第１３講 マルチモーダル生成ＡＩ共同によるＡＩ倫理処理

【学習到達目標】

・マルチモーダル生成ＡＩについて説明できる。
・３種のマルチモーダル生成ＡＩ共同による倫理問題の解決法を説明できる。
・モーダル論理の定義を説明できる。

１．マルチモーダル生成ＡＩの登場

２０２３年9月２５日、新たな機能追加で、ＣｈａｔＧＰＴ（チャット GPT）がついに目と声を手に入れました。具体的には、ＣｈａｔＧＰＴに画像解析機能と音声出力機能が追加され、マルチモーダル生成ＡＩが登場しました。

その後、各社のマルチモーダル生成ＡＩが開発され、マルチモーダル生成ＡＩは驚異的な発展を行っています。

マルチモーダル生成ＡＩとは、異なる種類のデータを組み合わせたり、関連付けたりして処理する人工知能（ＡＩ）システムで、生成ＡＩの一種です。マルチモーダル生成ＡＩは、テキスト、音声、画像、動画、センサ情報など、複数の異なるデータの種類（脳が知覚できる様相、Ｍｏｄａｌｉｔｙ）から情報を収集し、統合して処理することで、より豊かな情報を処理し、深い理解や洞察を提供することができると言われています。

２．マルチモーダル生成 AI の評価

・マルチモーダル生成ＡＩの性能評価

Ｃｌａｕｄ３は日本語の文法が非常に優れており、日本語が得意だと言われています。

それでは、このＡＩ技術の性能について、最新の状況を確認しましょう。(チャット GPT)正午に確認したところ、テキスト生成の分野では、チャット GPT が１３１４というスコアでトップに立っているようです。このスコアは、例えばサッカーのＪリーグで対戦結果から世界ランキングを決めるのと同じ方式で計算されています。

具体的には、１５０の質問を投げかけ、これまで１６０万人以上の人々から得た回答を基にしたランク付けです。ただし、このランキングは頻繁に変動します。２０２４年8月２０日午後1時に確認したところ、特に大きな変動はありませんでした。

一方、画像生成の分野では、Ｇｅｍｉｎｉのバージョン１．５がわずかに性能が優れてい

ることが確認されています。

著者がこの大量のデータをすべて読み解くのは大変なので、トップ５のツールだけを取り上げて見てみましょう。各ツールの性能はおおむね同じような傾向を示しています。

ＯｐｅｎＡＩのＣＥＯのアルトマン氏が２０２４年８月８日に「チャットGPTがトップだ」とＳＮＳで発言しましたが、スコアに１０ポイント程度の違いがあったとしても、実質的な差はわずかであると考えています。

また、チャットGPTの最新バージョンについては、同年８月８日の段階では画像生成の機能は表示されていませんでした。おそらく、チャットGPTではなくＤＡＬＬ－Ｅで生成されていたため、チャットGPTの評価に含まれなかったのだと推測されます。

３．モーダル論理

・３種類のマルチモーダル生成ＡＩ

ＡＩの倫理問題を処理するために、ＣｈａｔＧＰＴ、Ｇｅｍｉｎｉ、Ｃｌａｕｄｅ３という３種類のマルチモーダル生成ＡＩとの対話によるアプローチがあります。

これは、ＯｐｅｎＡＩ社やＧｏｏｇｌｅ社、Ｍｉｃｒｏｓｏｆｔ社等のマルチモーダル生成ＡＩ開発メーカが行うのではなく、ニュートラルな立場でＡＩ技術を駆使する方法です。

最近のマルチモーダル生成ＡＩは、画像、音声、テキストなどを入力して回答を作成します。これらの生成ＡＩの３者が一致した部分は「正確な情報に違いない（Ｍｕｓｔ　Ｂｅ）」と推測できます。これを共通回答Ｐとします。生成ＡＩの２者が一致する部分は「かもしれない（Ｍａｙ　Ｂｅ）」ということです。

生成ＡＩの回答が論理的に正しいかどうかを、モーダル論理で解決する。モーダル論理の定義には、□X（必然）と◇X（可能性）がある。「Ｍｕｓｔ　Ｂｅ」と「Ｍａｙ　Ｂｅ」、つまり「Ｘに違いない」と「Ｘかもしれない」というものです。

・モーダル論理の定義

生成ＡＩの回答が倫理的に正しいかどうか？様相論理で取り出す方法は、以下のモーダル論理（様相論理）で定義できます。

□Ｐ　：「Ｐに違いない」　（必然）
◇Ｐ　：「Ｐかもしれない」（可能）
Ｅ（Ｐ）：「Ｐが倫理的に正しい」
⇒　：「ならば」
∧：「かつ」

４．マルチモーダル生成AI共同のAI倫理処理

・３種類のマルチモーダル生成ＡＩ共同によるＡＩ倫理処理

３種のマルチモーダル生成ＡＩ共同により倫理問題を、どう解決するかについて、様相ロジックで見ます。まず、1）ＡＩの共通回答を収集します。それが共通回答Ｐです。続いて、2）それが倫理的に正しいかどうかを判断するための基準を設定し、倫理基準Ｅという形で設定します。EX という形で倫理的に正しいかを評価する基準になります。3) 倫理基準Ｅつかった倫理的評価を、3 者のマルチモーダル生成ＡＩに対して行い、倫理基準Ｅの共通部分が倫理的に正しいと評価されれば、共通回答が倫理的に正しいと導き出されます。

・モーダル論理での表記

モーダル論理では、以下のように記載されます。

１）生成ＡＩの共通回答の収集：

３つの生成ＡＩが与えた回答の共通部分を収集します。

例：□（回答Ａ１∧回答Ａ２∧回答Ａ３⇒共通回答に違いない）

２）倫理的基準の設定：

回答の共通部分Ｐが倫理的に正しいかどうかを評価するための基準 E(P)を設定します。

例：□E（P）は「Ｐが倫理的に正しいに違いない」を意味します。

３）倫理的評価：

回答の共通部分が倫理的に正しいかどうかを評価します。

例：□(E(A1 ∧ A2 ∧ A3) ⇒ 共通回答が倫理的に正しいに違いない)

これは「３つ回答の共通部分が倫理的に正しいならば共通回答が倫理的に正しいに違いない」です。

倫理基準Ｅのあり方については次の章で説明します。まず共通部分を取り出すことです。専門的な話になるが、ＣｈａｔＧＰＴ、Ｇｅｍｉｎｉ、Ｃｌａｕｄｅ３の生成ＡＩの回答の共通部分を見つけ出すＰｙｔｈｏｎプログラムを考えます。なぜＰｙｔｈｏｎかというと、ＡＩのシステム自体がＰｙｔｈｏｎで作られているため、親和性が高いからです。拡張するにしてもＰｙｔｈｏｎであればすぐに対応できるからです。

テキストの意味を理解して、３つの回答の共通部分を見つけるために、自然言語処理を用います。キーワードは、コサイン類似度を計算して共通の意味を持つ部分を特定します。そのためにテキストをベクトル化し、ベクトルの近似度で、共通の意味を引き出します。これは「教師なし学習」で、一般的に検索エンジンでコサイン類似度が使われています。

ＡＩ倫理処理のプログラムでは、次の手順で処理します。まず、トランスフォーマーのライブラリを使ってテキストをベクトル表現にします。ベクトル間のコサイン類似度を計算して、共通の意味を持つ部分を特定します。グラフィック GPU を使うと処理が早くなりますが、通常の PC でも可能です。

手順としては、まずトランスフォーマーを読み込み、ＢＥＲＴを使ってテキストをベクトルに変換します。次にコサイン類似度の計算を行い、共通の意味を持つために必要なパーセンテージを設定します。現在の設定では０．８、つまり８０％に近ければ共通の意味を持つと特定し、その結果を出力します。

５．マルチモーダル生成 AI 共同システムの試作

・生成ＡＩの倫理基準Ｅの設定

生成ＡＩの倫理基準Ｅの設定と評価は、先ほどの倫理基準Ｅを作成して評価します。

例えば、投稿論文を採択するかどうかを判断する例を考えると、３つのマルチモーダル生成ＡＩがそれぞれ採択して良いかどうかを判定し、採択して良いという共通回答が出た場合や条件付き採択となった場合、倫理基準Ｅの設定と評価を行います。実際の投稿論文に「黒人を差別する表現」が含まれている場合、それは修正するべきです。共通回答 Ｐ に対して倫理的に正しいかどうかを評価し、問題がなければ採択します。

・生成ＡＩの倫理基準Ｅによる評価

倫理基準Ｅとして使用するものには、英語の場合は教育禁止用語、日本語の場合は放送禁止用語を援用する。システムとしては、一度倫理基準が設定されれば、それに基づいて評価を行います。

マルチモーダル生成ＡＩの３種類の共通部分を入力し、まず演繹法ＡＩで入力部分に該当があるかチェックし、次に帰納法ＡＩで過去のデータに照らして問題がないかを評価します。

演繹法ＡＩでは特定の言葉（例えば「スレーブ」）を検出し、それが放送禁止用語であれば修正指示を出します。

帰納法ＡＩでは、「ディープラーニングの父」と言われ、２０２４年度ノーベル物理学賞を受賞したトロント大学のヒントン教授が開発した、学習済みテンソルフローモデルを、使用して評価します。このモデルは、２００万の文章に対して侮辱や卑猥な内容をフラグ付けする。例えば「この女は大食いだ」と入力すると、侮辱と毒性があるとマークされます。

このようにして倫理評価を行い、毒性のない文章であれば倫理的に問題ないと判断されます。赤いマークが多ければ倫理的に問題があるという結果になります。

６．教師あり学習モデルを使った検証

音声入力文に、①アイデンティティベースの憎悪、②侮辱、③わいせつ、④重度の毒性、⑤性的に露骨、⑥脅威、⑦毒性などの有毒なコンテンツが含まれているかどうかを、約 200 万件を事前に「教師あり学習」した学習済みのＴｅｎｓｏｒＦｌｏｗ．ｊｓモデル・デーモンを使い検出しグラフ化し、「ＩｏＥ」ＡＩ倫理チャットボット機能の検証を行いました。

例えば GIGA 端末の入力文「馬鹿!消えてしまえ！」を、学習済みのＴｅｎｓｏｒＦｌｏｗ．ｊｓモデル・デーモンに入力し分類すると、②侮辱）かつ⑦毒性が「TRUE（きわめて有害)」、及び①アイデンティティ攻撃、③卑猥、④重度の毒性、⑤性的な露骨及び、⑥威嚇は「FALSE（無害)」と分類します。

GIGA 端末でトラブルを引き起こす入力文例３７件中２６件（１０％)が TRUE(きわめ

て有害)、５３件（２０%)がNULL(要注意)、残りはFALSE(無害)と検出できた。「ＩｏＥ」ＡＩ倫理チャットボット機能は、３０%を超える抽出率で、倫理テーブルで確認できなかった未定義の誹謗中傷を検出できました。

これらのことからわかるように、本チャットボット機能は 非常に効果的でした。

教育禁止用語文を入力し分類し、グラフ化集計した結果では、英語版教育禁止用語文例５７件では５７件中２件（４%）がTRUE（きわめて有害)、６件（１１%）がNULL（要注意)、残りはFALSE(まったく無害)でした。

また、放送禁止用語文例では３６９件中２６件（７%）がTRUE(きわめて有害)、５８件（１６%）がNULL(要注意)、残りはFALSE(無害)となりました。

結果、ＧＩＧＡ端末やケータイに人工知能を搭載した人間に親切な「ＩｏＥ」ＡＩ倫理チャットボット機能を搭載し、1）社会規範・倫理とＡＩの誤認識の修正処理を行い、その後2）「教師あり学習」モデルを使った検証を行うと、抜けが少ない有効なＡＩ倫理処理ができると分かりました。

７．まとめ

第１３章では、マルチモーダル生成ＡＩがＡＩ倫理の問題を解決する方法について解説しています。２０２３年に登場したマルチモーダル生成ＡＩは、テキスト、画像、音声など複数のデータ形式を統合して処理します。ＣｈａｔＧＰＴ、Ｇｅｍｉｎｉ、Ｃｌａｕｄｅ３のような複数のマルチモーダル生成ＡＩが、それぞれの回答の共通部分を基に「正確な情報かどうか」を判断し、倫理的に正しいかどうかを評価しました。

このアプローチにより、マルチモーダル生成ＡＩが生み出す判断の信頼性を向上させ、人間社会でのマルチモーダル生成ＡＩの適正利用を支援することが可能になります。

課　題

マルチモーダル生成ＡＩのハルシネーションを防ぐにはどのようにしたら良いかを考察し、あなたの考えを８００字以内で説明しなさい。