<参考2> Synthetic Data_慎重派の論考

ESOMAR “RESEARCH WORLD”: 2024/08/15

合成データ： 参加した上で、賢くやろう！

著者 Crispin Beale、Simon Chadwick、Finn Raben、Mike Stevens　8 月 15 日

(Crispin Beale、Simon Chadwick、Mike Stevens、Finn Raben による書面討論)

人工知能 合成データ　　19 分で読める　　共有

第一に、観察： 調査の専門職は、常に 3 つの基本原則の基に成り立っている。 -i) 厳密さ (三角測量／検証／文脈化)、ii) 客観性、iii) 透明性。しかし、私たち専門職の「新しい」アプローチの多くは、これらのストレステストを受けることなく、私たちのツールボックスに同化されてしまうことがあまりにも多い。その結果、重大で意図しない結果を招いている。

一例として、私たちの多くは 10 年以上（たぶん 20 年！）前に、調査参加率と回答者の質の低下のリスクについて警告していた。これらの懸念を軽減するために、私たちはアクセスパネルを採用し、コストと重要な品質の面で「底辺への競争」を不注意にも開始してしまった。私たちは今もまだ、これらの問題に苦慮している。私たちの業界には、新しいコンセプトを採用したり、真っ先に飛び込んだりした「歴史」がある。利益を得られることはあっても、長期的な影響を常に評価しているわけではない。

合成データは、同じような機会と課題を私たちに提示している。過去の過ちを繰り返さないようにしようではないか。

歴史的なアナロジー

今から 100 年も前に、私たちは衣類に革命をもたらす画期的な新しい合成繊維であるポリエステルのことを耳にした。1960 年代には、「アイロンをかけずに 68 日間そのまま着ても見栄えのする奇跡の繊維」として宣伝された。

一般的にポリエステルとして知られるポリエチレンテレフタレート (PET) は、1950 年初頭に米国で商業生産が開始されて以来、その耐久性、収縮や伸びに対する耐性、手入れの容易さなどから、現在では世界で最も人気のある繊維の一つとなり、何千ものさまざまな消費者および産業用途に使用されている。

しかし今日、私たちはこの「奇跡」が意図せず、予期せぬ結果をもたらしたことを知っている。

私たちの海底には 1400 万トン[1] のマイクロプラスティックが存在すると推定されており、その数は毎年 50 万トンずつ増加している。この汚染の約 1／3 は、洗濯された合成繊維の衣類から発生している。マイクロファイバーが放出され、地球を汚染し、環境への影響をもたらしていることが、ようやく理解され始めたところだ。

人類は、有意義な方法での対応が遅れており、今になってようやく、パタゴニア、サムスン電子、オーシャンワイズなどの企業が協力して、この汚染物質の進行中の蓄積を減らす方法を模索している（例えば、サムスンの洗濯機用フィルターは、海洋に到達する前にこれらの汚染物質の半分強 (54%) を除去する）が、この問題に対処する効果的で世界的な解決策はまだ存在しない。

合成データ (または合成回答者)

「ビッグデータ」やそれ以前の他のトレンドと同様に、合成データは現在、専門家の想像力を捉え、さまざまな方法によって業界全体で探求され、採用される可能性がある。合成データは一枚岩でも孤立したイノベーションでもないが (結局のところ、それは多くの異なるイノベーションの構成要素の一部である)、最も一般的には、独自の課題を伴うことがわかっている「本物」または「サーベイ」データに対立する二者択一として提示される。とはいえ、この記事では、その二者択一を前提として合成データについて議論する。

第一に、合成データは「新しい」ものではない。私たちは何年も前から使用してきたのだから、その限界を十分に認識しているはずである。しかし、新しい技術や「おもちゃ」に投資し、採用することへの情熱に対して、過去の過ちを繰り返さないためにも、3 つの基本原則を適用することを止めるべきではない！

ChatGRT による合成データの定義の 1 つには、次のようなものがある。

市場調査における合成データとは、個人を特定できる情報 (PII) や機微な情報を含まずに、実際のデータの特性やパターンを模倣して人工的に生成されたデータを指す。

実際のデータセットに見られる統計的特性や関係を刺激するために、数学モデルやアルゴリズムを使用して作成される。合成データを使用すると、リサーチャーはデータ不足の問題を克服し、プライバシーコンプライアンスを維持し、限られた、あるいは機微な実世界のデータソースだけに依存することなく、さまざまなシナリオを調査することができる。

これにより、プライバシーと機密性を保護しながら、より堅牢な分析、実験、モデルのトレーニングを行うことが可能になる。

非常に簡単に言えば、合成データはモデル化され、重み付けされたデータであり、データ参照情報の「ギャップ」を埋めるように設計されている。元データは、静的な定量データ、標準的なアルゴリズム出力、または LLM データなどであるが、本質的にはモデル化されたデータであることに変わりはない。

1980 年代、アイルランドの AC ニールセンの小売監査は、国内最大の小売業者の一つである Dunnes Stores の協力を得ることができなかった。そのため、彼らは同じような規模、売上高、および顧客数の競合店を選び、特定のブランドが提供できる「ハード」な販売データを追加し、それによって Dunnes の販売をモデル化して、全国および地域の「合成」推定値を提供した。このアプローチは、普及率が高く、購入頻度の高いブランドでは非常にうまく機能したが、浸透度が低く、購入頻度の低いブランドや、

Dunnes 独自のプライベートブランド商品の推定では、非常に不安定だった。この不安定さは、時として誤解を招く方向性のガイダンスや、クライアントとの非常に興味深い議論につながることもあった。

合成データは、「難しい」サンプル割り当てを (より簡単に) 満たす手段として提示されることが多くなっている。.... MRX が登場して以来、リサーチャーは、合成データが解決すると主張する「リーチするのが難しい」聴衆にコスト効率よく迅速にインタビューすることに苦労してきた。個人的には、金曜日の夜、サッカーの試合が行われていた（！）中で政治的な世論調査のインタビューを行い、18〜21 歳の男性を見つけようとしたことを覚えている。干し草の山から針を見つける方が簡単だっただろう。最終的にはうまく完了したが、時間がかかり、商業的な観点からは多額の費用がかかった。

したがって、これらの新しいアプリケーションを、サンプルサイズを達成し、結果をより良く、より速く、より安くクライアントに提供することを可能にする奇跡のソリューションや、「聖杯」(または、商業的な売上増、利益の向上、コスト削減を実現できる) と見なす誘惑は、明らかに大きい。しかし...油断は禁物だ。人生は、物事が真実であるとは思えないほど良く見える時には、多くの場合、リスクを伴うことを教えてくれる。

厳密に制御されたパラメーターと独自の環境を持つ特定の分野や状況においては、合成データは歓迎すべき発展である (例えば、「合成」がん患者は既存のデータに基づいてモデル化／作成することができるため、個人データやプライバシーの問題なしに、完全に匿名化された合成ペルソナを使用して、服薬遵守、副作用、疾患症状などに関する調査を実施することができる)。

しかし、より広範で複雑な状況 (2023 年 8 月に ESOMAR が開催した AI 会議で、Annelies Verhaeghe 氏が発表した「Land Rover perceptions」に示されたように)では、合成データによって提供される回答は実際の回答者データと一致しておらず、「モデル」は、実行可能で関連性を維持するために、「実際のデータ」によって継続的に更新される必要がある。

ここで、厳密さ、三角測量、および検証の話に戻る。市場調査で合成回答者を使用すると、さまざまなリスクが生じる可能性があることは (既に) 明らかだ。このトピックは、多くの著者によって取り上げられている。Jason Dunstone の論文 (https://www.researchsociety.com.au/news-item/16368/embracing-synthetic-datas-potential-while-valuing-real-people) もその一つだ。

私たちの観点からは、次の 5 つの重要な懸念事項があると考えている。

1.バイアスと代表性の欠如:

- **固有のバイアス**: 合成回答者は、トレーニングデータに存在するバイアスを誤って反映する可能性のあるモデル／アルゴリズムに基づいて生成される。これにより、ターゲット母集団を正確に表していない歪んだインサイトが生成される可能性がある。
- **限定された多様性**: 合成データは、実際の回答者の完全な多様性を完全には捉えることができない可能性がある。特に、アルゴリズムが広範な人口統計と行動をシミュレートするのに十分に洗練されていない場合に。

2.品質と信頼性の問題:

- o **データ品質**: 合成応答の品質は、その生成に使用されるアルゴリズムとデータに大きく依存する。適切に設計されていないモデルは、信頼性の高いインサイトを提供しない、低品質のデータを生成する可能性がある。
- o **検証の課題**:現実世界のデータに対して合成データの精度と信頼性を検証することは困難な場合がある。これにより、導出されたインサイトがデータの信頼性を保証することが困難になる。

3.倫理と透明性の懸念:

- o **透明性の問題:** 企業が利害関係者にその使用を透明性をもって伝達しない場合、合成回答者の使用は倫理的な問題を引き起こす可能性がある。これは、クライアントや一般の人々との信頼問題につながる可能性がある。
- o **倫理的な影響**: 合成データの作成と使用には倫理的な考慮事項が必要であり、特に望ましい結果を生み出すためにデータの誤用や操作を行う可能性がある。

4.意思決定への影響:

- o **誤った意思決定:** 実際の消費者行動を正確に反映していない合成データに基づく意思決定は、誤った戦略や行動につながり、金銭的損失や評判の低下につながる可能性がある。
- o **テクノロジーへの過度の依存**: 合成データやアルゴリズムへの過度の依存 (そしてそれが常に「正しい」と思い込む) は、人間の複雑な行動を理解する上で不可欠な、人間の判断や定性的洞察の重要性を低下させる可能性がある。

5.規制およびコンプライアンスリスク:

- o **規制上の課題**: 合成データの作成と使用は、特に規制の厳しい業界では、すべての規制基準に準拠しているとは限らない。合成回答者を使用する場合、データ保護とプライバシーに関する法律の遵守を確保することが複雑になる可能性がある。
- o **法的な影響**: 合成データが法的基準に準拠していないことが判明した場合、法的な問題や罰則につながる可能性があり、市場調査におけるこのようなデータの使用はさらに複雑になる。

その他にも、十分に認識しておかねばならない 3 つの問題がある。

1) 中国での開発は非常に不透明なままであるが、合成に関わる開発作業のほとんどは西洋の英語圏で行われている。そのため、他の国、言語 (LLM)、文化を含めることが困難になっている。

2) 今日利用可能なあらゆる技術開発があったとしても、マイノリティや多様な社会経済グループの意見を完全に代表することは、より困難なままだ。したがって、合成マイノリティデータの作成は、非常に綿密にレビューされ、チェックされる必要がある。

3) 最後に、そしておそらく最も懸念されるのは、この (潜在的に代表性がない、偏っている、または誤った方向に導かれている) 合成データがデータレイク（データの湖＝貯蔵庫）に行き着いた場合にどうなるのかということだ。これらのデータウェアハウスやデータレイクでは、データが他の複数のソースと

組み合わされ、将来他のプロジェクトのためにこれらのレイクから引き出されると、抽出されるサンプルの 100%が合成データで構成され、「本当の」回答者がサンプルに含まれていないということが十分にあり得る。これは現在進行中のプロジェクトに有効かもしれないし、そうでないかもしれないが、結果を解釈する (そして利用する！) 人々がデータの出所を知り、理解することがこれまで以上に重要になっている。

合成回答者を適切かつ一貫して特定しなければ、合成繊維が海を汚染したのと同じようにデータレイクを汚染するリスク（懸念）がある。マイクロプラスティックのように、過去に遡ってフィルタリングして限定的な成功を収めるよりも、今のデータセットでこれらの回答者に「タグ付け」し、将来的に特定できる (必要に応じて削除する) ようにした方がはるかに良いだろう。

ハワイとカリフォルニアの間に位置する「太平洋ゴミベルト」は、現在誰でもが知っており、目にすることができる。...データレイク (データセット) のさまざまな領域は、将来的には「ゴミ」になるか、目に見えなくなるのだろうか? 無防備なユーザーは、抽出されたデータが合成データで汚染されていることに気づかないのだろうか? (良くても) 無関係な合成データの回答者は、最悪の場合、実施されている分析から積極的に除外されるべきなのだろうか?

最後に、合成データを使って「正しいこと」を行おうとしている企業がないと言っているわけではない。それどころか、LivePanel のような企業がその好例であり、間違いなくトレンドに逆らおうとしている。しかし、彼らの努力は、上記で指摘したリスクを軽減するためにグローバルに連携しなければ、私たちに降りかかるはるかに大きな落とし穴から目をそらすものではない。同様に、Ipsos のナレッジセンターで公開されている合成データに関する見解の記事をダウンロードすることもできる。これは、合成データの役割に関する、別のバランスのとれた見解である。
(https://resources.ipsos.com/Get_Synthetic_Data_POV.html) 。

私たちの願いはこうだ。確かに、新しい技術を受け入れる必要がある。しかし同時に、リスクにも留意し、意図しない結果を予測する努力を怠らないようにしなければならない。今こそ、業界の標準規格 (ISO 20252)、倫理ガイドラインまたは行動規範 (MRS 行動規範・ESOMAR 行動規範)、または標準化されたグローバルなテストスキームを通じて、これらの課題を認識し、軽減するようにしよう。このスキームでは、データについて情報に基づいた意思決定を行うために十分な情報を、サプライヤーがバイヤーと共有することを義務付けていまる (Ray Poynter 氏によるウェビナー (2024 年 6 月)) 。現在、そして将来にわたって、合成データを確実に識別できるようにするための対策とプロセスを導入しよう。理想的には、グローバルに連携し、一貫したアプローチが必要である。

反論: 落ち着いてデータをテストせよ！

人々は合成データについて冷静になる必要がある。拒否、制限、規制を求める声が調査業界の一部で高まっているようだ。

主な議論は、このようなものは最悪の場合、捏造されているか、良くてもインターネット上の最も恐ろし

いもののすべてを含む不透明な学習データのスープから得られているということだ。一方で、一次調査やインタビューから得られたデータは、より高次の真実、つまり本物の、有機的な人間の信憑性を表していると。
この二項対立の構成はナンセンスだ。

Peter Thiel 氏のように聞こえるかもしれないが、強力なイノベーションの真の特徴を理解する前に、リサーチャーがそれを中絶してしまう危険性がある。

私たちはこれを「合成された」データ、つまり複数のソースから生成され、ブレンドされたものと考えるべきだ。「合成された」という意味ではなく、模倣的で安っぽくて意地悪な「合成」と考えるべきだ。

次の例を考えてみよう。

Glimpse には、リサーチャーが個々の参加者やサーベイデータセットのセグメントに対して質問できる機能がある。「合成された」回答は、サーベイで得られたデータから合理的に推論できる場合は決定論的に生成され、推論できない場合には、大規模言語モデルの助けを借りて確率論的に生成される。

Vurvey は、何百人もの調査参加者からビデオベースの回答を収集する。回答はテキスト化され、さまざまなペルソナを代表する「エージェント」のトレーニングに使用される。リサーチャーは、これらのペルソナのエージェントと対話して質問をしたり、仮説を検証したり、アイデアを生み出したりできる。

Market Logic Software の **DeepSights** は、Retrieval Augmented Generation (RAG) などのメソッドを使用して、何百、何千もの調査およびデータソースからインサイトを合成し、それに関連する質問に対する回答を組織内の利害関係者に自然言語で提供する。

インサイトのデータを「合成」するこれらの創造的な方法は、すべて「合成」されたものなのだろうか?
このトピックの面白いところは、学べば学ぶほどその定義が難しくなることだ。

「合成」という言葉が怖いとか、きちんとした定義ができないという理由だけで、調査やインサイトのための創造的な使用例を拒否すべきではないと言いたい。

以前にも同じようなことがあった。オンライン調査、ソーシャルリスニング、ビッグデータ … これらはすべて、「正しい」調査を守るという名目で、同じように恐れられ、正当性を否定されようとした。

第一に、合成データは一枚岩ではない。広範な技術(LLM、**GAN、RNN、SMOTE**)と、はるかに広範なデータソース (プライマリ調査、論文、書籍、ポッドキャスト、ブログ投稿、機密文書) を使用して生成されたデータのカテゴリに対する広義のラベルである。

第二に、新しい形式の合成データと、一次データの不十分なサンプルを扱うための確立されたアプローチ (補定、重み付け、拡散モデリングなど) との違いは、それほど大きくない。

第三に、これらの新しい技術を、アプリオリな仮定ではなく、意味のある客観的な基準で評価できないものだろうか。それらは機能するだろうか。どこが足りないのだろうか。どのようなメリットとリスクがあるのだろうか? それは、確立された調査の手法によって決定された「真実」を反映しているのだろうか? それは、根底にある社会的なデータセットのバイアスを永続化または緩和するのだろうか? **Yogesh Chavda 氏、Ray Poynter 氏、Joel Anderson** 氏らは、これらの問題について、思慮深く、偏見なく、可能な限りエビデンスを用いて書いている。

まだまだ先は長い。合成データの一部は、必然的に蛇の油（偽薬）になるだろう。合成データに依存した人々によって、非常に悪い決定が下されることもあるだろう。合成データに基づく「インサイト」の多くは、当たり障りのない、刺激に乏しいものになるだろう。

しかし、最後の段落で使われた「合成データ」という用語はすべて、今日の「市場調査」に置き換えることができるだろう。

したがって、私たちは客観的で科学的、そして証拠に基づいた考え方を持って、これらの新しい手法を受け入れるべきだ。どこで、どのように機能するのかを学ぶ。操作、誤解、モデルのバイアスなど、付随するすべてのリスクに注意してほしい。すべての報告において、データソースと手法の責任ある開示を徹底してほしい。

しかし、まだ完全には理解していないからといって、生まれたばかりのこの首を絞めるのはやめよう。

どちらも的外れ

2013 年、私はシカゴで開催された MR 業界の会議に出席していた。そこで、市場で認知され始めた新しい技術製品に関するワークショップに参加することになった。この新しいアプローチに内在する危険性について議論が交わされたとき、ふさふさとしたオレンジ色のひげを生やした若い男性が立ち上がり、次のような不朽の言葉を口にした。

> 「リサーチャーの厄介なところは、あなた方が技術のことなら何でも知っている（と思っている）ということだ。我々技術者の厄介なところは、我々が調査のことなら何でも知っている（と思っている）ということだ。今こそ、我々はお互いのことを学ぶべき時だ」。

その Dave Carruthers 氏は VoxPopMe の創業者兼 CEO であり、約束を守る若者だった。彼は業界について学び、周囲の人々にテクノロジーについて教育し、業界のガバナンスと成長に関与することで、その後退と前進の両方の対価を得た。

しかし、Dave 氏は ResTech の少数派だった。彼は他の技術者を業界のインフルエンサーやリーダーに紹介するという点で優れていたが、業界で活躍していても業界の一員ではない人は他にもたくさんいた。そして、私たちが注意を払うべきだったのは、これらの人々だったのだ。

投資の波に乗っている多くの起業家にとって、成功の鍵は参入しようとしている業界を破壊することだ。結局のところ、(『イノベーションのジレンマ』の）Clayton Christensen は、破壊者は多くの場合、最初は既存のリーダー企業が提供するものよりも劣るが安価な製品を持ってパーティーにやって来て、彼らが気づかないうちに彼らを追い越したと言わなかっただろうか?

2010 年にそうなった。ただし、気づかなかったリーダーとインフルエンサーは、業界とその行動規範や倫理を管理し、擁護することを任務とする国内外の協会団体であった。ResTech 革命の主要メンバーが協会階層の正当なメンバーとして受け入れられるようになるまでには､実に 10 年を要した｡その時点で、**「底辺への競争」は順調に進んでいたのだ。**

今日、私たちはまた同じ話を繰り返してしまう危険性がある。多くの傑出した調査会社が、生成 AI と「合成」データを使って実験をしている。そして世界中の会合で実験結果を公表し、その厳密さ、透明性、客観性の度合いについて議論することだろう。

しかし、こうした価値観は、おびただしい数のベンチャーキャピタリストから資金提供を受けている技術系起業家の多くが共有しているものではない。

そして、そのような投資家に関して私を「他力本願」だと非難する前に、彼らが 12 年間で ResTech に 600 億ドル以上を投資したこと、その中には真の価値とイノベーションを生み出したものもあれば、調査になりすましたクソゲーのような形で破壊をもたらしたものもあることを忘れないようにしよう。

今、私たちは調査のすべてを知っている技術系起業家と投資家の別の波に直面している。誰が彼らを迎え入れるのか? 「厳格さ、客観性、透明性」の必要性を教えるのか? そして、将来的に彼らを頼りにする世界に彼らを紹介するのだろうか?

この新世代のインサイト・プロフェッショナルを迎え入れ、巻き込み、教育することに成功すれば、合成データを含む AI のすべてのメリットは、ビジネスパフォーマンスに影響を与えるという点で真価を発揮することになるだろう。しかし、今後 10 年間、彼らを野放しにしておくのならば、代表性と品質の欠如は、誤った意思決定が行われ、業界の倫理が問われることになるだろう。
そこでは、規制当局からの望ましくない (しかし、おそらく正当な) 注目を浴びることになっているだろう。

あるいは、別の「底辺への競争」か？

だから、私たちの愛する業界のすべてのリーダー、インフルエンサー、そして協会に、私は言いたい。腕を伸ばして、これらの新しいプレーヤーをテントの中に連れてきてくれ。今すぐに。

[1] オーストラリアの連邦科学産業研究機構 (CSIRO)による