回のミーティングとSlack上での議論
- ◆ **2026年5月21日にお披露目会（350名が参加。写真）**
- ◆ 2026年7月30日に運営者全体会議 & 懇親会（対面）
- ◆ 2027年3月に最初の評価を実施することを目標として活動
 - ・ 評価対象の同定、評価指標の設計、ベンチマークデータ構築、評価プラットフォーム構築



[AISII HPでの報告](#)



jai-trust.org



AISI

Japan AI Safety Institute