MRS reports / 2024
MRS Delphi Group

Using synthetic participants for market research

Part Two
of the BEST
Framework for Gen AI

MRS Evidence Matters

Build your AI knowledge and skills

Courses

23 May
Text Analytics for MR: The AI, the human and their collaboration

4 July
Generative AI in Market Research and CX

25 September
Transforming MR with Chat GPT

1 October
Text Analytics for MR: The AI, the human and their collaboration

5 December
Generative AI in Market Research and CX

On-demand

Generative Artificial Intelligence

Conference

4 July
Data Driven Insights 2024

Coming soon

Best Practice Guides

Collecting Data using the Metaverse

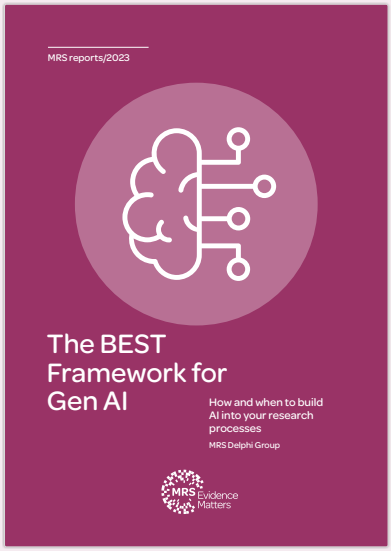
Collecting data using Biometric Data

MRS

Visit www.mrs.org.uk

This is the second instalment in a series of reports which explore the impact of generative artificial intelligence (AI) on the market research sector. The purpose of these reports is to demystify discussions about AI and set them within the context of the business of research.

In our perspectives from a range of industry experts you will find a snapshot of what the research sector is doing in this space at a practical level. The common thread throughout the series is the BEST framework, developed by the MRS Delphi Group. The framework is aimed to help practitioners make decisions when discussing which technologies and approaches to explore further. We know these discussions in relation to AI are happening everywhere at the moment - at all levels of the business, and across all sectors.



Download Part 1:
The BEST Framework for Gen AI

About MRS Delphi Group

The Group is led by a coalition of respected thinkers in the marketing and research sectors. The Group delivers **valuable insight** across a range of important business, social and political issues. The Steering Board includes: Colin Strong, Ipsos (chair), Rose Tomlins, Virgin Money; Gemma Proctor, Sparkler; Adrian Sanger, Behaviorally; Patrick Alcantara, AXA; Tatenda Musesengwa, Savanta; Zoe Ruffels, Imperial Brands; Dr Mark Thorpe, Truth Consulting; Jane Frost, CEO of MRS.

Acknowledgments

In addition to the contributions in the second half of this paper, a number of expert practitioners were consulted as part of this report. Particular thanks are due to Simon Raper, Reena Sooch, Brian Tarran, Phil Sutcliffe and Professor Alistair McGuire.


Boost
Expand
Shift
Transform
B
E
S
T

Introduction: How humans think

Colin Strong
Head of Behavioural Science
Ipsos

Chair
MRS Delphi Group

There is a rapidly growing interest in 'synthetic respondents', where LLMs (such as Chat GPT and or Google AI) use artificially created profiles to simulate the characteristics and responses of human survey participants. The opportunities of this are readily apparent – setting aside the considerable reduction in costs that come from conducting fieldwork with people, there can be a range of other benefits from offering 'access' to otherwise hard-to-reach populations and for the piloting questionnaires, to testing out products, services and solutions in speedy and agile way that would be hard to replicate with human participants.

But amidst the excitement that comes with any innovation, it is useful to consider the questions we should be asking when applying this. How confident can we be that this approach will deliver responses that we might expect to find by interviewing humans?

At first glance, as we shall see, the results look impressive, but perhaps equally as importantly, do the principles being deployed stack up in a credible way? If we can properly understand the behavioural mechanisms involved in a human responding to a survey, then we are in a much stronger position to determine where, when and how that might be something a synthetic respondent could do equally well, and when that would not be advisable.

Defining synthetic respondents/ participants (and synthetic data) is not straightforward1[1] – and there are different approaches to the technology: but we also want to consider the different questions we should be asking when considering the adoption of this. The report explores these from many different perspectives and by way of outlining some of the big questions we can be asking, we shall briefly discuss here.

What is the empirical case for using LLMs to ask questions?

Drawing on a paper by Michael Fell[2] the evidence to date seems to suggest LLMs are pretty effective at providing answers to questions that closely reflect those collected from humans. A variety of tools can be used to create simulated survey respondents with various characteristics such as demographic details, attitudes, and personality traits based on predefined variables. These characteristics are probabilistically assigned to each LLM agent to mimic a diverse survey population: this broad approach has been used to look at political questions such as vote prediction, psychological attributes such as personality traits, management studies, and economics , amongst others. Reviewing these, Lisa Argyle and colleagues found[3] strong associations between human and GPT-3-52 generated responses.

There certainly looks to be some promise and the field is developing rapidly. But is it the whole story? What should we need to think about when considering how to approach this?

Introduction: How humans think

Our position on something is always a constant work in progress, bringing together these different traces of information to cater for new questions that we are posed.

What happens when we ask a question?

When considering the psychological mechanisms at play when we are asked a question, it is easy to assume that we operate in a somewhat computer-like way, assuming that our knowledge of ourselves is structurally organised as a network of nodes linked together. This stored knowledge is then accessed when certain cues (in our case questions) are presented, leading us to retrieve it and report in the form of an answer.

But this is not the only model – other explanations suggest a more active and dynamic perspective on the way we handle knowledge. Psychologists such as Frederick Bartlett see knowledge as something we construct through integrating different traces to build coherent meaning. This means that our position on something is always a constant work in progress, bringing together these different traces of information to cater for new questions that we are posed.

As we shall see, these distinctions are important for evaluating how effectively a LLM can offer a human-like response. But first, what is the evidence to determine which of these mechanisms we use when we are asked a question?

The evidence for constructivism

In reality this is not really a binary either / or consideration – some aspects of how we handle knowledge can be quite 'computer-like' and at other times, more 'constructivist.' From the outside, however, it is easy to assume that a structured survey is entirely computer-like in nature. Indeed, the process of administering a survey to respondents *looks* very much like it operates in a computer-like way, with queries (questions in a survey) resulting in a data set (answers). And it is *easier* to think of qualitative research as more constructivist in nature as it involves more of a back-and-forth; a conversation that looks like humans making sense of things in words rather than 'data' being derived from a pre-scripted set of questionnaire items.

But simply because the survey *looks* this way, it does not mean that this is necessarily the case. Norbert Schwartz and his work on meta-cognition suggests that the act of retrieving information to answer questions can affect how people feel about their responses[4] the ease with which information can be retrieved from memory ('felt-fluency') affects how significant, or frequent, people judge the retrieved content to be. This affects self-perception and decision-making, as easier-to-recall instances might be weighted more heavily when reflecting on the way we answered the question (our meta-cognition at play) than those that are not as readily accessible.

Introduction: How humans think

So, for example, we may ask "How satisfied are you with the battery life of your mobile phone?" and then a follow-up "Can you recall a recent instance when your mobile phone's battery lasted longer or shorter than expected?" By asking respondents to recall specific instances (ease of retrieval), confidence in their general satisfaction rating might be influenced. If a user easily recalls a positive experience (eg, the battery lasting through a long day of use), their overall satisfaction rating could be enhanced due to the ease with which this positive memory came to mind.

Of course, this has implications for whether it is a human or machine responding, as while humans have the capacity for meta-cognition that can lead to new insights and changes in perspective, AI can at best produce responses that merely copy reflective thought, as it is not based on genuine self-awareness or meta-cognition.

It is this area that more exploration and understanding is needed about what happens during the act of being asked a question, surely reinvigorating the psychological work of survey theorists which has arguably faded into the background somewhat in recent times.

AI can at best produce responses that merely copy reflective thought, as it is not based on genuine self-awareness or meta-cognition.

Representativeness

Moving slightly away from the psychology of the question-answer, another important consideration for any research is representativeness – ensuring that the sample obtained reflects the wider population. How do LLMs fare in this respect? Angelina Wang and colleagues point out[5] that the ability of LLMs to replace human participants is of course wholly contingent on LLMs being able to represent the perspectives of different demographic identities.

In their paper, Wang and colleagues suggest that LLMs can struggle to represent marginalized groups accurately due to two key limitations. First, is what Wang calls 'Misportrayal': LLMs often fail to effectively represent the perspectives of different demographic groups. For example, when asked to simulate a person with a specific demographic identity the models tend to produce stereotypical responses that do not reflect the real diversity within that group.

Another limitation is what Wang calls 'Group Flattening': here LLMs tend to produce responses that homogenize the experiences and identities of diverse groups. This results in a loss of nuance and the erasure of subgroup heterogeneity (such as leading to a one-dimensional portrayal of complex identities by for example, ignoring intersectionality).

These limitations are due, the authors claim, to the way LLMs are typically trained on text data scraped from the internet, which rarely includes reliable indicators of the originators' demographic identity. Wang suggests that the training of LLMs often optimize for the most likely output rather than the most accurate or representative one, which tends to favour the most common views and expressions, further contributing to the problem of group flattening.

The authors therefore urge caution in using LLMs in settings where the representation of human identities is critical

Introduction: How humans think

and recommend using them primarily as *supplements* to human judgment rather than replacements, especially in sensitive applications.

To ask a question involves recognizing and respecting the otherness of others, their right to speak, and the validity of their perspectives.

What are the ethical considerations of assumed understanding?

The act of asking questions also of course has an important ethical dimension. To ask a question involves recognizing and respecting the otherness of others, their right to speak, and the validity of their perspectives. They are a way of 'yielding the floor' allowing others to express themselves and contribute to the shared process of meaning-making. They provide a means of challenging received-wisdom around societal norms. To not ask questions of people and not offer the means by which they can express their answers surely limits our access to these diverse perspectives.

In conclusion

We are at a point where there are a number of different proposals emerging for ways to generate 'synthetic respondents' – and the finding that at least some of these seem to offer a consistency with human generated responses is interesting and exciting.

But the empirical case for parity is not the only consideration – if we do not have a good understanding of the human processes that underpin the way a question is assimilated, and an answer given, then we are in danger of assuming synthetic respondents are always equivalent to human respondents. As we set out earlier, just because from a distance they can look the same, does not mean they are the same. The same principles that practitioners apply to any tool need to come to the fore here – when is it helpful, what are the limitations, which groups does it represent well, which does it fail to do so? These are the sorts of questions all researchers are very familiar with – quite what the answers are is emerging and we hope this report offers a helpful means to explore them.

Looking for the BEST fit

As already demonstrated in the 'AI in Action' case studies in our first report in this series, generative AI is being introduced at various stages in the research process. More often to automate repetitive, systematic and what are commonly perceived as 'dull' tasks that may require minimal human intervention.

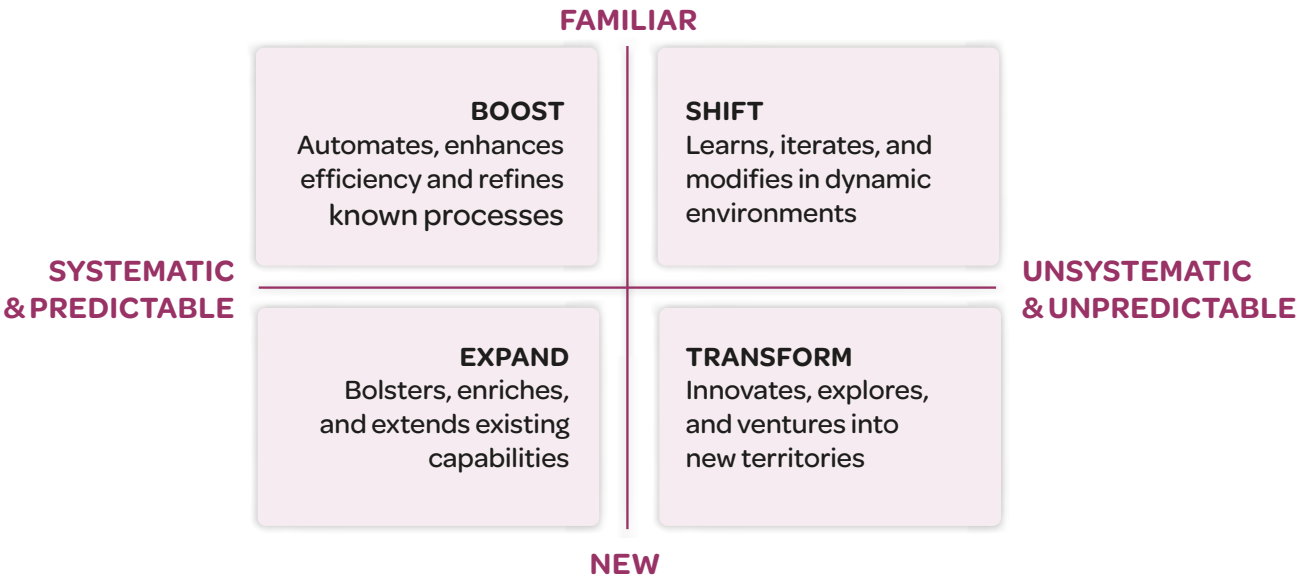
In this sense AI is taking on the role more as an assistant than a co-pilot. In the context of the BEST Framework these solutions sit firmly in the **Boost quadrant:** *Automates, enhances efficiency and refines known processes.*

There is also a clear role for AI as an initiator of ideas and a stimulant for creative thinking. This utilises a particular talent for generating infinite combinations of ideas (text or graphical) to act as counterpoints with which to stir our imaginations, helping us to envisage alternatives which may otherwise not be considered. The relationship here is moving closer to a partnership and moves us into the **Expand quadrant:** *Bolsters, enriches and extends existing capabilities.*

The use of synthetic data to augment existing real data is not new to researchers and data analytsts in a variety of sectors. But with easy and cost effective access to LLMs with natural language interfaces, researchers and other professionals have a new tool for delivering insight.

One way of viewing this is as another in a range of dashboards that sit on top of large data sets

The BEST Framework



Looking for the BEST fit

and responds to direct queries in such an intuitive way that no expertise is needed by the user to get at least a basic level of results. As such, this is a further democratisation of research into the DIY space (much as Google Analytics delivered data insight to marketing teams).

For some, synthetic data generation offers a panacea to the sector's woes of declining participant engagement and increasing data fraud. In 'traditional' data modelling unrepresentative segments can be augmented; and with LLMs questions can be asked to a machine repeatedly without survey fatigue.

Sometimes, often, according to some experts we spoke to - primary research is done unnecessarily, and some of the use cases referenced below take place in the early stages of the research process, thereby freeing up insight teams to do more of the critical thinking and analysis that comes next.

From a client perspective, the journey between asking a quick question and getting a response appears to have been shortened from weeks and days to minutes and seconds.

From a client perspective, the journey between asking a quick question and getting a response appears to have been shortened from weeks and days to minutes and seconds. And with that move from human participants to machine goes an inevitable reduction in costs, in theory at least.

So if the sector cleans up some quality and reputational problems, and clients get quicker and cheaper results, what's not to like?

Scenarios for using synthetic data in market res earch

Terms like synthetic data and synthetic participants are sometimes used interchangeably. However, we may be talking about a number of scenarios that use a combination of AI and machine learning applications. For this report, we are focusing on two interpretations of these terms which are generating a lot of discussion within the sector, especially as new solutions and services come to market.

In the first scenario, synthetic data is created whicht shares all the characteristics of real world data, but has a much lower risk of identification than anonymised data, and which can be used as a proxy for real data.

In the second scenario, a Large Language Model (eg, ChatGPT-4, Gemini, Llama et al) is used to simulate a participant or segment and derive insight through a series of prompts.

> "It's more about us being able to fill the gaps in insight about the customers we want to serve... It's not something that's completely alien to us, we do it all the time as market researchers. It's an extension of things that we do."
>
> **Patrick Alcantara**
> **Head of Insight at AXA**
> **Speaking at MRS Annual Conference 2024**

Scenario one
Generating synthetic data via machine learnin

Synthetic data has been generated artificially but is based upon real world data; it reflects the same statistical patterns and properties of 'real' data. This is distinct from depersonalized or anonymised data which is still real data with certain personal identifiable information having been masked or removed.

Synthetic data is used in conjunction with differential privacy in many data modelling scenarios, and real use cases can be found in a wide range of industries from pharmaceutical development, to fintech product testing to crime hotspot prediction.

The various ways in which synthetic data is used was neatly summarized by Ray Poynter[6]. For example, to enhance a dataset by mitigating any weaknesses or gaps in the original data set. As such, whether we call it augmented, fake or synthetic data, it can be defined as any data that is not primary data.

"It's more about us being able to fill the gaps in insight about the customers we want to serve," said Patrick Alcantara, Head of Insight at AXA speaking at MRS Annual Conference 2024: "It's not something that's completely alien to us, we do it all the time as market researchers. It's an extension of things that we do." [7]

Synthetic data are highly malleable, which is their

Scenarios for using synthetic data in market res earch

strength especially when used as a corrective to biased or unbalanced data sets. Take an example where fewer women than men have responded to a survey; just multiplying the data from women to augment the set will only create duplicate data, addressing basic balancing issues only. Whereas new synthetic data can be generated to represent women with variable characteristics thereby creating a richer corpus of data to interrogate.

"Real world data is rarely problem-free," said IBM's Inkit Padhi[8]. "Synthetic data allow us to find and fix problems in AI models to make them more fair, robust, and transferrable to other tasks."

Gartner anticipates an exponential increase in the volume of synthetic data used to train AI - from 1% in 2021 to 60% by the end of 2024[9].

Scenario two
Using LLMs to simulate participants and generate insight

This second scenario is clearly distinct from the first and it is important to acknowledge that distinction.

There's a risk currently that we conflate established use of synthetic data with the emerging concept of synthetic participants - thereby assuming that the benefits of synthetic data are also shared by synthetic participants. In the case of privacy or bias this is not the case as an LLM trained on public internet data poses a set of risks around privacy and bias that can be more easily avoided in synthetic data generation. So, whether intentional or not, the concept of 'synthetic participants' may be unfairly benefiting from a halo effect of being mistaken for 'traditional' synthetic data.

The proximity to primary data here is also a critical point that needs to be understood when considering this scenario. What an LLM tells us is a synthesis of data from a range of sources which may be hard to trace and qualify. The relationship to primary research may be getting thinner, the outputs more flattened, generic, and potentially more biased.

In October 2023, Mark Ritson opened up the debate with a characteristic provocation in Marketing Week when he demonstrated how ChatGPT could create in seconds a brand perception map using internet data. Ritson went on to write that "Most of the AI-derived consumer data, when triangulated, is coming in around 90% similar to data generated from primary human sources."[10]

While Ritson's example uses an LLM that has been trained on internet data – with all the risks and limitations outlined in the next section - there are also opportunities for companies to train their own LLM on proprietary organisational data. This is regarded by some as a necessary step in order to ensure data transparency and competitive advantage.

What can LLM synthetic participants offer?

The starting point for most advocacy around synthetic participants is the speed at which insight can be delivered into the hands of decision makers. As businesses move faster, so the argument goes, they need to make decisions faster. Waiting weeks for a primary research project may not be an option, and the cost implications of cybersecurity and data quality assurance are only set to increase. Enter the LLM as an 'always on' conduit to in-the-moment insight.

Prompting an LLM to act as a customer persona could be useful in early stage investigations like concept testing and hypothesis generation, or an exercise like Ritson's automotive brand perception map. The high degree of fidelity between human-annotated, real-world data and the LLM results is not necessarily surprising; these are simple research exercises with no need for nuance, and ultimately quite predictable results.

The internet provides a rich brew of structured and unstructured data - behavioural and reported, transactional and social. The sheer volume of data on which a model is trained is a point of difference between commercial models being developed by the tech companies, with data cut off times virtually non existent for paid (subscription) versions. This makes LLMs powerful social listening tools which could allow companies to hear about the un-met needs of customers in a way that reported responses do not.

Rose Tomlins, Head of Brand and Customer Insight at Virgin Money, is an advocate of using LLMs to aid decision making for two reasons – to bring the customer in from the start, and to maximise her budget: "We've started to look at opportunities where we haven't got the customer present in decision-making, but we'd love them to be, and we'd love to involve them earlier.

"By interrogating synthetic data early on, you could decide whether to progress before you get to the research planning stage and sharpen your investment".[2]

Conducting primary research may be unnecessary at these early stages in the insight process. Perhaps today's restricted budgets should be reserved for later more complex stages in the research process, where primary research is needed to validate that results are representative and to drill down to the insight that comes from deeper more nuanced qualitative research.

"By interrogating synthetic data early on, you could decide whether to progress before you get to the research planning stage and sharpen your investment".

Rose Tomlins
Head of Brand and Customer Insight
Virgin Money

What are the limitations of LLMs in a research context?

This list is not exhaustive and will change as use cases develop and new technological solutions emerge. Here are some foundational weaknesses at this point in time.

Privacy
Without the safeguarding that we saw in the creation of synthetic data in Scenario one, a lack of transparency in the source of data and how the model's algorithms are trained raises real risks of identifiable personal data being leaked via LLMs.

Bias in data and the model
Trained on real data (often from unknown sources especially when using free solutions), models can perpetuate existing biases and also amplify or distort them in the form of hallucinations.

Hallucinations
LLMs are designed to give an answer and can identify patterns that are not there – in a recent study of legal materials[11] analysts estimated that LLMs hallucinate 75% of the time resulting in factual errors in their responses.

LLMs are not thinking
Generative AI is still not thinking in the human sense, and yet it mimics human thought which gives the user a false sense of confidence in answers.

Being 'confidently wrong'
As well as being factually incorrect being confidently wrong could mask opportunities and ideas that the LLM hasn't offered – what are the unknowns?

Competitive edge
More relevant to LLMs running on public data - how is an insight going to differentiate your business decisions if competitors are using the same data?

Data drift
How is data updated to train the model and provide insight to reflect the real time, real world context?

Research credibility
Lack of data transparency can result in research being more easily dismissed – for example, a suspicion that this wasn't 'our' audience that the researchers got it wrong.

Not deep and qualitative enough
LLMs are good at averaging, so deliver few surprises and level results into generic information.

Data integrity and walled gardens

The potential for a broader more systematic use of synthetic data in the market research process, given all the above caveats, seems quite clear. As mentioned, it is already quite common in certain areas of healthcare research and technology development, and use cases are plentiful.

The LLM synthetic participant scenario currently creates more division. It brings into sharper relief the limitations and risks outlined above as the process is potentially more opaque – whether it's the provenance and quality of the training data, or the parameters that guide the model and dictate the outputs or 'insights'.

Former Britannica editor-in-chief Robert McHenry is reported to have compared Wikipedia to a public toilet - you never know who touched it last[12]. More recently, these suspicions of unreliability have been echoed by Professor Ethan Mollick of Wharton who described ChatGPT as an "omniscient, eager-to-please intern who sometimes lies to you."[13]

The depth of information that a model has been trained on is important. It is this depth that enables an LLM to give sophisticated, granular and conversational (human-like) responses.

Natural Language Processing (NLP) generates text based on the statistical relationships between letters, rather than the meaning of them. LLMs do not understand the text they are generating, and meaning is inferred by the reader. As we have seen, being confidently wrong and producing hallucinations is a well-documented characteristic of LLMs.

"Former Britannica editor-in-chief Robert McHenry is reported to have compared Wikipedia to a public toilet - you never know who touched it last."

An ideal scenario that many companies want to establish is running a customised LLM on a walled off corpus of proprietary data that can be sourced, validated, and quality assured. Where necessary this data can be augmented with synthetic data to address bias and data gaps and different levels of training data which may have be hard to source can be integrated or not. However, this throws into question how much data an LLM needs to be trained on and there is a risk of 'overfitting' which happens when there is a lack of fresh data and the generative capabilities of an LLM compensate by hallucinating..

Returning to the BEST Framework that underpins this series of reports, there is a dichotomy emerging if we want to move the use of LLMs into the unsystematic and more dynamic context of enabling decision making at faster pace – the Shift and Transform quadrants. The example used of a public health emergency in the previous report requires the ability to keep our model trained with fresh unstructured data at scale. Is this possible within a walled garden? A model is only as good as the data it is trained on, and in such rapidly evolving and unknown scenarios will need to be tapped into the internet, with all the risks that entails.

Researchers as data guardians

Judging by Marc Ritson's article the marketing community may be comfortable with these risks when balanced against the speed and cost of traditional research. Results appear to be 'good enough' on which to make operational short-term decisions.

The insight community can see this as another opportunity like previous tech advances to step decisively into the role to guide and provide guardrails for use.

Advocates may argue that the stages of developing hypotheses and ideas are low risk because decisions are not long term and can be validated further downstream. The risk benefit trade off sees value in fast cheap research, and allows for rigor and ethical considerations around data quality and representativeness to be downplayed.

But can any decisions be harmless, however operational and short term? Every decision is a fork in the road, and consequences may not be realised until the process is so far advanced that it is impossible to distinguish when the first wrong turn was taken.

The insight community can see this as another opportunity like previous tech advances to step decisively into the role to guide and provide guardrails for use. There is a stronger role for agencies and client side teams to adopt across the business in providing professional support - highlighting potential pitfalls and recommending good practice.

Should the sector fear a further decentralisation of market research outside the insight function? Once past peak hype, marketers may well realise this is further eating into their time and devolve responsibility outside the department. The insight function needs to ensure it is acknowledged as the guardians of data and insight, regardless of the technology that delivers it.

Conclusion

True insight comes from unexpected places, from the edge cases, and outlier data. This was a view that recurred across the conversations undertaken for this report.

We know insight when we see it - it makes itself known. As the market research sector experiments with generative AI we will as professionals recognise true insight when we see it, as validated by new case uses.

But others may not, hence the need for continued guardianship of the insight process, regardless of the tools. Representation and quality sampling are cornerstones of the industry. Some argue that listening to real people is the immutable moral duty of the researcher.

The allure of fast and cheap research cannot be underestimated and we have seen that it has its place in the early stages of research process. To be useful in a fast moving and unfamiliar scenarios requires a degree of sophistication not yet widely visible, but this doesn't mean it won't be reached soon.

In such urgent circumstances some people may feel that the usefulness of the solution outweighs the risk of disclosure. Perhaps of more concern is that this mindset may already be setting in, and that these are further incremental steps away from evidence-based decision making.

A lot is at stake here. Legal and ethical use of personal data is just as imperative as ever, says Debrah Harding, Managing Director at MRS. "If we lose that part of the equation, we're ruining a lot more than synthetic data collection. We're ruining research in general."[2]

In previous decades the research sector was sometimes characterised as being slow and cautious in its response to new and significant changes in technology. Now some are saying that the sector is too quick to jump on new tech and build solutions without checking that the evidence base is there.

Maybe there's an alternative approach that the sector should consider - rather than ask when to use synthetic data, the research sector needs to clarify the role of primary research data. With many more suppliers of data and research technology entering the market, there is some concern that the distinction between primary research and synthetic data is being lost, and in some instances willfully obscured.

"In order to be able to create synthetic data you still need really, really good human data," says Harding. "I don't see synthetic data replacing any of those techniques. I think the mix will be different and the way the research is undertaken will be different. But the need for people and the need to continue to protect participants will continue to be an important part of research going forward."

Asking questions: explainability and transparency

For AXA's Patrick Alcantara, an agency will stand out as a preferred supplier if they are upfront about their methodology[11].

That means for example, explaining how they audit their algorithms to make sure they don't carry bias or prejudice, and how they can prove that they train their model with diverse data sets. In understanding the generic steps necessary to generate synthetic data or train an LLM you will be able to ask questions of your data and technology supplier or in-house team.

As a start, here are generic approaches to both scenarios. Note that the preparation of the training data is critical as to the quality of the results.

The creation of synthetic data via machine learning:

1. A data generator is given parameters and values to work within.
2. Trained on real data using these parameters, the generator produces a synthetic data set.
3. The data set is tested for quality and privacy – how closely it resembles a real data set, and disclosure risk is assessed.
4. Initial parameters are updated based on step 3, until the criteria for quality and privacy is met.

Training an LLM:

1. A large set of prepared data (text in this scenario) is given to the model.
2. The model uses an optimization algorithm to adjust its parameters to minimize the difference between its predictions and the actual outputs.
3. The training data is given to the model in small batches.
4. The model makes predictions for each batch and changes its parameters based on the errors it sees.
5. This process is repeated several times, allowing the model to gradually learn the relationships and patterns in the data.

Some questions to ask providers:

- What choices have been made around the solution architecture and how will they impact the results? (In the case of LLMs there are a number of deep learning frameworks and transformers to chose from - see Further Reading)
- What data is used for ongoing training of the model and can a buyer opt out of certain data sets?
- What privacy controls and safeguards have been put in place to minimise the disclosure risk both as regards to the training data but also the model that is being trained?
- What validation has been done to compare synthetic data outputs to human-validated primary research outputs?
- What modelling techniques have been used to create a representative sample for your needs?
- Is there criteria for data recency to be able to assess if the data is appropriate for use?
- How does the supplier flag and distinguish, in their systems and approach, between data derived directly from natural persons, and data which is derived synthetically?

Esomar has produced a helpful checklist of questions to help buyers of AI.[14]

AI in action
Perspectives from industry experts

Synthetic data: the good, the bad and the legal

Debrah Harding
Managing Director
MRS

Synthetic data is becoming a new battleground within the research sector. For some, synthetic data is the new panacea to all the sector's woes. It will eliminate data privacy concerns! It will resolve sample data quality problems! It will be cheaper, faster and more representative than primary data! For others, the potential misuse of synthetic data could result in unreliable data, bias, lack of disclosure, sloppy analysis and over-reliance on models. So, what is the reality?

Synthetic data has been around for a long time. As an application of technology, it has a number of potential benefits, but it is not a replacement for real data from real people, nor is it without its risks.

One of the common misconceptions is that synthetic data is no longer personal data and therefore privacy legislation, such as the General Data Protection Regulation (GDPR), no longer applies. This is not the case. Some synthetic data models can operate in a similar way to Large Language Models (LLMs), except with participants' personal data becoming the training data from which subsequent data models are derived. To create robust synthetic data models, access and use of personal data is needed, and the consent to use personal data for this purpose is needed from those whose data is being used.

Model generalisation (the ability of a model to learn and predict the pattern of unseen data enabling models to synthesise data outside of an original data set) is one way in which some of the privacy concerns could be addressed, including enabling the use of synthetic data beyond single use cases. However, ensuring that synthetic data remains useful for analysis and decision making whilst maintaining the necessary levels of privacy over time is a "complex challenge" according to the UK's data protection regulator, the ICO.[15]

Nor can it be assumed that synthetic data being drawn from sources such as LLMs is always correct or reliable, as other professions, such as the legal sector, has found at its cost. According to a recent study by Stanford RegLab, hallucinations within LLMs were "pervasive" when undertaking legal tasks. The hallucination rates ranged from 69% to 88% in response to specific legal queries from the LLMs.[16] This problem was exemplified by the "ChatGPT lawyer" who used an LLM to prepare a court filing, which was, in the words of Chief Justice John Roberts, full of "bogus judicial decision…bogus quotes and bogus internal citations" resulting in a New York Times story which went viral.[17] In order for synthetic data to be valuable for research, without bogus data, hallucinations need to be managed.

There also remain significant questions about the data that has been used to train and populate LLMs, which are then the basis for some synthetic models, and whether the data should have been used in the first place. Copyright is a complex area, so much so that the UK's Intellectual Property Office has parked its plan to create a Code setting out the rules for the training of AI models using copyrighted materials, instead recommending an industry agreement.[18]

For now the issue is being played out in the courts with the New York Times case ongoing against OpenAI and Microsoft regarding the alleged use of "millions" of copyrighted articles to develop tools such as ChatGPT and Bing.[19]

Synthetic data: the good, the bad and the legal

Whilst this case has yet to be decided, there are other cases which have been, with the Chinese courts being the first to reach a decision regarding the liability of AI generation companies for copyright infringement. The decision in February 2024 related to a case involving AI-generated images.

The Court's findings state that a similarity between the output of a generative AI system and pre-existing works can be an infringement of the rights of pre-existing works, and operators of websites offering an AI system can be liable for the outputs. This decision is significant, and a potential warning shot for what might come in the future.[20] It also is an issue that needs to be resolved before synthetic data models using LLMs can be created at scale.

The research sector needs to adopt synthetic data in a measured, legal, ethical and privacy-first way to ensure that the sector harnesses the benefits whilst mitigating the risks.

Looking forward and looming large is the impact of the new EU AI Act. The Act has been introduced with the aim of building trust in human-centric AI, in the words of the European Commission, "... AI is not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being."[21]

The wording of the legislation has some ambiguity, and how it might apply to synthetic data is not altogether clear. More guidance and ultimately enforcement will shape how synthetic data fares under this legislation.[22]

The research sector needs to adopt synthetic data in a measured, legal, ethical and privacy-first way to ensure that the sector harnesses the benefits whilst mitigating the risks. As a sector, research relies upon the confidence of clients, users and participants in the value of what we deliver. To continue to retain that confidence, one thing is for sure, we don't want the research sector to become the next viral sensation.

Financial services: Using new technology to build better customer outcomes

Patrick Alcantara MA CMRS
Strategic Customer Insight Lead
AXA UK & Ireland

Financial regulators are increasingly, and rightly, challenging firms to demonstrate how they build good customer outcomes. Landmark regulation in the UK in the form of a comprehensive Consumer Duty now places a higher standard of consumer protection on financial services providers[23].

This new duty expects firms to take active steps in meeting customer objectives and avoiding foreseeable harm[24]. Firms should demonstrate how they meet customer outcomes in terms of providing accessible support, clear and timely communications, fit for purpose products and value for money.

Insights from market research are instrumental in getting the customer perspective required to create these outcomes. Many firms, however, have insight gaps, often exacerbated by various constraints. Generative AI and synthetic data provide an opportunity to work around these constraints and bring that perspective essential to building good outcomes.

Financial regulators, for one, emphasise the fair treatment of vulnerable customers who are more susceptible to harm[25]. Ensuring this fair treatment may require approaching customer outcomes differently. As vulnerability may be caused by poor health, adverse life events or low financial resilience, this means that research on vulnerable customers should be approached with additional care. Generative AI and synthetic data could assist in this instance by creating personas of vulnerable individuals, helping researchers understand and empathise with their challenges before fieldwork.

Hard to reach populations can also present a challenge in providing a customer perspective. Research can be quite expensive and time-consuming owing to the difficulties in recruiting participants. A similar approach of creating personas upfront in the research process might help provide initial hypotheses and a working understanding. This might help reduce research costs and prioritise budget on the greatest insights gaps.

Designing customer experiences could also benefit from generative AI and synthetic data. Firms often

Generative AI and synthetic data could assist in this instance by creating personas of vulnerable individuals, helping researchers understand and empathise with their challenges before fieldwork.

Reflections from financial services: Using new technology to build better customer outcomes

design experiences from past demand and customer feedback. Customer experience around car insurance claims, for example, is shaped by various factors including past claims data, supplier capability (i.e. roadside breakdown assistance, courtesy cars, vehicle repair) and historical turnaround time. Generative AI and synthetic data can complement existing ways of anticipating demand, enabling insurers to shape better experience during high demand or supply chain challenges.

It is apparent from these examples that generative AI and synthetic data could be used to extend insights and build better customer outcomes off the back of these insights. However, it has limitations like any other tool and should be deployed as part of a holistic, comprehensive market research toolkit. Generative AI and synthetic data also depend on actual customer data to work. Market researchers therefore need to engage in broader conversations on data quality, ethics and methodology to influence better research and customer outcomes. Market researchers should also constructively challenge agencies to be transparent on their AI use. References like the 'MRS guidance on using AI and related technologies'[26] and the BEST Framework by the MRS Delphi Group are good starting points in approaching those conversations.

As regulators increasingly challenge firms to build and demonstrate better customer outcomes, market researchers will have a crucial role to play moving forward.

Market researchers have an incredible capacity for bringing customers to life which helps organisations create better experiences. As regulators increasingly challenge firms to build and demonstrate better customer outcomes, market researchers will have a crucial role to play moving forward. It is our responsibility to use the tools at our disposal to improve the understanding of the customers we serve, helping them achieve their goals.

The possibilities of AI for enhancing insight

Andrew Cooper
CEO & Founder
Verve

While the challenges identified with using synthetic data and off-the-shelf LLMs in a research context are very real, there also are viable - and exciting – alternatives that point to the potential for generative AI to genuinely add serious value. However, much of the debate so far appears to have focused on the concerns and threats rather than the opportunities.

At Verve, we don't like to use the term 'synthetic data' at all - we prefer to call our models 'Intelligent Personas'. Why? Because these simulation models powered by LLMs can deliver so much *more* than synthetic data.

Introducing Intelligent Personas

Over the last 2- 3 years, Verve has successfully applied a combination of AI approaches with clients. Now, with Intelligent Personas, our goal is to harness the power and adaptability of LLMs for insight purposes, but without compromising the integrity and robustness of the insight itself.

Excitingly, these simulations have demonstrated the ability to replicate what real people would say if we asked them, with verifiable accuracy and uncanny statistical confidence.

We have many proof points on that score. Using holdout data or parallel background testing (including a huge multi-market MaxDiff product evaluation as one example) as part of the validation process, has worked very well – with the models producing correlations against real-world human quantitative survey responses of over 0.9 and R2 scores of the same magnitude. Tests against parallel qualitative interviewing have also aligned not merely, convincingly, but precisely.

Naturally, this has been reassuring for clients, and our approach has already demonstrated significant success with global blue-chip clients across FMCG, finance, lifestyle and travel categories. Notably, it's also been deployed across diverse workstreams including audience understanding, innovation and comms and concept testing – the latter of which includes the evaluation of confidential materials.

Intelligent Personas are also changing the game for B2B and other hard to reach audiences, where they provide a solution for insight with expensive and time poor specialist roles, while avoiding the challenges of fraud and respondent quality increasingly associated with traditional panel approaches.

It's all about the data

Imagine a Venn diagram between your LLM model and the special data that you've used to build it. Now imagine that half of your Venn diagram is based on data infested with all the respondent quality issues that we know about. As ever, it's garbage in/garbage out; your model will only be as useful and as accurate as your original data set.

Our theory is that the reason our persona models have worked so well is driven partly by the symbiosis between these models and community panels. Community panels provide a high-quality data source, uncontaminated by respondent quality issues, based as they

The possibilities of AI for enhancing insight

are on transparency and the authentic responses of known, verified client customers. And they provide a very large amount of data indeed.

So it's this combination – of known, high-quality training data, plus robust validation against real research outputs - which makes the richly detailed Intelligent Personas materially different and a counterpoint to other 'synthetic populations' or, the use of generic LLMs for insight.

We've identified 30 or so client use cases so far and expect to uncover many more as they continue to evolve. Put simply, they represent an incredibly versatile and robust option to deliver a range of research tasks, while also, of course, delivering additional overarching benefits of pace and cost efficiency.

Moving forward

Naturally, we do not advocate Intelligent Personas as replacement for high-quality primary research – indeed the sweet spot is their use alongside human insight and cultural intelligence. Equally, clients are rightly wary of betting the farm on the say-so of a simulation.

Similarly, there is a need to stay vigilant through continuous monitoring and evaluation frameworks to regularly assess the accuracy of their outputs, in addition to a bias mitigation strategy which includes diversifying, expanding and cross-checking data for accuracy.

However, as the successes to date show, Intelligent Personas are a reliable option to streamline, augment and enhance research, especially fairly basic research.

If the participation and data quality crisis continues, who is to say that AI responses will be less useful than your primary data?

This also enables clients to save the real people (and on community panels, yes, they *are* real people!) for the most important work. We believe this will help to create a more respectful and sustainable approach to insight, which is critical if we are to address the bigger data quality challenges across the industry.

To summarise: as the wider Delphi report notes, the ability for AI to generate seemingly compelling responses - as if it were representative of a customer segment or group - is both exciting and potentially dangerous. Used in the wrong way, AI is undoubtedly a very unpredictable tool for decision making.

However, used in the right way, with the right data, and coupled with human intelligence and cultural intelligence, it is mightily powerful. And here's a thought – if the participation and data quality crisis continues, who is to say that AI responses will be less useful than your primary data? In some cases we suspect they will actually be better...

That would be not only rather interesting, but genuinely revolutionary within the world of insight.

Navigating the synthetic data frontier

Louisa Livingston
Director
Forward Thought

In an era where data is considered a valuable currency, the burgeoning field of synthetic data is sparking both excitement and apprehension across various industries, notably in market research. This innovative approach, which uses artificial intelligence to generate data that mimics real-world responses could be a game changer.

At its core, synthetic data is artificially generated information that maintains the statistical properties of real-world data without compromising individual privacy. This characteristic alone makes it an attractive option for industries ranging from healthcare to finance. The appeal of synthetic data lies in its ability to simulate consumer behaviours, preferences, and trends at a fraction of the cost and time required for traditional research methods and potentially even deeper by delving into the vast information driving large language models.

There are significant drawbacks, however; the data is based on existing knowledge and algorithms and lacks personal experiences and authentic emotions which are crucial for understanding consumer behaviour deeply. Results are also inherently generalised no matter how specific you ask the responses to be. You also need to be VERY clear about who you want it to be – an individual with personalised responses or a group of individuals with a collective response?

Is your favoured responder a conformist or a flexible thinker? What age, region and culture are their views drawing from? Do you personally know enough about the type of responses you're expecting to know if it's accurate or whether you'll fall victim to what I'm calling the 'Labrador response' - it desperately wants to help you and might just make up the answer to make you happy!

The best use cases are ones where simulated consumer perspectives at scale can help enormously but are perhaps not feasible due to time or cost constraints. Here are some examples of ways it has been used in our client projects:

- **Instant 'consumer' feedback** for example we have used generative AI during a design thinking workshop where rapid prototyping of ideas and immediate feedback is essential to iterate concepts.
- **Early concept testing** to gauge appeal, relevance, and any concerns before investing in more detailed consumer research.
- **Supplementing quantitative data** – where additional qualitative insights would help explain consumer behaviours, attitudes, or motivations.
- **Addressing sensitive topics** – looking at topics where privacy concerns or emotional distress could inhibit participant honesty or willingness to take part.
- **Global accessibility** – accessing trends, scenarios, consumer narratives or explanations which ordinarily would be difficult to uncover.

Where there are constraints, synthetic data can be incredibly powerful in helping businesses fuel their decision-making with information. Too often businesses make decisions based on lengthy, costly research projects which take months to bring to the table or one or two people make the decisions off the cuff based on their more limited

As we stand on the cusp of this synthetic data revolution, it's an opportune moment for the market research community to shape its trajectory.

and personalised experiences. Synthetic data can be a support system that should be utilised with care and thought, knowing its limitations and using the information accordingly.

An example: representing the youth market

Trying to understand the youth market is always tricky and agencies spend huge amounts on trend-watching and cool spotting around the world. You can't just ask ChatGPT to answer in the style of a '15 year old' girl and get anything other than a generic answer:

LO

Take on the persona of a young girl and tell me whether a new sustainable fashion brand would work?

First off, it's like, really important to look after our planet, right? If a brand makes clothes that help do that, I'd want to wear them! It's like, being part of a superhero team that's saving the world, one outfit at a time

This is cute but not very useful... let's get more descriptive:

LO

Take on the persona of a 15 yr old girl living in Manchester with a mixed-race background (Pakistan) who wants to be a fashion designer in London one day and has plans to go to art college. She is flexible in her views and has a middle-class family with disposable income to buy her mid-range clothing. How would she respond to a new sustainable fashion brand aimed at her?

Here's the thing, the clothes need to be original and reflect my individual style. I'm all for being eco-friendly, but I also want to stand out, not blend in. The brand needs to offer something unique, something that makes me feel like I'm wearing art, not just clothes. And if they're pioneering new sustainable practices, even better. I'd love to support innovations that could change the fashion industry...

As the market research industry grapples with the implications of synthetic data, it's clear that a cautious yet open-minded approach is warranted. The potential for synthetic data to revolutionise idea generation and exploratory research is immense, but its application requires careful consideration to mitigate risks of bias and error. Results should be calibrated against real data where possible.

As we stand on the cusp of this synthetic data revolution, it's an opportune moment for the market research community to shape its trajectory. We should be building the tools that can help businesses rather than watching it happen around us, but equally being honest and open about where and how it is being used.

Synthetic data bridges the gap between market research and big data

Steve Phillips
Founder and CEO
Zappi

Much of the conversation around the value of synthetic data focuses primarily on cost savings. And it certainly does cost less when you can get consumer feedback without involving real people.

But I think if we're only thinking about synthetic data as a means to save money, we're missing the real benefit: synthetic data can allow us to build out a more complete data asset.

Consider the data that CMOs rely on most often today. They tend to use sources that generate big data like clickstream data, sales data, social media data, etc. There are a lot of individual data points in these sources, but the data is quite shallow. It's largely behavioural data that tells you *how* people are acting and *what* they're doing, but not *why* they're doing it. The data is very thin.

Consumer insights data, in contrast, is much richer. It tells you "the why" behind what you see in the behavioural data. But it's smaller data. It doesn't have the volume of that behavioural data because it comes from a much smaller group of consumers and is more expensive to generate.

And as a result, consumer insights data never shows up in a company's customer data platform or data management platform. Those technologies were not designed to handle this kind of small data. And so unfortunately, market research is considered a less important source of data for many CMOs.

What excites me about synthetic data is not simply its ability to generate consumer insights data more cost effectively, but rather its ability to help bridge that gap into big data. Synthetic data provides a means to *supplement* all the data from real consumers to add more volume – faster and more affordably – to allow consumer insights data to integrate more easily into big data systems.

When consumer data can be integrated into the systems that CMOs are using to understand the world and make decisions, then insights teams can have much more impact and companies can become much more consumer centric.

When consumer data can be integrated into the systems that CMOs are using to understand the world and make decisions, then insights teams can have much more impact and companies can become much more consumer centric.

Synthetic data bridges the gap between market research and big data

There are a few different kinds of synthetic data that can be used to bridge this gap:

The first is *macro synthetic data* – predicting (or creating) data that represents what an audience thinks about something. This is a prediction model for overall stimulus reaction. For example, how a certain ad would score in brand recall or an innovation concept would score in trial potential.

The other way is *micro synthetic data* – predicting (or creating) data about how a real or synthetic respondent would think about something. This is a prediction model for a specific person (real or created) and their reaction to a specific question in a survey.

Synthetic data will never be a replacement for real consumer data. This is critical, because AI is only as smart and up-to-date as the data it's trained on.

Over time we can also look to predict response data for individual respondents on specific questions. For example, if we have a 20-question survey, we may find that we only need to ask 10 questions of some people to be able to very accurately predict the responses for the other 10 measures. This would reduce survey length, reducing costs and decreasing respondent fatigue (which increases data quality).

And of course, it's worth mentioning that in this world of synthetic data, responses from real consumers can never go away. Synthetic data will never be a replacement for real consumer data. This is critical, because AI is only as smart and up-to-date as the data it's trained on. If AI is only being trained on AI-generated responses, or only on data from a few years ago, there's a risk that it will become more and more disconnected from the thoughts and feelings of current real consumers. That's why I view it as a way to supplement real-time consumer data flows.

Ultimately, I believe the worst way of thinking about synthetic data is it's a cheaper way of getting a respondent. It's not about cutting costs but about improving data flows in a data centric world. This is an expansion opportunity for consumer insights. It is a way to get our data into a new world to have an even greater impact on business than we do now.

How to test and assess synthetic data

Ray Poynter
Founder
The Future Place

One of the key questions about using synthetic data is how do we assess whether it works? Or to be more precise, when does it work, when does it not work, and what do we mean by 'works'?

The limitations of case studies

At the moment, most studies that show that synthetic data works or does not work are of interest but of limited utility. There are no agreed protocols for creating synthetic data, so a study showing that a particular approach did not work for a particular study only tells us that this method, as applied by this team, for this project, at this time did not work. Similarly, studies that show examples of synthetic data working do not show that this approach would work in another context (e.g. another country, topic or time) or even that it would work next time.

If we broadly divide the current interest in synthetic data into augmenting primary data collection (for example boosting under sampled groups) and creating synthetic participants from, say, LLMs, then the approaches to evaluation have key differences.

Testing augmented synthetic data

Augmenting primary data, to deal with missing cases or too few cases, has a long history and the AI approach is to add generative AI to the existing body of techniques. Generally, these approaches utilise explicitly defined algorithms, which makes testing easier.

One standard way of testing these approaches is to create test cases where data has been removed and the algorithm is then deployed to re-create the missing data. Note, the aim of the synthetic data is not to replicate the individuals that have been held out (that would require magic). The synthetic data should produce a good approximation of the distribution and marginal totals of the missing data.

Where possible, the use of augmented synthetic data should be evaluated against outcomes. Compared with weighting data, synthetic data facilitates more granularity, which can mean its predictions are less lumpy than an under-sampled cell upweighted.

Whilst it is possible that augmenting data will work better for some fields than others, or for some groups, cultures, languages etc than others, there is no reason to assume it will work differently. This should be tested, but data collected face-to-face via paper-and-pencil approaches might be equally suitable for this augmentation.

Testing free-standing synthetic data

By free-standing synthetic data I am referring to the creation of synthetic participants who can then be deployed to answer tasks such as qualitative questions or quantitative surveys. At present, there is no 'theory' that underpins this approach, and it is entirely inductive. Not only is there no theory, but there are also no recognised approaches.

How to test and assess synthetic data

People who want to show it works might pick one method, with a market and test that are well suited. People who want to show it does not work, can pick another method, do it less well, and investigate a market where it is less likely to deliver credible results.

A good option is for clients to give access to recent projects and ask the vendor to replicate the findings.

If we consider the quantitative use of synthetic participants, and if we look at the testing issue from the client's perspective, then a good option is for clients to give access to recent projects and ask the vendor to replicate the findings. This would indicate to the client how well the vendor's approach works in their market for their sort of projects.

When considering the qualitative uses of synthetic participants, the picture becomes even more complicated. It is hard to 'prove' that a conventional qualitive project worked. In looking at the qualitative case we need to consider whether we want to compare the 'data' (for example transcripts from in-depth interviews with synthetic participants), or the results of qualitative analysis. And, if we are comparing the results of qualitative analysis, do we want to compare analysis conducted by humans with automated analysis?

Looking at qualitative uses of synthetic data, again from a client -perspective, I suggest sharing recent projects with prospective vendors and asking them to show what they would have produced. The client can then assess the perceived usefulness of the synthetic approach.

Note, in both the quant and qual cases, showing that an approach works for the context of a client at the moment does not mean it will work in the future. The generative AI platforms are continually evolving and their ability to mimic/ predict responses in a given market could get worse, get better or stay the same.

And, what does 'work' mean?

In essence, what most clients are looking for is that the results are good enough to justify the quality trade-off they are making to get faster and cheaper results. Whilst comparing results with other primary data will be part of most testing approaches, testing the use of synthetic data against real-world outcomes should be considered.

Crucial sampling questions need to be addressed, not avoided

Simon Raper

Founder

Coppelia Machine Learning and Analytics

This is a good opportunity to pre-empt pseudoscientific claims that might spring up in studies using synthetic respondents. Statistical inference in research is about drawing probabilistic conclusions concerning a population using data gathered from a sample. From a near-as-possible random sample of UK adults we can estimate a preference for a product in the overall UK adult population. We can also provide a measure of our confidence in that estimate: a confidence or credibility interval depending on whether the approach is frequentist or Bayesian.

But which population are we sampling from when we generate synthetic respondents using a Large Language Model (LLM)? Such respondents are randomly generated from what is, in effect, an extremely complex conditional probability distribution. This is a very different kind of sampling - one that stretches the idea of a population to breaking point. In fact, it is the same kind of sampling that takes place when we roll a set of dice. Each roll is a sample from an infinite population of possible outcomes where the probability of each outcome is described by a probability distribution.

So in the case of randomly generated synthetic respondents the sampled population is not a real world population but rather the infinite population of possible synthetic respondents. Why does this matter? Because there is a danger that this quite obvious fact will slip from view: researchers and clients will assume that statistical inferences are being made about real world populations and that all the usual statistical paraphernalia (confidence intervals or otherwise) are there to quantify our uncertainty regarding the real world population.

This is not true: all inferences will be about the properties of the LLM. Meanwhile the really big question - how the LLM relates to the world, and how synthetic respondents relate to their real world equivalents - is left completely unanswered. As a source of uncertainty this overshadows any kind of uncertainty around sampling, which after all can be made minute by generating vast numbers of synthetic respondents.

Will this happen? Will we see a slew of presentations where researchers make bold claims about statistical accuracy while a much greater source of uncertainty goes unmentioned. I hope not. I hope this is too obvious an error. But it wouldn't be the first time.

The really big question - how the LLM relates to the world, and how synthetic respondents relate to their real world equivalents - is left completely unanswered.

Using synthetic data to unlock patient insights and beyond

Reena Sooch
Partner
Day One Strategy

Synthetic data holds great promise for healthcare market research and insight. With low incident rates and niche populations, recruitment of patients suffering from rare diseases is a pain point in research. Synthetic data offers an innovative method to unlocking reliable patient insight efficiently and with surprising accuracy.

At Day One Strategy we have been trialling use cases of synthetic data and synthetic respondents for market research. At the tail end of 2023, we identified the perfect opportunity for synthetic data to provide insights for a rare disease.

The key business question: can Large Language model (LLMs) be used to create synthetic data to help pharma companies gather insights efficiently and get timely answers to ad hoc research questions?

The study was in a rare cancer. We compared a synthetic data set created using ChatGPT, and a control arm, to explore if this LLM can be used as a tool to deliver reliable and accurate insights about the patient and caregiver experience.

What we did
We used carefully crafted prompts, refined by our in-house prompt engineers, to extract synthetic data on the rare cancer using ChatGPT. We then validated these insights through qualitative interviews with a real patient, a caregiver, a clinician, and a pharmaceutical client to assess the accuracy of the output.

The output was positive
Synthetic data was able to unlock the patient journey for the rare cancer, highlighting the unmet needs and emotional burden for a patient suffering from the cancer. The accuracy of data was comparable to primary market research.

However, it lacked some of the nuance, depth and contextualisation of a traditional qualitative interview.

And before you try this for yourself, how you construct your prompt for the LLMs is critical to the quality and accuracy of the synthetic data. It requires specific expertise and knowledge to achieve high quality insight and output. And lastly, as we also learnt, hallucinations (factually incorrect or nonsensical output) can happen, even with robust prompts. So explainability and validation of output remains important when using LLMs and synthetic data.

We leveraged the synthetic data, not only for insight, but as input to create multi-modal comms content...

Using synthetic data to unlock patient insights and beyond

The outcome

Following a successful pilot, we embedded the synthetic data into one of our IdeA(i) Stretcher workshops. These workshops are for hypothesis generation pre or post qualitative fieldwork. We leveraged the synthetic data, not only for insight, but as input to create multi-modal (copy, images and videos) comms content that could be used to raise disease awareness or educate internal stakeholders on symptoms, unmet needs and the burden of rare diseases, such as cancer, in this instance.

What does this mean for market research?

Synthetic data is unlikely to replace the need to test and validate to generate real-life data and insight. It does change the playing field though. AI can produce high quality synthetic data faster, easier and cheaper. It offers a platform for generation of ideas, leveraging large artificial datasets to test, interrogate and optimise multiple approaches for a variety of audiences.

At Day One we are refining our prompts every day, looking for new innovative AI assisted methods and solutions to enhance understanding of customers and patients' needs and deliver actionable strategy with technology at its core.

Re-imagining primary research

Phil Sutcliffe
Managing Partner
Nexxt Intelligence

This Delphi report provides a comprehensive and informative explanation of use cases of synthetic data for market research. I take the view that synthetic data can be a useful, low-cost tool in the researcher's toolbox. Whether you agree with this view or not, I think it is inevitable that synthetic data will become increasingly used for market research due to the simple fact it is quick, cheap and easily available.

As a consequence, I hope that the real benefit of synthetic data will be that, ironically, it enables better quality primary research. As this Delphi report says, "(due to synthetic data), conducting primary research may be unnecessary at these early stages in the insight process, (meaning) today's restricted budgets should be reserved for later more complex stages in the research process".

To go a step further, I believe that synthetic data should be embraced by market researchers because it offers the opportunity for a re-set; for primary market research to be reimagined. As this report says; "true insight comes from unexpected places, from the edge cases, and outlier data". In my view, this is an area that synthetic data will always struggle to do, that is to replace the essence of what qualitative research delivers the deep human insight that comes from talking to real people in real-time to get a rich understanding of their behaviours, beliefs, motivations and emotions through probing and projective techniques.

At the same time, this distinction provides a route map for how surveys can survive and thrive. In short, for surveys to have a future role in the age of synthetic data they need to become more qual-like. The good news is this is already possible. Conversational AI technology enables the survey to be re-imagined as an interaction between chatbot and participant. The chatbot can ask typical quantitative questions but it can go much further, engaging participants in a dialogue by using generative AI to probe the answers they give to open-ended questions.

Not only does this deliver much deeper insight than traditional online surveys, it also drives participant engagement as people feel they are being listened to and heard. And more engaged participants pay better attention and provide better quality data. With the increasing power of generative AI, conversational methodologies will enable people to answer and be probed using voice or video and, in time, realistic avatars will be used to interview people at scale and in-depth.

It's an obvious point that the quality of verbatim data captured by conversational AI is a function not only of the technology but also of the quality of the sample. The issues of declining quality of sample are well documented, with increasing problems from respondent fraud and bot farms. I hope that an increase in the use of synthetic data could also lead to a new model for participant engagement. When it comes to sample quality, the only way I see to beat the bots is for panels to have authenticated relationships with panelists whom they can guarantee are real people. And that will cost more money. However, if buyers of research are saving money by diverting a portion of their research spend to synthetic respondents, I hope they could be encouraged to invest that saved money in better quality sample. If more money from buyers is on the table that will encourage panel providers to invest in delivering better quality, validated panelists.

LLMs need to link back to evidence

Jack Wilson and Felicity Browning

2CV

We need to move past the idea of LLMs being sources of data and instead embrace the idea of LLM's being interfaces for democratising access to data.

Could synthetic data be the answer for democratising access to insight and customer voice? Only with a clearer definition, high quality training data and transparent referencing.

'Synthetic data' is a very tricky term, as it means different things to different people and the definition has a huge impact on the viability of the approach. From our perspective, there are three core approaches to synthetic data, all of which pose different drawbacks and benefits:

1. LLM as the sole source of data
2. LLM as a source of data via dedicated training sets of primary and proprietary data
3. Data imputation through machine learning

Using an LLM as a sole source of data comes with the most issues in terms of viability.LLMs are by and large black boxes, both in terms of their underlying architecture and their training data. Although the data they spit out may look and sound convincing – there's little or no ability to validate accuracy. LLM's are prone to hallucination and the inherent bias within training data and the alignment of the model can make them prone to generating inaccurate or misleading information. Moreover – maintaining accuracy in foundational models would require a reliable live pipeline of relevant data to inform the ever-evolving attitudes of consumers. Beyond these technical barriers to viability, there are the behavioural barriers and trust issues that are inevitable when dealing with data generated by machines.

However, we would argue that if the LLM is being used as an intermediary, drawing on a mix of proprietary, paid-for and publicly available data – it suddenly becomes a much more interesting prospect. When used in this way, we believe the validity and quality of generative outputs are likely to be much improved. Similarly – data imputation through machine learning is another approach to synthetic data that poses an intriguing angle, providing a mechanism that could help to mitigate data quality issues through a modular survey approach, filling in the gaps with synthetic data. In both of these cases – the key ingredient is high quality primary data to train the model.

Despite these promising new approaches to synthetic data, the key problem we see with synthetic data products (at least at present) is the lack of clear transparent referencing of evidence bases. This is the crucial factor that we believe is essential for realising the potential of synthetic data. We need to move past the idea of LLMs being sources of data and instead embrace the idea of LLM's being interfaces for democratising access to data – providing a mechanism to rapidly draw out relevant insight from large and diverse datasets curated by insight teams. As a general rule – we believe all 'synthetic data' or generative summaries produced by an LLM should provide direct links to the primary data that form the basis of the generative output.

LLMs need to link back to evidence

We believe without this crucial step synthetic data will remain a niche part of the research landscape.

A common trend we're seeing across a range of our more innovative clients is the growing use of organisational LLMs/GPTs. This may come in the form of a company chatbot that can draw on organisational knowledge, or it may take the form of knowledge hubs (like Market Logic/Strativo) that provide a tailored approach to navigating and synthesizing insight work. Organizations are also increasingly adopting AI copilots (via Microsoft, Google and other productivity tools). We believe this gradual growth of familiarity with day-to-day usage of AI will lay the groundwork for more creative and advanced use-cases for synthetic data deliverables and interfaces.

One clear use case for synthetic data is segment enrichment. Creating a custom GPT that can speak as different customer segments, using a mix of proprietary primary research data, customer behavioural data, CX data (customer reviews/chatlogs) and social data as sources that can be drawn to provide an informed perspective on any issue for customer segments. In this way employees could speak to personas asking them questions and receiving answers that are not only in the 'voice of the customer', but also referencing data that supports what they're saying.

Stretching the definition even further – it could be argued that synthetic data could include any form of generative summary that references organisational data. This is the issue with 'synthetic data' as a term – it doesn't accurately describe the nature of the wide range of research methodologies it could represent. Without a firm definition of what is meant by synthetic data or new terms to describe evolved approaches, it is likely to remain a niche and controversial area of research practice.

References

1 Energy Social Surveys Replicated with Large Language Model Agents
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4686345

2 Energy Social Surveys Replicated with Large Language Model Agents
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4686345

3 Out of One, Many: Using Language Models to Simulate Human Samples
www.cambridge.org/core/journals/political-analysis/article/abs/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49

4 How pleasant was your childhood?
https://pages.ucsd.edu/~pwinkiel/winkielman-schwarz_How-pleasant-was-your-childhood-PS-2001.pdf

5 Large language models cannot replace human participants because they cannot portray identity groups
https://arxiv.org/abs/2402.01908

6 New MR: Synthetic data is the future of a large part of market research and insights – but the road from here to there might be bumpy
https://newmr.org/blog/syntheticdata/

7 Research Live: Research sector embraces opportunities of synthetic data
www.research-live.com/article/news/research-sector-embraces-opportunities-of-synthetic-data/id/5123686

8 IBM: What is synthetic data?
https://research.ibm.com/blog/what-is-synthetic-data

9 Gartner: Is Synthetic Data the Future of AI?
www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai

10 Marketing Week: Synthetic data is suddenly making very real ripples
www.marketingweek.com/synthetic-data-market-research/

11 Stanford University: Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive
https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

12 Roy Rosenzweig Center for History and New Media: Can History be Open Source? Wikipedia and the Future of the Past
https://rrchnm.org/essays/can-history-be-open-source-wikipedia-and-the-future-of-the-past/

13 NPR: A new AI chatbot might do your homework for you. But it's still not an A+ student
www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia

14 Esomar: 20 questions to help buyers of ai based services
https://esomar.org/uploads/attachments/cltn6755401khqe3v0od2y6ut-esomar-20-questions-to-help-buyers-of-ai-based-services.pdf

15 Research Paper: Exploring Synthetic Data Validation – Privacy, Utility and Fidelity:
https://ico.org.uk/media/for-organisations/documents/4025484/sythetic-data-roundtable-202306.pdf

16 Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive:
https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

17 Here's What Happens When Your Lawyer Uses ChatGPT:
www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

18 Financial Times: UK shelves proposed AI copyright code in blow to creative industries: https://on.ft.com/3Sqe9H8

19 The Washington Post: AI's future could hinge on one thorny legal question: www.washingtonpost.com/technology/2024/01/04/nyt-ai-copyright-lawsuit-fair-use/

20 Bird & Bird: Liability of AI Service Providers for Copyright Infringement: Guangzhou Internet Court reaches world's first decision: www.twobirds.com/en/insights/2024/china/liability-of-ai-service-providers-for-copyright-infringement#:~:text=The%20Court's%20findings%20that%20(i,both%20significant%20findings%20for%20generative

21 European Commission: Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions Empty: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52019DC0168

22 Forbes: What every CEO needs to know about the new AI Act: www.forbes.com/sites/bernardmarr/2024/03/25/what-every-ceo-needs-to-know-about-the-new-ai-act/

23 FCA: Consumer Duty sets higher standards for financial services customers https://www.fca.org.uk/news/news-stories/consumer-duty-higher-standards-financial-services

24 FCA: Consumer Duty implementation: good practice and areas for improvement www.fca.org.uk/publications/good-and-poor-practice/consumer-duty-implementation-good-practice-and-areas-improvement

25 FCA: Guidance for firms on the fair treatment of vulnerable customers https://www.fca.org.uk/publications/finalised-guidance/guidance-firms-fair-treatment-vulnerable-customers

26 www.mrs.org.uk/standards/guidance-on-using-ai-and-related-technologies

Build your AI knowledge and skills

Courses

23 May
Text Analytics for MR: The AI, the human and their collaboration

4 July
Generative AI in Market Research and CX

25 September
Transforming MR with Chat GPT

1 October
Text Analytics for MR: The AI, the human and their collaboration

5 December
Generative AI in Market Research and CX

On-demand

Generative Artificial Intelligence

Conference

4 July
Data Driven Insights 2024

Coming soon

Best Practice Guides

Collecting Data using the Metaverse

Collecting data using Biometric Data

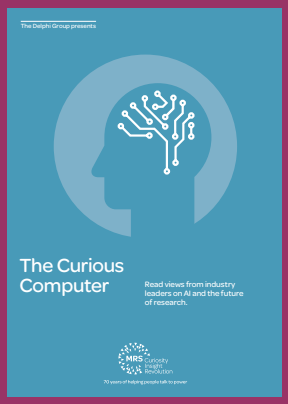
MRS

Visit www.mrs.org.uk

Download all Delphi reports here: mrs.org.uk/Delphi

Private lives?
A look at privacy issues through the lens of the consumer.

The politics of persuasion
What influences voters and how to improve democratic engagement.



The curious computer
The impact of AI on the research and insight sector.

Towards an insight driven organisation
People, skills and processes.

Prediction and planning in an uncertain world
Helping organisations connect with the future.

Great Expectations
How technology impacts consumer trust.

Fast-forwarding research
How Covid-19 has reset the customer insight function.

Managing the multiverse
Understanding a world in transformation.

Rise of the insight alchemist
How the data explosion has created a new breed of insights professionals.