第 8 回特別講義　レポート

日時	2019 年 1 月 11 日(金)　10:00 ～ 12:00
会場	(一財)日本科学技術連盟・東高円寺ビル　地下 1 階講堂
テーマ	IoT・AI 時代のテスティング・検証技術の最前線
講師名・所属	石川　冬樹　氏(国立情報学研究所／本研究会 研究コース 5 副主査)
司会	栗田　太郎　氏(ソニー／本研究会 研究コース 5 主査)
アジェンダ	＜AI・機械学習×品質＞ １.イメージをもつための事例 ２.本質的な違い:振る舞いの帰納的な構築 ３.改めて:要求と仕様(と環境) ４.「IoT・AI 時代」の品質？ ５.機械学習における品質保証のための原則・思想 ６.機械学習に対するテスト・検証技術の追求 《メタモルフィックテスティング》 《サーチベースドテスティング》 ７.おわりに
アブストラクト	機械学習を用いるシステムの動作は、プログラムコードにより演繹的に(規則的に)定まるのではなく、帰納的に(データに基づき)定まる。このため、システムから思いもしない出力が得られることがある。これに限らず、AI や IoT といった言葉で総称されるようなシステムにおいては、実世界や人の感覚に深く踏み込み、その要求やテストオラクル(成否判断基準)、扱う外部環境が不確かで変わりやすい。本講演においては、この「不確かさ」の問題に対するアプローチや、テスティング・検証のための技術を紹介する。

講義の要約

石川　冬樹　氏
ソフトウェア工学および自律・スマートシステムに関する研究・教育に従事。
特に、形式手法、自動テスト生成、最適化、機械学習といった技術の活用や、サイバーフィジカルシステムなど先端システムにおける品質保証に興味を持つ。
国立情報学研究所トップエスイープログラムおよび日本科学技術連盟 SQiP 研究会での活動を中心に、産業界向けの教育・応用研究も行う。

＜AI・機械学習×品質＞

1．イメージをもつための事例

・(技術の進化の一例)画像の自動生成
⇒特定のキャラクター画像に対し、指定した姿勢の画像や動画を自動生成できるようになった(これをどうテストし、保証するか)。

・(使い方を問う例)2016 年のテスラ自動運転車の事故
⇒運転者がハンドルを握っておらず警告音も鳴っていたが、自動運転を過信して、警告を無視していた(現在は、警告に運転者が反応しなかった場合は、自動運転を使用不可にするよう仕様変更)。

・(技術的限界の例)Google フォトの画像認識
⇒黒人に「ゴリラ」とタグ付け(現在もなお、直せておらず、ゴリラを禁止ワードにして対策)。

・(よく知られた課題)画像認識における敵対的サンプル
⇒パンダの画像に特定のノイズを加えることで、ソフトがテナガザルと誤認識。
⇒停止の交通標識にテープを物理的に張り付けることで、別の標識に誤認識。

・(攻撃に関する例)Microsoft の AI ボット「Tay」
⇒差別や放送禁止用語を「教えた」ユーザがおり、不適切な発言を連発した(リリース後、数時間で停止に)。

・(倫理的要求に関する例)Amazon の AI を活用した人材採用システム
⇒女性を差別する欠陥があった(運用を取り止め)。

・(実装のブラックボックス性に関する例)どこを見て画像認識しているか
⇒ハスキー犬の画像の背景に雪があるのを見て、オオカミと誤認識。

・これらは全てバグなのか、潰さなければならないのか、そもそも潰せるのか
・どうやってこれらに気づくのか、同種や類似の状況は列挙できるのか
・うまくいかないケースがあるのを、顧客やユーザにどうやって受け入れてもらうのか

2．本質的な違い:振る舞いの帰納的な構築

・従来のシステム開発は、演繹的
(演繹:一般原理から特殊な原理や事実を導くこと)
⇒計算や判断を行うための知識・規則を人が決めてプログラムという形で書き下す。

・機械学習は、帰納的
(帰納:事実や命題の集まりからそこに共通する性質や関係を取り出し、一般的な命題や法則を導くこと)
⇒計算や判断を行うための知識・規則を訓練データから獲得し生成する。

・機械学習(帰納的システム開発)のすごいところ
⇒人が明確に規則として書き出せないことも、判断・予測などの計算ができる。
⇒訓練データを更新していけば、新しいことに対応できる。

・機械学習(帰納的システム開発)の大変なこと
⇒原則として機能は不完全。
⇒どの程度の性能が出るか作ってみるまで分からない。
⇒大量かつ「適切な」訓練データが必要。

・機械学習が従来のシステム開発と異なると思うところをアンケートした(MLSE 研究会)
⇒今までと根本的に異なる考え方が必要として、一番多かったのは「顧客との意思決定」だった。
(実際に作れるか、どのくらいの精度で作れるかが保証できない)
二番目に多かったのは「テスト、品質の評価・保証」だった。

・機械学習は保守の面でも問題を抱えている
⇒実世界のデータを元に学習している為、世界の変化に常に対応しなければならない。
(例:コンビニの各商品の売上予測にて、ある日突然隣に別のコンビニができたとしたら、これまでの予測方法では当たらなくなる)
※機械学習ではこういった特徴や、グル―コードが乱立しやすいという特徴がある為、技術的負債を負いやすく、「技術的負債の高利息クレジットカード」と呼ばれることがある。
(Sculley et al. , Machine Learning: The High Interest Credit Card of Technical Debt、2014)

3．改めて:要求と仕様(と環境)

・Zave / Jackson のモデルでは「要求」と「仕様」を別の概念として扱う。
⇒「要求」は、実現したいことがらであり、実現しようと決めたもの。
⇒「仕様」は、システムの振る舞いが満たすべき制約を定めたもの。

・Zave / Jackson のモデルでは「要求」と「仕様」の間には「想定環境」があることを言っている
⇒マシンの「仕様」は、「想定環境」の下で、「要求」を満たす。
(例えば、改札機の仕様として、二人同時に入ることは想定しないなど)

4．「IoT・AI 時代」の品質？

・ソフトウェアが実世界・人の感覚により深く踏み込むようになり、「要求」と「想定環境」と「仕様」の範囲が広がってきた

・「要求」と「想定環境」については、完全・網羅的な列挙が不可能に(オープン・不確かなものへ)
(例:「様々な」歩行者を識別する　⇒自転車を押している人、着ぐるみを着た人、など)
(例:「適切に」運転する　⇒安全、快適、マナー遵守、って具体的には？)
(例:「適切に」給与判断する　⇒男女や人種の差別はしないとして、何はすべき？)

・「要求」と「想定環境」がオープン・不確かなものになると、「完全な機能」や「リスクゼロ」は無理
⇒継続的な修正・更新が前提になる。

・機械学習を用いた場合、自分たちが構築した「仕様」までオープン・不確かになる。

5．機械学習における品質保証のための原則・思想

・機械学習で得たソフトウェア部品の性能(精度)において、何を 100%の基準とするのか
⇒そもそも顧客が決めるのか、エンジニアが決めるのか。
⇒自動運転のための画像認識にて、「霧の日、山道、逆光もテストした」といっても、それらの区分が機械学習で作ったモデルには意味がないかもしれない。

・不完全なものをどう受け入れ役立てるか
⇒天気予報のように、誰もが確実なものでないと認識しているような、そういった利用者側の意識も重要。
⇒目的や用途と照らし合わせて、受け入れる。
(例:人より悪いが人件費よりずっと安い)

・(先端企業から出ている原則・指針の例)Martin Zinkevich，Rules of Machine Learning: Best Practices for ML Engineering，2016
⇒機械学習におけるルールが計 43 個記載されている。
(例:データの統計を追跡するなどして、問題として顕在化しない失敗を見張れ)

・(先端企業から出ている原則・指針の例)Eric Breck，What’s your ML Test Score? A rubric for ML production systems，2016
⇒機械学習に対するテストをどれだけできているかをスコア化して評価する仕組み。
(例:訓練データが開発時の想定からはみ出ていないかをテストしているか)

・(先端企業から出ている原則・指針の例)Sculley et al. ，Machine Learning: The High Interest Credit Card of Technical Debt，2014
⇒機械学習では、n 個の入力データのうち 1 個の傾向が変わると、他のデータの重要度や利用法が変わってしまう(CACE：Changing Anything Changes Everything)という性質を持つ為、各モデルを独立させて分析するなどの対応が必要である。
⇒機械学習では、フィードバックループが隠れていることがあり、それが後に大きな影響を及ぼしてくる為、可能な限り取り除く必要がある。

・石川先生の個人的思想(研究者として)
⇒実行時に想定外のことが発生することが原則になる。(実行時のモニタリングが必要)
⇒テスト入力を多数用意しても成否判定が困難。(疑似オラクルを使ってテストする)
⇒何が実際実現できているのか把握・説明できず、問題の原因特定・修正も難しい。(部分的な検証が必要)
⇒要求の実現可否・実現コストや変更の影響を事前に把握しての分析・意識決定が難しい。(探索的・試験的な評価を繰り返す)

・実行時のモニタリングの仕組みとして、Models@run.time といったアプローチが注目されている
⇒要求や設計に関するモデルをシステム側に持たせようとする思想。

6．機械学習に対するテスト・検証技術の追求

・機械学習はそもそも「テスト不可能」
⇒画像分類など、正解を与えることのコストが大きい。
⇒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、未知の正解を求めることが目的で、出力の期待値は存在しない。

・機械学習の結果がおかしかったとしても、性能限界による場合があり、バグを見つけることに直結しない。

・機械学習の実装は大きなブラックボックスであるため、単体テストができない。(概念がない)

・機械学習の場合、「要求」や「想定環境」がオープンで不確かなことが多く、要求カバレッジを完全に満たすことがない。

・機械学習の場合、分岐ではなく数値が出力を決めている為、分岐網羅といったテストにあまり意味がない。

・今は機械学習に対する確立したテストが存在せず、既存のテスト・検証技術を機械学習に適用してみるなど、研究を進めている段階

《メタモルフィックテスティング》

・推薦のような正解(期待値)が明確でない場合や、画像分類のように正解を与えるコストが高い場合にテストする手法

・入力にある種の変換をかけた場合、出力が想定通りに変わることを検証する
(例:お勧めの本を表示するシステムで、1番目に表示される本を入力から抜くと、出力はどう変わるか
【変換前の結果】　1番目:「A」という本　2番目:「B」という本
【Aを抜いた結果】　1番目:「B」という本　2番目:「C」という本
※「A」を入力から抜くことで、「B」が1番目に繰り上がることを確認する)

・入力にある種の変換をかけても、出力が変わらないことを検証するケースもある
(例:画像のRGBを入れ替えて海の色を赤くした場合でも、船を船として認識するか)

・メタモルフィックテスティングは、要求に直結するテストではないため、どこまでテストするのかを決めることが難しい

《サーチベースドテスティング》

・目的に沿った最適なテスト(スイート)を見つける手法で、テストを生き物と仮定し、テスト同士を戦わせることで、最終的に最適なテストが生き残るというもの
⇒生き残ったテスト同士をさらに交配して、より最適化するといったことも行う。

・これはコンピュータの処理能力が高くなったことによる力業的な手法

・テストをスコア評価して、最大を選んでいる仕組みのため、スコアリングの仕方を変えることによって、汎用的に使用できる
⇒「車が人に最も近づくような危ういテスト」ケースを生成。
⇒「高カバレッジで小さい」回帰テストスイートを生成。

・(機械学習に適用したケース)敵対的サンプルを見つけるテスト
⇒自動運転にて、画像から次の進路を判断する場合、画像に雨を加えると次の進路が変わってしまう場合(敵対的サンプルとなる場合)がある。サーチベースドテスティングを用いることで、そういったケースも最適に探すことが可能。

・そもそも誤認識を探してもきりがなく、システム全体の要求を踏まえて、どういった誤認識だと問題があるかを探して、狙ってテストすることが重要
⇒自動運転にて、遠くの停止標識を誤認識しても実質的には問題ない。一方、近くの停止標識を誤認識するのは問題がある。

・サーチベースドテスティングの研究はこの 10 年ほど盛んで、2018 年の Java Unit Testing Competition では、サーチベースドテスティングを使うことで、人が作ったものよりよいテストスイートを 10 秒で生成した実績がある
⇒ただし、理解容易性や自然さという点においては、課題を抱えている。

・Facebook では、すでに実用段階に入っている。(Sapienz)

・Facebook では他にも様々な自動化に取り組んでおり、自動バグ修正といったことにも挑戦している。
⇒自動バグ修正といっても、中身は力業であり、四則演算のプラスをマイナスに変えるといった典型的なバグ修正を色々試してみて、それで全てのテストが通るかどうかを確認するもの。ただし、これはテストがしっかりしていないと成立しないし、典型的な修正では直らないものについては対応できない。

7．おわりに

・機械学習における品質保証の問題は、実はこれまでも存在していた問題がほとんど。これまでと違うのは、より実世界に踏み込んだシステムになってきたことで、逃げられない問題になってきたという点である。
⇒これまでの時代とは異なり、本気で取り組んでいく必要がある。

～質疑応答～

《イメージをもつための事例》

■黒人さんとゴリラが識別出来ない、とありましたが、人は画像だけ見ていても識別出来るのに、データ量の問題なのでしょうか？

⇒データ量なのか、学習方法(ニューラルネットワークの構造など)なのか、いろいろな原因があると思います。根本的に人と同じことができるか、というとそうでもないので、人と比べてしまうと難しいかもしれません。

■敵対的サンプルの「敵対的」というのは、誰が判断するのでしょうか？

⇒「敵対的サンプル」の「敵対的」は、政治的に対立しているとかそういう言葉かと思いますが、「敵対的サンプル(adversarial example)」は、実世界における意味として「攻撃者がいる」「敵がいる」ということは意味しません。単に、機械学習モデルにとって「イヤな入力の例」というくらいにとらえて下さい。

《本質的な違い:振る舞いの帰納的な構築》

■訓練データから規則や法則を導き出すためのプログラム自体は人が作っているのに、データ内の何を判定材料とするかは全然想定できないものなのでしょうか？

⇒特に深層学習の場合は、データのどの部分を判定材料とするかまで学習させるので、その部分は想定できない部分があります。「全然想定できない」ではなく感覚はあると思いますが、人の感覚・期待が裏切られることはあると思います。挙げた一例ですが、「訓練データのうち、オオカミの画像にはたまたま雪がよく写っていた」とします。その場合、「雪を見てオオカミだと判断する」と正解率は高くなるのでそういう学習をしてしまいます。訓練するときに「画像の中心だけ見なさい、背景は見ないで」とかそこまで訓練の仕方をチューニングすれば、この例は設計により避けられるかもしれませんが、そんなことまでやっていられませんし、これは事後だから気づいているという点もあります。

■世界が変わるとシステムが劣化するというのは、学習データが漏れていたということでしょうか？

⇒いえ、学習したときとはデータの傾向が変わるということです。「となりに別のコンビニがなかったときの売上げデータ」で、売上げ予測・在庫調整などの方法を学習したとして、となりに別のコンビニができてしまうと、学習したものの有効性は下がります。

《改めて:要求と仕様(と環境)》

■Zave / Jackson のモデルについて、例えば改札機に対して「2 人同時に入らない」「2 人同時に入れないスペース設計」は同じ内容で環境(仮定)と仕様のどちらとも捉えられそうですが、この境界はどう判断したら良いでしょうか？

⇒自分たちの制御下にあるかどうか、だと思います。スペース設計の余地がある、つまり自分たちの責任範囲として、物理的な機器のデザインも入るのであれば、それは仕様かと思います。

《「IoT・AI 時代」の品質？》

■機械学習において、想定環境はどう扱うのでしょうか？

⇒機械学習は、環境と向き合って要求を実現するシステムの一部品となる計算を実装するためのものなので、環境だろうが何だろうが、傾向や判断基準を学ぶ訓練データや、入出力に過ぎません。自動運転のための物体識別であれば、確かに歩行者や様々な運転状況が環境になりますが、これは物体識別であれば入力データになります。私が全く話さなかったこととして、強化学習というやり方は、もしかしたらご質問のイメージに合うかもしれません。

《機械学習における品質保証のための原則・思想》

■機械学習で得たモデルの性能を評価する何パーセントというメトリクスは、無限の条件からの偏ったサンプリングなので意味がないのではないでしょうか？

⇒究極的にはまさにその通りです。「リスクゼロにはできない」のですが、「意味がない」とはならないように、サンプリングの仕方、サンプリングされたデータの特性(霧の日はあるか、など)を検討し、リスクを軽減し価値を高めることを目指していきます。偏っている可能性はやはり否定できないとして、運用・継続的な監視を通して改善していくしかありません。

■100%完全と言うのは普通のソフトウェアでも無理なのは変わらないと思います(そんなプログラムが組めるなら正常系のテストだけすればよくなる)。 それらのソフトウェアの品質保証との考え方の違いは何なのでしょうか？

⇒機械学習を使っての正常と異常の境界は誰にも言葉では表現できない、確信が持てないものです。通常動くはずのデータでも突然異常な振る舞いをすることがあります。

■精度が上がらないなら、複数のモデルを用いて判断するということもありでしょうか？

⇒はい、それは十分にありです。複数のモデルを組み合わせて一つの判断器を作るような手法はよく研究されています。例えば、メールのスパム判定は、そもそもスパムというものの多様性もあり、複数のモデルの判断結果を合わせる方法がはまることが知られています。

■Models@run.time アプローチの「モデルをシステムに持たせる」に興味(具体的な方法)があります。

⇒簡単な例として、システム側の状態遷移図と環境側の状態遷移図を持たせて、設計時の想定と実動作との差を見つけるなどがあります。

《機械学習に対するテスト・検証技術の追求》

■「バグなのか」「確率的に外れたのか」と言うのは「バグの定義」が曖昧だと言う事とは違うのでしょうか？

⇒一つの解釈としては、同じだと思います。バグの定義が曖昧であるため、「テストを実行したときの結果(出力)を見て、バグの存在を検出する」という行為が難しい・不可能になるということです。
「バグの定義」が曖昧な中でも、「これは『バグ』と言っていいだろう、開発側としては責任が問われる問題点と言えるであろう」という種類のものがあると思います。例えば、ループの終了条件を誤っており、行列の最後の列が計算に反映されていない、といった、「設計の意図に合っていないコーディング」という人為的なミスです。人為的なミスは避けられないが、品質としては問題になるので、テストで見つけてつぶしたいわけです。ただそれが従来と同じ方法(テストの Pass/Fail を見る)ではできないということがお伝えしたかった点です。

■精度限界とバグに違いはあるのでしょうか？バグの結果として精度限界が低下する、という理解は間違いなのでしょうか？

⇒これは「バグ」の定義の問題で、この点私の話し方も不正確だったかもしれません。例えば、ループの終了条件を誤っており、行列の最後の列が計算に反映されていない、といった、「設計の意図に合っていないコーディング」という人為的なミスがあったとします。これにより精度は下がるかもしれません(与えたテストデータセットで下がるとは限りません)。そういうケースもありますし、いかに今の技術で可能な設計をいろいろ試し、上記のような明らかなバグ(ミス)はないとしても、精度が出ないということはあり得ます。製造責任を果たすためになくす努力をする「バグ」というものと、できないことを責められてもどうしようもない「技術の限界」と、後者も精度が下がる要因かと思います。後者を「バグ」と呼ばれて、責められると辛いかなと思います。

■サーチベースドテスティングとは、テスト設計をするものでしょうか、それとも、テストケースを選択するものでしょうか？

⇒テスト設計もしています。最初はランダムにコマンドを適当に打つことから始まり、学習進化し続けてカバレッジを高くしていきます。或いは、回帰テストのために数が少なく効率的に冗長性のないケースを選ぶこともできます。

■サーチベースドテスティングにて、テストが戦うというのはどういうことでしょうか？

⇒申し訳ございません、これはあくまで比喩です。実際には例えばテストスイートを実行してカバレッジを計る、テストスイートの総コマンド長を計るなどして、テストスイートに対して点数付けをする、その点数により勝者を決めるということです。

■サーチベースドテスティングにて、「欲しいテストケース」(不具合が見つかるテストケースが欲しい)に対してどのようにスコアをつけていけばいいでしょうか？

⇒実際何をやりたいかという問題依存性がありますが、似た問題の事例(論文しかないかも)を見てみるとよいかと思います。

■サーチベースドテスティングの具体的なやり方が知りたいです。テストケースまで自動で作ってくれるものなのでしょうか？(仕様書とか何か Input に入れたりするのでしょうか？)

⇒Facebook の例は、クラッシュを探すだけなので、仕様書などを使っているわけではなく、いろいろな入力コマンドを試せばいいというものです。様々な入力に対して、それぞれ出力が期待したものになっているか、まで確認しようと思うと、それでは済みません。機械処理可能な仕様(任意の入力に対して出た出力に、ちゃんとテスト正否を付けるためのアサーション)があればよいですし、今ではある程度期待値を推測するようなツールもあります。具体的な一歩としては、EvoSuite というツールが一番代表的なので、まずそれを使ってみるのがよいかなと思います。

■Facebook での自動バグ修正の事例紹介にて、強い、しっかりしたテストでないと成立しないとありましたが、それはどういうテストでしょうか？

⇒バグ修正してテストをパスすれば修正できたとの判断もできますが、テストが貧弱だとあまり意味がないものになるということです。期待するところを正しく表現するテストケースでなければなりません。

《全般・その他》

■機械学習では今までの品質保証が使えないというのは、どういうことが使えないのでしょうか？

⇒申し訳ございません、これは講演内容そのもので、例えば、今までの標準的なテストができない、といったスライドはあるので、もう少しどこが捕らえられなかったかご確認いただければと思います。
説明の対象としては、毎回の出力と、モデル(学習結果)がありますが、後者の話として回答します。「人に読める IF-ELSE 形式で学習する」とそもそもやってしまえば、学習結果は人に読めて論理的に説明できるものになります。ただ、深層学習の方が、精度が高くなるようなことは多いと考えられます。

■訓練画像の品質問題、そして、その判断情報の曖昧さ(情報の品質)が誤りに影響するのではないでしょうか？
■学習データの品質、判断情報の品質が、機械学習の要求問題ではないでしょうか？

⇒結果が誤っている・期待と異なっているときに、確かに訓練データの品質に問題がある可能性が一つ大きなものとしてあります。また「結果が正しいか誤りか」を判断するための判断情報(と解釈しました)もあいまいとなる場合、そもそも誤りとは何か、誤りとして検出した出力が本当に問題なのか、ということに影響します。その通りかと思います。

■これはエンピリカルなモデルと言えます。経験(学習)してブラックボックスのモデルをつくりあげていると言えます。ブラックボックスなので、品質としてコントロールできるのは、学習データの品質と判断情報の品質だけになります。その学習データの品質と判断情報の品質についての議論はどこまでできているのでしょうか？

⇒はい、しっかりご理解いただいていると思います。これらの品質についてはある程度議論されていますが、あまり一般的な指針として、しっかりとした言葉で整理されていません。例えば、訓練データのラベル付けが誤っていない(「歩行者」に対して「サイクリスト」というラベルを誤って付けてしまっていることがない)といった正確性であれば、一般的な品質特性としてあげることができ、標準はなくても「常識」としての感覚はあると思います。ただ、「偏りがない」「社会的に不適切な出力を含まない」といったことになってくると、アプリケーション依存性が出てきます。それらについての議論はまだ始まったばかりで、ガイドラインなどの初案がこれからどんどん出てくるという状態です。

■非 AI の製品も、一般市場向けの製品と、特定顧客向けの製品では品質保証で必要な方法は変わってきます。各社の AI は特定顧客向けの製品が多いでしょうが、世の中の色んなニュースや説明記事で一般市場向けの製品しか例に出てこないです。 本日の説明の「顧客との意思決定」と言っても「顧客」って誰だろうと思いました。

⇒アメリカにおいては、例えば Google のように、自社で機械学習を使ったシステムを構築し、それをエンドユーザ向けに提供する企業が目立っていると思います。これに対し、製造業や農業、サービス業などの企業が、ICT を得意とする企業に開発を発注するというケースがあると思

います。ICT を得意とする企業でも子会社での下請けなどがあるかもしれません。これらのケースでは、開発を行うチームだけが「価値を産む」ための意思決定を自由にできるわけではなく、依頼・発注元(これを「顧客」と呼びました)との対話や合同での意思決定が必要になります。

■医療のようなミッションクリティカルな分野に DL を用いるのが良いことなのか、分からなくなりました。

⇒DL の実行結果を医者が判断するような想定であれば、出力内容や想定するユースケースにおいて、「医者という人の存在も含めたシステム」として十分に信頼でき、有用なものになる可能性は十分あると思います。また、判断のぶれやブラックボックス性があったとしても、人間の医者にはそれがないというわけでも必ずしもありません。少なくとも DL というかソフトウェアには、「その日気分が悪くて雑に見てしまった」というようなことはあまりないかと思います。リスクゼロは無理だというのは、人間がやってもそうだと思うので、うまく折り合いが付く使い方を追求できればと思います。

■アサーションというのは、どのようにかけるのでしょうか？機械学習では中でどのようになっているかわからないのに？

⇒ご指摘の通り、推論過程(どこを見て歩行者と判断しているか)などの途中でアサーションをかけることはできません。例えば訓練を行うプログラムにおいて、「統計的にこういう計算をしているはず」というアサーションをかけることはできます。エンジニアとして直接送り出すのは、理論上・経験上適切とされる数学的・統計的計算を行うことで、これは直接製品となる部分ではありませんが、この部分を適切に作っているか確認するということが重要な場合や、それにより問題の切り分けが楽になる(「そこは大丈夫」と言える)場合はあると思っています。

■アシュアランスケースにおいて、「自動運転に対して、検証内容は妥当だ」というゴールを設定するには、どうすればいいでしょうか？

⇒これ未解決なのですが、Bosch の人たちが以下の論文などで試みています。私は不十分だと思っており、自動車業界の方々と追求をしています。
Structuring Validation Targets of a Machine Learning Function Applied to Automated Driving, SAFECOMP 2019

■自動運転だけではなく色々なモヤッとした不確かさを具体的にしていくノウハウや技術はあるのでしょうか？(仕様を作ったり読んだりするときに役立ちそうです)

⇒これは「モヤッと」という言葉をしっかり定義し切れていなかったかもしれません。少なくとも３つの軸があると思っています。
・あいまい 対 厳密: 1 つの言葉が、人によって複数の意味にとらえられてしまう状態。

・抽象 対 具体:「現時点では最低限の制約だけ決めて、詳細な点は決めない」といった意識(暗黙的かも)により、詳細を捨てて本質だけ表現している状態
・不確か 対 確か:現時点で検討を尽くしても明らかにできないこと、運用時など後になってから初めてわかるであろうことがある状態。

今回の話題である機械学習を用いたシステムの話ではなく、一般的な仕様に関しては、最初の 2 点は非常によく話題に出ます。現場の問題としてよく聞くのは 1 つめです。原則・指針・手法としては 2 つめが重要となりますが、意識し使いこなすのは難しく、意識していない人も多いかもしれません。SQiP 研究会第 5 コースでの主題の 1 つです！(去年の栗田さんのご講演)

■内閣府が 12 月に発表した指針「人間中心の AI 社会原則」にある「AI の説明責任」にはどうすば対応できると思いますか？
■学習結果の論理的説明は可能なのでしょうか？

⇒今回私の方ではほとんど取り上げませんでした。阪大原先生の資料(総務省の Web サイトにおいてあります)がよいと思います。
http://www.soumu.go.jp/main_content/000587311.pdf
技術的にできることに限りがあります。説明ができない深層学習を使うことで精度が上がったところもあります。無茶なことを求められてもどの企業も対応できないので、「指針」からの現実的な落とし所を決めるのだと思います。政府の動きに惑わされず(気にしないわけには行かないかと思いますが)、「説明」とは誰のために、何が必要なのかなど、ぜひ考えてみていただきたいです。

■テストの定義ができていますか？

⇒これはとてもよい質問です。私の資料は、いろいろな種類の「テスト」をまぜこぜに話してしまっていたと思います。非常に広義には、国語辞書にあるくらいの「性質や力などをためすこと、または検査すること」ということまで指す話をしてしまっていると思います(Google のガイドラインなど)。別の文脈では、実行結果に対するアサーションを設けて Pass/Fail を判断するような狭義のテストの話をしていました。そもそもテストとは、というところから、もう少しうまく整理して話せるようにしたいですね。(という質問だと解釈しました。)

■人間の能力値のデータはありますか？例えば、画像認識の精度、年齢に応じたデータ入力量や、判断に必要な説明変数など。

⇒申し訳ありません、具体的に調べていませんが、きっとあると思います。認知科学や、人工知能・機械学習と認知科学をつなごうとする方々が、きっと出していると思います。エンジニアリングというより、知の追求としてとてもおもしろい話題だと思います。

■画像の認識について素朴な質問ですが、平面画像で奥行きは認識できるのでしょうか？

⇒例えばカメラを 2 台使えば奥行きは求められるかと思います。専門ではありませんが、1 台でも推定する技術はしっかりあると思います。

■機械学習は、時間的な情報(例えばコンビニの季節による判断)を学習できるのでしょうか？

⇒時系列変化だろうがそれはただのデータなので、教えれば学習できます。そばの売上げが 12/31 に高い、ということが毎年起きていたら、その傾向は学習して、予測に使うことができます。ただ、そのときに、季節とか大晦日とかそういう概念を理解してくれるわけではないです(12/31 という日付と売上げデータだけを訓練で教えたとするならば)。12/31 が大晦日ということを教えれば、もちろん「大晦日にはそばが売れる」ということは学習できます。

■自動運転で必要なのは「目の前に何が出てきたら止まる」ではなく、「目の前に何かが飛び込んできたら止まる」ではないのでしょうか？

⇒私の言い方のどこにひっかかったかを把握できていませんが、後者は必要な一つの要求です。ただ、停止標識で止まる、渋滞で止まる、といった、必ずしも「飛び込んでくる」とは呼ばないことはあると思います(日本語のニュアンスの問題かもしれませんが)。