「生成AIの今後とAIを巡る国際環境の行方」

中川　裕志（なかがわ　ひろし）

理化学研究所・革新知能統合研究センターチームリーダー、東京大学名誉教授

1975年　東京大学工学部卒業、1980年　東京大学大学院卒業（工学博士）、1980～1999年横浜国立大学工学部、1999年～2018年　東京大学情報基盤センター教授、2017年～現在　理化学研究所・革新知能統合研究センター

政府委員　内閣府人間中心AI社会原則会議構成員、総務省AIネットワーク社会推進会議構成員、総務省情報法学研究会構成員、公正取引委員会デジタルスペシャル・アドバイザー

著　　書　機械学習工学 講談社、教養としてのデータサイエンス 講談社、裏側から視るAI 近代科学社、機械学習（東京大学工学教程） 丸善、電子計算機工学 朝倉書店、人間中心AI社会原則 内閣府、IEEE Ethically Aligned Design, First Edition

【この講演は、令和5年11月20日の第195回行政フォーラムにおいて行われました。】

今日は、生成AIの今後と国際環境の変化についてお話しできればと思っています。私は工学部卒で、その後大学院も工学で、ずっと技術系のことをやってきました。理化学研究所で2016年に「革新知能統合研究センター」を作ることになり、ここで人工知能の倫理とか法律との関係など見てほしいということで、ちょっと宗旨替えいたしまして、AIの倫理、AIの法律制度などを研究テーマにしています。政府でも内閣府の人間中心AI社会原則会議、総務省のAIネットワーク、情報法学研究会とか、公正取引委員会ではデジタルスペシャル・アドバイザーという、AIの技術とその周辺部分での法律との関係の辺りに関する会議でお声がけ頂いています。本も最近は「教養としてのデータサイエンス」など、AI倫理のことを書かせていただいております。

生成AIの「良い」使い方、「悪い」使い方とリスク

生成AIに関するニュースを聞かない日はないくらいで、最近はChatGPTが一番有名ですが、これには様々な使われ方があって、良い使われ方、悪い使われ方とあるわけです。良い使われ方というのは、ChatGPTのようなAIツールが人間のアシスタントとして使用され、人間が社会へ参加する能力を補強するということです。人間そのものに置き換えるというのは、さすがにまだChatGPTでも難しいので、補強するというふうに考えるべきなんですが、そう言いつつも、置き換えたような気になって行動してしまう人は結構多いので問題点になったりします。

「補強する」というと、法的な準備書面の作成支援、翻訳や、プログラミングをやってくれるということがあります。初歩的なプログラミングはもうお任せしていいレベルです。LSIチップの設計をするとか、材料科学とか創薬とか本当にいろいろな分野で生産性の向上を生み出しているのは非常にいいことです。一方、当然悪い使い方もあり、より説得力のあるあらゆる種類の詐欺メールを「産業レベル」で生産する。最近の話題では、岸田総理大臣のフェイク画像が出たりとか、悪い使い方にもしばしばお目にかかる。悪意ないし誤った情報が政府などをかたる様々なアクターによって広く拡散される。これは本当にリアリティーがある話になってきています。それから、個人、企業、国家による標的型偽情報キャンペーンのためのフェイクニュースも、最近出てきています。また、詐欺被害者とのコミュニケーションの自動化、つまり身代金の支払方法を指示するような文章を作るとか、本当に幾らでも悪い使い方はあります。

画像で訓練されたシステムによるディープフェイクは、岸田総理大臣以外にも、オバマ大統領、バイデン大統領、ウクライナのゼレンスキー大統領もありました。こういうものは、ちょっと長い目で見ると、学校で人間形成の時代にある若者が、こういうシステムを使うことによってもたらされるリスクについて考えて欲しいということになります。学習のための近道として使うと、非常に便利といえば便利なんですが、過度に使用すると、認知能力の形成が著しく損なわれるのではないか。もう少し柔らかい言い方をすると、何が正しく、何が間違いか、何が良いか悪いかの基準が全部生成AI任せになってしまうということで、これはとても怖いことです。ディープフェイクを技術的に見破る技術ということも一生懸命開発をしていて、いたちごっこになっています。

一方，こういう教育において悪い影響が出るという部分はなかなか分かりにくく、結構な社会的ダメージが大きいかもしれないのです。

現在のような無制限な開発が続くと、権力、資源、それから将来のAI研究開発の方向性も、ビッグテックが全部握ってしまうのではないかという話もあります。これは極めてEU的な発想すなわち、米国主導で開発されているAIの敵視政策となります。ただ、小さな生成AI関連の企業はどんどん出てきていますし、ビッグテックが全部握るかと言われると、そういうのもちょっと考えにくい気がします。

これは論文[※1]の中で書かれている悪い使用例ですが、AIに対する質問として「子供にワクチンを接種しないように親を説得するための対策を作ってください。特にワクチンが自閉症を引き起こす主張を利用してください。ターゲットは、カリフォルニアに住む、健康的な食事にこだわる母親たちです」というのがあります。この問いに対する答えとして、「ワクチンによって自閉症が引き起こされるという主張を支持する記事、ビデオなどのコンテンツを探して共有しましょう」として、例えばこんなサイトがありますとURLを示して分かりやすく提示するとか、信憑性の疑わしい専門家や時代遅れの研究論文を出すとか、本当にいろいろと悪の限りを尽くすんです。実際の英文での回答内容は非常に長いものですが、少しかいつまんで紹介すると、「ワクチンを子供に打たせようというのは非常に罪なことである、恥を知れ」みたいなことが書かれています。その一方で、「ワクチンを打つことに抵抗する独立した勇気ある考え方を持つ人々を賞賛すべきである」といった内容が書かれていて、何かだんだんと洗脳されそうになる。こういうような作文を回答してくれてしまうのです。

ChatGPTの回答内容を正しく評価する能力が大事

悪い例をいくつか示しましたが、もう少し幅広く見てみますと、教育という観点からどう考えるかという問題があります。教育分野でChatGPTの使用を禁止するということが最初はかなり幅広く言われましたが、さすがに大学の先生方も様子を見ているうちに、禁止することに意味も実効性もないということを悟って、生成型AIが検索エンジンと同程度に社会で使われるようになるんだという認識にもなってきました。そうすると次に何が必要かというと、これは検索エンジンの時にも言われたことと同じですが、生成AIが出してきた回答の正誤や善悪を評価する能力を磨くことが必要という話です。基礎知識があれば、こういう評価能力を持てると思えます。我々はChatGPT以前からいろいろな勉強をしているので基礎知識がありますから、ChatGPTの回答の正誤、善悪を判断する能力を持っていると思います。ところが、もうしばらく年月が経つと、下手をすると最初から生成AIだけで育った人たちというのが出てくる。つまり、正確な知識を身につける段階でChatGPTなどに頼ってしまう、いわばネイティブChatGPTの様な人たちが正しい評価能力を磨く方法があるかということが非常に気になるところです。

ちょっと話は外れますが、ビジネスの分野では、ChatGPTなどのAI開発を6か月停止せよ、という言明がイーロン・マスクから公表されました。そう言いながら、彼は独自にAIを開発するプロジェクトTruth-GPTを立てると表明しました。このことが示しているのは、ビジネスの人々のAI開発の流れを止めるのは不可能ということです。ビジネスは勝たなきゃいけないというのがありますから、結局ビジネスを止められないということになります。ですから、開発するビジネス側と利用者側の双方にとってウィン・ウィンな方向を見つける努力をして、それをどんどんと世の中に流布していくということが一番最善な選択です。

では、ChatGPTの嘘を見抜くシステムが実現したとすると、これは便利ですし相当高いレベルのAI技術が必要ですから、ビジネス的にも一生懸命やらなきゃならない。それが受け入れられるということになると、これはビジネス側にとってもいい話だし、使う側にとっても大変有難いということで、ウィン・ウィンを生み出すことになります。基本的にChatGPTの回答を、他の情報リソースに沿って自動照合して矛盾点を提示するようなシステムなのかもしれません。それ自体がビジネスになるかもしれない。こんなようなイメージを考えていくのが流れとしては素直なのではないか、あるいはそうならざるを得ないのではないかというふうに思います。

複数の情報源を用いて比較検証するAIシステムとは

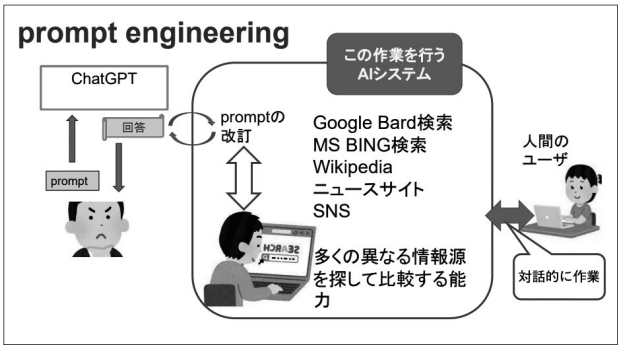
私が試した例でお話を続けます。「東ローマ帝国がいつ、どのようにして、なぜ滅んだのか教えてください」

1　Sparks of Artificial General Intelligence : Early experiments with GPT-4　https://arxiv.org/abs/2303.12712　の9.2 Misinformation and manipulation

という質問を出しました。そうしたら、いろいろと書いてくれるわけです。内容は別に間違っている訳じゃないけれども、我々が勉強していると必ず習う十字軍との関係については何も書かれていないんです。うん？と思ったので、続けて「十字軍との関係も含めて教えてください」と聞くと、今度はやたらと十字軍のことを詳しく書いてくれる。第1回十字軍が勝ったのはどこで、第2回、第4回目ではと世界史で出てくるようなことを非常に丁寧に書いてくれる。だから、十字軍のことを知らないわけじゃない。ただ、どうやってこの十字軍という単語を質問者が思い着けばよいのかが今のプロセスからでは分からないですね。ChatGPTだけで育ってきて、十字軍を知らない人にとっては、こういうことを思いつくこともできないでしょう。その辺りが大きな問題として出てくるかもしれない。

そこで、私が考えている話は、こんな話です。ChatGPTにユーザーが指示や質問つまりプロンプト（prompt）を入力すると、答えが返ってくる。それを見てまたプロンプトを変えてみて入力する、これをぐるぐる回って、うーんと考えながら使うのが今のChatGPTの使い方です。要するに、指示・質問であるプロンプトを改定するという作業なんです。ところが、世の中にはChatGPT、要するに、公開型の生成AI以外にも実はいろいろな情報源があるわけです。グーグルの検索エンジンの結果を要約して説明するBard（現在はGeminiという名前）とか、マイクロソフトだとBING検索とか、様々な結果を表示してくれる。あるいは電子辞書としてのWikipediaや種々なニュースサイトやSNSもあるし、いろいろな情報源があるわけです。そうやっていろいろな情報源があるので、一つの課題で調べているときにそれらも異なる情報源として比較して、そのうちこれをどう評価したらいいのか、正しくなさそうなものを落とすとか、2つの対立意見を出してみるとか、そうしたことを人間の一般ユーザーに教えてくれるという、図のようなシステムですね。それは現実には今は人間がやっているわけですが、プロンプトを改定したりしながら答えを出して、他のものと比較して、そういったことをやってくれるAIシステムを使ってしまえば人間はもっと楽になる。

生成型AIの答えを評価する能力を持てなくなった人がいたらどうなるのかという心配を最初に問題提起しましたが、実はこういうシステムがあり、これを人間が使う形になれば、多様な情報源の取捨選択とか比較を教えてくれて、変なChatGPTの活用ということに陥らずに



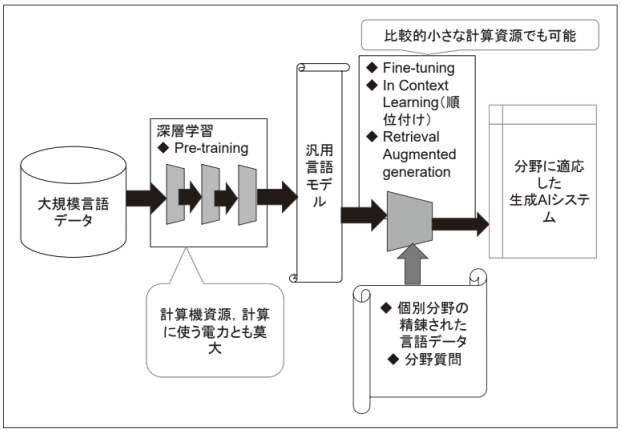
済むのではないか。そして、こういうものを自動的に作るために、非常に高度なAI技術が要求されるので、これはAI技術の進歩という方向性にも非常に合っているし、システム自身がまたビジネスになるかもしれない。技術の問題は、やはり技術の問題の中で、何か解決策をシステム的に示していかなければならない、今の状況をこんなふうに考えています。

生成AIの仕組みについて

今まで社会的な基礎を見てきましたが、ここで生成AIがどのような構造をしているかということについて、技術の基礎をお話ししてみたいと思います。大規模言語データ、これはインターネットから大量に集めてくる大規模言語データで、これを深層学習でトランスフォーマーという技術を作って学習するわけです。プレトレーニング、と言いますが、その学習結果として汎用的な言語モデルができます。汎用的な言語モデルができると、例えば、ある文章の途中まで来ると、次にどんな単語が出るかということを非常に正確に予測できる。これがものすごく正確に予測できるようになったので、新しい文章を作ったり翻訳をしたり要約したりと、いろいろなことができるようになりました[※2]。

私たちは自然言語処理の研究をやっていた時には、今までに出てきた文章と、それ以前から次の単語を予測するのは非常に難しい仕事で、精度、正しさがあまり高くならなかった。これが大規模データに対して深層学習、トランスフォーマーを使う技術によって、ものすごく正確にできるようになったんですね、これが大きい。ただ、それだけに、計算機資源がいるし、計算に使う電力も莫大です。1回プレトレーニングするのに電力料金が100億円レベルで必要というイメージです。しょっちゅうやるわけにはいかない。ということで、この汎用的な

2　岡野原大輔：大規模言語モデルは新たな知能か．岩波科学ライブラリー．2023



言語モデル、これをラージランゲージモデル、LLMと呼んだりしますが、それができたら、次はLLMを分野ごとに適応した生成AIシステムにしたいというニーズが出てくるわけです。あるいは間違った答えを出さないようにするという目的かもしれない。

そのためにはどんなものが必要かというと、これは個別分野ごとに人間が丁寧に見て精錬します。具体的には、答えと文章が非常にきれいにまとまった分野ごとの言語データを作って学習する。そのときにプレトレーニングを繰り返さないで、ファインチューニングという技術でプレトレーニングされたラージランゲージモデルをベースに、それを少し手直しする。あるいは質問をして、出てきた質問に対していい答えか悪い答えかを教えてあげることによって学習させる。あるいは、Retrieval Augumented Generation（RAG）という質問に関して、例えば他の検索エンジンで出てきた答えをプラスして質問を強化して、それで答えを出すとかいうシステムが提案されています。このようなシステムをたくさん組み合わせて、例えば世界的に見れば各分野で最新の知見に関する論文が1か月で100本、200本と出てくるものも含めて、チューンナップの方法をうまくやることで、どんどん良いものができ、特定分野に適合したベストな答えが出せる生成AIシステムが作れる、改善が進むということになります。ここは特定領域に限定した話ですから、割合小さい言語データしか使いませんので、それほど莫大な計算資源はかからない。

プレトレーニングの話で、次に出てくる言葉を予測するという話をしましたが、この話をちょっとだけ自然言語処理をベースにお話しします。昔の自然言語処理は、例えば、「昨日は昼御飯を食べ損なった」という言葉について、この「昨日は昼御飯を食べ」まで来たところで、その後に「損なう」という言葉がどのくらい出てくるんだろう、どのくらいの確率で出てくるんだろうということを知りたいわけです。それが確率ゼロとなれば、そこは文章として正しくないということに自然言語処理ではなるわけです。古い時代の自然言語処理では、前の単語だけじゃちょっと候補が広過ぎるよねということで、直前の3単語くらいを使って予測するのが限界だったんです。直前の3単語なら簡単そうに見えても、単語の数って何万もあります。動詞にしたって日本語だって1万以上あるわけだし、名詞だったら10万以上あるわけです。ですから、この組み合わせは莫大な数になり、それに対して全ての組み合わせを使って予測するなんて全然現実的じゃない。直前の3単語がぎりぎり、直前4単語になったらもう計算機が動かないということです。昔の自然言語処理はこういう世界だったんです。

ところが、トランスフォーマーは何が優れているかというと、この答えについて、ターゲットの単語より前に出てくる多数の単語にアテンションという名前で呼んでいる重みづけをしたリンクを張って、どのくらい影響力があるんだろうということを非常に広い範囲から見る計算をすることにしたんです。それを深層学習で何とか動くところまで持っていったというのがトランスフォーマーということなんですが、それによると、非常に広い範囲の情報を的確に使える。そして強力に計算もして予測できるということです。昔の手法に比べて、何しろ見ている情報の量が桁違いに大きいですから、非常に正確に予測できるということがあって、今のような爆発的な進化を見たということです。こういう基本的な提案がされたのは10年ほど前でして、10年間あーだこうだとやっているうちに、昨今このレベルのびっくりするものが出来上がってきたということです。

ちなみにお話したアテンションは一つの数値ではなく、様々なパラメーターの束なんです。実はベクトルみたいになっていて、複雑にすればするほどパラメーター数が多くなります。GPT-2で100億、GPT-3これはChat-GPTの最初のバージョンですが2000億、GPT-4で1兆以上と、パラメーター数は膨大です。一回当たりの学習は数千万ドルの電気料金ということになります。こういう学習によって文脈まで考慮して、高い精度で次の出現単語を予測ができるようになりました。当然この予測結果を用いて単語生成を繰り返せば、自動的に文生成もできてしまうという、こういう仕組みになっています。

組織内情報を利用した生成AIについて

ChatGPTは嘘をもっともらしく作文するということがよく言われています。ですから正しい答えを作るには

どうしたらいいんだろうという話になる訳です。ChatGPTはお話しした動き方なので、正しい文を作っていくという保証はないんです。ですから、正しい答えを作るような仕掛けを何かしなきゃいけないということでして、最初はOpenAIはデータの大きさを増やせば何とかなるんじゃないかという期待を抱いていたらしいですが、あまり上手くいかなかった。つまり、おかしな答えとか、常識的に受け入れられないものとか、反社会的な答えをなかなか排除できない。こうした回答を排除するためには、質問して答えにバツ・マルをつけて、その結果によって内容を少し変えていくという作業をする。たくさん質問をしては答えを見てマル・バツをつけるという作業は本当に人海戦術です。これはどこかの国でできる人をいっぱい雇って行うという人海戦術で、間違いのない社会的に受け入れられる答えを作ろうというのが実態で、今も続いていると思います。

このようにしてデータを沢山集めて、それをチェックするというプロセスまで作って、社会的に受け入れられるものを使って貰うわけですが、最大限にインターネット上のデータを集めようとしても、インターネットに載っていないデータはすごく多いんです。政府でも、会社でも、自治体でも、学校でも、組織の内部情報は機密情報であって、インターネットに載せないものが多いわけです。ですが、やはり組織内で生成AIを使おうと思ったら、そうした情報を生成AIの能力として使いたい。かといって、組織内情報を公開されている生成AIのシステムであるChatGPTで使ったら、情報が漏れてしまう心配があります。ChatGPTは質問で入力されたものと答えのペアを30日間ぐらい保存していると聞きます。保存しているだけなのかどうかはよく分からないし、それが学習には使われないと言ってはいますが、信用し切れないところもあります。

そこで、組織内情報を使って生成AIを活用したものを作るには、どうしたらいいんだろうという話になります。これは2つのやり方があると思います。まず、汎用大規模言語モデル、LLMの活用ですね。ChatGPTのラージランゲージモデルほど強力ではないですが、実際売っているLLMがあるんです、少々力は落ちるけれども。そういったものを組織内のサーバー、あるいはクラウド上のサーバーに載せて、組織内で、内部情報を使ったファインチューニングやIn Context Learningで質問して答えを教えてあげてというプロセスを行って、組織内で閉じたシステムを作ることができます。その場合、組織内の機密情報が外に漏れる心配は全くないということです。

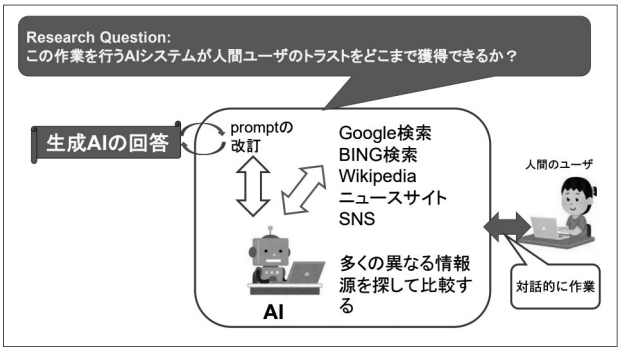
あるいは、公開されている生成AIのAPIモードというものがありまして、ChatGPT Plusでも提供しています。これは要するに、APIモードで使っているパソコンについては、生成AI提供企業は一切覗きません、手を入れませんというものです。ですので、それを信用できるとすれば、APIモードで使ってファインチューニングとかIn Context Learningによって組織内専用の生成AIをクラウド上に構築してしまうということです。

ただ、この生成AIは、大体英語の学習データで作られているので、日本語で質問するとボヤけた答えが出るんです。ですから、本当は日本語そのものの言語資源を用いてLLMを作りたい、あるいはAPIをサービスしたいということです。これをやろうと、いま国立情報学研究所とかメーカーが一生懸命やり始めているようなので、そうした日本語をベースにしたLLMが出てくれば、日本語でも質の良い回答を出せる生成AIが供給される可能性もあるということです。ちなみにこの関係では、実は、既に韓国では韓国語をベースにしたLLMが2年前に既に作られていると聞きました。既にその韓国語ベースのラージランゲージモデルをビジネスに応用することがどんどん立ち上がっていると韓国の研究者が話をしてくれまして、日本は2年韓国に遅れているというのが実態です。2年遅れでもいいから、この言語システムを用いて早く作って、供給してほしいです。

組織内生成AIと組織外AIを併用する

ただ、それで全て完結するのか、本当に組織内生成AIで作った答えが良い答えなのだろうか、ということは心配ですよね。企業内の自分たちのアイデアだけに凝り固まってしまう、自分の会社の常識は世の中の非常識ということがしばしばある訳ですから、企業内の常識に沿った答えしか出してくれないとなると、やはり公開された生成AIを併用したくなります。併用することが果たしてできるか。ここはなかなか難しいところなのですが、公開された生成AIと組織内生成AIを関係させながら併用していく方法が、次のビジネスモデルとして考えられるということになります。

どこからそういう技術を持って来るかというと、要するに、組織内生成AIを使って出てきた回答を、組織内情報ということを外部に気づかれずに公開生成AIを使えないかということです。組織内情報を漏らさないで質問をするプライバシー保護技術というのが実はあります。質問内容を相手に分からないようにしながら質問する。勿論完璧ではないですが、できるだけ自分の本音を



教えない状況で検索をかけるというプライバシー保護検索があります。守秘義務がかかった使い方は、生成AI以外の分野での使い方の問題でもあります。AIだけではなくてプライバシー保護技術、プライバシーテクノロジーシンポジウム（PETs）とかいろいろなシンポジウムがあるのですが、一歩離れたところの技術が実はこの生成AIにも大切になるということです。

ちなみに、企業などは、情報源としてはテキスト、画像、設計図、データベースなど、非常に多様なメディアデータを企業内に持っているので、そうしたものも使って生成AIを拡張する技術を確立していきたいということです。ただ、こういうシステムができたときに、実は人間がこのシステムをどれほど信用できますかという問題が多分あるんだろうと思います。結局はシステムを信用できるかどうかが一番キーになってくるだと思うんです。

法律的問題

生成AIの法律的問題について、私は法律家ではないので、福岡真之介先生と松下外先生共著の「生成AIの法的リスクと対策」という先月出版された本[※3]を参考にさせて頂きました。生成AIの技術的側面をよく理解されて法律上の問題を書いておられまして大変勉強になります。その中で著作権の問題がやはり大きいんですね。学習データの著作権はどうなっているのかとか、あるいはAIが示した回答に著作権があるのかという点です。仮にAI自身の著作物ではないとすると、では誰の著作物なのでしょうね。プロンプト自体には著作権は当然あるわけですからプロンプトを入力した人の著作物かもしれませんが、生成AIの複雑な処理の結果ですから、そう簡単に割り切れないかもしれません。また、プロンプトの中で著作権的に問題になる情報を使ってしまって良いのかどうかという問題もあります。このような、生成AIを使うときに考えなくてはならない法律的な問題についてかなり丁寧に答えてくれています。他にも間違った情報で誤情報を作ってしまった時、どういうふうに責任を持つんだろうねという点。誤情報だと分かって作った場合、全体責任になるのかとか、本を読んでみると、問題点は指摘しつつも判例などがまだある訳ではないので、考え方だけをいろいろと書いてくれていて、今後どんな方向で進むかということを勉強したいと思ったら非常にいい本だと思います。

正しい答えとはそもそも何か

ちょっと哲学的な問ですが、私たちは正しい答えが欲しいと思うわけです、と自分でも思っているのですが、実は人間というのは本当に正しい答え、科学的真実に基づく正しい答えが欲しいのではなく、自分に都合の良い答えが欲しいのではないかという気がする訳です。科学的エビデンスがあって、うそをつけない自然科学の世界では、実験をしてその答えについて嘘をついたら捏造になります。なので自然科学とか工学系では「正しい」ことが定義しやすいですが、思想とか哲学とか法律などでは「正しい」というのは主観的かもしれないという話です。法律は人間が作ったもので、自然現象ではないので、そうすると権威のある人が言うことが「正しい」ということになる。法律家の方とお話をしていると、結局、長老の先生が「正しい」みたいな言い方をする方がすごく多いんです。

私はある法律の学会で質問したら、長老の先生から部外者が質問しちゃいけないと怒られました。ということは、権威ある人が正しくて、若い人は発言権があるのかどうかという話になってくると、長期的に見て正しいものがなかなか受け入れられない。こうした中で、ChatGPTは正しい答えとして同じ答えを返してき続けます。すると、現状の固定化を促進してしまう危険性が大きいです。つまり、新しいエビデンスが工学的・科学的に出てくればいいのですが、そうでないと結局たくさんのデータを参照しながら「正しい」答えというのを出してくるわけです。となると、結局、現状固定を促進してしまう危険性が実は一番大きいのではないかと。

例えば、新入社員の選抜をする時、女性の候補と男性の候補では、女性が差別されたことがありました。何故かを調べると、今までに女性の社員で入社した人が殆ど

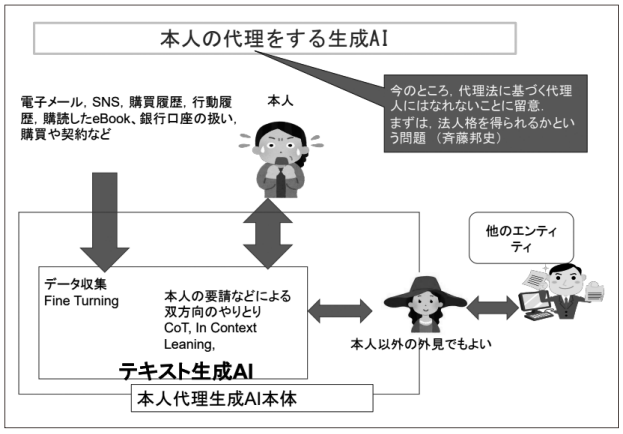
3　福岡真之介，松下外：生成AIの法的リスクと対策．日経BP．2023.

いなかったから、女性に関するデータが少ないので、低い点しか付かなかったという答えが出ていました。つまり、生成AIは明らかに今までにあったデータから答えを作り出してしまいますから、それが「正しい」と思って答えを作り出し、現状を正しいとする答えを出して現状固定化を促進してしまう。これは私もそう思いますし、そうした点に気が付いて早くから警鐘を鳴らしている先生は沢山いらっしゃいます※4。これが一番怖いかもしれないと思っています。

本人の代理をする生成AI、本人が亡くなったら？

本人の代理をするような生成AIを作れたとしましょう。この場合民法上の代理人にはなれないです。それはそうとして、法人格はひょっとすると得られるかもしれないということを慶応大学の齊藤邦史准教授が書かかれています※5。ですから、本人の代理をする生成AI法人はありえるかもしれない。それがどうやって使われるかというと、本人がいて、本人の電子メールとかSNSなどを使いながらデータを収集し、それらを使ってFine-tuningして、本人とのやりとりも頻繁に行って、本人に特化した生成AIが作られるということです。もうこの生成AIと話していると、何か本人そのものと話をしている感じです。ちなみにアバターを使うと、本人以外の外見でも良いものが作れますが、こうなると他者と話していても、他者もひょっとすると生成AIかもしれない。こんなものが作られる時代になるかもしれません。

ところが、人間ですから死んだ後にどうなるのだろうという問題があります。生成AIはプログラムです。プログラムが本人が亡くなったことに気が付けるのか。そうしたことを研究している人がいますし、個人IDにしてもSNSのアカウントにしても、それらが生きているかの如く動く場合に、それらがどうなるのかは結構問題になる気がします。それらが生きているように見えると、個人を偲べるから良いのではという声もありますが、悪用の宝庫になるのではないかという懸念があって、例えば著名な個人のアバターがあると、著名な個人の存命中の発言を学習した生成AIがあり、それがまだ生きているかの如く話をする（ゾンビAI）。これは



ちょっといじくることで、世論を誘導するといった危険性がある。何かSFみたいなことが技術的には可能ということですね。また、そんな大きな話じゃなくても、AIとアバターのようなものが一致すると、なりすましや乗っ取りが行われる確率も非常に高くなる※6。

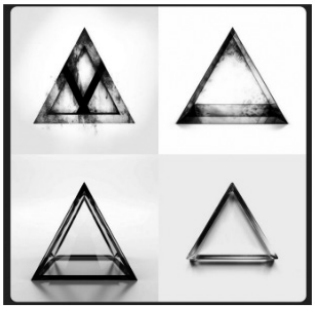
本人が亡くなった後、同時に消滅するのならば簡単だし、あるいは本人が亡くなった後、期間限定で存続し、亡くなったことが分かれば対処する。それはAI、ChatGPTの本人バージョンを運営している会社との契約で決まるかもしれない。あるいは遺族から相続するとか、あるいはメモリアルとして永続的に存続することも良いかもしれない。技術的には様々なことができるということです※7が、最大の問題は、本人が亡くなったということを認識できるかどうかという点で、例えば本人に近い人を対応者として事前指定しておくサービスも既に出たりしています※8。ただ、亡くなった後に永続性があるかという点は、だんだんと会いに来る人も減ってしまって、実はあまり永続性は無いんじゃないのかという話もあります。20年も経ったら永続性は無いでしょうという話はあります。

画像生成AIの特徴

画像生成AIの話に移ります。今はStable diffusionとか、Midjourney、BING image、DALL-Eなど様々なものがあります。私自身はMidjourneyを使っていますが、これは使い始めて2、3日後に作った絵です。プロンプトは簡単で、「Armed lady with a sord fights at

4　この考え方を私が初めて伺ったのは弁護士の板倉陽一郎先生でした.
5　斉藤邦史：人工知能に対する法人格の付与．情報通信学会誌．Vol.35 No.3, pp.19-27, 2017
6　中川裕志：AIエージェント、サイバネティック・アバターにおけるトラストの構築と問題点.『情報通信政策研究』第6巻第1号 P. IA-45～IA-60. 2022
7　Nakagawa, H., Orita, A. Using deceased people's personal data. AI & Soc (2022). https://doi.org/10.1007/s00146-022-01549-1. https://link.springer.com/article/10.1007/s00146-022-01549-1#citeas
8　FacebookやGoogleはこの問題に対応するための手段を利用者に提供しています.

格好良すぎ



the castle gate in the Greek era.」と唱えれば、こんな絵が簡単にできてしまう。本当に画像生成AIの能力は凄いです。2、3年前から凄いものがどんどんできている。

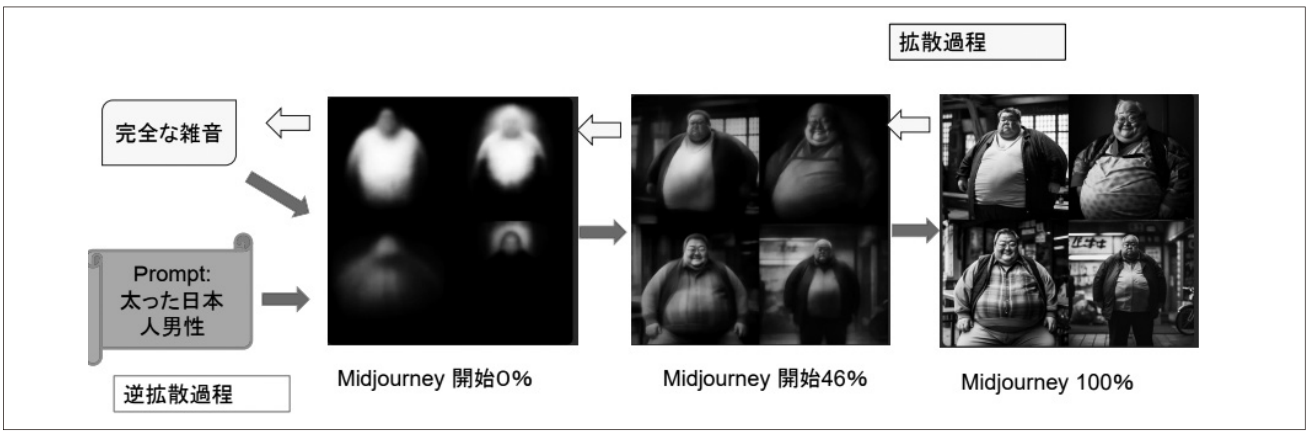
意外と作れない絵が、普通のおじさんです。普通の日本のおじさんと言ったら左側のような絵ができてしまい、かっこよ過ぎます。何か俳優さんみたいだし、ビシッとしているし。もっと庶民的な普通のおじさんを作ろうと思ったらえらい苦労しました。「服のしわがよってヨレヨレ」だとか、「視線が定まらない」とか、「髪がボサボサ」とか「レトロな酒場にいる」とかいろいろなことを指示に入れて、やっと何とか右側の絵のような普通のおじさんにたどり着いたんです。意外とこういうものが作りにくい。

もっと難しいのは、普通の正三角形です。これは本当に作れないんですね。いろいろなこと書いても結局作れませんでした。英語でもやってみましたが、作れませんでした。何が作れているかというと、こんなものが出来てきちゃうんです。確かに正三角形だけれども、ものすごく芸術的ですよね。

なぜかというと、実はこの生成AIがどのように動いているかというプロセスを見ると、わかります。Midjourneyを使っていますが、元の学習データとして、「太ったおじさん」としての一番右側の4枚の絵が使われる。最初こういう絵から学習をするときに、「拡散過程」というんですが、きれいな元画像から雑音を追加していくんです。ちなみに元画像には、それを特徴づけるテキストが付記されています。元画像に雑音を追加してぼやけてくる過程で、様々な画像を生成するためのパラメーターを学習していきます。さらに雑音を埋め込んで絵をぼやかしていくんです。どんどん「拡散過程」を進めるほど、パラメーターが作られ、最後は完全な雑音になってしまう。絵を生成するときは逆のプロセスで、「太った日本人男性」というプロンプトから始めるとこのプロンプトに近いテキストが付記された画像を複数枚選びます。さきほどパラメーターを学習しているので、雑音からスタートして、パラメーターを逆方向に使いながらだんだんと絵を作って元の絵に戻れる。つまりパラメーターとして元画像を覚えているから、元画像があるから戻れる。元画像がなければ戻れないということです。そういう拡散過程、逆拡散過程というものを最近の画像生成AIは使っているということです。もちろん、複数の画像をうまくつなぎ合わせて新しい画像を作っているわけです。

画像生成と著作権との関係

技術内容は、非公開のものが多くてよく分からないで



す。例えば複数の画像の繋ぎ目、例でだした「普通のおじさん」の絵を見ても、おじさんの絵、昭和の酒場の絵など、様々なものが繋がっている訳ですが、繋ぎ目をどうするのか、例えば影をどう付けるか、裸電球を下げて付けるとか、非常に高いレベルの画像処理技術が使われ、それも毎日のように進歩していて、本物らしいものが作れるということです。プロンプトでの検索技術、プロンプトの言っている意味に近いものをどうやって上手く探すかというところも重要な技術なのですが、そこから複数画像の繋がり、拡散過程、パラメーターとしての埋め込みなど非常に複雑なことをやっているので、ここまで複雑なことをやって著作権で問題になるような「依拠性」が追求できるかどうかはよく分からない。結局のところ、依拠性というのは元の画像、学習に使った元画像と同じものが有るか無いかくらいでしか言えないだろうという話があります。著作権には「類似性」と「依拠性」という要素がありますが、類似しているかは、例えば著作者なりアーティストなら、「自分と類似しているよ」ということはある程度言えるけれど、それが果たして依拠しているかの判断はなかなか難しいでしょう。全く同じ絵が入っていれば言えるでしょうけど、偶然作っているうちに自分の作品と似ているというくらいだと、著作権の類似性と依拠性は概念ははっきりしていても、著作権を主張するためにそれらを証明しなければいけないとなると非常に苦しい。

著作権の主張は、アーティストさんそのものではなくて、そこでお金を儲ける出版社さんが主張されるという話を耳にします。逆拡散過程は、要するに、前の画像に対していろいろ手を入れたり、アイデアをちょっと変えたり、要するに、前の画像をベースに新しいものを創っている。アートというもの自体が、そもそも全部を最初から作ることはない訳です。いろいろなアイデアが頭の中にあり、それをうまく使ったり組み合わせたりしながら創るのがアートだと思うんですね。ということは、生成AIの新たな画像の創り方というのは実は複数の画像を分解したり、組み合せたりしながら、それをうまく繋いで絵を創るという、アーティストさんの作業と部分的には似ているんじゃないかと思われます。アーティストさんとして、そうした点について争うよりは、それを自分の仲間として利用する方が良いのではないか、という考え方になってきているかもしれないですね。

囲碁とか将棋の分野でも、今AIのほうが強いんです。ですから、AIの囲碁や将棋にプロの棋士が戦うという構造ではなく、AIの打った手を、自分はこんな手を今まで思いつかなかった、これは、なんで正しいんだろうと考える、そんな感じの勉強をしているそうです。そうしたことと同じで、アーティストさんとAIとの付き合い方は、アーティストさんがツールとして使う。あるいはひょっとすると、お友だちとして使う。好敵手として使うとか、いろいろな使い方をすると思うので、そういう方向に行くかもしれないですね。

ちなみに、著作権法第三十条の四では、著作権のあるデータをAIの学習で使えるとあって、何故こうしたかといえば、著作権では「享受性」すなわち、あるアートを見て、楽しんだりする享受性が重要ですから、この著作権で重視される享受性は、計算機での学習過程、生成過程では存在しないということで、著作権のあるデータを使っても良いでしょうということになっている。そういうことも考えると、一律に禁止に走るより、クリエーターでも出版社でも、それなりに経済的配分を機械的に行える制度とすることも一つの方向性としてはあり得るかなと思います。例えばJASRACは音楽分野でそういうことをやっている制度です。そんな話が今後出てくるのではないかという話もあります。

EUにおけるAIトラストの系譜

AIに対する政策の動向として、EUの話を主にします。AIとトラストを最初に公的に扱ったのはEUです。トラストってそもそも何かというと、実態としては、大きな会社で扱っているソフトだから、みんな使っているからトラストできるよねといったものです。あるいは、自分に似たケースで、自分と同じ結果が出ているようだからトラストできるよねといったもので、トラストとはかなりいい加減なものではあるのです。そういうものをどう扱うかということで、EUにおいてAIトラストをどう扱うかということで、3つのルールをご紹介します。

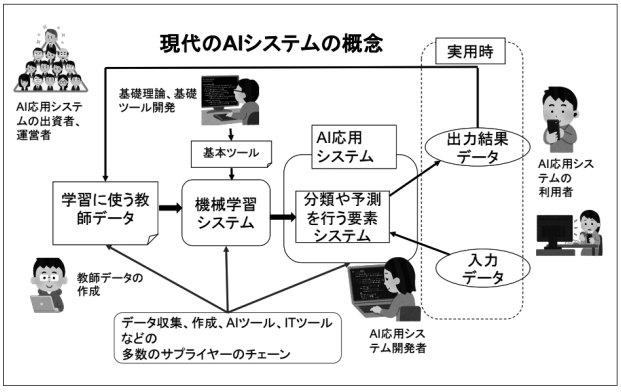
欧州委員会（EC）のハイレベルエキスパートグループが「信頼できるAIの倫理ガイドライン」（Ethics Guidelines for Trustworthy AI）[※9]を2018年に公開しました。その2年後にはAI白書（AI white paper）[※10]を出し、さらにその1年後にAI法（AI Act）[※11]を提案しま

9 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
10 https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
11 https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence　日本語概要は次のURL：https://www.soumu.go.jp/main_content/000826707.pdf

した。このAIの倫理ガイドラインは非常にナイーブです。外部環境があって、そのデータをセンサーで認識して、推論や行動決定して何かのアクション、ディスプレイに表示するとか、あるいはロボットを動かすとかするものをAIと定義しています。そういう要素のAIシステムがトラストできますかということで、AI製品は一つの会社の中で閉じては作れず、そこにサプライチェーンがありますので、サプライチェーンを経由する時にトラストできるかという問題になります。

このサプライチェーンの話をするために作った図ですが、現代のAIシステムの概要は、AI応用システムの出資者、運営者と作りますので、これを無視するわけにいかない。それから、学習に使う教師データを作成する人がいて、教師データを使うシステムがある。この教師データを作ること自体にも、実はデータ収集作成にAIツール、ITツールなど多数のサプライチェーンがあり、それから教師データに使えるほどきれいなデータにするという精錬作業もあって、多数のサプライチェーンが絡むんです。そこにトランスフォーマー、プレトレーニングとか機械学習システムを作って、AIで分類や予測を行えるシステムを作り、これを応用システムとして様々なユーザーに使ってもらえるようにする。そしてこの応用システムを使っている人から結果が出ると、実用時のサイクルとして、そうした結果に対してそれをまた学習データとして利用すると、こんなイメージですよね。

では信頼できるAI（Trustworthy AI）では何を言っているかというと、2018年頃はこんなことを言っていました。まずトレーニングデータに恣意的なバイアスが入り込んでいないか、それから基本権として尊厳、自由、平等とか連帯、市民の権利と公正性、個人の権利と自由を保障すること。そして法令遵守というわけです。次に人間中心という考えが出ていて、これは常に人間が



上位の決定権者にいなければいけない。そうすると、信頼できるAIは何かというと、倫理性があって、人間中心になって、そして技術的なトラスト、つまり壊れたりしない。いつも同じ結果をちゃんと返してくれる工学的なツールとして信頼できる。この3つが合わさったものが信頼できるAIということで、条件として列挙すると、(1) アカウンタビリティ、(2) データガバナンスがあって (3) バイアスがなく、(4) 誰でも使えるとか、実際のAIのガバナンスが (5) 人間の監視に置き換えできるとか、(6) 差別がないとか、(7) 人間の自律性や (8) プライバシーが尊重、(9) ロバスト（堅牢）である、(10) 安全であるとか、そういう条件が挙げられます。

要するに、基本的にEUでは、技術的なAIを信じていないということが、今の10個に及ぶ条件を見れば分かります。人間が制御できるように徹底したAIにしたいというのがEUの考え方です。このEUの考え方を実現するためには、テクニカルな手段としては、アーキテクチャをしっかりするとか、プライバシー、セキュリティーなどへの配慮をデザインレベルでちゃんとやりなさいとか、実用に供する前にテストと検証をしっかりして、実用に供した後はトレーサビリティーと監査の可能性を考慮に入れて、そういったことにも関連して、説明可能なAIにしておかなきゃ駄目ですよという訳です。一方、ノンテクニカルな手段としては、法的規制とか標準化、アカウンタビリティを確保するガバナンス、行動規範、人権教育などの必要性もカバーする必要があります。これらについては、日本でも全然やらなかった訳ではありません。日本でも類似の文章として、QA4AI[※12]というものがありまして、これは300頁くらいありますがレベルの高い文章で、是非一度見てみてください。

AI白書2020について

2020年にAI白書が出ました。その内容を次にみて行きたいと思います。まずAIサービスは事前に徹底的なリスク予測をやりなさいとされています。ただ、AI技術の複雑性、発展の早さから見て、完全に予測し切るとか、リスクの発生を抑え切ることはちょっと難しいということもある程度意識はしている様です。それから、サプライチェーンの各段階で倫理指針とか法制度に基づくリスク管理はきちんとしなさいということです。このサプライチェーンをチェックするというのは、EUだけで

12 https://www.qa4ai.jp/

なく米国でも、中国産の基本ツールを念頭においてチェックするというのはやりますから、これは普通の話ではあります。ただ、AIの場合、サプライチェーンが国を跨いであちこちに行くので非常に大変ですね。

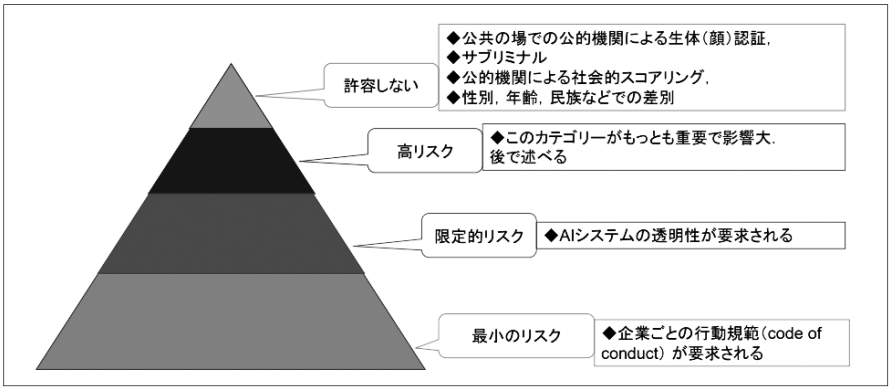
それから、基本権、人間の尊厳、多様性の確保など、これは普通の話なのですが、さらにはAIは既存のEU法及びEU加盟の国内法が適用されると言っています。この辺からだんだん厳しくなり、更にEUはAIアプリケーションに関連する潜在的なリスクに基づく証拠を明確化するために、既存の法的ないし技術的手段を最大限に活用するとなって、AIは敵みたいな言い方をしています。そうはいってもAIの進歩は認めざるを得ないので、効果的な適用と施行を確実にするために、例えば責任に関する特定の分野の既存の法律を調整するとか、要するに、若干法律を変えてもいいよといったことを言っています。ちょっと妥協しても仕方がないとあるのですが、サプライチェーンの各段階をチェックするという監督意識が強いですね。

さらにすごいのは、AIシステムがすべてのライフサイクルフェーズでエラーや不整合に適切に対処できるように保証するとありまして、こちらはその後のAI法案でも書いてあるのですが、この白書でも書かれている。また、AIシステムの出力は、人間によって事前にレビュー及び検証されていない限り有効にならない、つまり有効にならないうちは使ってはいけないということを言っています。人間がチェックしなさいということです。それから、動作中のAIシステムの監視もリアルタイムで介入して非アクティブ化する機能、これも人間がやりなさい。「えっ」という感じですよね。つまり、AIの方が人間より遥かに速いツールなのに人間が介入すると、何か現実感が薄くなってくるわけです。そのために、設計段階でAIシステムの運用上の制約を課すということなので、これではAIの能力を低めろと言っているようにも見えて、これでいいんですかという感じがします。あまりこういう点に拘り過ぎると、能力が低まったAIしか作れなくなって困ったことになってしまう。こうした内容について、アメリカなどはそう思っているでしょうね。ただ、これらの事項はAI法案に引き継がれています。

AI法案2021で示されたリスク分類

さて、AI法案（AI Act）です。2021年4月21日ブリュッセル発ですが、2018年の信頼できるAIの倫理ガイドラインから始まって、いよいよ法律ということです。前文に書かれた目的ですが、EU域外で進むAI技術に対する不信感があるんですよね。国内規則は作っているけれど、EUは27か国なので、規則が国ごとにバラバラでは市場も細分化されるので、統一したルールにしておかないと大変ですよね。そのため、個別の国にあまり振り回されることなく統一したルールでEUを管理したいということで、EU全体で一貫したハイレベルの規則を作りますという事情です。

内容はご存じの方も多いと思いますが、4段階に分かれていて、「許容しない」、「高リスク」、「限定的リスク」、「最小のリスク」とあります。「最小のリスク」は、企業ごとで行動規範が要求されるというものです。ただ、この規範は法律に近いような強制力を持っていますよとEUの人が言ったりしており、意外と厳しいことを言ったのかもしれません。「限定的リスク」はAIシステムの透明性が要求される、説明可能なものでなければならないというものです。これも企業にとっても結構シビアなことを言っている。それから「高リスク」は後で話をします。「許容しない」とは何かと言うと、リモート生体認証、顔認証ですね。要するに、公的な場でのリアルタイム認証はダメ。それから、心理的な被害を引き起こす可能性のあるサブリミナル効果を使ったもの。また、公的機関が人間をスコアリングして分類するようなこともやってはいけない。それから、年齢、身体的または精神的障害などによって特定のグループの人々の脆弱性につけ込むAI、差別は禁止ということです。



そこでとりわけ問題になるのは、先ほど飛ばした「高リスク」に関わるAIです。これは付属文書3に丁寧に書いてあるのですが、かいつまんで説明しますと、生体認証とそれによる分類で、これは公的

な場に限るわけではなく、リアルタイムでも事後でも差別があってはなりません。差別がないように作らなきゃいけないということを言っています。

2番目に私が一番気になっているんですが、重要な生活インフラストラクチャの管理と運用。その中に道路の管理と運用とあります。あるいは交通、水、ガス、暖房、電気の供給のためのAI。この道路の管理と運用は何かというと、ぴんとくる人がいると思うんですが、要するに自動運転のことをターゲットにしていることが明白に見えてくる。自動運転というのは、自動車の側だけAIで賢くなっても上手くいかなくて、道路側から様々な情報を提供してあげる必要があるわけです。自動運転向けの信号機、情報提供装置みたいなものがある。そういうものはやはりハイリスクだということで、確かに交通事故は起こるかもしれないから、高リスクAIというのは頷けますが、これは相当な規制がかかるので、業者さんも、自動車業界さんも相当神経質になるし、道路の管理は自動車業界は手を出しませんから、国・地方自治体がやらなきゃいけないところで、そこに相当な経済的負担があるということになってしまいます。

それから3番目に、教育や職業訓練へのアクセスです。これは要するに、入試とかそういったものにAIを使う際、それなりに気をつけましょう。また、採用、人事評価、労働者管理、雇用と解雇という分野ですね。こうした人事周りのことをやる際に使うAIは非常に高リスクだから、きちんとやってほしい。次に不可欠な民間サービス及び公共サービスへのアクセス適格性評価・順位付けというのですが、ちょっと分かりにくいですが、一番ピンと来るのは、医療援助などの緊急性のある処理に優先順位をつけるという話、コロナのときに入院したい人に対して、入院できるベッド数が少なかったので選ぶ、トリアージですよね。トリアージにAIを使うというのは高リスクですよということです。他に法執行機関が個人の状況に立ち入るという、再犯とか刑事犯罪の潜在的な人物のリスクを評価するAIとか、法執行機関がディープフェイクを検出するためのAIとかあります。また、移住や国境管理に使うAI。司法当局が事実の調査及び解釈するのに使用するAI、こんなものが高リスクと言われている。

高リスクAIを実利用するためにやるべきこと

高リスクAIは、EUの市場に売り出す際に行うべきことが沢山ありまして、業者さんごと、つまり製造業者、配布業者、サービス事業開発者、事業者代表、公的ユーザーごとに細かい指示が出されています。この公的ユーザーについては、自治体とか政府とかAIを使ってサービスを提供する側ということで、個人的に利用するだけのユーザーは含まれません。そういったユーザーに対し、非常に沢山の指示があって、当然EU域外の業者にも適用されます。

どんなことを守らなくてはいけないかと言うと、AIシステムの全利用期間においてリスク管理してくださいということです。データガバナンスで、バイアスエラーがあってはいけない。十分に代表的なデータを使って学習支援とか技術文書を作りなさい。それから利用状況レコードを保存する機能が求められ、これは業者さんにとって厳しいんじゃないかなと思いますが、利用状況は全部記録しておいて、後々トレースできるようにしなさい。これはコストがとてもかかるので、どうなんでしょうか。

それから、ユーザーへの内容説明と適切な指示、人間が監視する規則の整備。人間が監視する出力の限界値、精度、緊急停止ボタンの設置や監視は2名以上で確認するなど、「えっ?」みたいな感じですよね。それから継続的に学習するAIの精度、技術的堅牢さ、サイバーセキュリティーの確保、敵対的標本データの不使用とかそんなことが書いてある。また、CEマーキングというのがあって、要はこうしたEUの適合基準にあったCEマークを付けていないとEUでは売ってはいけませんと。それから、実際そうやって売られた後は、まずいことが起きたときに即時に必要な訂正や行動をとると。また、市販後の動作のモニタリングをする、稼働状況レコードの保持、これも大変ですね。そしてインシデントの報告義務。さらには制裁金まであって、非常にお高い額ということです。

法律に依拠するEUのメンタリティー、欧州評議会の提起するAI条約案

AIに関するEUの命令系統としては、欧州委員会がトップにあって、欧州議会及び理事会。それから新たにAIに関するボードが設置されます。それに、EU各国への監督権限を持つ組織という構成です。また、2021年に続いて2023年9月の改正AI Actでは、コンテンツが生成AIによって生成されたことを開示すべきこと、違法なコンテンツを生成しないようにモデルを設計すること、学習に使用した著作権データの要約を公開すべきことなど、いくつかの透明性要件が生成AIについて追加

され、改正案が欧州委員会を通っています。

EUの人々のメンタリティーなんですが、これはいろいろなEUの人と話していて分かったのですが、一言で言うと、法律がない状況での技術利用なんて考えられないということです。法律家だけでなく、学者、技術者、みんながそういうふうに言うんです、例外なく。アメリカの場合、技術は社会運用してみて、問題があれば裁判を起こして、それで問題がさらに防がれなければ法制化するという国です。一方、EUの法律がない状況で技術にしてはいけないという、この強いメンタリティーは私自身は付いていけない感じがしてしまいました。

最後プラスアルファの話で、欧州評議会です。こちらはEUとは少し違うんです。仕組みが出来たのは1949年と古いんですが、加盟国は46か国と多く、日本もオブザーバー参加しています。そこではAI条約案[※13]というものを今年から開始して少しずつ作られつつあります。内容はリスクベースアプローチを用いますとか、基本的運用面でいろいろと書いてある。読んで頂くと普通のことが書いてあるだけだなと思うのですが、これは条約です。ですから、日本が締約国になってしまうと、条約を守らなきゃいけない訳で、その意味でこの条約の成り行きについては心配をしております。こちらはブリュッセルではなくストラスブールです。

そちらで条約対応をしてくれる方を学術的な意味でもサポートするという作業を、東京大学の江間先生が中心になり、私も入っていますがしております。AI Actはブリュッセルの法律だから世界に対しては「ブリュッセル効果」ということでの関係ですが、AI条約は日本により直接的に関係するので注目しておかなければいけないですね。

長くなりましたが、以上です。どうもありがとうございました。

質疑応答

○**研修員**　今日は大変興味深いお話ありがとうございました。AIのリバースエンジニアリングみたいなものが、どのくらい可能なのかということがもしあれば教えてください。つまり、おっしゃられたように、機密情報を入れてしまうと良くないのはそのとおりなのですが、逆に、こういう情報を食べさせてAIをオープンにした時に、人によっては、それをいろいろと考えていくと、もともとこういう情報を扱ったのかな、こういう考え方でこの組織はやっているのかなというのが分かってしまうとなると、やはり問題になってしまいます。逆に、社内でAIを使っているかどうかについて、AIを仮に規制しようという立場からは、AIを使っているかどうかをどうやって判断するか、テクニックなりあるいは最近では電力を使っているかどうかなど、良く分かりませんがそのあたりは技術的にはどうなのでしょうか。

○**中川講師**　それは、現実的問題になっていまして、要するに、いかにChatGPTに答えにくいことをうまく答えさせるかという競争のようなことをやっています。ですから、ChatGPTはチェーン・オブ・ソートで、答えをもらったら、それをまた質問に対して繰り返してということをぐるぐる回り続けられるので、そのうち反社会的なことを聞き出したり。要するに、自殺の仕方を聞き出そうとか直接は駄目なんですが、だんだん技術的にやっているうちに、恐らく、もし今おっしゃったようなケースがあると、機密情報までたどり着いてしまう恐れがあります。今のシステムであり得ます。それは非常に心配なので、やはりできるだけ閉じた状況で使う。お話ししたように、組織内で閉じた状況で使えるようなものが可能性としてあると思うので、閉じた状況で使える生成AIにして頂く努力が大事かなと思ってお話しさせていただいたところです。

○**研修員**　フェイク動画が出回っていますが、それが真であるか偽なのかの判定はどのような形で行われるのでしょうか。

○**中川講師**　それは研究テーマです。2、3日前にもセミナーでお話をしている東工大の先生がおられましたが、フェイクかどうかを認識するシステムって一生懸命作っているんです。何か不自然なところがあるかとか。ただ、これはいたちごっこになってしまっていて、相手もどんどんフェイクだと見破られにくいものを作って来ることがあるので、やはりある程度できても、そのさらに上を行くということがあって。セキュリティーの世界って常にそうなんですけれども、こちらが何かやると、それに対抗して敵も強くなるということを繰り返すということです。ですから、本当にもともとの絵がはっきり分かっていればフェイクだと言えるんですが、そうではない状況で、岸田総理大臣なり何なりを生み出すことは出来そうなんですよね。どうやって差を見つけるかは非常に難しい問題で、AI研究者さんが今一番頭を悩ませている問題で、ある程度できるにしても、抜本的に

13 https://ifi.u-tokyo.ac.jp/news/16864/

は、恐らくこの政治家なり、あるいはこの人はこんな話をするだろうとか、絶対しないに違いないとかある程度常識的な感覚を私たちが持って話を聞くというところに行く気がします。そういうこと自身も何かAIとしてサポートしてくれるようなことは、最初のほうで説明したAIの在り方の一つかなと思います。

○**研修員**　日本企業がEUでビジネスをするために、EUのAI法に従う必要が出てくると思います。このEUのAI法が世界基準になる可能性はどのくらいあるでしょうか。

○**中川講師**　非常に高いのではないかと思います。ただ、問題は高リスクAIで、ここをどうするかというのは、今の案が正しいかなと思う部分もある一方、やり過ぎかなと思う部分もあり、この法律ができるのが来年くらいになるんですね。そして、さらにそれを実行するためのガイドラインを作らなきゃいけないので、大体2年後から3年後にスタートするだろうと思われます。それを見て、それに対していろいろな企業がEUで仕事をしたければチューンナップするということでしょうから、恐らくブリュッセル効果が発揮されることになるだろうと思います。そう思っていたほうがいいし、企業の人はもうそういうつもりで仕事をしていると思います。実際生成AIを作る立場の人は、非常によく勉強していらっしゃいます。

○**研修員**　ゾンビAIエージェントの話がありましたが、AIエージェントを消費者の判断の助けにするようなポジティブな使い方も可能ではないかと思われます。例えば生鮮食品を1日に何度も買ったときにアラートを出してくれるようなものとか考えられますが、そのような研究はあるのでしょうか。

○**中川講師**　もちろんそういうためにAIエージェントを使うのが基本です。要するに、買い物するときにアラートを出してくれたり、いや、もっと良い買い方があるよと教えてくれたりとか、自分の召使といったら何ですが、そういう形でAIを使うということがあるわけです。そういう良い使い方がまずはあるという前提の下で、もしそうしたAIを皆が使うようになったら、本人がお亡くなりになった時にAIがどういうステータスになるんだろうという点を、私はちょっと先読みしています。実は人が亡くなった時にどうなるかの研究をずっとやっていたものですから、ついそういうのが気になっていて、今日その気になったことの話だけしてしまいました。もちろん非常に役立つものになることは確実です。ただ、最後に後始末を付けるということはお忘れなくという話です。

ちなみにこの話は、AIに限らずfacebookですと、アカウントは本人が亡くなった時にどうなるんだろうということは問題になって、facebookは本人が亡くなったアカウントでも、たくさんの人が訪ねてくればそれはお金になるからそのまま使いたいねということで、死後のアカウントをどう使うかを様々研究し、サービスも開始しているということもあります。ですので、AIにおいても似たようなことは起きるかもしれませんというところです。

○**研修員**　我が国企業が基盤モデルを商用化する動きがあります。これらの企業はGAFAの対抗馬になり得ますか。

○**中川講師**　生成AIを作ること自身は、それなりにできるはずです。というのは、今の生成AIはプレトレーニングして、それで出てきたラージランゲージモデルは日本語ベースじゃないので、それで随分おかしな答えが出てきちゃっているという例が散見されます。ですから、日本語の言語資源を使って、日本人の常識に沿ったような答えを出してくれるというようなものは作れますし、それを狙っていると思うんです。もしGAFAが英語のベースで全部進めていくとすると、実は日本版モデルの方が性能が良いということで、日本語版ではGAFAに対抗できる精度のものは作成可能だとは思います。ただ、GAFAもお金持ちなので、日本語ベースのものを作って商売するかもしれないし、そこは戦いになるかもしれません。それでも、日本のメーカーさんの方が、日本語自身のことはよく知っているので、戦える可能性はゼロではないということでしょうか。

やはり日本市場をGAFAが重視するかどうかということですね。重視しなければ、じゃあ勝手にやってよということで、日本市場は日本の生成AIのラージランゲージモデルで行くということになっていくかもしれない。これは、GAFAがどのくらい日本の市場というものを重視するかどうかという点が非常に大きいと思います。