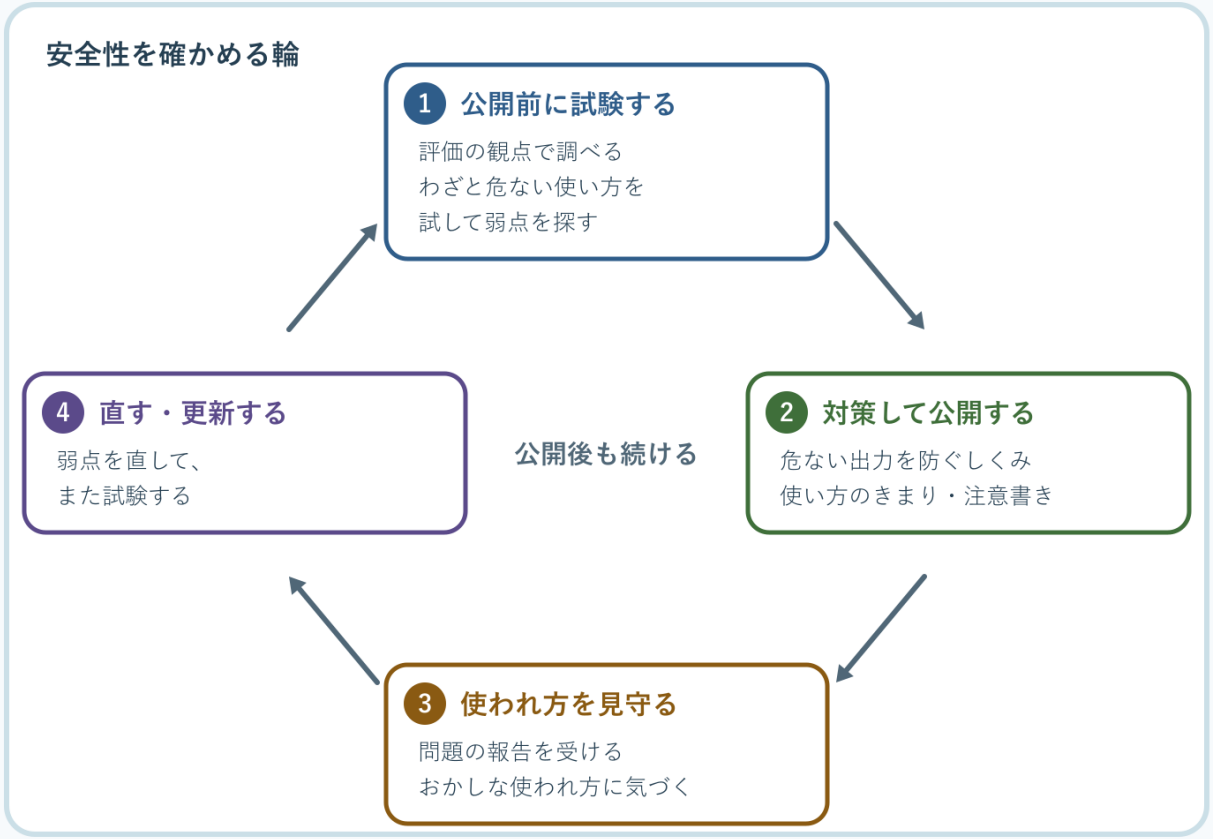
AIの安全性を確かめるしくみ

AIは、公開の前と後に、どのようにして安全性が確かめられているのだろう



だれが確かめる？

開発する組織

公開の前と後に、自分たちで
試験し、直す

外部の専門家

開発した組織とは別の立場から
試す（第三者の評価）

国の機関（日本の例）

AIセーフティ・インスティテュート
（AISI）：2024年2月に設立。
評価の観点や、わざと攻撃して試す
手法（レッドチーミング）の手引きを
公表し、ほかの国の機関と連携する

確かめる観点の例

- 有害な情報を出さないか
- うその情報・かたより
- 個人の情報を守れるか
- 乗っ取り・悪用への強さ
- 判断の理由を説明できるか

※ 模式図。確かめ方は、AIの種類や組織によってちがう。観点は例。