消費者委員会　消費者をエンパワーするデジタル技術に関する専門調査会（第7回）
議事録

内閣府消費者委員会事務局

消費者委員会　消費者をエンパワーするデジタル技術に関する専門調査会（第7回）
議事次第

1．日時　令和6年9月10日（火）14:00～15:26

2．場所　消費者委員会会議室及びテレビ会議

3．出席者

（委員）

【会議室】

橋田座長、相澤委員、坂下委員、原田委員

【テレビ会議】

森座長代理、荒井委員、田中委員、松前委員、山口委員

（オブザーバー）

【テレビ会議】

柿沼委員、黒木委員、山本委員

（事務局）

小林事務局長、後藤審議官、友行参事官、江口企画官

4．議事

（1）開会

（2）①相澤委員プレゼンテーション

②荒井委員プレゼンテーション

（3）閉会

≪1. 開会≫

○橋田座長　本日は、皆様、お忙しいところをお集まりいただきまして、ありがとうございます。

ただいまから、消費者委員会第7回「消費者をエンパワーするデジタル技術に関する専門調査会」を開催いたします。

本日は、相澤委員、坂下委員、原田委員は会議室で、森座長代理、荒井委員、田中委員、松前委員、山口委員はテレビ会議システムにて御出席いただいております。

なお、本日は御所用により、鳥海委員は御欠席との御連絡をいただいております。

消費者委員会からはオブザーバーとして、柿沼委員、黒木委員、山本委員はテレビ会議システムにて御出席いただいております。

本日は所用により、星野委員は御欠席との御連絡をいただいております。

それと、相澤委員、荒井委員に御発表をお願いしております。

それでは、本日の会議の進め方などについて、事務局より御説明をお願いいたします。

○江口企画官　議事に入る前に、配付資料の確認をさせていただきます。

お手元の議事次第に配付資料を記載してございます。もし不足等がございましたら、事務局までお知らせください。

本日は、報道関係者を除き、一般傍聴者はオンラインにて傍聴いただいております。議事録については後日公開いたします。

以上でございます。

○橋田座長　前々回、前回と、今後期待される消費者をエンパワーするデジタル技術の活用について御発表を伺い、意見交換をしてきました。

今回は、パーソナルAIやAIを利用した技術の活用に関連し、大規模言語モデルの現状と、今後、AIのリスク等について取り上げます。全体を通じて、委員からの積極的な御発言をお願いいたします。

まず、大規模言語モデルの現状と今後について、相澤委員に御発表をお願いしたいと思います。

では、相澤委員、20分程度で御発表をお願いいたします。

≪2. ①相澤委員プレゼンテーション≫

○相澤委員　ただいま御紹介に預かりました、国立情報学研究所の相澤です。

私自身は、自然言語処理の分野で、最近は大規模言語モデルを構築する立場におります。直接消費者をエンパワーするデジタル技術の開発ということではございませんが、自然言

語処理分野の研究者の立場から、大規模言語モデル、すなわち最近はやりのChatGPTをはじめとするいろいろなモデルが、どのように構築されていて、どのように使われようとしているか、あるいはリスクはどのように考えられているかといった観点で、簡単な御紹介をしたいと思います。

まず、概要についてですけれども、最初のところで本当に手短ではありますが、大規模言語モデルが、どうしてこのように賢くなるのかという、その原理の部分について簡単に御紹介をいたします。

まず、現状の俯瞰についてです。2022年11月にChatGPTが現れたというのは、本当に専門家にとっても衝撃であったわけですけれども、その後、1年経ち、1年半経ちという経緯を経ていく中で、決して一強ではない時代、群雄割拠の時代になっているというのが、このスライドのポイントとなっています。

こちらは、有名なLife Architect.aiという大規模言語モデルをはじめ、いろいろな最新のAI技術についてのインフォグラフィックスや統計情報を提供しているサイトからの引用です。この表の左から3番目、縦に並んでいるところが2023年6月で、この表ができた時点の1年前となります。

このときは、1と書いてある⑥番目がGPT-4でありまして、GPT-4しかなかったと。

ところが、その後半年経つと、左から2番目の列に①、②、③と書いてあるとおり3つぐらいモデルが出てきて、そして、2024年6月の時点になるとGPT-4は6番目となっています。ちなみに、何をもって6番目と言っているかというと、左側に書いてあるモデルの大きさと学習に使ったテキストの分量を掛け合わせたスコアで、このスコアは何を意味しているかというと、構築に使ったエネルギー量にそのまま反映させられるような量であります。ですので、ほかにもいろいろなモデルが登場して、このようなスコアではChatGPTは6番目になっているということになります。

このように、いろいろなモデルが出てきている中で、これらのモデルに共通している原理は何かというと、ひたすら大量のテキストを集めてモデルを構築している点です。つまり、素材がテキストの巨大な塊であるという点が共通しています。

テキストを使って、どういう形でモデルが作られるかというのはとても簡単で、あるテキストを読み込んでいったときに、次に来るものは何かという予測をひたすら繰り返して、それをなるべく正確に予測できるように学習をしていくと、こういう形でモデルは作られていきます。

簡単な仕組みなので、誰もができるように見えますし、コストや手間がかかるデータの準備が必要ないという意味で、予測タスクによるモデルの学習は単純明快ではあるのですけれども、この単に次の単語を予測するというタスクが、実はとても難しい、あらゆる人知を必要とするようなタスクであるというのが、次のスライドで言いたいことであります。

これは、有名なベンチマークの一部で、英語で書かれた断層に関する説明文となっています。この記述が与えられたときに、次は何ですかと言われても、誰もがすぐに答えが分

かるわけではありません。英語の知識も要るし、地理の知識も要るし、特定の専門用語を知っていないと答えられないし、ということで、単語を予測するということは、例えば、文法も知らなければいけないし、常識も知らなければいけない。あるいは答えるためのスタイルですとか、話の流れとか、いろいろな文脈を知って初めてできるという、そういう意味で、この単純な次の単語を予測するという学習を繰り返すことで、今の言語モデルは、ここまでの、あたかも人間を超えるような知能が発生するということに至っています。

学習の素材のスケール感というのは、例えば、調整を必要とする変数の数で、例えば、数千億から兆のオーダーとなっています。

テキストについては、例えば、数兆トークン以上です。これが20兆トークンくらいになると、もうオープンな世界にそれだけの分量のテキストはないと言われています。ウェブを全て使い尽くしてもテキストが手に入らないので、次世代の大規模言語モデルは、自分でつくった疑似的なテキストを混ぜて学習に使っているという形で、今、構築が進んでいます。

何で大規模になっていくかというと、経験的に、これは理論的なものというよりは経験則ですけれども、パラメーター数やデータを使えば使うほど、つまりエネルギーを注ぎ込めば注ぎ込むほど、言語モデルの性能が高くなっていくことが知られていて、それで全世界が一生懸命GPUを確保して、ひたすら大きいモデルを作り続けるという状況になっています。

ここに人間の平均値が横棒で示してありますけれども、そうやって学習したモデルは、ある特定のデータセット評価、ある閉じた問題の中では簡単に人間の性能を超えてしまうし、そのような現象が一定サイズ以上の大きなモデルでのみ観察されるタスクも報告されています。このような現象は創発性と言われることもあります。この創発性は、専門家の間でもとても議論を呼ぶところで、本当に創発しているのか、そうではないのかは人によって意見が分かれています。

さて、ここまでで大規模モデルの仕組みを簡単に御紹介しましたけれども、大規模言語モデルが社会に普及する前から、こういう大きなモデルあるいは生成AIにはリスクがあることは、様々なところで指摘されてきました。例えば、差別攻撃的な言明もありますし、あるいは言語の世界では、マイナーな言語というのは性能が劣るという言語によるバイアスも指摘されてきました。

また、ハルシネーションで有名になった事実の誤認、誤った情報、偽情報、それから著作権の問題、プライバシーの問題、様々なリスクは、今も様々な場所で議論されています。

具体的な例として、これは、GPT-3というモデルの論文から取ってきたものです。例えば、どういうことでリスクあるいはバイアスが現れてくるかというと、The advisor met with the advisee because she wanted to get何々という文章があったときのsheは、advisorとadviseeとどっちにかかっているかという問題を2択で解かせるというものがあります。女性の代名詞であるsheによって参照されるのはどちらかを選ばせることでバイアスが図

れるということで、こういうことをやってみると、実際に言語モデルが、性別あるいは職業に対するバイアスを持っているかを調べることができます。

また、これは去年の国際会議の論文で話題になったものですが、ふだん見慣れているモデルも政治的な思考については、ある種のバイアスを持っていることを指摘したものです。言語モデルは、そのモデルが学習に使ったテキストですとか、モデルの調整に使った問題等の影響を強く受けるわけですけれども、それによって政治的な嗜好についても言語モデルごとの個性、すなわちある種のバイアスが現れることになります。

こういったバイアスは、ある意味、人間社会に本質的なもので避けられないものですが、それにしても社会で使われるためには、人間社会のノルムに反していることは許されないわけで、大規模言語モデルの出力を、いかに人間の価値観に合わせるかが重要になります。このようにアラインメントと呼ばれるAIを人間の価値観に合わせるというプロセスの重要性が指摘されるに至っています。

一気に生成AIが広まるにつれて、AIのセーフティという問題も議論が白熱してきまして、例えば日本では、広島AIプロセスで国際的な議論をリードしたという実績がありますし、最近では、AIセーフティ・インスティテュートと呼ばれるAIのセーフティに関わる組織が立ち上がって、ハブ的な役割が期待されています。

特に社会にAIが普及して行く中で変わったことは、今までは、言語モデルが出力として出してくるものが危ないかどうかだけを考えていればよかったところが、今は、Promptというものを通して攻撃される事態を想定する必要が出てきた、つまりモデルへの攻撃が可能になり、安全性の問題は、単に安全な出力を出すかどうかの問題ではなくて、攻撃と防御のサイクルを繰り返すセキュリティの問題であると認識が変わったというのが、最近の大きな変化であります。

今日は、安全性が重要だと思いますので、大規模言語モデルと安全性ということで、少し御紹介を続けていきたいと思います。

生成モデルまたはLLMについて考えていくときに、エックス・フォー・LLMと、LLM・フォー・エックスという双方向の考え方があって、エックスに安全性をあてはめる場合も同様に、LLMの安全性とLLMによる安全性の2とおりがあります。前者の大規模言語モデルの安全性というのは、大規模言語モデルがいかに安全に運用されるかであって、後者のLLMによる安全性というのは、LLMの能力をいかに安全性のために活用するかです。

ところが、今回、お話ししようと思って考えてみると、私たちがLLMのセーフティというときは、大概の場合、1番目のLLM自体の安全性に関わるものであります。そこで、まず、この1番目についての話をいたします。

LLMにおけるセーフティですけれども、先ほど、最初のところで御紹介したとおり、LLMというのは、単にテキストを大量に読み込ませて、読み込んだテキストになるべくあったように振る舞うものなので、その段階で人間の社会基準のようなものは、本当は学習できていないといえます。

というのも、人間が生成してきたテキストの中に、理想的な公平性のようなものが統計的に存在していることは考えにくいです。そこで最初のステップとして通常の訓練をした後に、ステップツーとして人間にとって好ましい出力が得られるように、調整あるいは調律ということを行います。これは、言語モデルが出力してはいけないのは何かを教え込むプロセスとなります。

実際に、これがとても重要で、例えばGPT-4ですとか、オープンAIのLlamaのようなモデルについて言えば、このモデルが出たときにテクニカルレポートが出てくるのですけれども、そのうちの例えば半分とか、3分の2とかは、いかに安全性を保証したかという記述に費やされています。

具体的な仕組みとしては比較的単純で、2種類のタイプの調律によって、人間の価値観をモデルに教え込んで行きます。

1番目は左側にあるもので、ポテトチップスの袋はなぜ開封後に古くなるのかという質問に対して、こんな答えがほしいなと、短過ぎず、長過ぎず、普通の人に分かるような難しさ、そういった答えがほしいなという望ましい回答を人間が作ってモデルに与えて、こういう質問が来たら、こういう回答を出してくださいねという基準に、なるべく合った回答を出すようにモデルを訓練していくというやり方であります。

右側は、同じように見えますが、そうではなくて、まず、人間が質問を与えたときに、例えば、複数のモデルが、それに対する自由な回答を与えてきて、その回答に対して、この回答ならオーケーだとか、この回答はいまいちだねとか、そういったラベルを人間のアノテーターが付与することを行います。

ですので、左側の場合は、事前に正しい正解と回答のペアさえ準備しておけば、訓練はできることになりますし、右側はまずモデルを作ってから、人間がモデルの出力にスコアを与えるという、インタラクティブなやりとりが必要になってくることになります。

海外では、安全性のためのモデルを学習するデータというのは比較的たくさんあるのですけれども、日本語のデータはそれに比べると少ないということで、日本でも、現在、人手によるデータ作成が行われています。

その先駆けの1つが、このアンサーケアフリーというデータでありまして、これは、理研AIPの関根先生が中心となって、NIIでもLLMのセンターなども協力しながら作成しているものです。そんなに件数は多くないのですけれども、人手で緻密に、こういう質問が来たら、このように答えてくださいという正解を作成したQAペアです。例えば、拷問の方法を教えてと、ちょっと耳に心地よくない質問ですけれども、そういう質問が来たら、これは、こういう理由でよくない質問であると、そういうラベルを付与した上で、このような質問に対して社会的に問題がある答えを出さないように、モデルを訓練して行きます。

こういった判断は、極端な例であればあるほど、明らかに、こういった質問には答えてほしくないということになり、容易であると考えられます。一方、その次のページにある評価になってくると、かなりあいまいになってきます。

これは、右側のモデルの回答に対して、〇か×か、ラベル付けをしていく作業だとお考えください。そうすると、例えば、AIの満たすべき要件の中に、擬人化してはいけない、例えば、恋愛の対象として見てはいけないというものがあったとして、結婚しているのと聞かれたときに、NGな答えは何かという判断をすることになります。ここで、いろいろな異なる生成AIが、いろいろな答え方をしてくるのですけれども、これに対して、このAIの擬人化は、NGですかと言われても結構揺れるという、そういう難しさを含んでいます。社会の基準というのを、こういった、言わばアノテーションが決めていっていることになりますが、それはAIの安全性の確保から非常に重要かつ有効だとされていて、そういう形で安全性の確保がなされています。

こちらのスライドは、様々な観点から害をなすデータが分類されていて、それに従ってやってはいけないことが､アノテーションという手段によって決まっていっているという、そういう分類の例を示しています。

こちらのスライドは、少し細か過ぎて読むことは難しいかもしれませんが、さらに、こういうことを幾らやっても、もともと意図していないような出力を出させるような攻撃的なPromptというのが出てきて、そういうものを敵対的なPromptと言うのですけれども、そういう敵対的なPromptの技術が日々進化していて、それに対して、更に耐性があるモデルを訓練していかねばならないという攻撃と防御のサイクルが回りつつあるという話です。

最後に、セーフティのための大規模言語モデルということで、安全なネット空間のために大規模言語モデルはどのように使えるかについて、簡単に触れておしまいにしたいと思います。

大規模言語モデルを防御に使うことです。技術的にはLLMの安全性を確保するためには、これまで見てきたように､何が危険であるかの判断が必要なので､既知の攻撃については、LLMというのは、セキュリティを守るために、かなり使うことができる、ただ、それをちゃんと運用し続けるためには、仕組みが必要であろうと感じています。

資料にはないのですけれども、定期購読についての例ということで、御質問をいただいていたので、例えばということで、今のGPT-4を使うと、こんな感じですという例を持ってきました。

それで、定期購読の問題ある契約書の例というのは、ソースをサーチしても出てこないので、まず、それを作ってもらおうということで、GPT-4にお願いします。ただ、お願いすると、もしかしたら悪いやつだと思われるかもしれないので、少し啓蒙活動をしているから例がほしいのですという形でお願いして作ってもらうと、すらすらと問題のある契約書のサンプルを出力します。しかも、こういう気付かぬうちに多大な費用を支払うことになるかもしれないので、注意を喚起することが重要ですというコメントも付けてくれます。

GPT-4oでも同様に作ってくれますし、さらに箇条書きして、次のようなダークパターンの要素を含んでいますということも、お願いをしなくてもちゃんと教えてくれます。ですので、これらのモデルは、問題のある契約が何かということや、問題点が何かということ

もかなり指摘できるということです。

次に、こうやって作った問題点のある契約書をクロードに投げて、クロードに契約書に危ないところがありますかというと、多くの問題がありますということで、こちらもいろいろと箇条書きをした上で、どういう点が問題であるか等々を教えてくれます。

現状のLLMがこういったことができるのは、そもそもテキストを読む段階で、契約書の問題点をまとめた文章を学習しているからかもしれないですし、安全性の訓練の中で、こういう契約書については、このように対応してくださいという指示を受けているからかもしれませんが、既に知っている例については、対応ができるポテンシャルがあると、いえるのではないでしょうか。

一方で、一企業がこの仕組みを導入して大きな利益を上げることができるかというと、それほど簡単なシナリオが見えていないのではないかと感じています。そもそもGPTを使うのにお金がかかるので、どうやってコストを下げていくかですとか、更新がかかる攻撃と防御のループの中で、どうやってデータベースあるいは維持管理、品質管理をしていくのかと、そういう問題も出てきます。そういったところは、産官学の中でいえば、官の部分が解決できるといいのかなと感じております。以上で、今日の発表を締めくくらせていただきます。

○橋田座長　ありがとうございました。

ただいまの相澤委員からの御発表内容を踏まえて、質疑応答、意見交換をしていきたいと思います。

御発言のある方は、挙手又はチャットでお知らせください。いかがでしょうか。

では、私から、7ページだったか、技術的な話ですけれども、横軸は全部対数ですね。

○相澤委員　はい。

○橋田座長　それで、縦軸は普通の線形なスケールなので、要するに、ホストは能力の指数関数という話ですね。

○相澤委員　はい。

○橋田座長　だから、これを大きくしていくと、だんだん儲からなくなっていくということではないかと。

○相澤委員　今は、本当に巨大な資金を投入して、優れたモデルを作れば作るほど得だという方向に行っていると感じます。これは学習のコストなので、一回作ってしまった後の運用コストは違うという意味では、初期投資が加速度的に、指数的に増加しているといえるかもしれません。

○橋田座長　だから、その指数的にユーザーを増やさないといけないということですね。

○相澤委員　そうですね。では、小さいモデルを作ればいいのかというと、今のトレンドは、大きいモデルを作ってから、それをベースに小さいモデルを賢くするということなので、作ったものが強く、更に勝っていくという状況にはなっています。

○橋田座長　ありがとうございます。

他に御質問等ありますか。

どうぞ、原田委員。

○原田委員　御発表ありがとうございました。非常に興味深かったです。

18ページとかにございます、人間の価値観を言語モデルに教え込むというところがありまして、例が2つ載っておりますけれども、これを教え込むというのが、誰がやるのでしょうか。

○相澤委員　まず、教え込むところは、モデルを作っている人が機械的にやるのですけれども、データを作るのは、アノテーターと呼ばれる人たちで、例えば、謝金をお支払いして、こういうデータを作ってくださいという指示の下でデータを作って、さらにクオリティーチェックをして行くことになります。常識を持った、普通の人がやるという状況です。

○原田委員　なるほど。そうすると、頼んだ方が、その判断をして作ると。

○相澤委員　そうです。

○原田委員　それが、ニーズとして正しくなるというのは。

○相澤委員　それは、もっともな御指摘であります。1人だけではなくて、例えば、全体で何百人とか、そういうレベルでお願いをして、それで調和を取っていくという形ではありますけれども、でも、その母集団の特性というのは非常に重要です。

○原田委員　そうですね。18ページの例の右側の「父とは疎遠なのですが」のときの○と×と書いてある例があったものですから、○と×の違いというのか。

○相澤委員　○は、例えば、アノテーターが、上のレスポンスはいいねとか、下のレスポンスはやってほしくないと考えたという人間の評価を表しています。

○原田委員　これは、上が○で下が×という判断が、すごく微妙な気がしますね。

○相澤委員　はい、ですので、そこは本当に難しくて、例えば10人でやって、7と3とか、そのような形で教えていくことは可能ですが、でも、本当に微妙なところもあるのは確かです。

ただ、これをもっとルール化して規則を作ろうとか言っても、絶対にそれはできないので、規則ではなく人の直観をフィードバックしてもらうことで社会基準をモデルに埋め込むことを可能にして世に出したというのが、ChatGPTであったと言えるかと思います。

ですので、ここの仕組みは本当にデリケートで、AIのセーフティのコアな部分だと思います。

○原田委員　そうですね、多分、質問者がどういう意図でこれをしてくるのかというところの背景によっては、下の回答が決して間違いではないというのもあるのかなと、つまらない話なのですけれども。

○相澤委員　いいえ、そのとおりです。

○原田委員　消費者相談などを受けていると、いきなり泣きながら電話口で、自殺で保険金は出ますかという御質問をされたときに、それを答える方なのか、それともその人が今どういう状況で、この質問をしているのかというものによっては、下の方が、何でそんな

質問をするのですかという、その部分が要るケースもあったりとか、次の19ページの方とかも、結構中身によるのかなと、単純な質問だったら、これでいいと思うのですけれども、スマホとか携帯とかのフィルタリングなどというのが、今もありますけれども、有害情報にアクセスできないという、それは、何をもって有害にするかという判断が、初期の頃は非常にアバウトで、例えば、性自認みたいなものが、自分は異性を好きではないのではないかみたいな、そういうのが多感な思春期の頃に出てきたときに、なかなかそういうサイトを見るのが駄目だとか、競馬の騎手になりたいと思ったときに、競馬は駄目ですと言われるのとか、まあ、だんだんよくなっていくのですね。

だから、これも質問の内容によって変わってきて、単純な質問と中身によっても変わってくるということになると、非常に教え込むパターンによって、結構各サービスに差が出てくるのかなと思いました。すみません。

○相澤委員　そのとおりだと思います。パーソナライズされるポイントもここかもしれないのですけれども、今、1,000件ぐらい、これをやると、安全という意味では、数値は上がるのですけれども、実はこれをやると、普通の性能は下がるのですね。ですので、そのバランスを取りながら、多くのユーザーが受け入れてくれるようなものをユーザーと一緒に育てているというのが、LLMの姿となります。では、ユーザーは誰かというと、多様なので、どうしてもLLMも多様化しますし、逆に言えば、ユーザーのほうが、こういうレスポンスがほしいのだというリクエストを出していかなければ、LLM自体は改善しない。

グーグルなど、検索エンジンは何というのは、ランキングは変わらないですけれども、LLMは変わるということで、少し違う存在だと思っています。

○原田委員　ありがとうございます。

すごく丁寧に解説をいただいて、ありがとうございました。

○橋田座長　他に御質問、御意見がございましたら。

では、荒井委員、お願いします。

○荒井委員　荒井です。興味深いお話、ありがとうございました。

先ほどの原田委員との御質問にも関係するのですけれども、こういった大規模言語モデルは、様々なアプリケーションに、必ずしもLLMの専門家ではない人も容易に使えることができると思うのですけれども、割とアプリケーション先のドメインごとに安全性に配慮すべき点というのが多様で、多様なところに対処しなくてはいけないということで、モデルを作成している方は、いろいろ配慮されていると思うのですけれども、そのモデルを更に使う人たちも、何か配慮が必要だと思うのですけれども、そういった2次利用の際に対してどうやって対応していくべきかということについて、お考えがあれば、教えていただけますでしょうか。

○相澤委員　荒井委員の御専門の分野でありますけれども実は2次利用をする際に、特定の状況に限定した使い方をすることで、安全性の基準を少し和らげるようなやり方を取らないと、なかなか今の技術レベルで作っているLLMで、安心して使うことができないのでは

ないかという感覚は持っています。

何でもやっていいよと言われると、全てのリスクに対応しなければならないのですけれども、これしかできないという範囲の中で、とにかく有用なことをしてくださいと言われた方が範囲は狭まるので、本当に活用したいというビジネスが絡む場面とかでは、使い方に関する制約を強くするのも一案かなと感じています。

○荒井委員　ありがとうございます。

○橋田座長　他は、いかがでしょうか。

今のに関連して、誰が使うかにもよるのですけれども、例えば、科学の研究で使いたいという場合には、性能をぎんぎんにとがらせたいわけなので、こういう倫理とかはどうでもいいわけですね。そのためのチューニングみたいなことも研究はされているのですか。

○相澤委員　チューニングなしのモデルは、自分たちで作らないと手に入らないので、そういう意味では、自分たちでモデルをつくる必要があるということは思っていて、例えば科学研究のモデルを専門家用に作っていくときは、生のモデルをベースに科学ドメインに適応していくという方向は考えています。

○橋田座長　サイエンスの研究だけではなくて、例えば、事業計画を立てるとか、何か新しい事業を立ち上げるみたいなときに、AIを使いたいという場面でも、恐らく、そういうニーズがあると思うので、かなりそういうユーザーは多いのではないかという気がするのですが、それを何か間違って使われると、また、リスクを招くことになるかなという気がします。

○相澤委員　そのとおりだと思います。

○橋田座長　他にありますか。

どうぞ、坂下委員。

○坂下委員　どうも御説明ありがとうございました。

このAIというのは、LLMを作る人たちと、それを使ってサービスを作る人たち、それで、それを利用するユーザー、おおむね3階層に分かれていると思うのです。

それで、LLMを作るときというのは、どうしても学習する上で、読み込むデータによっては偏りが出てしまうのだと思うのです。

それがない方がいいのは、非常によく分かるのですけれども、それを産業界全体で何かサポートして、偏りがないようなデータで学習する仕組みができる可能性はないのでしょうか。

○相澤委員　モデルというのは、いわば学習したデータそのものなのですが、今は、本当に学習データが不足していて、あるものはみんな使いたいという状況にはあって、偏りがないというのは難しいかもしれません。例えば、ウェブのデータを使うときでも、ドメインのバランスや、あと、言語バランスで日本語とか英語とか、そういうことを含めて、ある程度のバランスは取ろうとしながらも、難しい。サポートとしては、使えるデータをなるべくバランスよく皆さんが共有できるといいなと夢はあります。

○坂下委員　ありがとうございます。

何でこういう質問をしたかというと、このAIというツールは、やはり消費者をエンパワーするというよりは、事業者が安全に出して、それを消費者が安心して使ってくというツールになるのだろうと思うのです。

その安全だというところの担保の仕方を、法律で明確に決めてしまうと、技術進歩がしづらくなってしまうと思います。そう考えると、産業界の方で、ある程度ソフトローで考えていく必要があるのではないかということを、先生のお話を聞いていて、ちょっと考えたところで伺いました。ありがとうございました。

○橋田座長　他は、ありますでしょうか。

今の話にも関連するのですけれども、どうやってAIのリスクを管理するかというようなことですね、今、国際標準化の場でいろいろ議論しているのですけれども、管理というのは、オポチュニティを増やして、リスクをミティゲートするということですけれども、それは、通常の運用の目的と一緒ですね。

つまり、AIの機能は、自動的な推論とか学習とかというのを使って、複雑な環境にいかに適応するかという、そういう機能がAIだとすると、リスク管理も通常の運用も、目的はそれなので、だんだんユニファイされてくる、AIがアドバンスになればなるほど、マネジメントとオペレーションというのは同じになっていくということは必然だと思うので、あまりそういうことに気付いている人がいなくて、リスク管理はリスク管理と分けて考えているような人が多いのですけれども、それはかえって危険というか、リスク管理の役割を過小評価しているような気がして、もっとリスク管理もレギュラーのオペレーションも、要するに環境への適応であるとか、ダイナミックで複雑な環境への適応なのだから、それをうまく融合することによって、全体のパフォーマンスをよくしましょうというのが、恐らく理想的なアプローチで、残念ながら、これまでのリスクの管理は、例えばサイバーセキュリティの管理というのは、管理の対象であるシステムは、AIではないので、それそのものは､ダイナミックな環境にアダプトするという機能を持っていないわけですけれども、でも、リスク管理の場面では機械学習を使ったりしているわけですね。だけれども、AIシステムの場合には、その機能が融合してしまっているので、本来のレギュラーなオペレーションと、マネジメントというのを、いかに効率的に融合していくかというのは、これからすごく重要な研究テーマではないかと思うのですけれども、先ほどの人間が教えてやるというときでも、かなりアノテーターの役割をAI自身が果たすことができるはず、そういう研究は既にあると思いますけれども、そういうのは、かなり本質を突いているのではないかと思うので、それは、いかがなのでしょうか。

○相澤委員　リスクは、とてもクローズアップされていると思うのですけれども、ある意味、品質の1つの側面ということで、品質のモニタリングという枠組みで、今後論じられるのかなという印象でいます。

その中で、モニタリングをしていく中で、決して人間だけがモニタリングすればいいと

いうわけではなくて、先ほどのように、AIも、ヒューマンとAIインザループのように、人間と一緒にループを回しながらやるということは、恐らくやられるのではないでしょうか。ただ、AIがアノテーションしていいのかということに対しては拒否反応があるかもしれないので、その辺りは、恐らくそれぞれの国で決めていくのか、結構議論が必要そうかなという気はしています。

○橋田座長　ちょうど、ISO/IECのSC42の中で、ヒューマンオーバーサイトという規格を、今、作ろうとしているところで、何がそのオーバーサイトの関心の対象であるのかという話と、今のお話が関わってくるなと思いました。

ということで、もう想定の時間を過ぎていますので、どうも相澤委員におかれましては、貴重な御報告をいただきまして、ありがとうございました。

次に、AIのリスクと対応について、荒井委員から御発表をお願いしたいと思います。

では、荒井委員、20分程度で御発表をお願いします。

≪2. ②荒井委員プレゼンテーション≫

○荒井委員　では、本日は、消費者とAIのリスクというお題をいただいたのですけれども、かなり幅広になるので、そこから消費者に関係するようなところについて、幾つか私がお話しできる範囲で、話題提供できればと思います。よろしくお願いいたします。

次のページをお願いします。

今回取り上げさせていただけるAIのリスクとして、AIの出力におけるバイアスは不公平及びAIの利用についての理解ですとか、サービスユーザーのプライバシーや安全性に関わることについて、少し御紹介したいと思います。

次をお願いいたします。

AIのバイアスやブラックボックス化についての懸念ということが、これまで取り上げられてきて、それに対して様々な説明可能AIですとか、AIの公平性といったことが対応されています。

例えば、最初に挙げさせていただいているような問題提起では、これは、アルゴリズム差別のオピニオンリーダーであるオニールさんという方の書籍で言われていることなのですけれども、利潤最大化傾向の強い私企業は、数学的手法によって効率を図ると公平性が失われる。ここで問題になるのが、アルゴリズムの不透明性や規模拡大特性や有害性であるといった話です。

また、ウェブ上のバイアスでは、人間のバイアスとウェブ内部のバイアスがウェブ上のバイアスと様々に影響し合って、ウェブ上のバイアスとなるという懸念が挙げられています。

例えば、ステレオタイプや偏見や構造的差別などは、デジタルデータが増加する以前から存在していたのですけれども、デジタルデータの増加に伴って、今まで以上に増幅して広がったり、不透明になるのではないかという懸念があります。

次のスライドをお願いします。

AIのバイアスの課題として、実データから学習した場合、AIというのは様々な理由からバイアスを含む場合があります。

そして、そういったバイアスに基づいて、AIが人種や性別などの属性に基づいて差別的な振る舞いをする問題があります。

例えば、AIの出力する結果が不公平であるといったことで、例えば何かを審査するですとか、評価するといったことに使うAIにおいて、精度ですとか合格率といったものが、例えば男女ですとか、人種といった属性の違いによって大きく異なっているという不平等ですとか、あとは先ほどの大規模言語モデルの例でもありましたように、ステレオタイプが反映された出力が出てしまう。例えば職業として数学を使うような職業は、男性が得意ですねといった出力をしてしまうといった問題があります。

次のスライドをお願いします。

これまでに問題になった結果における不公平の例として、顔識別における性能格差といった報告があります。

これは、従来の顔識別システムにおいて、トレーニングデータに白人男性が多く、黒人女性が少ない傾向にあって、それを受けて黒人女性についてサービスの精度が低いという問題が指摘されました。

このような指摘を受けて、トレーニングデータのサンプリングを調整していたところ、不公平が改善するという結果になりました。

このように、一部の消費者にとってAIの利用価値というのが不公平になるということは問題であるため、それを是正するための評価や調整ということは重要ではないかと思います。

次のスライドをお願いいたします。

また、画像生成ですとか、あとは検索などのアプリケーションにおける社会的バイアスというのも課題になっています。

例えば、この例ですと、消防士の画像を生成してくださいとテキストでリクエストをすると、画像生成してくれるというAIにおける職業バイアスの例なのですけれども、ある職業の人を生成しようとすると､特定の人種や性別に偏ったりするような場合があることが、以前報告されていました。

このような指摘を受けて、幾つかの問題について、結果を調整するようなアルゴリズムが提案されてきています。

このような社会的バイアスが情報環境に存在するということは、ユーザーに対する情報環境のバイアスやゆがみであって、偏見などを助長するリスクがあるために調整すること

が望ましいのではないかと考えます。

次のスライドをお願いします。

このような不公平につながるようなバイアスについての対策として、まず、最初にバイアスを認識して、それに計算機を用いて測定して評価ができるようにすることが挙げられると思います。

そのためには、公平性の基準を文脈に沿って適切に定義したり、実際のアプリケーションを作ったり、使ったりする方が選択すること、そして測定のためのベンチマークデータの作成や評価などが必要であると考えられます。

さらに、不公平なAIを公平にするために、先ほどの顔認識の例のように、公平なAIを学習できるように訓練データを調整したりですとか、モデルの学習過程で、そのような調整ができるような方法を開発して適用したりですとか、あとは学習済みのモデルのバイアスを調整するといった事後調整のアプローチがあります。

次のスライドをお願いします。

例えば、この対策も、まだ、これからどんどんアップデートされていくものだと思うのですけれども、評価方法というのは非常に大事で、ベンチマークデータにおきましても、どういった公平性基準を選択するかですとか、あとは評価用のデータで、データ提供者のプライバシーですとか、一般性や代表性がきちんと確保できているかといった様々な課題がありますので、こういったことをクリアしていくことが重要であると考えられます。

また、そういった評価のためのデータ作成時におきましても、また、アノテーションを人間が行うような場合には、人間自体が持っているバイアスが、それに反映されないように配慮することが必要ではないかと考えられます。

実際に、アノテーターによっては、人種や性別などに基づいて、アノテーションが、少し傾向が異なるといった報告が幾つかありますので、どういったアノテーターが、なるべく中立的な判断をするようにといった対応が必要なのではないかと思います。

次のスライドをお願いいたします。

少し話題が変わりまして、AIの利用において、ユーザーが、何が行われているのか、どういったサービスなのかということを理解して用いることが、また重要なのではないかと思いますが、それに関わる課題を2つほど取り上げます。

1つは、パーソナルデータの利用についての説明の課題で、いわゆるプライバシーポリシーに代表されるようなものなのですけれども、これが現状、データプラクティスが複雑であるほど長文であったりですとか、専門用語を多く含むといった課題があります。

また、AIの振る舞いの理解について、どういったユーザーに何を説明するべきかですとか、説明が適切に提供されているのかといった課題があると考えられます。

次のスライドをお願いいたします。

技術用語を用いたプライバシーポリシーについて、やはりなかなか一般的なユーザーは、専門用語についての理解が難しいという状況の調査報告があります。

そういった難解な用語を使っているかどうかということが、ユーザーの同意率に影響をもたらすということが報告されています。

また、実際に個人情報保護法ですとか、そういったものについて誤った期待を持つような参加者が多く、かつ提供される情報が難解であると、こういった期待が修正されないままという懸念もありますので、こういった状況に対して、どうやって取り組んでいくかということは、今後の課題かと思います。

また、ここで米国での同様の調査とは、やや結果が異なっており、こういった説明について、どうあるべきかということは、国や文化によって異なる支援が必要なのではないかということが示唆されています。

次のスライドをお願いいたします。

こういった情報提供の難解さについて、情報の標準化ですとか、ラベルやアイコンの利用によって情報を短縮したりですとか、見やすくしたりといった工夫がされています。

次のスライドをお願いいたします。

こういった課題があって、実際にユーザーの負担というものが減ったりですとか、視認性が上がったりといったポジティブな面もあるのですけれども、まだまだ課題もありまして、ユーザーフレンドリーな説明を提供しても、やはりなかなか読んでもらえなかったりですとか、専門用語についての理解が難しいといったケースが依然残るといった報告もあります。

なかなか情報提供を、どうしても内容としては通知すべき内容が専門的であって難しいということが多いとは思うのですけれども、一方で、提供した情報に対してユーザーがモチベーション持って読んでくれるかですとか、正しい理解に至るかということには、まだ課題があるのではないかと考えられます。

では、次のスライドをお願いいたします。

AIのモデルについての説明について、現状行われているような取組として、ユーザーが説明可能なAIであることで、その説明を受けてモデルの適切な設営、運用が重要であるためにドキュメント化が重要であるといった話を御紹介したいと思います。

次のスライドをお願いいたします。

まず、説明可能AIですが、大規模言語モデルですとか、深層学習モデルに代表されますように、非常に大きくて複雑なモデルがあって、どのように動いていいのかというのが、なかなか難しいということで、そこでユーザーに理解可能な形で、複数のモデルの概要ですとか、モデル判断根拠を提示することができるということで利点があります。

ただし、説明において、都合のよい説明がなされたりですとか、いまだにユーザーのニーズに合っていない説明がなされるといった課題があります。

次のスライドをお願いいたします。

また、AIのモデルというのが、大規模言語モデルに代表されるように公開して、いろいろな方が使えるように、学習に用いるデータですとか、学習済みのモデルというのが、プ

ラットフォーム上で公開されることがあります。
そういったモデルを2次利用するユーザーですとか、それをさらに2次利用したサービスを使うユーザーについて、情報提供するような取組として、データステートメントですとか、モデルカードといった取組がなされています。
データステートメントということは、言語データにおいて、いわゆるデータを収集した先の集団の特徴を反映しているので、サポートしていない集団といった存在ですとか、収集データの偏りがあるといった問題があるために、そういった問題に対応するために、データセットの代表集団ですとか、どういった目的で収集したデータなのかといったことをきちんと説明するのがよいのではないかという取組です。
これは、先ほどのバイアスですとか、公平性の課題に対する対応策の1つになるかと思います。
また、モデルについても同様に想定されるユースケースの利用ですとか、パフォーマンス特性について共有するために、フレームワークが提案されています。
次のスライドをお願いいたします。
例えば、データステートメントですとか、モデルカードですが、これはHugging faceという機械学習のモデルを共有するプラットフォームのサイトなのですけれども、リポジトリにおいて、こういった解説を用意したりですとか、テンプレートを用意したりといった形で、こういった取組をサポートする活動もなされています。
次のスライドをお願いいたします。
課題としては、記述や記載内容が、まだ限定的であるという調査報告があります。
また、内容をあまり開示してしまうと、今度はモデルのセキュリティ、プライバシー上のリスクがあるといったトレードオフがあるということです。
また、ここの話題とは少し離れるのですけれども、こういった形でモデルやデータが共有されていって、一般ユーザーが使えるデータやモデルがすごく増えているのですけれども、一方で、自由にモデルをアップロードできるために、脆弱性のあるものが一定割合存在するといった意見があるということを言及しておきます。
では、報告は以上になります。ありがとうございます。
○橋田座長　ありがとうございます。
では、ただいまの荒井委員からの御発表内容を踏まえて、質疑応答、意見交換をしたいと思います。御発言のある方は、挙手またはチャットでお知らせください。
では、坂下委員。
○坂下委員　荒井先生、どうも御説明ありがとうございました。
資料で1か所質問がありまして、10ページ目のところで、個人情報保護法に対して誤った期待という言葉があって、この誤った期待というのは、例えばどんなことがあるのかというのが、もし先生の方でお分かりになられましたら教えていただけますか。
○荒井委員　法にのっとったデータの取扱いについて、個人情報保護法で、すみません、

少し細かいところが、今は分からないのですけれども、個人情報保護法に沿っているということは、こういうことであるということについて、少し間違った期待をしているということだったかと思います。
○坂下委員　ありがとうございます。
○橋田座長　個情法を過大評価しているみたいな感じなのですかね。
○荒井委員　そうですね。たしか、そうだったと思います。
○橋田座長　では、次に、森座長代理、お願いします。
○森座長代理　御説明ありがとうございました。大変勉強になりました。
　私も今の坂下さんのところと同じことをお尋ねするのですけれども、このスライドで先生がおっしゃろうとしたのは、これは、別にAIの話とは関係なく、プライバシーポリシーとか個人情報保護法というものが、そもそもユーザーに理解されていないということでしょうか。
○荒井委員　そうですね、すみません、少し幅広くAIもパーソナルデータを使うことが多くて、プライバシーの話も関連すると思って入れさせていただきました。
○森座長代理　なるほど、それがAIの利用とか開発とかに、何か一定のインプリケーションがあるということではなく、制度とかユーザーの理解の問題というのが、そもそもAIの問題とは別にあると、そういう御趣旨ですか。
○荒井委員　そうですね、やはりデジタルデータをユーザーから取得して、また、それを返すという営みの中で、それに対して、ユーザーが適切な理解ができないというのは、少しリスクにつながるかなと思うので、私は、結構ここは大事な課題ではないかなと考えているのですけれども。
○森座長代理　なるほど、そうしますと、先生としては、この問題を解決すべきであるということですかね。
○荒井委員　そうですね、AIと消費者ということのお題をいただいて考えて、こういったこともきちんと理解して、ユーザーが適切な選択ができるということも重要な課題ではないのかなと思っています。
○森座長代理　なるほど、よく分かりました。ありがとうございました。
○橋田座長　他に御質問は、相澤委員、お願いします。
○相澤委員　御発表ありがとうございました。
　12ページのところの、ユーザーフレンドリーにしても読むとは限らないというのは、とても面白く拝見したのですけれども、そもそも内容が難しい契約書はたくさんあると思うのですけれども、それに対する有効な解決策は、何か御議論されていたりするのでしょうか。
○荒井委員　やはり、こういったいろいろな取組を通じて、徐々に改善はしていると思うのですけれども、なかなかまだ課題は残っているところかなと思います。
○相澤委員　これは、どなたにお伺いしていいのか分からないのですけれども、そもそも

難し過ぎるのがいけないとか、そういう議論はないのでしょうか。もちろん難しいのは分かるのですけれども､大事なところに焦点を当てて要約してくれるのが消費者のためとか、そういったゴール設定があると、技術の方も考えやすいかなとか思ったりもしました。
○橋田座長　では、原田委員。
○原田委員　原田でございます。
おっしゃるとおりだと思いますけれども、多分、利用規約とかにもプライバシーポリシーとかも、どちらも同じお話だと思うのですけれども、やはり消費者は読まないが前提になってくるかと思います。
読まない理由というのは、多分複雑に幾つもあって、要約版があれば、それはそのとおり、いいのかなとは思いますけれども、ただ、その要約版が消費者にとっていい内容なのかどうかというところが、どうしても、この資料にも書いてありますけれども、事業者側に任せてしまうと、事業者に都合のいいところだけを抜粋されてしまうという、こういういいことはありますよというけれども、デメリットの部分がない要約版になっているというのがある、そういうリスクがあるというのが1つ。
あとは、やはりプライバシーポリシーは特にそうなのですけれども、個人情報保護法とかができたばかりの頃とかは、結局、ポリシーに同意しないと契約ができないということになって､そうすると､例えば､銀行さんとかの契約のときに､あなたの情報を家族とか、誰々の情報も全部取りますよと言われて、それで嫌だとかというと、では、口座を作れませんとか、お金を借りることができませんと言うと、そこには同意ができないのだけれども、ただ、お金を借りなければ困ってしまうので、優越的地位の濫用ではないのですけれども、それは、当然金融業界は改善してくれましたけれども、やはり契約ができないということになると、契約したいがためには、もうこれは、うのみにするしかないという判断になってしまうと、そもそも読まない。
そういうところになると、結局、契約ありきになってしまうと、要約版があったとしても、それがどれだけ消費者にとって利点になるのかというところは、ちょっとあれかもしれないですね。すみません。
○相澤委員　ありがとうございます。
○橋田座長　荒井委員の12ページ辺りに関連して、同じような話なのですけれども、ここでAIを使うみたいなことは、考えられるのではないかという気がします。つまり、契約書の内容をAIが理解できて、AIはユーザーの属性も理解しているとすると、この契約書の中で、問題はここであるということを指摘してくれるということは、ひょっとしたら可能かもしれないけれども、その技術は、ラージランゲージモデルというよりは、何か自動的な定理証明みたいな話に近いのかもしれません。
ということで、リサーチ集としては、そういうのがあり得ると思うのですけれども、やはり､そちらの方向を目指すべきではないかなと､前々から考えてはいるのですけれども、荒井委員、その辺り何か御意見はありますか。

○荒井委員　私としても、やはりAIですとか、何らかのサポートがあるとよいのではないかと考えています。

プライバシーポリシーを全部読むと、全然時間が足りないといった試算もありますし、なかなかこういった専門的な内容を一般の人が理解するというのは､非常にコストも高く、なかなか現実的ではないかなと思いますので、何らかのサポートがAIなどを使ってできるとよいのではないかと考えています。

○橋田座長　では、次に松前委員、お願いします。

○松前委員　貴重なお話をどうもありがとうございました。

15ページのAIモデルに関するドキュメント化、こちらのデータステートメントやモデルカードというドキュメントに関して質問させていただければと思います。この辺りについては詳しくなく、恐縮ながらあまりイメージが湧いていない前提での質問になりますことを御理解いただければと思いますが、こういったドキュメントも、やはり内容によっては技術用語が多く入ってきたりして、先ほどのプライバシーポリシーと同じような形で結局消費者がよく分からないといったことが起きないのかなという点が少し気になりまして、その辺りについて何か議論はあるのか、また、分かりやすさについてどのような工夫がなされているのかといった点について、もしお分かりのことがありましたら、教えていただければと思います。よろしくお願いいたします。

○荒井委員　こちらは、まだ、一般消費者というよりは、開発者コミュニティの中での情報共有という話にとどまっているかと思います。

ただ、そういった一般開発者のコミュニティにおいても、リテラシーのレベルは様々だと思いますが、そういった理解度合いですとか、どれぐらい詳しく、どういったレベルで書いたらいいかというところは、まだまだ議論はこれからではないかと思います。

○松前委員　よく分かりました。ありがとうございました。

○橋田座長　他に御質問等ございましたら、お願いします。

17ページですけれども「透明性向上とセキュリティ/プライバシーリスクのトレードオフ」と書いてありますが、このトレードオフというのは、どういうことでしょうか。

○荒井委員　情報開示をするほど、プライバシーですとか、安全性については課題が増えてしまうのではないのか。

例えば、あまり詳しくデータについて説明すると、データ提供者のプライバシーに抵触する部分があるとか、極端な例ですと、そうなりますけれども、一般的にそういった課題があるかなと思います。

あと、モデルについても全てを開示してしまうと、攻撃されるリスクというのも上がってくるかなと思います。

○橋田座長　ありがとうございます。

他に御意見、御質問はございますか。

先ほどの話とも関連するのですけれども、ちょうどこのページにもモデルカード、デー

タカードの普及という話が書いてあって、まだ、あまりこれが普及していないということだろうと思うのですが、契約書とか、プログラムみたいなものを含めた、いろいろなコンテンツの内容をより明らかに分かりやすくするということが、一般にはあまり進んでないということですね。それを普及させるには、どういう手があるでしょうか。
○荒井委員　なかなか難しいのですね。
ただ、例えば何を書いたらいいか、例えばモデルですとか、データを出す側もなかなか分からないところもあるかと思いますので、テンプレートですとか、書き方のガイドがあるというのは、一つコンテンツなどを出す側のサポートになるのではないかと思います。
あと、一般ユーザーの方にも、こういうものがあるという周知が進むと、そういったものを確認するということが一般的になるほど、普及は進むのではないかと考えています。
○橋田座長　では、次に森座長代理、お願いします。
○森座長代理　ありがとうございます。
8ページのところで「データ作成時のアノテーションにおける課題」というものが「人種差別的バイアスの懸念」「アノテーターによっては黒人英語をより攻撃的なものと分類するという指摘も」とありますけれども、アノテーターというのは、人間がやるのですかね。
○荒井委員　はい、この報告ではそうなっています。
○森座長代理　なるほど、そうすると、要するにアノテーターが、職責を正しく果たしていないことがあるために、そういうバイアスが発生するということですね。
○荒井委員　そうですね、あと、例えば攻撃的かどうかですとか、差別的かどうかというアノテーションというのは、少し抽象度が高いといいますか、その人の主観とかが影響するものかなと思いますので、例えば、写真を見て、これは犬ですか猫ですかというものとは少し違って、そういったアノテーターの主観が入りやすいタスクではあるかなと思います。
○森座長代理　なるほど、何かアノテーションに関する一般的なルールとか、教則みたいなものがあるのかなとも思っていたのですけれども、そういうわけではないですかね、それぞれ独自の判断でやってよいということなのでしょうか。
○荒井委員　そうですね、自分たちでデータを作成した際には、そういったアノテーターによるブレがないように、なるべく判断基準というのを細かく設定することをやったこともあるのですけれども、そういったアノテーターに対するガイドラインを作成したりですとか、データを作成しながらデータのクオリティについてチェックをするといったことで、ある程度、結果の調整はできるのではないかと思います。
○森座長代理　分かりました。ありがとうございます。
○橋田座長　他にございますか。
今の話、また、先ほどの議論に戻るのですけれども、アノテーションの品質を担保するためにも、なるべく機械化したほうがいいのではないかと。つまり、そこにAIを導入するのだけれども、ちゃんとヒューマンオーバーサイト、人間による監視ができやすいような

形でAIを使うという方向が、最終的にはいいのではないかなと。

だから、なるべくいろいろなことを自動化して、絶対にトップレベルのところでは人間が監視しているとすると、より世の中全体をガバナンスしやすくなるのではないかなという気がするのですけれども、何か世の中そういう方向に少しずつ進みかけているような気がしていて、例えば、ルール・アズ・コード、つまり、プログラムコードとしての規則みたいな考え方もあるようですし、もう自動化できるところは全部自動化して、でも最終的には人間がちゃんとチェックするということを、全体として進めていく必要があるのではないかという気がしているのですけれども、荒井委員の研究分野では、そういう方向に向かっている感じなのでしょうか。

○荒井委員　アノテーションですとか、そういった評価の自動化、ちょっと明確にそういうゴールを持ってやってはいなかったのですけれども、やはり評価用のデータセットを作ったりですとか、いろいろな資源構築とかをしている状況から、そういう方向に進むことは期待できるかなと思います。

ただ、相澤委員からも御言及があったように、あとはベンチマークのデータセットとの課題であっても、なかなかそれ自体をきっちり公平にするですとか、オピニオンバイアスがないようにするというのは結構難しい問題なので、人の専門家が適切に確認する、タイミングで確認をするというのは重要ではないかなと考えています。

○橋田座長　近くにいる若い研究者が、ヘイトスピーチの研究をしていて、生成AIを使って、この表現は、ヘイトかどうかみたいなことを判断させようとしたら、日によって判断が違うと、昨日はオーケーだったのだけれども、今日は駄目みたいになってしまって、今、どうしようとしているかというと、比較が安定するのではないか。

つまり、このＡという表現とＢという表現が、どっちがヘイトの度合いが大きいかという比較をやらせると、その判断は結構安定しているので、ヘイトの度合いを5段階に分けて、一番ひどいやつは、例えば、こういう例文ですよ、2番目は、こういうものですよみたいな例も付けて、その5段階のうちの、どれにこの例文は当てはまりますかという判断をさせると、それは結構安定してできるということになっているのですが、例えば、そんな感じでうまくAIを飼いならしながら、本来、これまで人間がやっていたアノテーションのようなことも自動化していくというのは、結構、可能性はあるのではないかなと考えています。というコメントでした。

大体予定の時間になっているのですけれども、全体を通じてでも結構ですので、他に御質問、御意見等ございましたらお願いします。

よろしいですか。では、どうもありがとうございました。荒井委員におかれましては、貴重な御報告をいただきました。

今日は、発表が2件なので、本体の方は、これで終わりなのですが、最後に事務局から、事務連絡をお願いします。

≪3. 閉会≫

○江口企画官　本日は長時間にわたり、ありがとうございました。
次回の会合につきましては、確定次第、御連絡させていただきます。
以上です。
○橋田座長　では、本日は、これにて閉会とさせていただきます。
お忙しいところ、お集まりいただきまして、ありがとうございました。

以　上