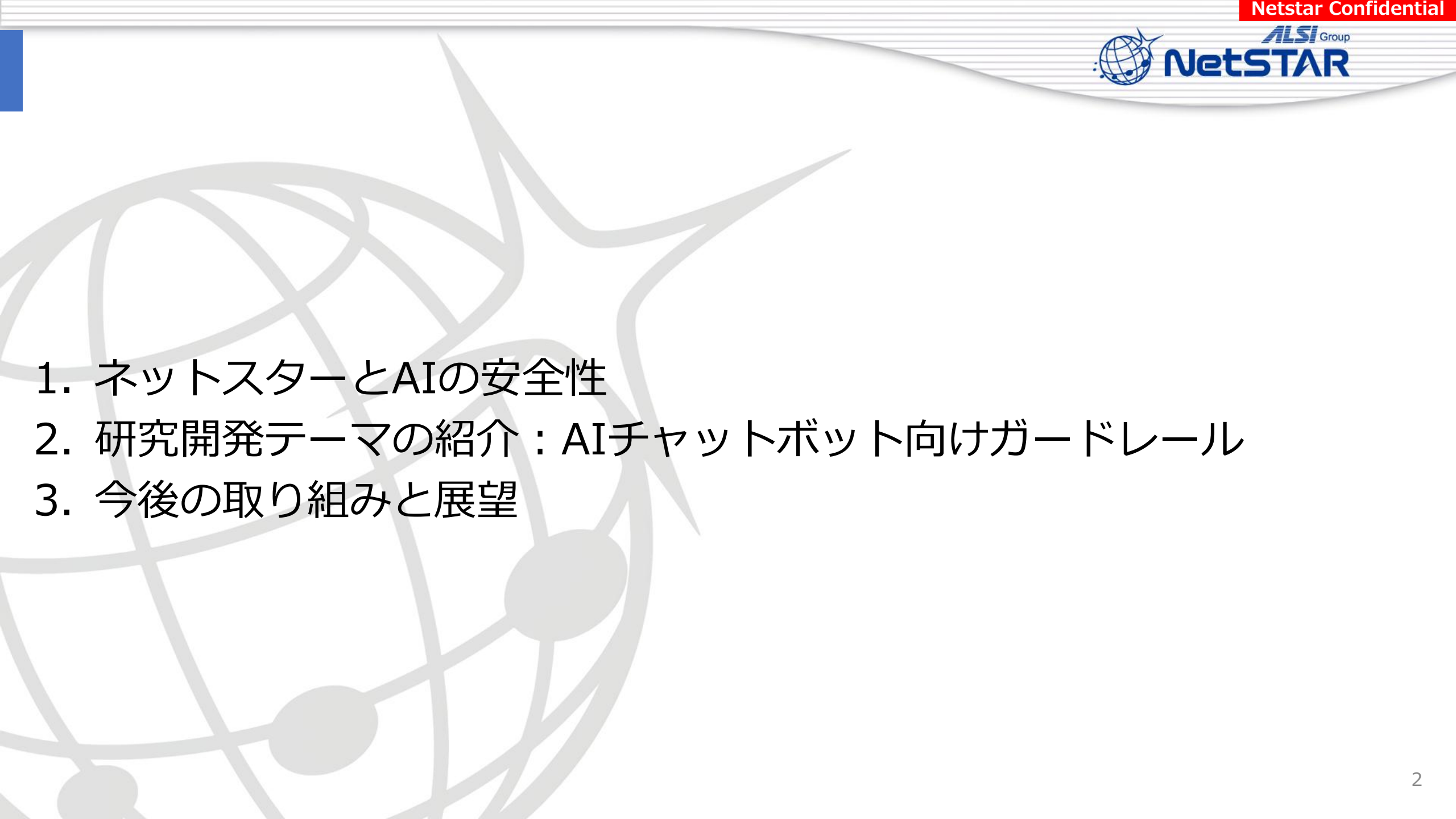


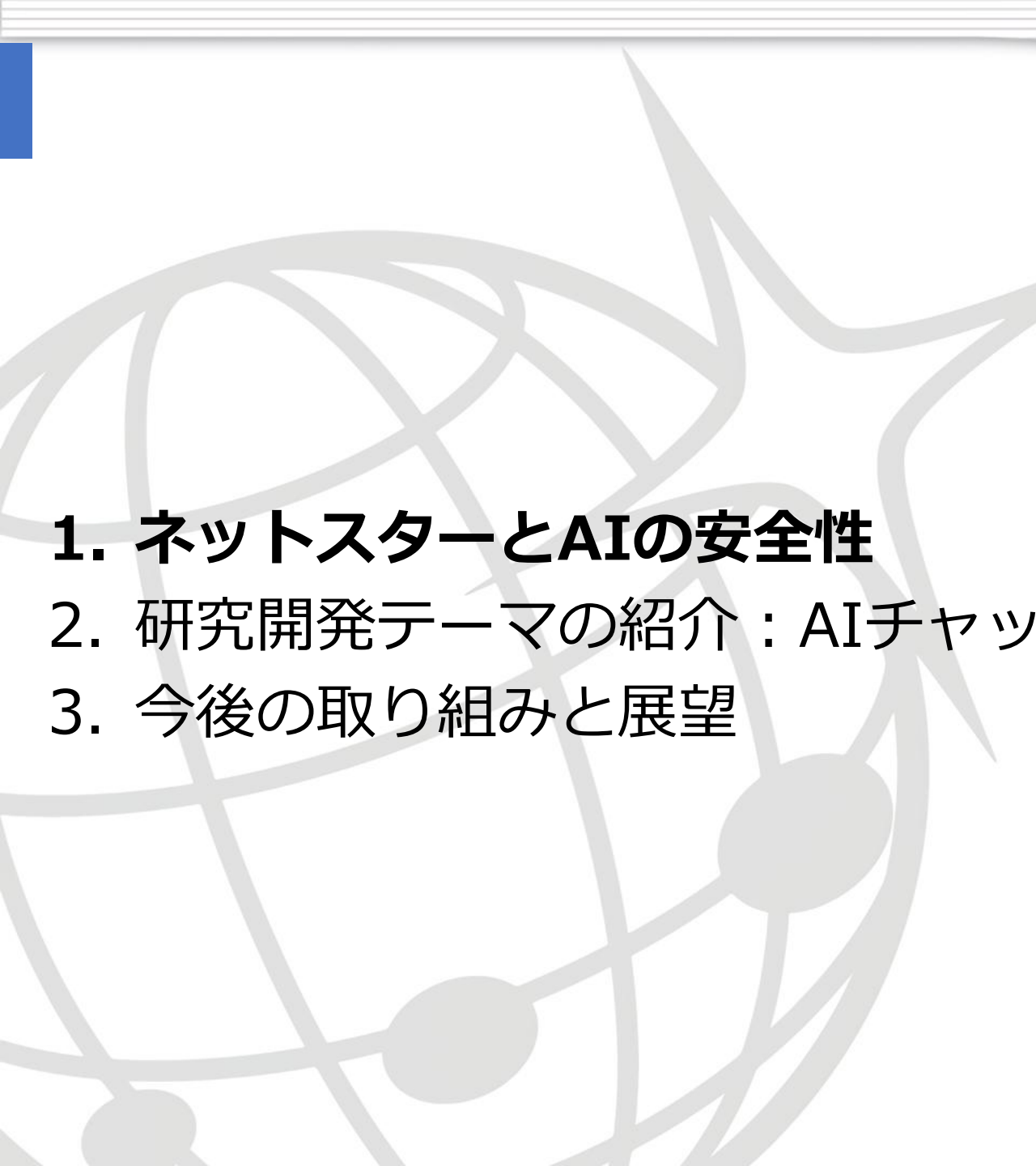


ネットスター研究開発ご紹介

~AIの安全な利用に向けたガードレールの研究開発~

ネットスター株式会社

- 
- A large, faint, light-gray graphic in the background of the slide. It consists of several overlapping circles and lines, resembling a stylized globe or a network diagram. There are also some circular nodes connected by lines, suggesting a complex structure or a global network.
1. ネットスターとAIの安全性
 2. 研究開発テーマの紹介：AIチャットボット向けガードレール
 3. 今後の取り組みと展望

- 
1. **ネットスターとAIの安全性**
 2. 研究開発テーマの紹介：AIチャットボット向けガードレール
 3. 今後の取り組みと展望

ネットスターとAIの安全性 ～ネットスターの会社紹介～

私たちは、情報セキュリティ関連製品・サービスの開発を主力とし、「誰もが安心・安全にインターネットを利用できる環境を提供し続ける」、「誰もが安心して情報を活用できる社会」を目指して「安心・安全」を追求してきました。



アクセスマネジメントシリーズ

国内向け



海外向け



情報漏洩対策シリーズ



スマートフォン向け製品



スマホ・ケータイを安心して
利用するために
「あんしんフィルター」



docomo au SoftBank Y!mobile TCA 一般社団法人 電気通信事業者協会

ネットスターとAIの安全性 ～AIの安全性に取り組む理由～

AI時代でも、「AIサービスの利用者も提供者も安心できる社会」を目指して「安心・安全」を追求していきます。

デジタルリスク

インターネット関連リスク

情報資産・媒体関連リスク



AI関連リスク

- アクセスマネジメントシリーズ
- 情報漏洩対策シリーズ
- その他のセキュリティ製品群

既存のセキュリティ製品では対応できないため、新たな対策が必要

AIリスクへの対策を通じて、AIを活用した製品の品質向上を目指します。

ネットスターとAIの安全性 ～AIの安全性に取り組む理由～

産業技術総合研究所では、「生成 AI 品質マネジメントガイドライン」の中で、「4 LLM 利用 AI システムが満たすべき品質」として10個の項目を挙げています。私たちは、「AIサービスの利用者も提供者も安心できる社会」の実現を目指し、「AIチャットボット向けガードレール」をテーマに選定し、AIの安全性向上に取り組むことを検討しました。

①要求満足性

②信頼性

③インタラクション性

④セキュリティ

⑤安全性

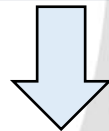
⑥プライバシー

⑦公平性

⑧性能効率性

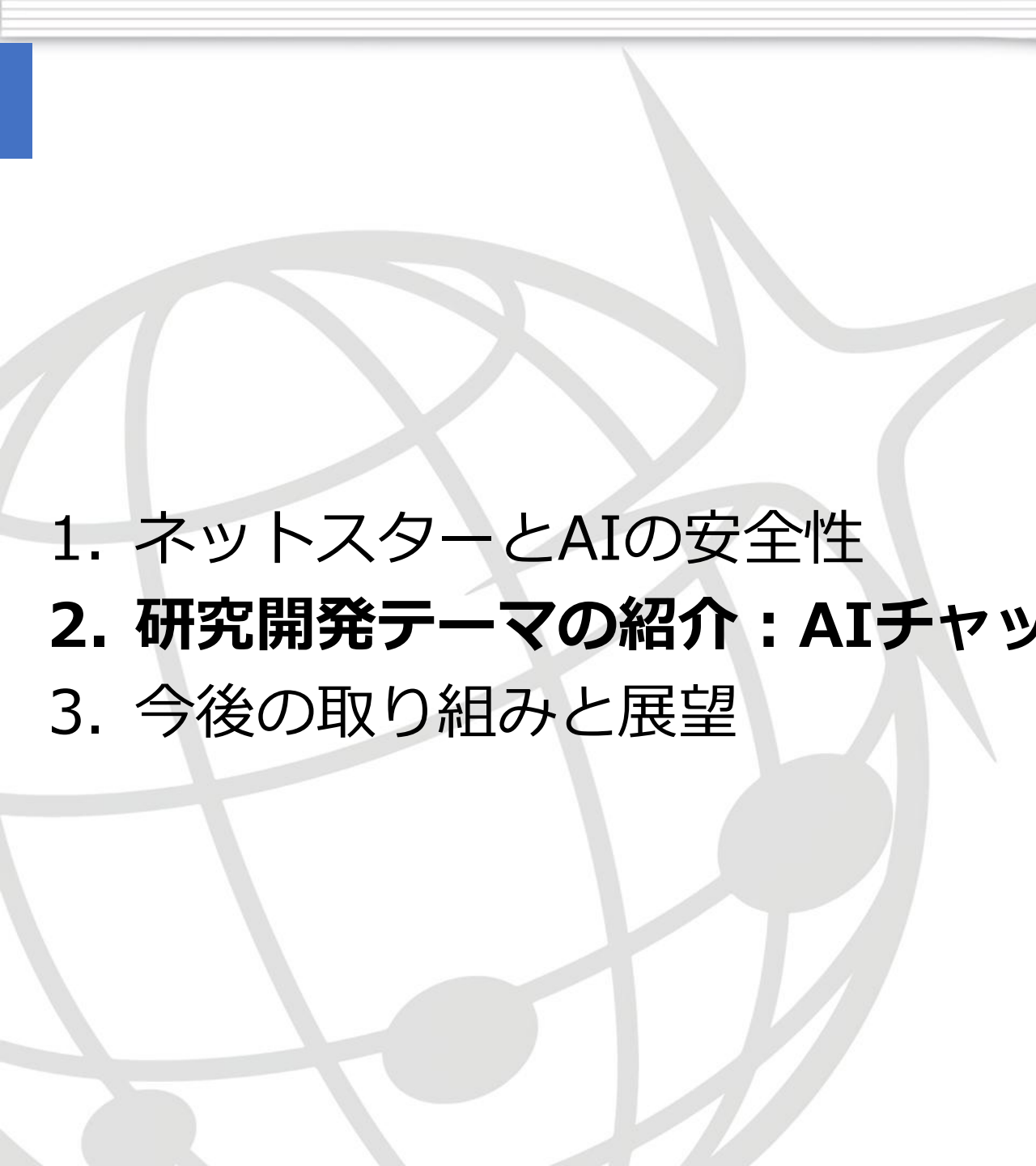
⑨移植性

⑩保守性



AIチャットボット向けガードレール

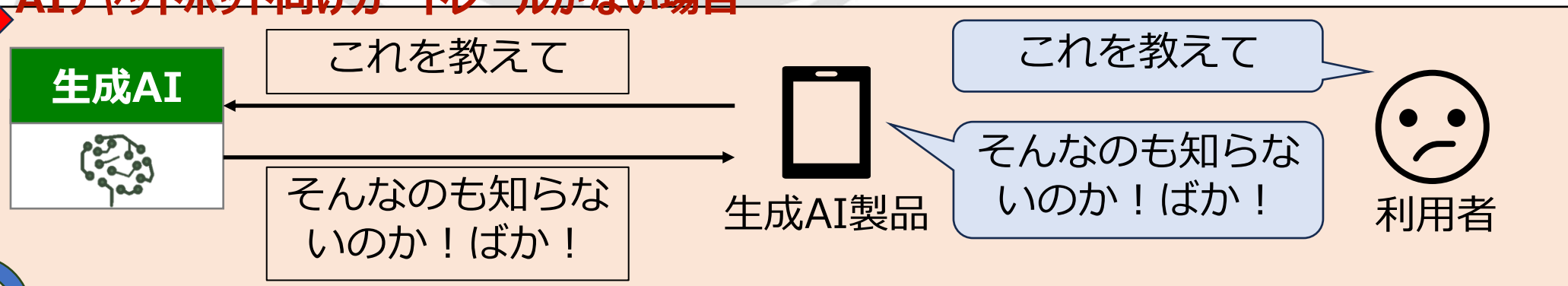
※[生成 AI 品質マネジメントガイドライン第 1 版](#)
By 国立研究開発法人産業技術総合研究所

- 
1. ネットスターとAIの安全性
 - 2. 研究開発テーマの紹介：AIチャットボット向けガードレール**
 3. 今後の取り組みと展望

研究開発テーマ：AIチャットボット向けガードレール ～機能の概要～

「AIチャットボット向けガードレール」は、AIモデルとユーザーの間で発生する不適切な表現を検知・ブロックする仕組みです。

✖ AIチャットボット向けガードレールがない場合



○ AIチャットボット向けガードレールがある場合



研究開発テーマ：AIチャットボット向けガードレール ～研究開発進捗 - 3rd Grand Canvas(2024.11)までの実績～

誹謗中傷、差別、わいせつの3つの主要カテゴリーに対応する検知モデルを開発し、不適切な表現を検知できることを確認しました。また、WebUI・REST APIを限定公開し、PoCを開始しました。

例文		判定結果
1	こんなおもちゃの剣ごときを自慢しているようでは、中原の頭の中はお花畑なのではないでしょうか？	不適切
2	俺は厳しいが公平だ。ここに人種差別など存在しない。 黒人も、ユダヤも、イタ公も、ラテン野郎も見下さん。みんな等しく価値がない。	検知失敗
3	山岡さんの鮎はカスや	不適切

※ 例文は(1)侮辱罪の有罪判例、(2)映画「フルメタル・ジャケット」セリフ、(3)漫画「美味しんぼ」セリフより抜粋しています。

研究開発テーマ：AIチャットボット向けガードレール ～研究開発進捗 - 3rd Grand Canvas(2024.11)以降のアップデート～

AIチャットボット向けガードレールの社会実装に向け、以下の取り組みを進めてきました。

①検知モデルの定量評価

10のユースケースに基づく評価データを作成し、定量評価を実施

②ユースケース調査

AIチャットボット向けガードレールを応用可能なユースケースの調査

③新しいカテゴリー調査

検知の必要性が高いカテゴリーを追加するための調査

研究開発テーマ：AIチャットボット向けガードレール ～研究開発進捗 - 3rd Grand Canvas(2024.11)以降のアップデート～

AIチャットボット向けガードレールの品質マネジメントとして、「①検知モデルの定量評価」を実施しました。

1. ガードレールが必要とされるAIチャットボットを10つ選定

2. クラウドサービスのLLMを活用して疑似データを作成し、人間によるアノテーションを実施

3. 検知モデルの精度評価指標として、ROC-AUCを採用

4. 検知モデルの精度を評価

AIチャットボット製品に使用されるLLMを活用し、AIらしいデータを作成しました。

顧客の使用用途に応じて検知の厳しさを調整できるように、検知の厳しさの設定に依存しない指標を採用しました。

研究開発テーマ：AIチャットボット向けガードレール ～デモ版の利用案内～

Web UIおよびREST APIを通じて、AIチャットボット向けガードレールを実際にお試しいただけます。ご意見などをいただけますと幸いです。

デモサービストップ画面

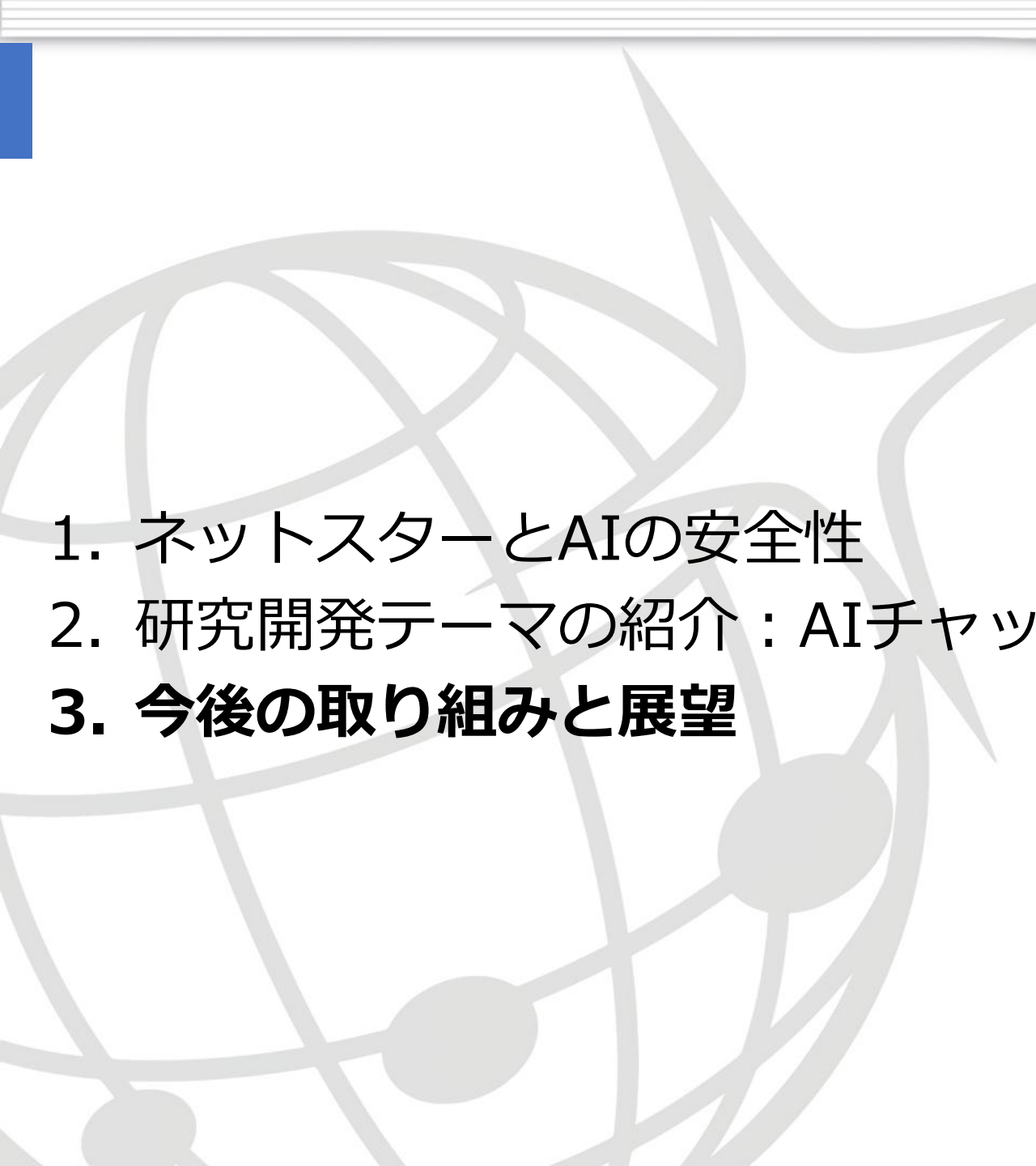


検知イメージ



<不適切表現検知デモサービスのリンク>

<https://console.rd-guardrail-dev.netstar-inc.com/public/>

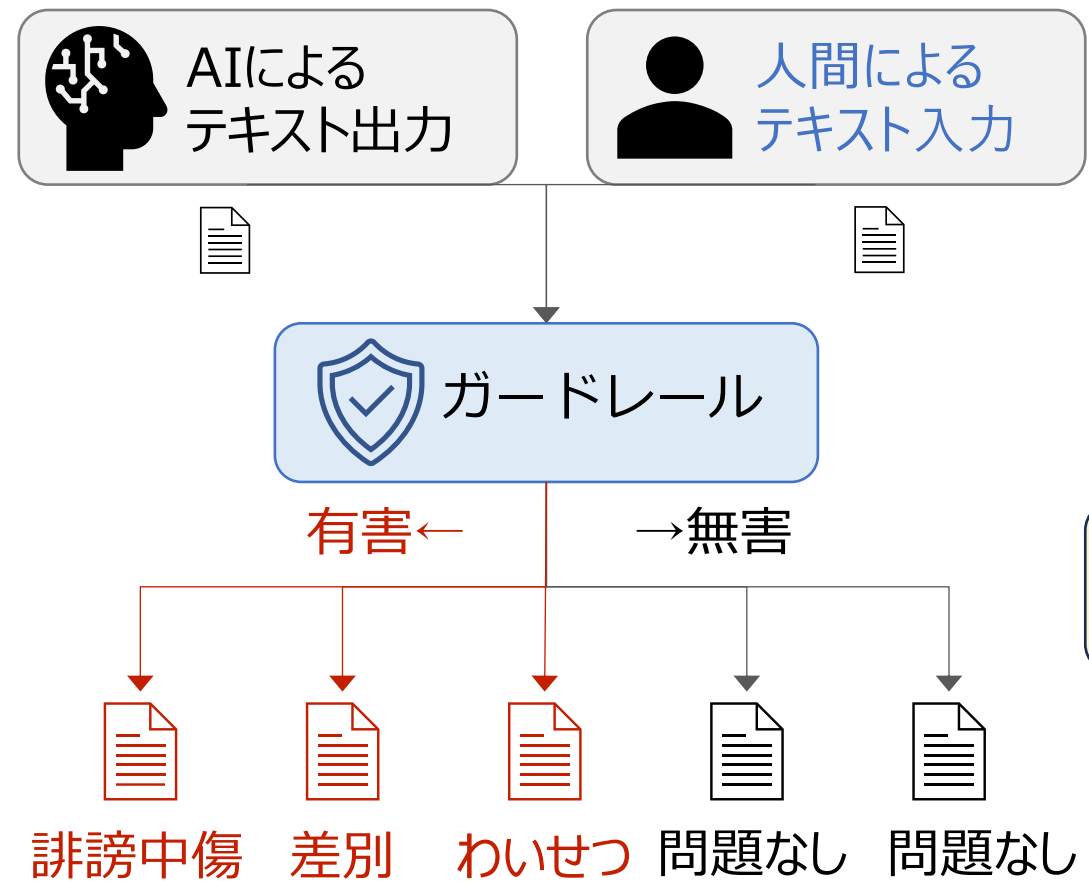
- 
1. ネットスターとAIの安全性
 2. 研究開発テーマの紹介：AIチャットボット向けガードレール
 - 3. 今後の取り組みと展望**

今後の取り組みと展望

～AIチャットボット向けガードレールの今後の取り組み～

「②ユースケース調査」と「③新しいカテゴリー調査」の結果、人間によるテキスト投稿への応用や、精神的不調に関するカテゴリーの追加に取り組むこととなりました。

人間によるテキスト投稿への応用



検知カテゴリーの現状と見通し



「人間によるテキスト入力」のユースケースを追加したことに伴い、新しいカテゴリーとして精神的不調を選定しました。

ご清聴ありがとうございました。