「人工知能の品質保証とは」

日本アイ・ビー・エム株式会社
東京基礎研究所　部長
インダストリー・ソリューション
セキュリティ＆サービス
リサーチ・スタッフ・メンバー
早稲田大学大学院　非常勤講師
細川　宣啓　氏

本日の講演タイトル「人工知能の品質保証とは？」には以下 4 通りに解釈できる曖昧さがある。

1. 人工知能で品質保証（方法の話）
2. 人工知能で品質保証なんてできるのか？（実現性の話）
3. 人工知能の品質保証はどうやってやるべきか？（要望・要求・理想レベル）
4. 人工知能の品質保証は可能か？（実現性の話）

■人工知能で品質保証

まず人工知能で品質保証が可能かという話からしたい。

例えば、「TODO というコメントが残存している」という未完成を示すバグを人工知能で検出することはできるか。

```
//     //TODO:ロギング処理など
```

というコードは「ロギング」処理が欠落していることをレビューを通じて「TODO」として指摘、つまり処理なしの処理になっており TODO の意味がない。

このプログラムの場合、何がソフトウェア欠陥なのか。”TODO”があることか、処理が空なことか、エラーハンドラになっていないことなのか?

さらに、

```
/// TODO3:フェールセーフ機能（現状異常にせず、ServiceTrace に出力するのみ、
/// 異常にする場合、ソースコード上のコメントを削除すること）
```

は、「コメントを削除すること」とは「コメントになっているところを普通の行に戻せ」なのか「コメントになっているところを行ごと削除しろ」と言っているのか二義性がある。しかも「異常にする場合」、「ソースコード上のコメントを削除すること」とあるので、異常にしない場合は削除するのかしないのかわからない。TODO を書いた人と実行した人の意識合わせをしなければならない。

こうした二義性のあるものに従うことは、誤った修正にミスリードする可能性がある潜在的なバグであり、実装した瞬間に実バグになる。実バグを生み出すタイプの TODO 文章といえる。AI は意味を理解しないので「二義性があり、潜在バグの可能性が高い」ということをどうわからせるかが問題だ。AI に単純にバグをとらせようと思っても難しい。せいぜい TODO の記述箇所を見つけるくらいしかできない。人間ならば修正順序まで大体推測できるものを AI に実装させるというのはかなり難しい領域である。

従来の品質管理は内在しているバグの数のみで保証すると思うが、コンピュータのバグには学問的な定義がない。「何が故障であり何がバグであるか」という定義が学習する元データになるが、AI では学習の対象となる悪いデータが集まっていない。

■人工知能の品質保証

人工知能搭載兵器や自動運転、医療分野などあらゆる産業面で人工知能システムの発展に伴い、パブリックセンサーやオープンなデータを検出し、上手く取り扱うことが現在の鍵になってきている。

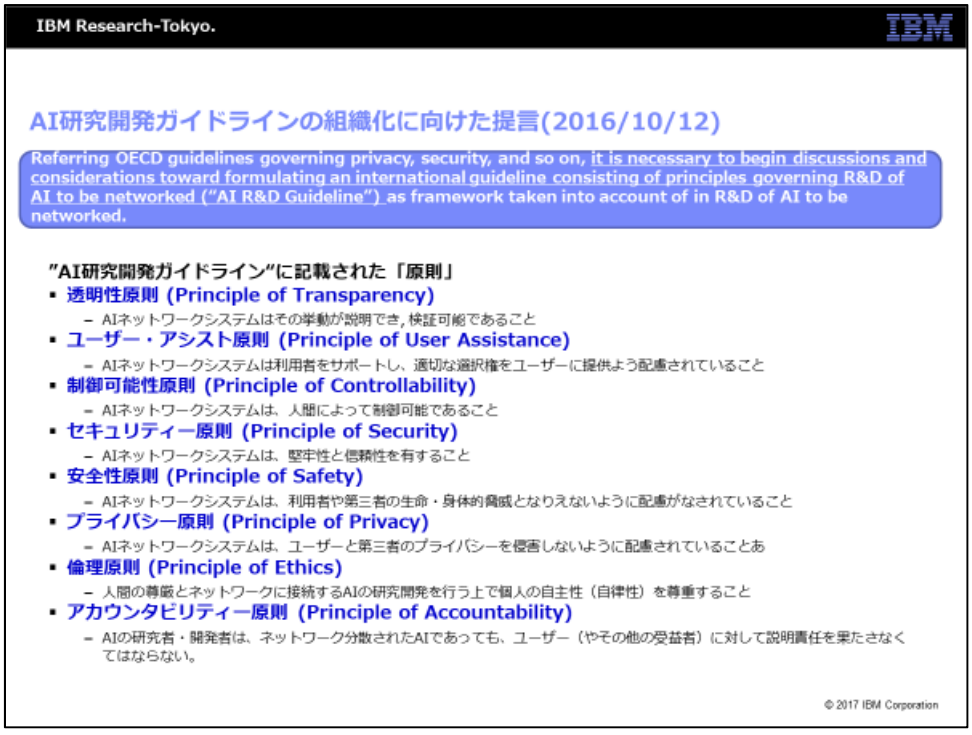
（図 1）

私たち人間はどうしてバナナとぶどうの違いを認識できるのか？「バナナ」＝「黄色くて細長くいもの」と定義すると図 1 にあるような唐辛子やキリンも「バナナ」といえる。開発者じゃない人はよく「大量のデータを食わせれば必ず人工知能が成立する、育っていく」と誤解しているが、AI はラベルを付けて分類したものでしか学習できない。

「プログラムのバグを AI で見つけられないか」とよく聞かれるが、バグの性質、挙動、被害がラベル付けされて初めて品質保証に役立てることができる。現在集まっている大量のデータを AI に学習させるために人間が認識、整理し AI に学習させるためのラベル付けをするところまで至っていない。AI はまだ意味を理解しておらず、あくまで与えられた特徴から似たものを探してくるところまでしかできていない。

AI はまだ照合や推論の域を出ておらず、本日のテーマ「人工知能の品質保証」というタイトルに多義性があり、「人工知能で」、「人工知能を」という二つの解釈ができるという意味論を理解できない。

2016 年 10 月にホワイトハウスから『AI 研究開発ガイドラインの組織化に向けた提言』が出された（図 2）。この中に「AI ネットワークシステムはその挙動が説明でき、検証可能であること（透明性原則）」とあるが、AI は意味を理解しないので「私（AI）は望まれていない答え、嫌な答えをしていないか」という自省をしたり意味を考えながら会話を進めることが現時点ではできない。



（図 2）

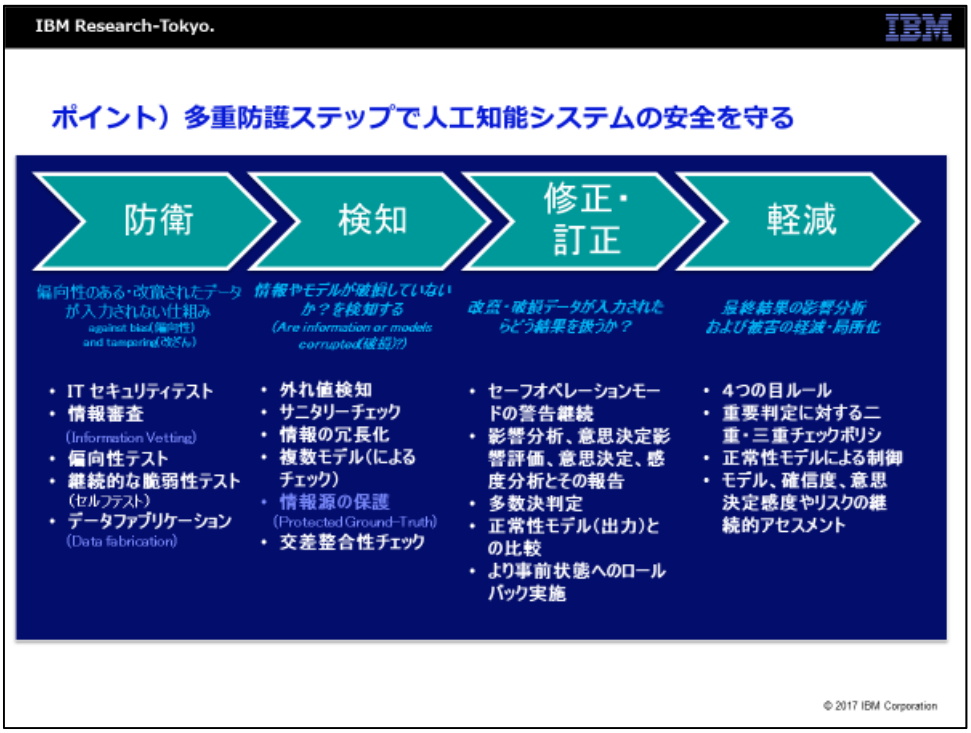
「人間の尊厳とネットワークに接続する AI の研究開発を行う上で個人の自主性(自律性)を尊重すること（倫理原則)」とあるが倫理というのは地域（regional）や宗教（religious）によっても異なり、A さんに不快なものが B さんに快適だったりするので尊重しようがない。また、いわゆる ChatBot に代表されるユーザー対話を伴う AI アプリケーションの場合も、学習元となるデータの品質に応じてユーザーに対する反応が望ましい・望ましくないといった快・不快問題を内包する。これは従来の開発におけるプログラムではなくデータの問題で、これをバグと呼ぶことは果たして適当といえるか。

「AI の研究者・開発者は、ネットワーク分散された AI であっても、ユーザー（やその他の受益者）に対して説明責任を果たさなくてはならない（アカウンタビリティー原則)。」についても、AI は毎分毎秒データを食わされて賢くなっていくものなので、例えば 1 秒間隔で 10 回テストすると 1 回目と 2 回目で同じ結果が出ない可能性がある。決定的（deterministic)な関数のファンクションテストではないので、どうテストしたらいいのか、どこまでテストしたらいいのかが明確でない。つまり、人工知能を品質保証しようとしてもテストができない、終わりがない。

AI の品質保証にはたくさんの課題がある。アメリカではロボットが人を襲わないよう、AI 自身が自分（AI）を殺す「アポトーシス問題（自分自身を止められない)」が話題になっているが自分がシャットアウトできたことを確認できないのでそういう機能は実現できない。

良いデータを食わせれば AI は賢くなると思われているが、私は「やってはいけないことだけ学習させれば良い」と思っている。例えば自動運転車が安全に左折するために、前走車、並走車、後続車、路面の摩擦係数、歩行者、自転車といったすべての組み合わせを掛け算でテストすると安全が保障されるのは数兆ケースの事例を検証したあと、つまり数千年後といわれる。そうではなくて、前後左右にぶつかったらアウト、人や自転車を轢いたらアウト、と条件を足していくと検出数が数十ケースにまで減る。人工知能のテストは背理法でテストすると検出数が極端に減る。シミュレーションしたらおそらく数分程度で済むだろう。ただし、それでその自動車を販売して良いかというのは法律学や倫理観の問題になってくる。

人工知能システムの安全を守るには、防衛、検知、修正・訂正、軽減という 4 段階のガードをかけていく AI のための AI を走らせるイメージで考えていただくと分かりやすい（図 3)。私も「悪魔回路」と「良心回路」から成り、良心回路の振る舞いを悪魔回路が常にチェックする AI を作りたいと思っており、少なくとも人間を殺したり暴走したりする状態を止められる安全装置をかけて作ろうというところまできている。



（図 3）

■まとめ

AI の品質の領域をきちんと捉えている国は日本を含めて今のところ存在しない。世界中の学会、組織で「テスト技法を作ろう」というレベルであり、品質に包括的かつ真正面から取り組めてはいない。もし、本分野に取り組んでいたり問題を抱えている会社さんがあれば是非呼んでいただきたい。AI と品質の問題は一社レベルではなく市場全体で取り組むべき課題だと考えている。今後とも是非皆さんと知見を共有していきたいと考えている。