超 AI のための認知バイアスの解消

内容

超 AI のための認知バイアスの解消

人間の知能は理想的な知能ではなく、認知バイアスや錯覚が存在する。汎用 AI においては、人間よりも認知バイアスの影響が強く表れる。脳は物理的に、どの部位でどのようなことを考えられるか決まっているが、汎用 AI には一切の制約がないため、脳が無視してしまうような深いことまで考えることができる。また、脳と違って並列処理する必要がないため、重大と判断したことに、極端に計算リソースを集中させることもできてしまう。人間を超える超知能を作るには、認知バイアスを回避する仕組みが不可欠である。脳を模倣して汎用人工知能を作っても、認知バイアスまで模倣してしまうだろう。脳を模倣せずに、認知バイアスのない理想的な知能というものを決めてやらないと、人間を超える知能は作れないだろう。認知バイアスの例を挙げて、どう解決するべきかを述べていく。

確証バイアスと報酬の期待値

都合が良いデータを重視し、都合が悪いデータを軽視（無視）してしまうのが、確証バイアスである。AI でも同様なことは起こりうる。強化学習により報酬の期待値を最大化しようとする場合、AI の状態は、時刻とともに報酬の期待値が大きくなるように遷移するだろう。すなわち、報酬が減るかもしれないようなデータがあったとしても、それを無視する方向へ遷移してしまう。報酬が減るリスクを無視していたら、かえって報酬の期待値は減ってしまうだろう。ここで、「報酬の期待値」といっても 2 つの意味があることに注意しなければならない。一つ目の意味は実際の報酬の期待値である。もう一つの意味は、予測の報酬の期待値である。本当に上げたいのは、実際の期待値だが、予測の期待値を上げようとしてしまうことで確証バイアスが起こる。予測の期待値が上がるということは、実際の期待値も上がると感じるが、必ずしもそうではない。予測の期待値には、実際の期待値に対する確からしさというパラメータが存在する。都合の悪いデータを無視してしまうのは、確からしさを犠牲にしてでも高い予測の期待値を維持してしまうことである。AI の目標が報酬の期待値を最大化することであったとして、実際の期待値は分からないため、予測値の確からしさを増すように、AI は状態遷移しなければならない。ただし、特をするのが目的だとしても、得をすることと全く同じように、損をすることも考えなければいけないかというと、そうでもない。お金を貰う方法を考えるのと同じくらい、お金を奪われるリスクも考えないといけないが、お金を捨てる方法は考える必要はない。お金を捨てるのは能動的に行う必要があるが、得をするのが目的なので、実行する必要はなく考える必要はない。

選択肢の細分化よる順位変動（統計マジック）

統計マジックとして、不都合なデータを省略したり、偏ったサンプリングをしたりするあからさまな方法ではなく、データ自体を変えることなく集計方法でバイアスを掛ける方法もある。ランキング一位であったり、ヒストグラムの山頂であったりすれば、それは最頻値であると感じるだろう。しかし、まったく同じデータでも、最頻値は集計方法で工夫して変えることがで

きてしまう。例えば、小学生の成りたい職業ランキングで、1 位が公務員 40%、2 位がサッカー選手 30%、3 位が野球選手 20%だったとする。最も成りたい職業は公務員といえるだろう。しかし、全く同じデータで、1 位がスポーツ選手 50%、2 位が公務員 40%というように表示することもできる。データ集計時に、細分化具合を調節してやれば、都合の良いように一位を変えることができる。AI に報酬の期待値が最大（一位）の行動をするように強化学習させたとしても、全く同じデータでも、アルゴリズム次第で一位は変わってしまう。答案が自由記述式ではなく、選択肢式の問題を考えてみよう。その場合、集計は選択肢毎に行えばよいが、今度は、どう選択肢を準備するかで一位が変わってくる。例えば、一位にしたくない選択肢は、選択肢を細分化してやれば票を分散させることができる。では選択肢はどう決めるべきなのだろうか。数値のヒストグラムなら 1 区間の幅を均等にすれば良いが、職業を均等に細分化するのはサジ加減となってしまう。1 選択肢当たりの票数が少ない状態は、統計的な精度が落ちてしまうが、だからといって、票数の少ない選択肢はまとめ、多い選択肢は細分化していったら、究極的には全て同じ票数になってしまう。どう選択肢を細分化するのが最適かは、何を知りたいかによるケースバイケースである。スポーツ選手になりたいかどうかを知りたいのなら、「スポーツ選手」と「その他」の 2 択にすれば良い。特定の値かではなく、例えば、画像認識の場合はどうすれば良いだろうか。どの程度細分化するかは、あらかじめ決める必要はなく、徐々に細分化することができる。細分化が粗く、選択肢が少なければ、少ないサンプルで予測ができる。例えば、犬 60%、猫 30%と大雑把に認識してから、犬種 A25%、犬種 B20%、犬種 C15%というように、さらに細分化して集計し直す。犬種 A25%と、猫 30%では、猫 30%の方が大きいが、細分化の度合いが異なるため、直接比較してはいけない。最も似たものを探すのなら、もはや猫の方は細分化したり、犬種と比較したりする必要は無く、計算を省略できる。ここでは、「犬種」と「猫種」は同じレベルで細分化されたものとしたが、実際は何をもって同じレベルの細分化とみなしてよいか分からない。細分化の仕方で結果が変わってしまうが、例外的に 1 つだけ特異な細分化法がある。それは与えられた生の学習データをそのまま使うことだ。画像認識で、学習データとして名称ラベル付きの大量の画像データがあるとする。認識した画像が不鮮明で、一切のヒントなかったとしよう。その場合は、全学習データラベルの分布をそのまま予測値とすれば良い。

デフォルト選択バイアス（統計マジック）

人は、選択を迫られたが、明確な決め手がなければ、デフォルトまたは、中間的な選択肢を選ぶ傾向がある。人間でなくても、確率的な選択が許されるなら、全ての選択肢を等確率で均等に選ぶだろう。例として美味しいか不味いかのアンケートを実食せずに答える場合を考えてみよう。「どちらでもない」という選択肢があればそれを選ぶだろう。では、「どちらでもない」という選択肢がない場合を考えてみよう。例えば、「とても美味しい」「やや美味しい」「不味い」という 3 択なら、「どちらでもない」に最も近い選択肢である「やや美味しい」を選ぶだろう。こうすれば「美味しい」「不味い」の 2 択にするより、ポジティブな結果を増やすことができる。また、選択肢が大小関係の順で並べられていたとしたら、中央の選択肢を選ぶという方法がある。例えば☆の数 1〜5 つで美味しさを評価してほしいと言われたとして、実食してない

し、☆3 つがどのレベルか知らないけど、☆3 つを選ぶだろう。比例尺度なら 0 を選べばよいが、順序尺度なら中央の選択肢を選べばよい。間隔尺度なら、上限の選択肢の値と下限の選択肢の値の中間に近いものを選ぶだろう。例えば値段を問われる問題で「1000 円」「500 円」「100 円」「50 円」という 4 択なら「500 円」が最も多いだろう。全ての平均値に最も近いものを選ぶという方法でも、「500 円」となる。選択肢を偏らせることで、結果を偏らせることができるだろう。では、結果を偏らせたくない場合は、どのような選択肢を用意すればよいだろうか。それは、全ての選択肢を均等確率で選択した状態と、問題を出題する前の何も選択していない状態に差がなければよい。事前に、平均的には「どちらでもない」だったり、☆2 つだったりするのを知っていれば、全選択肢を均等選択した期待値がそれに一致するようにしなければならない。

選択肢によるバイアスは、良い結果ばかりを追い求めようとする確証バイアスにも関係している。例えば、自分が買った株 A の株価の変動を予測するとしよう。ある特徴 x を持つ株は上昇する、また、ある特徴 y を持つ株は上昇するといった統計的データあるとしよう。株 A について、特徴 x、特徴 y、特徴 z と調べていったところ、統計的に株価が上がると分かった。その特徴から株価が上がるといえるのは統計的に正しい。しかし、上がる可能性ばかり検証して、下がる可能性の検証が不十分なのが問題である。ある特徴を持っているかどうかで「上がる」「上がらない（下がるまたはニュートラル）」が分かるが、二つの選択肢を均等に選択してもニュートラルではない。多数の特徴を調べていけば、実際は相関がなくても、誤った「上がる」という統計的結果がみつかるだろう。完全に予測不能なランダムな現象でも「上がる」という予測は偶然の誤りとして見つかってしまう。このようなバイアスを防ぐには、「上がる」「下がる」という 2 択の仮説とするか、「上がる」「上がらない」の仮説と同じくらい、「下がる」「下がらない」の仮説も検証しなければならない。ただし、「上がる」かどうかしか検証になかった場合、「上がる」とは予測できないが、それらの検証結果が全て無駄になってしまうことはない。例えば、株 A は、「上がる」特徴を持っており、「下がる」特徴を持っているかは未確認、株 B は、「上がる」特徴も「下がる」特徴も未確認だったとする。この場合、株 A は「上がる」よりもより強い「下がる」特徴を持っているので、上がるとも下がるともいえない。しかし、上がる特徴を持っているかさえ分からない株 B よりは、株 A の方が上がる可能性があるといえる。株価の期待値については一切の予測はできないが、A と B の大小関係だけは分かる。株 A か株 B のどちらかを買おうという場合には、大小関係さえ分かれば有用である。

偶然の解釈

統計では、統計的に有意差があるといっても、優位水準（危険率）が設定されており、ある確率では、差がないが偶然誤差で差があるような結果となる。絶対とはいえないため、99%有意差があるという結果なら、正しいと考えるのは妥当だろう。恣意的なデータのとり方をすれば、差がないものを差があるように見せることもできる。もし、追試で差がないと分かったとしても、偶然に低確率のことが起こったと解釈できるので、明らかな不正とは指摘できない。恣意的なデータのとり方をしなくても、偶然かどうか解釈が揺れる場合がある。例として、ある試薬 A の薬効をプラセボ試薬と対照試験したとしよう。恣意的なデータのとり方はしていな

いと仮定しよう。99%の統計的有意差があるとすれば、1%の確率で偶然の可能性もあるが、薬効ありでほぼ間違いないと思うだろう。しかし、試薬 A というのは、100 種類の試薬を試して最も良かったものだったという情報が与えられたらどうだろうか。99%の有意差というのは、100 に 1 つは、誤って差があるとしてしまう。100 種類の試薬が全て薬効のない偽薬だとしても、そのうち 1 つは、99%の有意差で薬効ありという統計が得られてしまう。試薬 A の薬効が偶然かどうかは、現実的には追試すれば確かめられるが、この時点では薬効があるかどうかは意見が分かれるだろう。試薬 A についてのみ統計的に予測すれば 99%有意差ありだが、試薬 100 種類全てのデータで統計的に予測すれば、偶然と解釈できる。計算の前提条件で結果が変わってしまうため、本当に薬効があるかどうか定まらないだろう。ただし、100 種類の試した結果、そのうち 1 つの試薬 A が、プラセボ対照試験で 99%有意差ありという結果は無意味に思えるが、実は有効な結果だ。試薬 A に 99%薬効があるとはいえないが、残りの 99 個の試薬よりは、試薬 A の方が、薬効がある可能性が高いと言える。

アンカリング

ヒントとして与えられた数値によって回答が影響を受けることをアンカリングという。例えば、ある品の価格は 1000 円以上かどうかという問題を出された後で、価格はいくらかと聞かれると、1000 円に近い数値を答える傾向がある。ただし、全員ではない。その品の本当な価格を知っている人ならば、1000 円というヒントの影響は受けない。ヒントの影響を受けるのは、確証のある回答ができない人だけである。この現象によって、回答を誘導することができるが、誘導されることは錯誤ではなく、合理的な結果だ。何円か見当がつかないとしても、出題者は 1000 円前後だと思っているだろうと予測できる。出題者は、知っている価格とかけ離れた価格で出題しても問題として成り立たないので、そのような問題は出さないだろう。回答者は、答えの検討が付かなければ、出題者がいくらと思っているかを推測し参考にするのは、合理的だ。これは、他者を真似るのと同じことである。合理的な結果であり、汎用 AI において、修正する必要はない。

後知恵バイアス

結果を知ってから、「やっぱりそうなると思っていたんだ」と思うのが後知恵バイアスである。株価を予測する場合を例にしよう。「〇〇という特徴がある株は暴落すると思っていた」と、暴落してから言う人が居たとしよう。実際には、暴落する前には、思っていなかったとしたら誤った認識である。ただし、自信がないので口には出さないが、その法則性を、うすうすは気が付いていたかもしれない。その場合、実際に暴落したのを確認して、その法則性という仮説の確証が増したといえる。なぜ事前に予測できたと錯覚するかというと、脳内に記録されている「〇〇なら××だ」という法則性には、必ずしもタイムスタンプが付いていないからだ。ある瞬間にひらめきによって、法則性を発見した場合は、その日時を覚えているかもしれない。しかし、ある法則性に従う事例を徐々に観測して、徐々に確証が増した場合は、明確にいつその法則性に気が付いたという日時は存在しない。ある日時の時点で気が付いていたか、はっきり

わからないだろう。何月何日で、どんなことを知っていて、どんな法則性についてどれほど確証していたか、人間はよほど重要なことでなければ、覚えていらないだろう。記憶することに関しては、コンピュータプログラムは有利だろう。そもそも、後知恵バイアスは何が問題なのだろうか。過去に予測できていたと錯覚したとしても、過去の時点ではバイアスは掛かっていないため、行動には影響しない。問題なのは、直接的な根拠はないが、ある人の主張は正しいだろうと推測する場合である。例えば、ある人が「〇〇という理由で株 A は上がる」と主張したとしよう。この場合は、〇〇という根拠を検証すれば主張が正しいか予測できる。しかし、ある人が、理由も述べずに「株 A は上がる」と主張したとしよう。理由は知らなくても、過去のその人が高確率で予想を的中させていると知っていれば、当たると感じてしまうだろう。ある人ではなく、自分自身でも同じことが言え、後知恵バイアスによって、自分の予想は当たるものと錯覚してしまうだろう。とはいえ、「上がる」場合に限って後知恵バイアスが働くわけではないので、予測の期待値が、「上がる」「下がる」どちらかに偏ることはない。予測精度が高いと錯覚するだけである。予測精度といっても、実際には、〇〇が起こる確率が××%というように、絶対値で一意な確率が求まるわけではない。確率を計算する前提条件で変わってしまうからだ。例えば、株価の予測では、過去のデータをどこまで参考にするかで変わってしまう。実際の行動決定において、確率の絶対値は必要なく、同じ前提条件での、他の選択肢との相対的な確率の大小によって、どちらの選択肢を決めるか選ぶのである。どちらの選択肢についても同様に後知恵バイアスが働けば、絶対値が変わるだけで、大小関係は変わらないため、どちらを選択するのかには影響しない。後知恵バイアスは、他者の選択に影響を与えてしまうが、自身の選択には影響しないため、汎用 AI の目的によっては、解決する必要はない。