

3rd Grand Canvas ショートトーク 「説明」はAI品質として管理できるか

2024年11月19日

情報処理推進機構

セキュリティセンター 専門委員

小川 隆一



Who am I?

AIよりは認知科学
に惹かれる

IPA

第二次AIブーム
を体験

小川 隆一 (r-ogawa@ipa.go.jp)

1983－2015 某ITベンダー研究所在籍

画像DB・ハイパーメディア・システムセキュリティ・
クラウド標準化・ユーザブルセキュリティなど

2016－2020 IPA ISEC 分析G GL

白書・セキュリティ調査分析・AIトラストなど

2021－現在 ISEC シニアエキスパート 退職 ⇒ 専門委員

白書・AIトラスト・AIセキュリティ

学会 情報処理学会・電子情報通信学会・人口知能学会・情報ネット
ワーク法学会 他



- 出発点：AIは学習不備により間違ふ。人は間違ふAIをどう信頼するか？評価にユーザの関与が必要ではないか？
⇒ 検知・防御が100%にならないセキュリティと同じ課題
Ex. データセットで検知精度95%、と言われたユーザはツール導入をどう判断すればいいか？
- 調査アプローチ
AI品質とAI利用行動の関係をモデル化、アンケートで仮説検証
情報処理学会・IISSEC等で成果発表、白書で啓発

AIセキュリティ調査活動（1）

- 当方担当：AI利用で先行する**米国の脅威・対策実態**を調査

2024年度調査

<https://www.ipa.go.jp/security/reports/technicalwatch/20241223.html>

米国文献調査・有識者インタビュー

選挙への脅威

AIガバナンス政策の有効性（AI-RMF, バイデンEO・・・）

ヒューマンインタフェース系国際会議（HCII2024）

利用者の**AI受容モデル化**に関する発表

最新**利用時品質**規格（ISO/IEC 25019：2023）を活用

AIと人間の協調に関わる研究動向把握

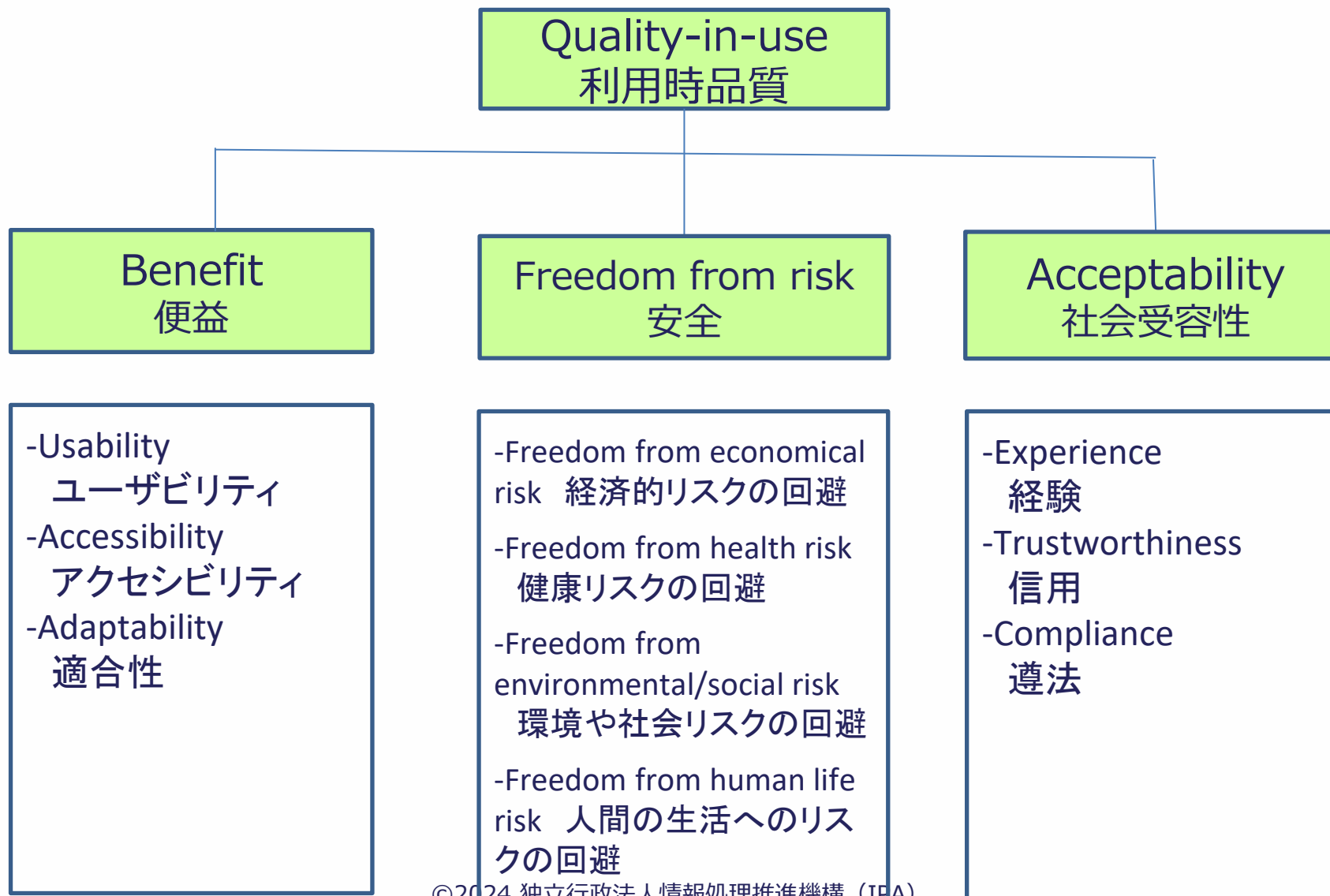
利用に関する**説明とAI依存**に注目

AIセキュリティ調査活動（2）

- 説明の重要性に関する示唆
 - 利用者の誤用抑止には**ベンダーの説明**が重要（有識者の意見）
 - 説明すれば**理解が深まる**とは限らない。説明の仕方によっては利用者は理解を放棄し、**何でもいわれたとおりにする（AIに依存する）**可能性がある（HCII2024の知見）
- ↓
- 説明のよさを**どう測れば**いいだろうか？
 - **説明品質**（利用者が理解し、実行できるか等）を管理することは可能だろうか？

ご意見を伺えれば幸いです

ご参考：利用時品質モデル (ISO/IEC 25019:2023)



- デジタル社会のトラストは多くの分野で検討され、定義は多様
- セキュリティ的な視点では二つの面がある
 - 1 情報の**真正性**・システムの**無謬性**の保証（既存方法で測定）
 - 2 製品のセキュリティ品質を人が**受容**し利用（受容の測定は難題）

「人はAIをどう受け入れるか（使ってもいいと思うか）」

測定方法を試行中 ⇒ 安全 から 安心 への橋渡し

Thank you!

IPA