正解がないシステムをテストする

国立情報学研究所　　石川　冬樹

2022/06/25　JaSST'22 Kansai

@fyufyu / http://research.nii.ac.jp/~f-ishikawa

自己紹介

- 国立情報学研究所　准教授
  - ソフトウェア工学，特にディペンダビリティ：形式手法，自動テスト生成，安全性論証など
- 現在の主な研究プロジェクト
  - JST MIRAI-eAI：機械学習システムのディペンダビリティ
  - JST ERATO-MMSD：自動運転システムの安全性
- 産業界向け教育・実践研究
  - トップエスイー，日科技連SQiP，電通大AISECなど
  - 機械学習工学コミュニティ（MLSE研究会，QA4AI）

7/22発売

eAI

ERATO MMSD

MLSE 機械学習工学研究会 MACHINE LEARNING SYSTEMS ENGINEERING

QA4AI
ＡＩプロダクト品質保証コンソーシアム

本日の内容

- 「正解がないソフトウェア」のテストについて考える
  - 「与えられたルール通りに処理」ではないソフトウェア
  - 「正解はあるが与えるコストが大きい」場合も
  - 様々な「AI」システムなど,
    幅広いエンジニアが関わる機会が増大
- 立場：アカデミック研究者
  - 昔から確立されているう原則・技術
  - 産学連携事例の紹介

テストで何をどう確認する？

自動運転の経路計画

教育コンテンツ推薦

Classi株式会社プレスリリースより引用（5/27）
[https://corp.classi.jp/news/2710/]

配達ロボットサービス設計

パナソニック株式会社ほかによる
プレスリリースより引用（5/26）
[https://news.panasonic.com/jp/press/data/2022/05/jn220526-3/jn220526-3.html]

画像生成・動画生成

株式会社ディー・エヌ・エーのサイトより引用
[https://dena.com/intl/anime-generation/]

最近広がっているシステム

AI，IoT，サイバーフィジカル，スマートXX，…

- 人間・社会や物理世界に深く踏み込む
  - 人事査定，医療診断，…
  - 自動運転，店舗内ロボット，…
- 実装が「動作ルール書き出し」ではない
  - 機械学習
  - 最適化

➡ *要求・想定環境の不確かさ ＋ 実装された挙動の不確かさ*

➡ ***何をもって『よいテスト』とするか？***

目次

- テスティングのそもそも
- オラクル問題に対する従来技術
- 機械学習型AIのテスティング
- 事例紹介
- おわりに

改めて，テストの定義・設計

入力側のキモ：テスト設計 / テスト自動生成

- 欠陥を見つけたい and/or 品質が十分高いと示したい

➡ 欠陥を起こす入力・事前状態を選べばよい（？）

- 通常は未知の欠陥に対しては無理

➡ 網羅性を定義し高める（同値分割やカバレッジ指標）

テスト設計者

テスト生成ツール

入力側の技術：テスト自動生成

- 自動テスト入力生成は10年以上盛んに研究
  - 「可能な入力の範囲」と「『よいテスト』の基準」さえ定義できれば，アルゴリズム・計算機のチカラを活かせる

例：Concolic Testing
SMTソルバーなど制約充足を活用

初期の一例 (DART)

[Godefroid et al., DART：directed automated random testing, PLDI’05]

例：Search-based Testing
遺伝アルゴリズムなど最適化を活用

初期の一例 (EvoSuite)

[Fraser et al., Whole Test Suite Generation, TSE’13]

最近の事例：Sapienz (Facebook)

- Facebookにおけるモバイルアプリテスティング
  - Continuous Integrationに乗せて，リリース候補に対し自動実行

テストスイートA
テストケース数：25
コマンド数合計：150
発見クラッシュ数：5

テストスイートB
テストケース数：80
コマンド数合計：500
発見クラッシュ数：63

テストスイートC
テスト数：90
コマンド数：520
発見クラッシュ数：20

＜進化ステップの例＞
淘汰：CはAとBと比べいいところがないので無視

交配：AとBが持っているテストケースを
半々で混ぜ合わせてみよう

最終的には
「コンパクト」な
テストスイート（レポート）
が得られる

[Mao et al., Sapienz：multi-objective automated testing for Android applications, 2016]
[Alshahwan., et al., Deploying Search Based Software Engineering with Sapienz at Facebook, 2018]

出力側のキモ：テストの成否判断

- 「テスト失敗」により欠陥に気づくような検査をしたい

➡ 機能仕様の場合，出力・事後状態の期待値を与えて比較
  - が，すべてを確認するのは困難（例：画面表示）

➡ 重要な部分・誤りそうな部分について期待値を与え確認

出力側の技術：成否判断の定義・表現

- テスト失敗により，欠陥の存在を検出
  - 個々のテストケースごとにアサーションを設定
    - 典型的には「出力が期待値と合致するか？」
  - 全テストケースにて成立すべき性質を定義，検査
    - コードに対する規範となるモデルや論理式を利用
      （Property-based testing, Model-based testing, 形式仕様記述など）
- 非機能要求などの充足度合いを評価することも
  - 例：レスポンス時間は，より短い方が望ましい
    （一定値以下であるという要求に加えて）

今日のメイン話題：出力側

- 機能要求に対する従来のほとんど
  - 基本はやはり個々のテストケースでの期待値設定
  - 計算・検索・予約など期待値が一意に定まっている処理が多い
  - ➡ *今日の話題：正解が定まらない場合が増えている！*

- 非機能要求などは定量評価（制約 ＋ 嗜好）
  - 例：レスポンス時間（上限はあり，より短い方がよい）
  - ➡ *今日の話題：より幅広い品質が重要となる場合が増えている！*

目次

- テスティングのそもそも
- <u>オラクル問題に対する従来技術</u>
- 機械学習型AIのテスティング
- 事例紹介
- おわりに

疑似オラクルの従来技術：N-Version Programming

- 同等な機能を持つ比較対象を用いる方法
  - （同じ誤りを埋め込まないように）別の人・方式で作る
  - コンパイラや数値計算など汎用的な機能の場合，
    オープンソースや他社製のものなどがしばしば利用可能

疑似オラクルの従来技術：Metamorphic Testing

先日 JaSST'22 Tokyo にて
ASTER松谷さんチュートリアル

- 既存テストから新たなテストを作る方法
  - 出力の期待値が定められない場合でも利用可能
  - 「入力を変えると出力はこう変わるはず」関係をテスト

疑似オラクルの従来技術：Metamorphic Testing

■簡単な例：sin関数

疑似オラクルの従来技術：Metamorphic Testing

- 様々なAIシステムに対する例
  - 推薦ランキング生成：1位商品を候補から抜いて再生成
    → 2位だった商品が1位になるはず
  - 時系列の異常分析：入力信号をすべて定数値ずらして再分析
    → 異常と検出されていた入力は，やはり異常と検出されるはず
  - ドローン探索プランニング：入力地図を90度回転させる
    → 出力されるルートが90度回転するはず

[Murphy et al., Improving the Dependability of Machine Learning Applications, 2008]
[Jarman et al., Metamorphic Testing for Adobe Data Analytics Software, 2017]
[Lindvall et al., Metamorphic Model-based Testing of Autonomous Systems, 2017]

疑似オラクルの従来技術：悩ましいところ

- 費用対効果の判断が難しい
  - N-Versionは，既存のものがある状況に限られがち
  - Metamorphic Relationは「テストのためのテスト」
    「バグ探しの便宜上だけ」に思えてしまうことがある
- Metamorphic Relationを得る難しさ
  - 指針は「追加・削除・並び替えなどのパターンから模索」程度
- 「言われた通りに動く」の先に価値があることが増えた

➡ *「そのシステムで検査すべきこと」から入るべき*

目次

- テスティングのそもそも
- オラクル問題に対する従来技術
- <u>機械学習型AIのテスティング</u>
- 事例紹介
- おわりに

再掲：最近広がっているシステム

AI，IoT，サイバーフィジカル，スマートXX，…

- 人間・社会や物理世界に深く踏み込む
  - 人事査定，医療診断，…
  - 自動運転，店舗内ロボット，…
- 実装が「動作ルール書き出し」ではない
  - 機械学習
  - 最適化

➡ *要求・想定環境の不確かさ ＋ 実装された挙動の不確かさ*

➡ ***何をもって『よいテスト』とするか？***

大ざっぱな図：Before

ここのギャップが
悩ましい

仕様化できない部分や
仕様化が不適切な部分を
人が補っていたり
使われなかったり

動作ルールを明記
計算，検索，予約，…

テストは主にここを
確認するもの？

ソフトウェア

例：アクセルを踏んだら加速
（前がコンビニでも…）

例：経歴情報を整理・提示
（判断はしない）

大ざっぱな図：After

例：安全になるように
アクセル・ブレーキを制御
（必要なら人の指示に反する）

例：経歴情報から採用是非判断

機械学習： ごくごく簡単に・・・

- ある会社での年齢 x のときの給与 y を予測したい
  - $y = ax + b$ と表現できると仮定すると，
    過去のデータと「一番合う」ように
    a, b を決めれば判定・予測プログラムが作れる！

- 実際は1次関数（パラメータ2つ）では無理
  - 深層ニューラルネットワークでは，
    何百万，それ以上のパラメータを使う
    - 画像や言語に対する性能向上，第3次AIブームの起因

機械学習：画像識別での例

テナガザル

35 24 210	20 121 24	122 81 20
211 54 42	12 222 90	88 79 116
24 36 98	98 181 31	66 31 198

13 83 33	13 45 94	75 74 111
111 8 73	192 1 221	237 31 1
74 35 122	93 76 244	73 211 45

0 245 210	20 12 114	84 99 100
11 86 99	121 88 91	18 0 77
46 87 121	70 76 122	122 14 94

画像 ＝
各点の色を表す
数値の集まり

パンダ

[画像： https://www.photo-ac.com/]

254 32 67	222 88 1	108 76 14
12 86 222	98 75 122	111 74 74
198 87 33	188 173 4	68 176 83

77 81 123	122 158 6	76 63 42
3 3 78	19 183 84	76 63 123
98 83 111	123 7 99	253 48 91

0 24 31	20 21 124	12 101 50
21 54 242	112 22 90	8 79 214
124 56 85	98 99 141	166 1 198

この線引きを訓練データから作る

機械学習：すごいこと

- 規則性を書き出すことが困難な機能であっても実現可能
  - 画像における物体の識別
    - 自動運転における歩行者や信号
    - 工場における不良品や機器劣化
    - 医療画像における腫瘍
  - 工場機器の制御
  - ローンや保険に関する判断
  - 音声や文書の分類，検索，翻訳
  - 画像や動画の色塗りや補完，生成

機械学習型AIに対する基本的評価：予測性能

■二値分類での例：陽性か陰性か（例：不良品検出）

■PCR検査など医療診断とも似た考え方

混同行列（Confusion Matrix）

正解＼AI	AIの判断：陽性 計45	AIの判断：陰性 計455
陽性 計20	True Positive (TP) （真陽性） 「正しい検出」 15	False Negative (FN) （偽陰性） 見落とし 5
陰性 計480	False Positive (FP) （偽陽性） 誤検出 30	True Negative (TN) （真陰性） 「正しい未検出」 450

メトリクスの一例

全体500件の製品のうち
良品・不良品の正解は465件

正解率（Accuracy）

$$\frac{\text{正解データ数}}{\text{全データ数}}$$

$$= \frac{TP + TN}{TP + FP + FN + TN}$$

$$= \frac{465}{500} = 0.93$$

機械学習型AIに対するテスティング？？

- テストオラクル（期待値の設定）？
  - 画像分類など，正解を与えることのコストが非常に大きい
  - 給与判断など，唯一の正解を決めがたい場合もある
  - 推薦やデータマイニングなどの場合は，本来正解はない
- 「テストでバグを見つける」？
  - 分類や予測が100%成功することは本来ないため，
    「結果が合わなければバグが存在」とはならない
- 単体テスト？
  - 一つの実装が様々な条件下で様々な対象を扱う（分割統治困難）

参考：機械学習型AIに対するテスティング

- 書籍「AIソフトウェアのテスト」
  - メタモルフィックテスティング（前述）
  - ニューロンカバレッジテスティング：
    ニューラルネットワークに対する網羅性指向
  - 最大安全半径・網羅検証：
    頑健性に対する検証の考え方

[佐藤ら、AIソフトウェアのテスト
– 答のない答え合わせ]

機械学習型AIのテスト：Metamorphic Testing (1)

■機械学習の訓練済みモデルへの適用

■頑健性（ノイズに対する耐性）のテストが大多数

画像に雨を加えた例

1.1 original　1.2 with added rain

[Tian et al., DeepTest：Automated Testing of Deep-Neural-Network-driven Autonomous Cars, 2018]

機械学習型AIのテスト：Metamorphic Testing (2)

■機械学習の訓練パイプラインへの適用

■実装バグ検出の可能性が上がるが，ここまでするか悩ましい…

テストケース01

入力データセット01 → 訓練パイプライン → 出力01

1. 全画像に特定の変化

2. 予測結果は変わらないはず

フォローアップテストケース01-01

入力データセット01f → 訓練パイプライン → 出力01f

(a) Original training data (RGB) (b) BGR (c) BRG

1. データセットを訓練用・評価用に分割
2. 訓練用データでモデル構築
3. 得られた訓練済みモデルで評価用データに対し予測実行

通常の世界で「船」とされた画像は，海が赤い世界で訓練したあとも「船」になるはず

[Dwarakanath, Identifying Implementation Bugs in Machine Learning, 2018]

機械学習型AIのテスト：幅広い品質

- ここまでの話：予測性能 ＋ 実装の正しさ
- 訓練データ・評価データの十分性も大きな課題
- より実世界・人間に踏み込んだ品質が求められることも
  - 公平性
  - 倫理
  - …

データの十分性に関する例 (1)

- 文字認識において対応すべき状況（QA4AIガイドライン）
  - プレ印字の有無・色
  - 株式会社や年号の異なる表記
  - 印鑑かぶり
  - ボックス区切り・網掛け
  - …

[http://www.qa4ai.jp/]

データの十分性に関する例 (2)

- 数字認識における
  いろいろな書き方のクセ
  - 下に横線を引くかどうか
  - 上に斜め線を引くかどうか
  - …
  - 言葉より具体例や分布可視化がよい
  - 中間的なものもある
  - 分け方は一意ではない

もともと64次元データで
表現された「１」の数字画像を
2次元の値で要約し座標配置した
（digits@sklearn → PCA）

公平性の課題：有名な事例 (1)

- 社会的に望ましくない差別的なバイアス（偏り）の発生
  - 人間の（ときには無意識な）差別を学習
  - 少数例外に対しては学習しきらず予測性能が低くなりがち

REUTERS ワールド ビジネス マーケット 外国為替 ビデオ

外国為替 株式市場 外国為替フォーラム 貿易摩擦 北朝鮮 トランプ政権 オピニオン 経済・政策 テクノロ

テクノロジー 2018年10月11日 / 15:30 / 1日前

焦点：アマゾンがＡＩ採用打ち切り、「女性差別」の欠陥露呈で

Jeffrey Dastin 2 分で読む

［サンフランシスコ　１０日　ロイター］ - 米アマゾン・ドット・コム(AMZN.O)が期待を込めて進めてきたＡＩ（人工知能）を活用した人材採用システムは、女性を差別するという機械学習面の欠陥が判明し、運用を取りやめる結果になった。

女性に不利な雇用判断
→
過去の差別を学習？女性が少数データ？

[https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN]
(access: 2021/09/27)

Ads related to latanya farrell

Latanya Farrell, Arrested?
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

Latanya Farrell
www.publicrecords.com/
Public Records Found For: Latanya Farrell. View Now.

アフリカ系の名前で検索すると
逮捕歴データベースの広告
→
ユーザの潜在意識によるクリックを反映？

[L. Sweeney, Discrimination in Online Ad Delivery, ACM Queue'13]

公平性の課題：有名な事例 (2)

- アメリカの医療システム
  - 「必要となる医療支出を予測」して治療優先度を決定
  - 収入差や教育差，医師による不当な扱い等により，
    黒人の医療支出が低い → 黒人の治療優先度が低くなる
  - 人種は入力にはないが，住所等から紐付いてしまう

[Obermeyer, Dissecting racial bias in an algorithm used to manage the health of populations, Science Mag'19]

※ センサティブな入力は直接使っていない！

- Googleの事例もそうであった

公平性の課題：様々な原因

- 不適切な訓練データ
  - データ作成者の偏見や認知バイアスを反映
    - 例：過去の雇用採否判断
    - 例：広告をクリックする大衆ユーザーの潜在意識
  - 予測対象の集団を代表しないような標本の選択バイアス
    - 例：「過去のローン返済成否データ」から「ローンを認可すべきでない人」を見抜きたい
- アルゴリズムによる帰納バイアス
  - 汎化のために，少数事例は例外扱いされる

補足：具体的な評価指標の例

	人種1			人種2		
	資格あり	資格なし	計	資格あり	資格なし	計
採用	60	3	63	12	15	27
不採用	40	7	47	8	35	43
計	100	10	110	20	50	70

補足：具体的な評価指標の例

	人種1			人種2		
	資格あり	資格なし	計	資格あり	資格なし	計
採用	60	3	63	12	15	27
不採用	40	7	47	8	35	43
計	100	10	110	20	50	70

Equalized odds 成立 （手続きが公平）

人種にかかわらず
資格があると採用率60%、資格がないと採用率30%

補足：具体的な評価指標の例

	人種1			人種2		
	資格あり	資格なし	計	資格あり	資格なし	計
採用	60	3	63	12	15	27
不採用	40	7	47	8	35	43
計	100	10	110	20	50	70

Demographic parity 不成立　（結果が不公平）

人種1の採用率63/110=57%，人種2の採用率27/70=38%

補足：具体的な評価指標の例

	人種1			人種2		
	資格あり	資格なし	計	資格あり	資格なし	計
採用	60	3	63	12	15	27
不採用	40	7	47	8	35	43
計	100	10	110	20	50	70

この例の場合：人種1は資格を持っている人が多い
（存在を知る機会・促される機会が多かった？）

それでも資格が重要なので手続きを平等にするか？
人種によって判断を変えて採用率を揃えるのか？
プロダクトオーナーの判断が必要！

さらに公平性と予測性能とのトレードオフについての判断も必要！

もともと訓練データ・評価データ自体が完全に公平でないとすると、「正解」を一部無視しないと公平にできない！

目次

- テスティングのそもそも
- オラクル問題に対する従来技術
- 機械学習型AIのテスティング
- <u>事例紹介</u>
- おわりに

事例 (1) 自動運転における経路計画

- マツダ株式会社提供のプロトタイプでの追及
  - 経路計画（Path Planning）：短い周期で定期的に計算，ハンドル角や加速度を最適化により決定
  - シミュレーターによるテスト・評価

他車や歩行者の位置・動き方
信号や標識
道路形状
…

安全性
快適さ
ルール遵守
…

事例 (1) 自動運転における経路計画

- 評価メトリクスは定義済みだがあいまいさもある
  - 例：衝突したら問題とは限らない（他車の動きが悪い状況も）
- 評価の観点が複数あり，優先度が異なる
  - 安全性のために中央線をやむなく越える状況はありうる
- 入力空間の膨大さ・特定状況でのみ問題発生
  - 入力の自動生成（自動探索）を考える

事例 (1) 自動運転における経路計画

■技術例 (1) 「避けられたはずの衝突」のみ検出

[https://www.nii.ac.jp/news/release/2020/0323.html]

事例 (1) 自動運転における経路計画

■技術例 (2) 多数の要求に対する違反を適応的に探索

[https://www.nii.ac.jp/news/release/2021/1115.html]

事例 (1) 自動運転における経路計画

■オマケ：開発した技術の一覧

テスティング	問題検出指向	「避けられたはずの衝突」のみ検出
		多数の要求に対する違反を適応的に探索
	網羅性指向	事故に至らない状況での急ブレーキや急ハンドル， それらの複合発生を探索
		「少し設計を変えると大きな差がでる」ように 反応性が高いテストを探索
デバッグ	要因分析	検出した衝突等に対し，どのパラメーターが影響したかを分析
	テスト単純化	検出した衝突等が起きるがもっと簡単なテストを探索
	自動修正	検出した複数の問題点に対処する修正案を探索

事例 (2) 教育コンテンツ推薦

- Classi株式会社との協働事例
  - 生徒の理解度に応じた問題推薦機能
    （アダプティブラーニング：適応型学習）
  - 典型的な仕組みのイメージ（一般論）

Classi株式会社プレスリリースより引用（5/27）
[https://corp.classi.jp/news/2710/]

1. *難易度Cの問題30正解*
2. *難易度Bの問題42不正解*
3. *…*

理解度0.6

問題105, 225, 320
がオススメ

部品レベルでもシステムレベルでも期待値設定は困難

※ オマケ：問題項目の難しさと正答率の関係については「項目反応理論」が広く知られている（TOEICなども利用）

事例 (2) 教育コンテンツ推薦

■Property-based & Metamorphic Testing の展開

出力検査に使えそうな
Property や
Metamorphic Relation
は多数挙げられる

テスト対象

理解度推定
or
問題推薦

ニーズ・ペイン
からの重要度
フィードバック

欠陥検出実績

試行により欠陥を
検出できた！

解けたのに簡単な問題が
推薦されるとやる気を失う！

出力の検査条件：
「正解すると理解度上昇」
を採用！

運用時の発生も記録

検査項目の選択・洗練
&
運用時監視にも展開

事例 (3) 実世界サービス設計

- パナソニックホールディングス株式会社との協働事例
  - 自動宅配ロボットのサービス設計
  - ステークホルダーと議論しながら
    シミュレーションにより追求

パナソニック株式会社ほかによる
プレスリリースより引用（5/26）
[https://news.panasonic.com/jp/press/data/2022/05/jn220526-3/jn220526-3.html]

設計はむしろ模索の対象

成否ではなく
品質の度合いを探る

事例 (3) 実世界サービス設計

■Design Space Explorationのための
自動テスト生成 ＋ 設計候補評価・説明

様々な設定試行

サービス設計

ロボットの台数
ロボットの最大速度
遠隔オペレーター数
…

シミュレーター

状況設定

配送需要パターン
時間帯
…

ランダム性もあり

評価値

各種のコスト・価値・リスク

評価を踏まえて
設定調整

ステークホルダーへの
途中報告・結果の解釈説明

事前表明が難しい
隠れたニーズ・ペインの
気づきとフィードバック

多目的最適化
＋ インタラクション・説明による
設計候補の探索

事例 (4) 画像・動画生成

- 画像生成や動画生成における品質
  - QA4AIガイドラインにおける活動
  - DeNA濱田さん，JAXA菅原さん

[https://dena.com/intl/anime-generation/]

[https://nvlabs.github.io/SPADE/]

事例 (4) 画像・動画生成

- 品質特性と評価方法を列挙
  - メトリクスの活用：自然さ，多様さなど
  - ルール評価の活用：主要な色・服の色が指定されたとおりかなど
  - ときには「評価用AIを作る」まで必要：
    姿勢が指定されたとおりか，社会的に不適切でないかなど

[https://dena.com/intl/anime-generation/]

事例 (4) 画像・動画生成

- テストの設計・実施例も
  - 自動運転用のテスト画像を系統的に生成するユースケースを想定

微少なスタイル変化で
出力が崩れた例
（特徴量監視で検出できた）

※ 実施したのは，機械学習を学んでいたエンジニアの皆さん

事例 (5) ゲームバランス調整

ゲームバランス
が適切か？

テスト対象の
ゲーム

テストプレイ
エージェント

強化学習，最適化など
複数のエージェントを追及
＋
人間のテストプレイヤー

よいテストプレイを
- 状態変化には
  ランダム性もある
- 何十ステップもの
  入力が必要

想定よりもスコアが低かった
→ ゲーム設定が難しすぎ？？

ゲーム設定のどの部分を変える？

エージェントが適切なプレイを
できていないしている可能性は？

高度なテストプレイ自動化技術が必要
＋
説明・解釈の技術も必要

株式会社コナミデジタルエンタテインメントとの協働
[https://cedec.cesa.or.jp/2022/session/detail/15]

補足：物理シミュレーションソフトウェア

- 複雑な計算をしてくれることに価値：
  人間には答えの計算ができない・大変すぎる
  - 14:15からの小川さんの内容にほんの少しお手伝い
  - 入力の探索・生成，Metamorphic Testing

S3) 事例セッション

14:15~14:45 (30分)
セッション 3

「物理シミュレーションソフトウェアのテスト事例」

小川 哲生 （JSOL）

セッションの内容

近年、製造業におけるものづくりの現場では、設計開発やその性能評価の手法として、物理シミュレーションの活用が進んでいます。
ここで用いられる汎用的な物理シミュレーションソフトウェアの開発では、多様な解析シーン、対象機器や物理現象への適用を考慮した動作保証が求められ、その複雑さからテストケースの設計や期待値の導出が困難な場合があります。
この発表では、物理シミュレーションソフトウェアの開発に伴うソフトウェアテストの取組み事例についてご紹介します。

講師プロフィール

小川 哲生 （JSOL）

紹介した事例のまとめ

	注目する品質	テストの出力評価	テストの入力設計	追加事項
自動運転の経路計画	安全性，快適さ，ルール遵守など優先度付きで複数	各要求の違反度合いを定量評価	複数の要求に関する違反を自動探索	要因分析や自動修正も
教育コンテンツ推薦	学習効果，ユーザの意欲付け，…	ユーザフィードバックも踏まえ成立すべき性質・関係を検討	ユーザペルソナも含め系統的に設計	テスト設計を運用時の監視項目へ
実世界サービス設計	サービスの効果，コスト，リスク	同左 ＋ ステークホルダーのフィードバック	自動探索	自動探索にステークホルダーとの相互作用も追加
画像・動画生成	自然さ，鮮明さ，社会的適合性，指示遵守，…	専用メトリクス，評価用AI，または評価ルール	系統的に設計	人間の感覚評価またはそれに類する大量データも重要
ゲーム	ゲームバランス	スコア	自動プレイエージェントを構築	プレイ記録の説明・解釈技術も

目次

- テスティングのそもそも
- オラクル問題に対する従来技術
- 機械学習型AIのテスティング
- 事例紹介
- おわりに

どう受け止めたでしょうか？

- 紹介した話：難しい？遠い話に聞こえる？
- 機械学習型AIのテスト・品質保証は広がっている
  - ある日担当になるかも（石川の立場ではたくさん会った）
- 「新しいことをやる」難しさ
  - 開発者側は乗り越えて，Python/統計・機械学習を身につけた
- 研究と開発の境界がなくなっていく風潮
  - 「AIを活用したテスト」にわくわくして楽しめる？
  - そういうことに進みやすい風潮かも？（会社による？）

まとめ

- 正解がないシステムをテストする
  - 疑似オラクルの技術ももちろんだけど
  - そもそもそのシステムで重要なことと向き合うべき
  - （複雑なので入力自動生成も活躍できる）
- 実世界・人間に踏み込むシステムに関わる機会が増え，扱う品質があいまい化・多様化している
  - 実世界での価値，自然さ，楽しめる度合い，様々なステークホルダーにとっての公平性や満足度，…

➡ *バグ探しに限らない価値を追求するテストへ！*