2019 年 6 月 5 日(水)
2019 年度 人工知能学会全国大会（第 33 回）企画セッション
「機械学習における説明可能性・公平性・安全性への工学的取り組み」

主な Q&A

Q: 公平性のチェックは誰が行うことを考えておられますか？ 開発者ですか？ ユーザーですか？ 国のような認証機関ですか？
A: 特に議論されているのを見たことはないです。規準が単純な独立性なので、認証企業でも、開発者の自主検証でも実施は可能と個人では考えます。

Q: FADM のコンペティションは実施されているかと思いますが、どんな事例・データが使われていますか？
A: まだコンペは知る限り行われていません。国勢調査、予測警備、再犯予測、ローンの可否などのデータが使われています。公平性配慮型分類アルゴリズムの比較調査論文 https://doi.org/10.1145/3287560.3287589 が参考になります。

Q: 個人情報保護法の要配慮個人情報のように、そもそもセンシティブ情報の入手に強い制限があることがあります。すると、現実には学習結果とセンシティブ情報とが独立/条件付き独立であることも分からないと思うのですが、なにか対策はあるのでしょうか。 差別を防ぐためには差別につながる情報を入手、分析するのはやむを得ないのでしょうか。
A: センシティブ特徴を予測するモデルを別のデータから作って、その結果と独立性を保つようにするといったことが考えられています。Sweeney さんの例では、人種は出生記録という別データから推定されています。

Q: Formal fairness を議論する際に、sensitive な feature 自体の必要性を議論するような話はありませんか？ 人種は、再犯判定にはフェアネスの問題があるが、病気の判定には、問題ないように思います。
A: センシティブ特徴に何を採用するかはドメイン依存で、技術的というより、法などに依存して決めるべきものとコミュニティでは認識されているようです。

Q: 統計的なエラーにより生じる倫理的な問題と、人間のバイアスに汚染されている状態により生じる倫理的な問題と 2 種類の問題がある気がしますが、結局倫理的な問題の解決は現状難しい気がします。
A: 前者は帰納バイアス、後者はアノテーションバイアスなので、そのどちらに対処するのかを決めれば対処は可能と考えています。

Q: モービルアイ社は、大量データによる機械学習で説明可能性が低いとして、シンプルなルールで説明可能性の高い自動運転ソフトを開発する方針と聞いたことがありますが、同社のこの開発方針は、その後どのように推移しているのでしょうか？
A: モービルアイが開発した説明可能性の高い自動運転ソフトにぴったり該当するものが見つからなかったのですが、シンプルなルールに基づいた自動運転技術という意味で、Responsibility-Sensitive Safety (RSS) モデルのことではないかと思いましたので、その前提で回答いたします。（引用：https://www.intel.co.jp/content/www/jp/ja/automotive/responsibility-sensitive-safety.html）
モービルアイが 2017 年に発表した RSS とは「ほかの車両がどのような動きをしたとしても自動運転車が事故の原因となることがないように」、「自動運転者を確実に安全な方法で運行できるように」するための「数学的公式モデル」です。RSS のイメージは、例えば、先行車との距離や自車速などとの関係からどの程度先行車に近づいたらブレーキをかけるべきかを数学的に導き出します。視界が悪い交差点では仮に飛び出しがあっても安全に停止できる速度まで落とすなど、人間ドライバーが常識的に行動するようなシーンも織り込まれています。このモデルの使われ方は、普段はデータから学習した AI の出力結果を基に自動運転しつつ、ルールベースの RSS がそれチェックし、AI の間違いによって重大事故に至りそうな場合は RSS によって回避する、いわばセーフガードとして働くものです。
RSS は自動運転の認知・判断・操作の 3 段階の機能のうち、判断において安全を担保するための有力な考え方の一つだと我々は捉えています。（その後の RSS の開発方針は同社のホームページ等を参照していただきたく、ここでは深追いはしません。）
一方で、RSS が有効に機能するのは認知段階で他車や歩行者など周辺環境を正しく認識していることが条件になると考えています。
つまり、認知側で対象物の見逃し（False Negative）や対象物でないものを対象物という見間違い（False Positive）があると、後段の判断側で対象物との距離や方向を正しく見積もれず、セーフガードが有効に働きません。現在、ディープラーニングベースの機械学習手法

が主流となっていて、ルールベースでは代替が難しい認知 AI に対する説明性・安全性は、依然として open problem であると我々は考えています。

Q: 性能の保証は不可能ではないでしょうか？ 従来通り、プロセスの保証でよいのでは？
A: 「必ずこういうときにはうまくいく」といったルールに従うことの保証は難しいですが、結果としてあるテストデータに対してこれだけ性能が出た、という評価は可能かつ重要です。一方で、テストも不完全（試していない入力に対する挙動は保証できない）ので、「適切に構築した」というプロセスの確認も必要です。このため、モデル、それを含むシステムについても、プロセスとともに保証・評価の対象になると考えています。

Q: 国際標準が道路交通法などに反映されたケースはありますか？ ドイツでは道路交通法を、自動運転を目指して改正したと聞いていますが。
A: ご指摘の通り、ドイツでは「限定的なレベル 3 相当の実用化を認める道路交通法を 2017 年 6 月から施行」しておりますが
（引用：https://www.kantei.go.jp/jp/singi/it2/dourokoutsu_wg/dai1/sankou5.pdf）、AI の文脈で法規に反映されたケースはまだないという認識です。
国連 WP29（自動車基準調和世界フォーラム）や各国政府で自動運転の議論が進められているように、AI システムの利用が想定される産業セクターごとに関連法規はあると思われますが、AI を利用するしないに関わらずアプリケーション側の視点で法規が整備されていくものと思われます。
AI を明示的に謳った国際標準については「ISO/IEC JTC 1/SC 42」や「IEEE P7000 シリーズ」の標準化が検討されていたり、各国の産官学から AI の倫理や品質保証という文脈でガイドラインが発表されていたりしますが、いずれにしても、これらが法規に反映されるかどうかは不明で、反映されるとしてもまだ時間がかかるものと思われます。

Q: 医療システムで利用される AI については、ガイドラインで言及されていますでしょうか？
プライバシーの扱いはガイドラインに入れなくてよいのですか？ GDPR では設計段階からプライバシーを考えるプライバシーバイデザインのルール化がされているようですが。
A: ともに重要かと思います。ガイドラインは専門家のボランティアでボトムアップにできていっているので、リクエストを投げていただくとともに、貢献についても是非ご検討いただければと思います。

なお、以下の 3 つについては、回答が難しく、調査中です。

Q: 説明の validation には、言及されていないのでしょうか。

Q: 心理学的／効果を重視ということは、説明を受けるユーザーが納得すれば、必ずしも正しい説明かどうかわからなくても良いということなのでしょうか？ ※正しい説明自体の定義は置くとして。

Q: 目的の最終段階にある trust の定義を教えてください。システムが安定して動くという技術な意味の trust ですか？(robust に近い？) あるいは、心理学的な trust ですか？両者はかなり違う方向を狙うと思いますが。。。

以上