HUMANIZING AI

Real human data to generate and predict real innovation success

December 2023

AUTHORS

Colin P Ho, PhD

Jiongming Mu

IPSOS
VIEWS

AI SERIES

GAME CHANGERS

#IPSOSHiAi

At Ipsos, we champion the unique blend of Human Intelligence (HI) and Artificial Intelligence (AI) to propel innovation and deliver impactful, human-centric insights for our clients.

Our HI stems from our expertise in prompt engineering, data science, and our unique, high quality data sets – which embeds creativity, curiosity, ethics, and rigor into our AI solutions, powered by our Ipsos Facto GenAI platform.

Our clients benefit from insights that are safer, faster and rooted in the human context.

Let's unlock the potential of HI+AI!

We are all uniquely human. As consumers, our decisions are complex, emotional, contextual, and often irrational. Although artificial intelligence (AI) makes new product development faster and easier than ever before, off-the-shelf generic models can distort or misrepresent consumers' realities.

In this paper, we discuss the practice of training AI models with real consumer data, to capture the essence of what drives real consumer behavior, and to generate and predict better innovations. Without connecting with real humans, even the most powerful algorithms will not be sufficient to guarantee innovation success.

The headline of a recent Newsweek article, "Artificial Intelligence: the 21st century Gold Rush",[1] perfectly captures today's sentiment: large language models (LLMs) like ChatGPT have turbo-charged interest in AI and inspired thousands of companies to dive headfirst into this space.

While nearly every industry has or is beginning to adopt AI, optimal applications for new product development are particularly distinct. The following pages delve into better applications of generative and analytical AI to power innovation success.

In a typical innovation process, an ideation phase is followed by an evaluation phase. AI can be leveraged in both phases:

- **In the ideation phase**, the divergent capabilities of generative AI can be leveraged to develop new product ideas.

- **In the evaluation phase**, the convergent capabilities of analytical AI can be used to predict their market potential.

AI presents an opportunity to improve the speed, and potentially, the success rate of new innovations, and how we go about doing this will determine whether we succeed. In both applications, the data used to train AI is critical.

New product ideas are more likely to succeed if these two phases are grounded in data reflecting consumers' intrinsically human needs and desires. This data needs to be timeless, or at minimum, up to date. As data is so central to AI, we start by explaining how training data determines the accuracy of its model.

BRAINS ARE LIKE SPONGES; THEY ABSORB WHAT THEY ARE EXPOSED TO - AI IS NO DIFFERENT

AI models become smart because they can learn. Their learning comes in two main forms:

1. **Supervised learning** requires a person to teach an AI exactly what to learn. For instance, if we would like to teach an AI to recognize positive and negative social media posts, we need to show the AI examples of each. In this way, the AI can learn features associated with positive and negative posts and develop decision protocols to classify new media posts that it was never exposed to.

   Think of supervised learning like a parent instructing a child. If a parent wanted to teach her child to recognize dogs and cats, the parent would show the child pictures of dogs and cats. A person is needed to provide examples and tell the machine what each example represents.

2. **Self-supervised learning** works differently. 'Training' involves feeding a large language model (LLM) a massive amount of text to generate predictions. In this type of learning, people do not need to spoon feed the AI what specifically it should learn. The AI can instead learn the structure within the examples fed (e.g., the word which usually shows up after the word "thank," or the word which usually comes before the words "love you"). With a tremendous amount of text and repeated exposure, LLMs become really good at predicting how words go together.

Think of self-supervised learning like a child learning a language on their own. As a child is repeatedly exposed to speech and text, she starts identifying patterns and making connections.

The commonality in the two forms of learning is the need to fuel AI with examples of what we would like it to learn. What makes AI smart is not the algorithms, but rather, the examples provided.

The process of building an AI is like developing the brain of a young child. By providing a child an abundance of learning opportunities, the child becomes intelligent. If the child is provided with few opportunities, or worse, incorrect information, the child's intelligence will be limited or biased.

In simple terms, the development of AI requires data, and the quality of the data determines the quality of the AI model.

The quality of training data can be evaluated in the same way we evaluate survey data quality. We look for data that is relevant to the task at hand, representative of the target population and product category, and timely (i.e., reflects the time period we are looking to understand). It is not possible to overstate the importance of training data: AI is only as intelligent as the data provided.
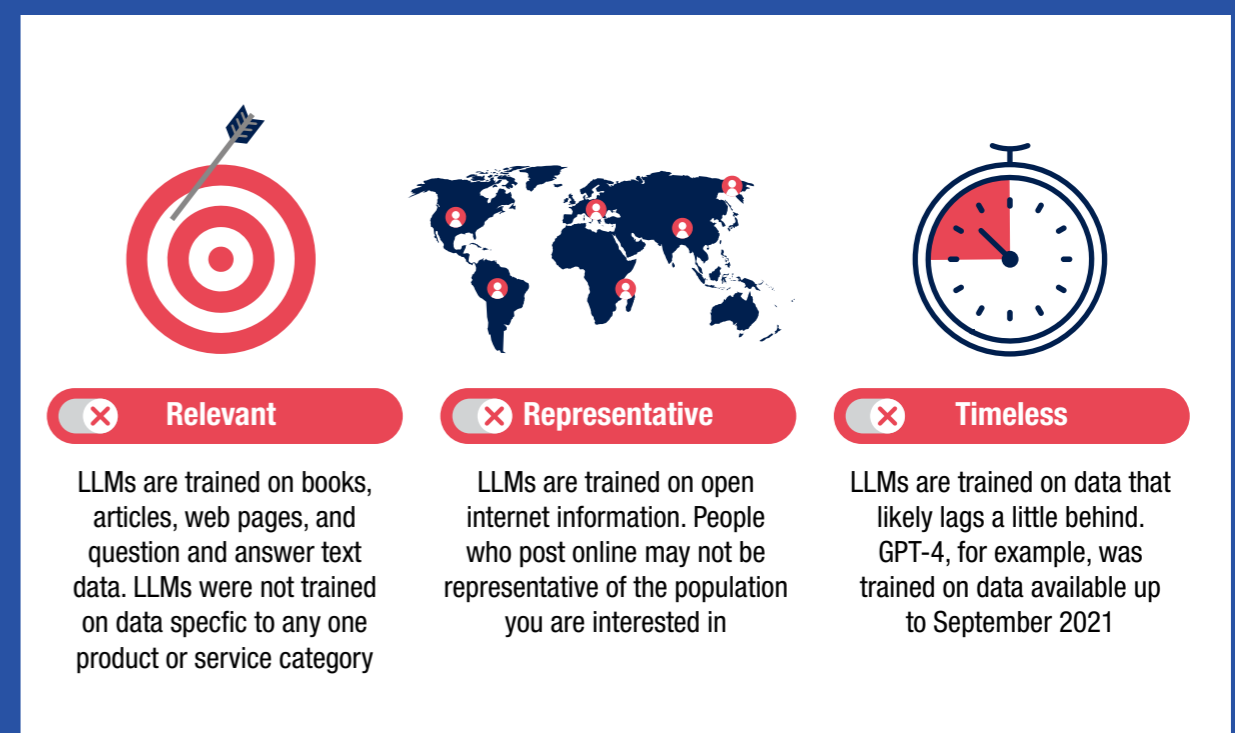
We share three principles that will be used throughout this paper (see *Figure 1*). While this framework was developed specifically to train an AI to generate or evaluate new product ideas, it can also be used to assess the adequacy of training data for any AI applications.

Relevant
The training data should capture the functional and emotional needs relevant to your specific product or service category

Representative
The training data should be from a robust and representative sample of your target population of consumers/customers

Timeless
The training data should be unaffected by the passage of time. Always true, valid, and applicable

Source: Ipsos

Figure 2 – Using off-the-shelf LLMs to generate new product ideas

Relevant
LLMs are trained on books, articles, web pages, and question and answer text data. LLMs were not trained on data specfic to any one product or service category

Representative
LLMs are trained on open internet information. People who post online may not be representative of the population you are interested in

Timeless
LLMs are trained on data that likely lags a little behind. GPT-4, for example, was trained on data available up to September 2021

Source: Ipsos

> **The online world is not a complete representation of the real world. Online content is posted by only those who are active online, with only certain topics deemed worthy of posting.**

USING GENERATIVE AI TO IGNITE FRESH IDEAS

The divergent capabilities of LLMs make them well suited for idea generation. Because LLMs have rendered it possible for any person to interact with an AI through natural language, any novice user can write a prompt to generate new product ideas (e.g., "please generate ten new product ideas for laundry detergent").

With prompt engineering, ideas generated by AI are typically articulate and thought-provoking. Given the large amount of text used to train LLMs, their excellence in expressing ideas fluently is not surprising. Although powerful, there are three good reasons to improve off-the-shelf LLMs (see *Figure 2*):

1. **LLMs learn from a vast and uncharted internet landscape**: We know LLMs learn from public internet information: web pages, reddit posts, online books, and other text sources. We do not know if these data sources cover a specific product or service category of interest.

2. **LLMs do not incorporate the unwritten realities of the offline world**: The online world is not a complete representation of the real world. Online content is posted by only those who are active online, with only certain topics deemed worthy of posting. For certain tasks which are meant to solve specific challenges, such as ideation, this can lead to knowledge gaps where real consumer knowledge in granular and specific areas are required, and online data alone is simply not sufficient.

3. **LLMs function in time capsules**: LLMs are often trained on data that lags. For instance, while ChatGPT-4 was first released in 2023, it was trained on data available up to September 2021 (see *Figure 2*). Although gaps in timing vary by model, and these gaps will fade over time, it is the researcher's role to ensure models continue to accurately reflect consumers' most recent realities.
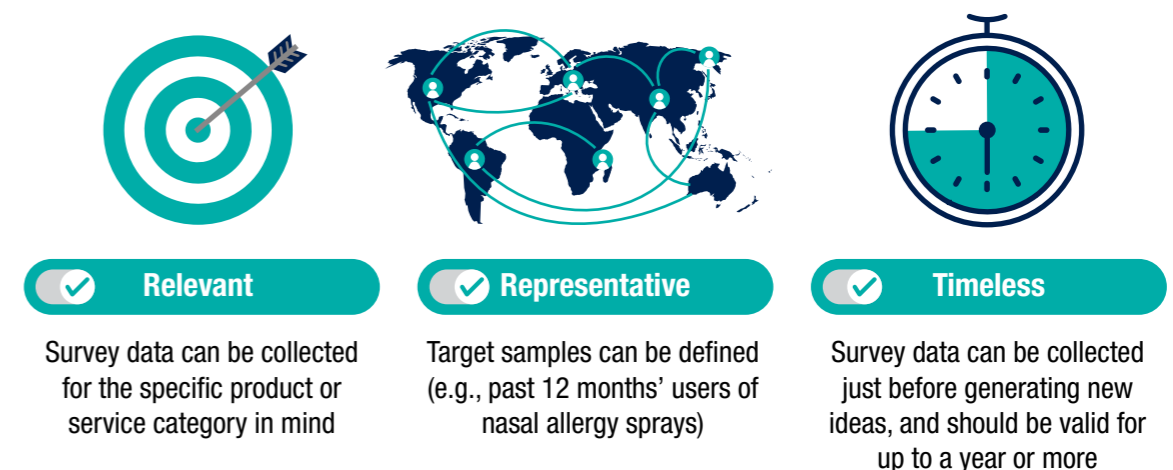
HUMANIZING GENERATIVE AI: USING REAL CONSUMER DATA FOR REAL INNOVATION

One way to improve LLMs is by introducing additional data to avoid generating ideas solely from public training data, especially when the goal is to address specific and granular consumer needs. Depending on the business challenge, this can be done using a variety of sources, such as surveys, social or search data. While Ipsos is simultaneously exploring multiple data sources for model training purposes, the following section looks more closely into using relevant, representative, and timeless survey data for innovation (see *Figure 3*).

To illustrate the advantage of humanizing AI models, we conducted a pilot by capturing consumers' unmet needs for nasal allergy sprays. Consumers were asked to, in their own words, express their problems, frustrations, or challenges with nasal allergy sprays. We then used a LLM to generate new product ideas from this data. The advantages were clear when we compared the ideas generated with and without survey data (*see Figure 4*). Overall, ideas generated without survey data were generic, functional, and completely devoid of emotions. In contrast, ideas generated with survey data reflected how consumers truly express their symptoms (e.g., "thick mucus," "ear popping") and emotions (e.g., "pleasant," "enjoyable," "discomfort").

Figure 3: Enhancing LLMs with relevant, representative and timeless consumer data

✓ **Relevant**	✓ **Representative**	✓ **Timeless**
Survey data can be collected for the specific product or service category in mind	Target samples can be defined (e.g., past 12 months' users of nasal allergy sprays)	Survey data can be collected just before generating new ideas, and should be valid for up to a year or more

Source: Ipsos

The second idea generated without survey data presented a strange result. The first sentence described a new product designed for people with "sensitive skin." Given that nasal allergy sprays are spritzed directly into the nose, it was odd for the LLM to suggest a term more commonly associated with skin or beauty products.

We believe this occurred because LLMs learn words by association. In the English language, the words "nasal" and "allergy" often go together, but so do the words "skin" and "allergy." In this sense, it was not surprising to receive an idea referencing sensitive skin. LLMs are designed to predict the next word and may do so inaccurately or inappropriately.

This does not mean there is no value in the ideas generated from publicly trained LLMs.

The massive amounts of data used to train LLMs can drive the discovery of valuable ideas. The black box nature of LLMs, however, means we will not know anything about the data used to generate the ideas — including its source, representativeness, or timelessness. As a result, LLMs may occasionally introduce ideas that do not make sense, even if they are sensible from a language perspective.

When generating ideas from survey data, the source of the data, the time of collection, and the reasoning behind the LLM's use are visible. The advantage of using survey data for generating new product ideas is 100% transparency. Understanding the reasoning behind decisions or predictions made by AI can provide reassurance on the validity of the AI responses.

LLM *without* **survey data**

A nasal spray that is designed to relieve both the symptoms of allergies and the underlying cause of allergies. This could be achieved by using a combination of ingredients that work to reduce inflammation, clear nasal congestion, and prevent the release of histamine.

A nasal spray that is specifically designed for people who have sensitive skin. This nasal spray would be made with gentle ingredients that are less likely to cause irritation or allergic reactions.

LLM *with* **survey data**

A nasal spray that combines a saline solution with an anti-inflammatory agent to help reduce inflammation and clear our thick mucus. The spray would also have a pleasant taste to make it more enjoyable to use.

A nasal spray that contains a combination of natural ingredients such as menthol, eucalyptus, and peppermint oil to provide relief from sinus pressure, headaches, watery eyes, and ear popping. The spray would also contain anti-inflammatory and analgesic properties to reduce discomfort and soreness in the head.

Source: Ipsos

TAILORING ANALYTICAL AI: CURATING DATA FOR CONCEPT EVALUATION

Before LLMs were available, analytical AI supported research needs. In the world of innovation, an AI that can instantaneously predict a new product's success is the promised land that many businesses seek. Unlike LLMs, however, off-the-shelf analytical AI models are not available. Because concept evaluation is mainly done in market research, customized analytical AI models are needed to predict innovation success. The criteria to build a successful analytical AI to predict a new product's success remains the same as those for generative AI: relevance, representativeness, and timelessness.

INITIAL EXPLORATIONS OF ANALYTICAL AI: COMPLETELY DEVOID OF HUMANS!

Inspired by industries in which the data used to train analytical AI were stimuli (e.g., pictures of healthy skin versus cancerous skin), some researchers advocated for the use of previously tested concepts as training data. That is, using text descriptions of successful and unsuccessful concepts tested in the past to train an AI model to predict the success of new product ideas. Such an approach eliminates the need to ask consumers for their responses to new product ideas and allows for instantaneous predictions.

This approach was indeed our first attempt to build an analytical AI to predict the success of new products. By feeding an AI model with past examples of successful and unsuccessful concept descriptions from Ipsos' concept testing database, we explored the features of previous concepts to predict market success. This first attempt failed. The AI model's accuracy fell short, because it thought there were more successful concepts than there really were.
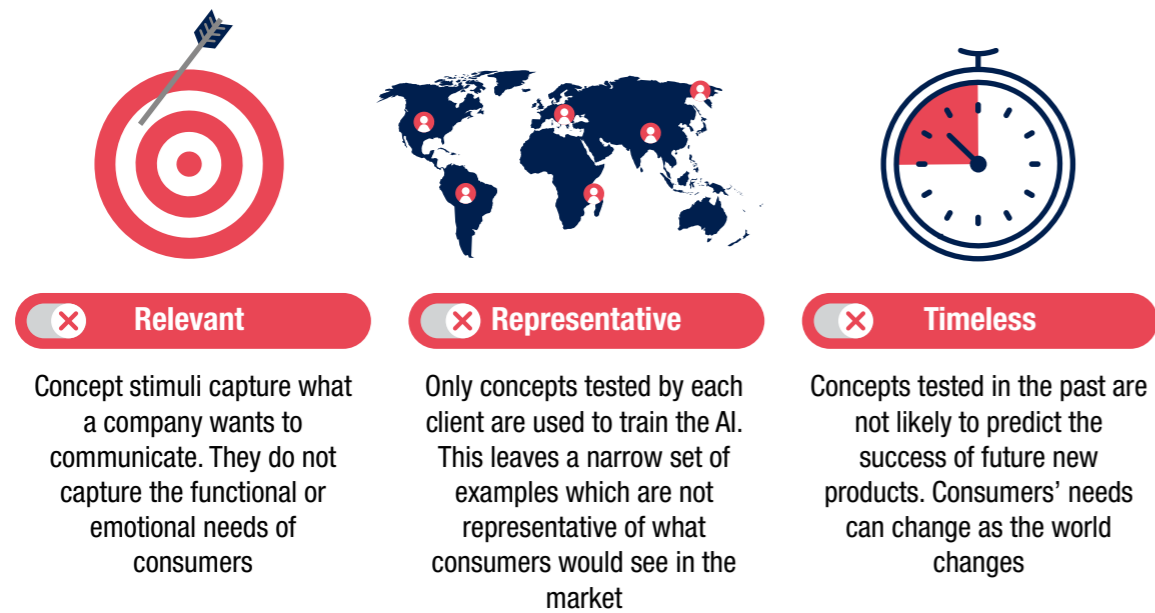
In diagnosing the problem, we found the analytical AI model to work in an overly simplistic way. The AI predicted a concept would perform well when certain benefits or phrases were present. The problem with this is that consumers do not simply respond to keywords or phrases. Consumers are far more complex than that; consumers react to entire propositions and not just individual elements.

We, as humans, evaluate whether a new product is priced reasonably, has desired variants, and meets our needs for our unique situation. In other words, a model trained on concept data is a gross oversimplification of how we think. As a result, using the information found in concepts alone does not capture the complex consumer decision-making process nor does it fulfill any of the criteria in our framework (see *Figure 5*).

> The criteria to build a successful analytical AI to predict a new product's success remains the same as those for generative AI: relevance, representativeness, and timelessness. **"**

Figure 5: Using past concepts to predict innovation success

Relevant	Representative	Timeless
Concept stimuli capture what a company wants to communicate. They do not capture the functional or emotional needs of consumers	Only concepts tested by each client are used to train the AI. This leaves a narrow set of examples which are not representative of what consumers would see in the market	Concepts tested in the past are not likely to predict the success of future new products. Consumers' needs can change as the world changes

Source: Ipsos

HUMANIZING ANALYTICAL AI WITH A PULSE OF VISCERAL REACTIONS

For some time now, we have been collecting consumers' top-of-mind reactions to new product concepts from a single open-ended question asked immediately after concepts have been viewed. As of 2023, we have accumulated about five million consumer responses from new product concepts across 60+ countries and seven mega-categories (human food, beverages, health care, homecare, personal care, beauty care, and pet care).

Consider what might go through your mind when you encounter a new product. If a person is responding to food concepts, for example, her immediate thoughts may include positive and visceral responses, like "yummy" and "looks tempting," or negative responses like "looks disgusting" and "too much sugar." In addition to responding positively or negatively, a person may react with skepticism or indifference because they already know a marketer's goal is to sell. A person may assess whether the product's price is within her budget and if the

new product is different from what is already on the market. In some cases, a person might want more information before deciding whether to try a new product. In short, consumers' responses to a new product can vary tremendously, and top-of-mind reactions allow us to capture it all.

We believe direct and visceral human reactions to new product ideas are inherently relevant and the ideal data to train an AI to predict innovation success. In fact, from our database of human reactions, we have made two observations:

- **First, reactions to new products do not change over time.** People will continue to be skeptical about new product claims, evaluate whether a product price is reasonable, assess whether they like the look of the product and so on, now and in the future. This timeless characteristic means that an AI trained on these verbatims will not become outdated.

- **Second, reactions to new products are similar across countries and cultures.** This universality enables the development of AI models as generalizable across related product categories or countries. If the way in which people respond to new products is similar across product categories and countries, it eliminates a need to build a model for every product category and country that we want to predict in.

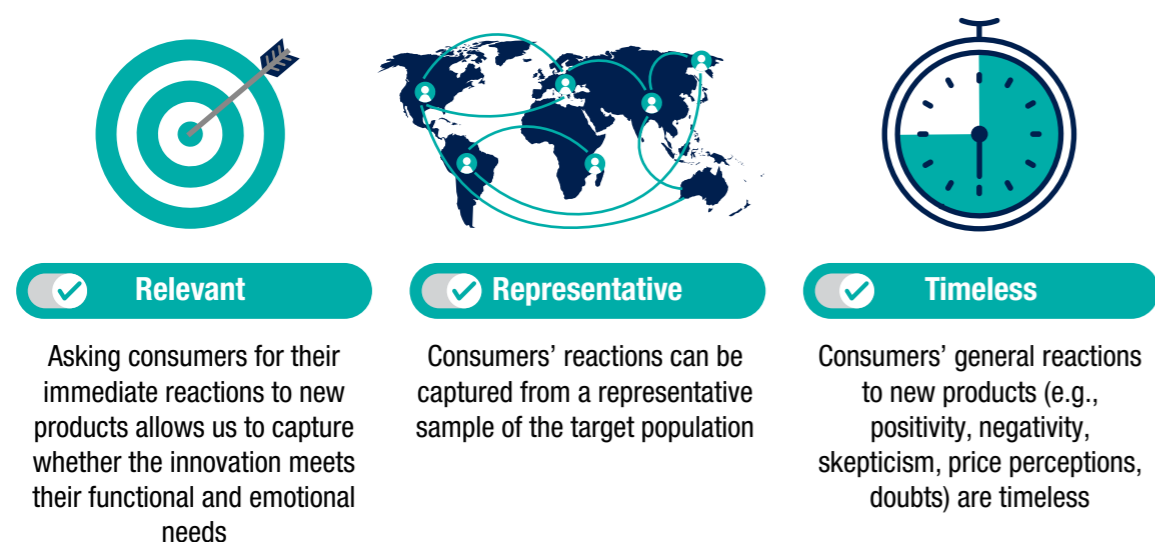At Ipsos, we use two sets of items from our database to build analytical AI models:

1. Consumers' visceral **top-of-mind verbatim responses** to new product concepts, and

2. Whether consumers choose the new product (or divert to their existing one)

on three **key metrics validated to real new product launches** – Relevance, Expensiveness and Differentiation

From these two elements, analytical AIs are built to predict if a person would choose a new product or choose to stay with their existing solution.

We have now built multiple analytical AI models to predict the success of new concepts across a variety of product categories (e.g., food, beverage, personal care, beauty, homecare). All the models have an accuracy of 70% or greater when predicting individual choice, and an accuracy of 80% or greater when predicting an aggregated trial metric. These accuracies are higher than those from our first AI models based only on concept text. Humanizing AI improved our model's accuracy towards predicting new product success.

Figure 6: Using consumer verbatims to predict innovation success

Relevant ✔	Representative ✔	Timeless ✔
Asking consumers for their immediate reactions to new products allows us to capture whether the innovation meets their functional and emotional needs	Consumers' reactions can be captured from a representative sample of the target population	Consumers' general reactions to new products (e.g., positivity, negativity, skepticism, price perceptions, doubts) are timeless

Source: Ipsos

TRUTH, BEAUTY, AND JUSTICE

Our focus in this paper has been on the quality of training data and how it impacts the accuracy of an AI model. If an AI model is accurate, then we have captured **"Truth."** Truth is important to all AI applications, not just innovation applications. At the risk of belaboring the point, we present three non-innovation examples to illustrate the threats of not incorporating quality training data:

- **Generic training can produce critical knowledge gaps.** *An AI for a self-driving Uber car hit a pedestrian because the AI was not able to recognize the pedestrian who was jaywalking (i.e., not using the crosswalk). When the AI was trained, the data did not include or did not have many examples of people who jaywalked. This meant the AI classified an object as a pedestrian only if the person was on or near a crosswalk.*[2]

- **Unwritten realities can breed bias and misrepresentation.** *AI systems designed to diagnose skin cancer are less accurate for people with dark skin because the data used to train these AIs have very few images of people with dark skin.*[3]

- **Time capsules can fail to capture quick-paced societal change.** *AIs trained on pre-pandemic human behavior did not work during the pandemic. During the pandemic, people's behavior changed a lot. One set of behaviors that did change dramatically was shopping behavior. These changes impacted AI models, causing*

problems for algorithms that were used in inventory management and marketing. AI models can decay over time, and the risks of model decay are exacerbated in times of rapid change.[4]

Truth, however, is only one of three pillars we use to evaluate AI applications. In our discussion of using survey data to supplement LLMs, we have also spoken about the need for transparency. At Ipsos, we call this second pillar **"Beauty"** – how explainable an AI is. For both generative and analytical AI, understanding how an AI arrives to an answer is important. Explainable AI allows us to check the face validity of an AI's output. Being able to see and explain WHY we got a certain result provides reassurance that the process and data that got us to the result is valid. Using survey data provides this transparency.

Our third pillar is **"Justice."** Given the broad use and practical implications, the fairness and ethical issues of AI and Generative AI are critically important today. For market research specifically, we take steps to ensure we do not use confidential or intellectual property in the training of an AI. There are other dimensions of justice such as the fairness of AI and the societal implications. While equally important, these are less likely to be of concern in market research. AI applications in market research are less likely to impact the livelihoods of individual consumers.

HUMANIZING AI: REAL HUMAN DATA TO PREDICT REAL HUMAN BEHAVIOR

While we are all fascinated with AI, it is relevant, representative, and timely data that will determine innovation success. These three criteria apply whether you build your own AI or use a pre-trained model from another company. Without this information, you cannot assess whether an AI model is suited for your specific business application.

The most valuable data of all is real human data, which tells us about who we are, how we live our lives, and what we need to make us safe, comfortable, healthy, and happy. We buy because we are humans. We have needs, insecurities, and care what others think of us. We make choices driven by emotions and are not always rational. AI applications in market research that fail to capture the true essence of being human will not perform well.

We started the paper by discussing how humans are teachers; we provide data for AI to learn. We are on the cusp of a technological revolution, and humans are still driving this revolution (at least, for now)! As we head into this brave new world, let us be good teachers of AI.

> People will continue to be skeptical about new product claims, evaluate whether a product price is reasonable, assess whether they like the look of the product and so on, now and in the future. "

REFERENCES

1. Chadha, P. (2023, May 30). Artificial Intelligence: the 21st century Gold Rush. Newsweek. https://www.newsweek.com/artificial-intelligence-21st-century-gold-rush-1802889

2. BBC News. (2020, September 16). Uber's self-driving operator charged over fatal crash. BBC News. https://www.bbc.com/news/technology-54175359

3. Davis, N. (2021, November 10). AI skin cancer diagnoses risk being less accurate for dark skin – study. The Guardian. https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study

4. Heaven, W. D. (2020, December 10). Our weird behavior during the pandemic is messing with AI models. MIT Technology Review. https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/

FURTHER READING

1. Beyond the Hype: Innovation Predictions in the Era of Machine Learning (2022). https://www.ipsos.com/en/beyond-hype-innovation-predictions-era-machine-learning

2. Exploring the Changing AI Landscape: From Analytical to Generative AI (2023). https://www.ipsos.com/en/chatgpt-and-rise-generative-ai-navigating-changing-landscape-ai

3. Conversations with AI: How generative AI and qualitative research will benefit each other. https://www.ipsos.com/en/conversations-ai-how-generative-ai-and-qualitative-research-will-benefit-each-other

4. Conversations with AI Part II: Unveiling AI quality in qualitative workstreams. https://www.ipsos.com/en/conversations-ai-part-ii-unveiling-ai-quality-qualitative-workstreams

5. Conversations with AI part III: How AI boosts human creativity in ideation workshops. https://www.ipsos.com/en/conversations-ai-part-iii-how-ai-boosts-human-creativity-ideation-workshops

HUMANIZING AI

Real human data to generate and
predict real innovation success

AUTHORS

Colin P Ho, PhD., Chief Research Officer,
Innovation and Market Strategy & Understanding, Ipsos

Jiongming Mu, Senior Vice President, Innovation, Ipsos

#IPSOSHiAi

www.ipsos.com
@Ipsos

GAME CHANGERS