IPSOS VIEWS

SYNTHETIC DATA

From hype to reality - a guide to responsible adoption

Michel Guidi
Benoit Hubert
Ciprian Sava
Rich Timpone, Ph.D.

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 Human Intelligence 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 Gen AI platform. Our clients benefit from insights that are safer, faster and grounded in the human context.

Let's unlock the potential of HI+AI!

#IpsosHiAi

> When using Generative AI to create synthetic data, remember that this technology is not magic – it is math.

Synthetic data, powered by AI, is poised to transform the market research industry. The question isn't if, but when and how.

Recognizing the potential, but also the possible pitfalls, of the issue, our clients asked us to provide Ipsos' trusted perspective on the topic. **In this paper, we demystify synthetic data and provide recommendations on when, where, how, and who to trust for responsible, safe, and value-adding implementation.**

When using Generative AI to create synthetic data, remember that this technology is not magic – it is math. It may appear magical when used correctly, but that's only when **it combines the best of human and artificial intelligence**: when experienced researchers combine proprietary analytics frameworks, select the right AI/model for the specific task at hand, inject fresh, purposeful consumer data from real people, apply prompt engineering from domain experts, tap into fine-tuned data science algorithms, and leverage norms databases and data assets.

Simply put, the quality and reliability of synthetic data is entirely **dependent on the real human data** used to create and update it, as well as the expertise of the people behind it all.

We'll also help you steer clear of the "snake oil salesmen" that have emerged in the wake of Generative AI's potential, who lack the proper controls, reputation, expertise, or validation of their claims, and can wreak havoc on brands and businesses.

This point of view aims to help you make sense of the current landscape and what the future may hold. We hope **it will help you form an objective opinion of synthetic data,** demonstrate both its potential and its risks, and refine the questions you need to ask yourself and your partners before you start considering it.

Synthetic data in a nutshell

Synthetic data refers to artificially generated data that does not directly correspond to actual events or people. Ideally, it would mimic the statistical properties and patterns of real-world data, although given today's hype and confusion, this is not always the case.

However, synthetic data is not a single method, but a broad domain that encompasses use cases as diverse as data enrichment and augmentation, imputation and fusion, synthetic populations, as well as new areas enabled by Generative AI such as AI Assistants, Persona Bots and Agents. Given this breadth, it is not surprising that this is not an entirely new development. Ipsos has extensive experience in many of the areas where analytics and machine learning solutions are applied. For example, these techniques have been key to combining surveys, filling in missing data and creating synthetic populations for years. They are battle-tested and methodologically validated by Ipsos and are good to use with trusted researchers and partners.

> Synthetic data refers to artificially generated data that does not directly correspond to actual events or people.

Amara's Law, and a parallel with self-driving cars

Roy Amara (Researcher, Scientist, Futurist and President of the Institute for the Future) coined what is known as "Amara's Law", which has become a guiding principle in technology:

> "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."

There is a lot of hype and confusion around synthetic data today, which makes it even more important to remember Amara's Law.

Even if a technology is "technically ready", that does not mean it is ready for real life. Let's draw a parallel with self-driving cars. Autonomous vehicles are theoretically superior to human driving, the technology has been "technically ready" for years and getting better every year. Yet, because of the potential for risk, need for legal protections, and low consumer desire, self-driving cars make up 0.0001% of the cars on the world's roads today. There are already great use cases for autonomous vehicles, and more will emerge over time, but adoption is currently limited and will take time[1].

Similarly, the rise of synthetic data in market research will also take time. We should see this as an opportunity to harness its enormous potential and create the means by which we can use it safely.

Enter Generative AI

So why all the excitement and confusion around synthetic data? Because of the transformative power of new methods in AI and Generative AI which have **democratized access to synthetic data.** Generative AI, such as deep learning models including Generative Adversarial Networks (GANs) and Large Language Models (LLMs), is taking synthetic data to new heights. These AI systems can learn from existing datasets and generate realistic and diverse synthetic data in volumes and varieties previously unimaginable.

A game changer for the insights industry?

The new era of synthetic data has transformative potential for market research and **presents numerous opportunities:**

01 **Faster data collection:** synthetic data could allow researchers to generate data in hours, as opposed to the weeks or even months it may take to collect real data.

02 **Cost-effectiveness:** it is available at a fraction of the cost compared to traditional data collection methods.

03 **Flexibility:** it allows for more fluid and interactive research experiences, such as the ability to chat with individual synthetic customers to gain deeper insights.

04 **Optimization:** you can use it to explore hypothetical scenarios, by running tens or hundreds of permutations on a given brief.

05 **Understanding:** it can enable us to understand consumers in ways that have not been possible before, providing a new impetus and solving some of the challenges we face in understanding real consumer behavior and attitudes using survey-based research techniques.

How can synthetic data be dangerous for brands and businesses?

Have you ever experienced "hallucinations" when using Generative AI; when a language model returns an answer that sounds very confident and consistent, but is completely wrong? Synthetic data can be like that, on steroids. **It can lead to biased or completely incorrect insights,** or even legal implications, with potentially disastrous financial, legal or reputational consequences for your brand.

For example:

Inaccuracy: Synthetic data may be an inaccurate representation of what sound, primary data collection would produce.

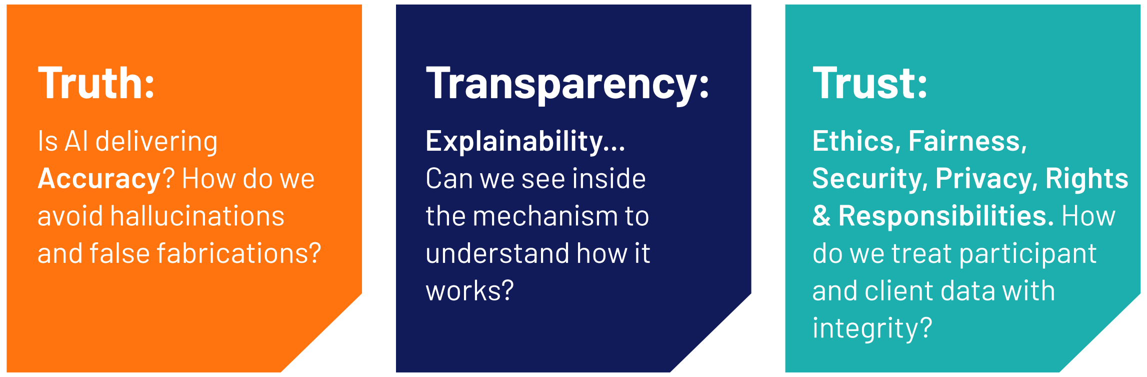
Bias: It can codify biases embedded in the data and methodologies from which it is built. Insights generated in this way may be contaminated.

Mediocrity: It may be less rich than actual human data, so it may find some main ideas but miss the much greater diversity of views that real people would produce. It may generate insights that are inherently recycled and derivative.

Inadequacy: Data validated for one purpose, such as the distribution of attitudes to identify trends, may not be appropriate for other uses of the data, such as segmentation or multivariate drivers of behavior. Also, it may be validated in one area, but not in others (Foundation LLMs have been tested in some cases where there was a lot of public data, for example). So it may not work as well for your area, or the dynamics may have changed since the training data was created.

These problems can occur if the methods are not properly evaluated. At Ipsos, we use the criteria of **Truth, Transparency** and **Trust** to ensure that AI applications are valid and sound[2], and we extend these to synthetic data.

Figure 1: Truth Transparency Trust Framework



Truth:
Is AI delivering **Accuracy**? How do we avoid hallucinations and false fabrications?

Transparency:
Explainability... Can we see inside the mechanism to understand how it works?

Trust:
Ethics, Fairness, Security, Privacy, Rights & Responsibilities. How do we treat participant and client data with integrity?

Source: Ipsos

Let's illustrate the potential risks with a few examples:

01 You run a concept test on a "sample" of 1,000 synthetic consumers of your category and get results that you put into production. **The result lacks robustness, specificity, variety and focus** compared to real data, even if it is well-founded, as found in our qualitative validations of answers from actual respondents and their paired "AI Twins".

02 You are tracking your brand equity and desirability using synthetic data, not realizing that there is a bias in the way the data is generated. **Bias in data generation**, particularly around race and gender, can seep into the results and damage brand equity if not addressed, like using a faulty altimeter when flying a plane over the Alps.

03 You create a new advertising campaign based on insights generated by synthetic data, not realizing that the underlying model is **using personal data illegally** – which can lead to lawsuits and regulatory issues, for example in relation to the EU AI Act. Campaigns based on such data may also lack novelty.

DOS AND DON'TS of using synthetic data:

✓ Do ensure the **quality** and representativeness of the synthetic data, ensuring it accurately reflects the characteristics, distributions, and relationships of the target population needed for the use case.

✓ Use synthetic data to **augment real-world data**, not replace it; use it as a supplement to the existing data to validate insights and make informed decisions.

✓ Prioritize **data privacy and security** and adhere to ethical and legal guidelines when generating and using synthetic data.

✗ Don't overlook the **importance of human expertise** in developing, interpreting and contextualizing the insights derived from synthetic data.

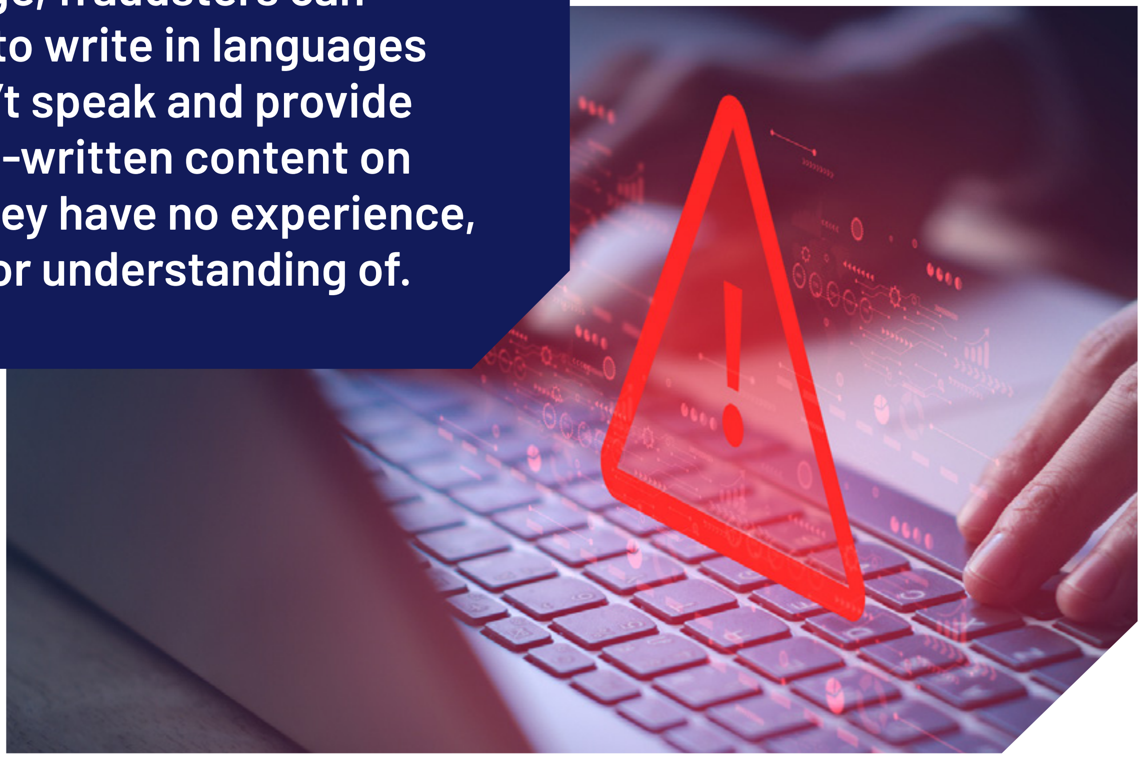
✗ Don't use synthetic data as a substitute for real-world testing **without proper validation**.

Fraud or synthetic data?

A growing concern for researchers is the **use of Generative AI to create fraudulent responses**: by using the technology to their advantage, fraudsters can pretend to write in languages they don't speak and provide rich, well-written content on topics they have no experience, context or understanding of. Ipsos and all the other players in the sector are working to counter this, including using Generative AI as a countermeasure to identify and eliminate fraudulent responses and block their authors.

But how do we distinguish fraud (whether from a fake respondent or a dodgy data provider) from legitimate synthetic data, especially if they are using the same tools? This is where having clear evaluation techniques helps. If the validation does not focus on **accuracy**, if it is hidden and not **transparent**, and if it disregards **ethical** rules and regulations, then it is fraud and violates each of the dimensions of **Truth, Transparency and Trust.**

> By using the technology to their advantage, fraudsters can pretend to write in languages they don't speak and provide rich, well-written content on topics they have no experience, context or understanding of.



Selecting the right partner to work on synthetic data (questions to ask)

In the past 12 months, we have seen a host of start-ups emerge who claim they can "do magic" with Generative AI.

To select the right partner and avoid the "snake oil salesmen" who have sprung up on the back of the technology's potential, we recommend you ask the following questions:

- How long have you been in the business of data science for insights? Not data science alone. Not insights expertise alone. Data science + insights experience together.
- How many data scientists did you have in 2022 (before Generative AI)?
- What data did you ground your models on?
- Do you own that data? If not, who owns it or how was it sourced?
- How did you validate your synthetic data models?
- What is the percentage of reliability of your proposed solutions?
- What are your standards and evaluation criteria?
- How frequently do you update the data models and the data itself behind your synthetic data?
- Is your business dependent on this specific particular or methodology?

Answers to look for:

When choosing a partner for Generative AI and synthetic data, look for a company with a reputation for **integrity,** a well-established **data science team,** access to vast amounts of **proprietary data,** robust intellectual property in **analytical frameworks** and methodologies, and a transparent evaluation and **validation framework.** The partner should also have mechanisms for updating data and models and not rely solely on synthetic data for their business.

The most critical factor is access to large volumes of high-quality, **curated proprietary data.** Companies that have this, along with skilled data scientists, proven methodologies, and expert knowledge, will be the best partners.

Be wary of companies that emphasize the cost and time savings of synthetic data without providing evidence of the **solution's reliability.** Generating synthetic data is highly complex, and "good enough" is not a sufficient evidence-based argument.

Measuring synthetic data's reliability: emerging principles

Until now, when using samples of a population, we could use statistics in the form of error margins and confidence intervals to understand how robust and reliable the data output would be. That changes considerably with synthetic data: **it's not just stats and probability anymore. It's all about how the models are created.**

However, the same concepts should apply to synthetic data, so we're working on defining new ways to determine how predictive and accurate synthetic data would be depending on the use case, based on factors such as the volume of real data used to create the synthetic data, its update frequency, and the robustness of the data science models in place. This is not an easy task, but it is a very important one.

If someone tells you their synthetic data's reliability is "insane" or "really f#%$ close", challenge them to give you the numbers and show you how they got there!

Evaluating synthetic data use cases

Ipsos believes that testing is essential to ensure that data is fit for purpose and enables real understanding. This applies to the entire data ecosystem, including synthetic data, if we are to avoid being misled by impressive but potentially questionable results. This is another area where we can use the **Truth, Transparency and Trust** framework. Not only can it help us evaluate predictive and Generative AI, but it is also suitable for assessing the quality of data sources.

The importance of each dimension varies depending on the specific data use case. Even for the same tool, different assessments may be required depending on the intended use. For example, a Persona Bot that generates product concepts will prioritize richness and fertility over accuracy, while one that reflects insights from specific segments will require greater accuracy.

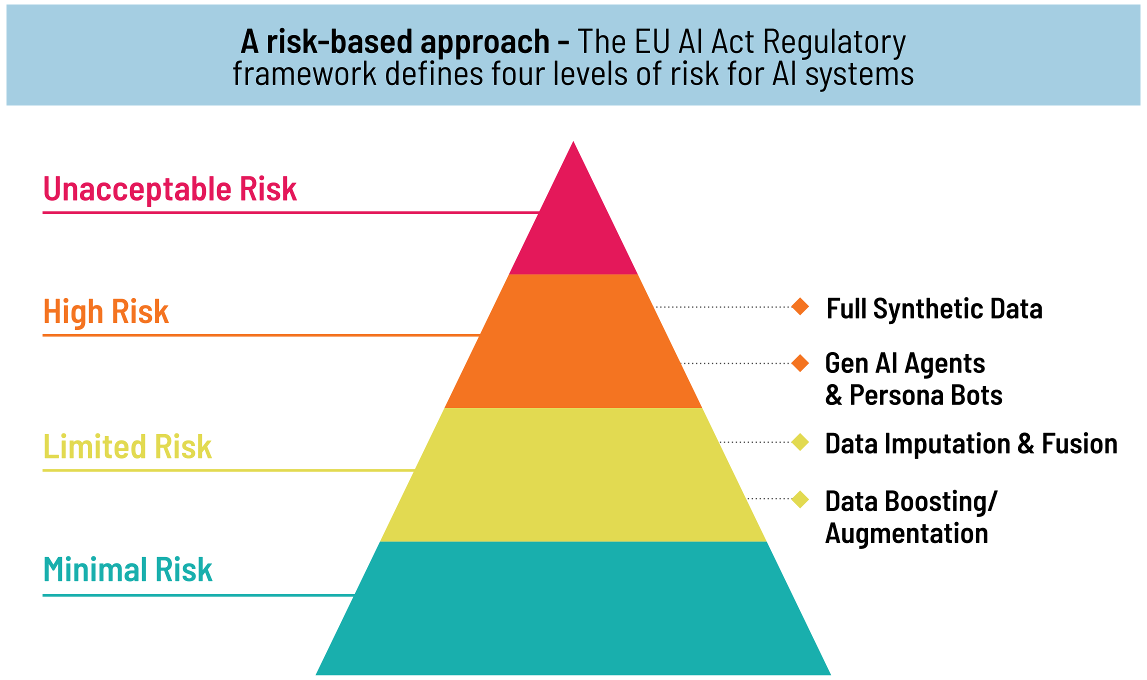
At Ipsos, evaluation is critical to sound and business-supporting solutions, and synthetic data is no exception. Our approach of using Foundation LLMs, grounding and training models addresses concerns about changes in accuracy over time and the risk of training on generated data. **Testing is essential** to ensure that the data is fit for purpose and enables real understanding.

Synthetic data use cases:

Here we highlight four major, current use cases for synthetic data: data boosting and augmentation, data imputation and fusion, Gen AI assistants and Persona Bots, and fully synthesized datasets.

In the diagram below, we have positioned them in the risk-based approach pyramid recommended in the recently published EU AI Act[3]:

Figure 2: Ipsos' view on the four major use cases vs. the EU AI Act Risk Pyramid


A risk-based approach - The EU AI Act Regulatory framework defines four levels of risk for AI systems
Unacceptable Risk
High Risk
Limited Risk
Minimal Risk
Full Synthetic Data
Gen AI Agents & Persona Bots
Data Imputation & Fusion
Data Boosting/ Augmentation

High Risk	◆ **Full Synthetic Data:** While some areas are improving, extensive validation is needed for general use and other areas are further from being ready ◆ **Gen AI Agents & Persona Bots:** Valuable for directional insights, need more rigor if treated as insights replacing human creations instead of hypotheses to test
Limited Risk	◆ **Data Imputation & Fusion:** Sound methods exist and have been used for years, but need to validate and use appropriately (i.e. aggregate tracking or modeling) ◆ **Data Boosting/Augmentation:** Sound methods exist but need experienced modelers and validation

Source:
EU AI Act Risk Pyramid
/ Ipsos

◆ Data boosting/augmentation

Definition: Data boosting is the addition of synthetic data to existing datasets to create a more comprehensive and representative sample. Examples include generating synthetic data in choice modeling, internet measurement products, and synthetic populations data while preserving statistical relationships.

Benefits: It enhances data availability and representativeness, reduces the need for extensive data collection, and supports better decision making.

Potential risks: The quality of synthetic data depends on the models and datasets used for training. Bias and amplified errors can occur if the synthetic data doesn't accurately reflect the target population. Validation for one use may not be applicable to others.

◆ Data imputation and fusion

Definition: Data imputation fills in missing or incomplete data points using available information. Fusion methods combine multiple data sources and complete all values for individuals.

Benefits: Imputation uses more collected data, avoids selection bias, reduces questionnaire length, and improves respondent experience. Fusion allows disparate data sources to be combined.

Potential risks: Accuracy depends on the validity of the models used for the intended purpose. Imputed data may be good for summary and descriptive statistics but not for modelling. Fusion data quality may break down as the number of sources increases.

◆ Gen AI Agents and Persona Bots

Definition: Customized Digital Assistants and Persona Bots emulate consumer segments or individual respondents, providing directional input based on synthesized responses from research data and specific topics we want the agent to be competent in.

Benefits: GenAI Agents and Persona Bots offer innovative ways to engage with and understand consumers, facilitating continuous and scalable segmentation research.

Potential Risks: Full synthetic data to replace real human data carries significant risks, including inaccurate insights and the potential for malicious behaviors, such as creating deepfakes or spreading misinformation.

Requirements: Persona Bots should be grounded in robust, validated research data, and transparency about their synthetic nature is necessary.

Full synthetic data

Definition: The use of samples that consist of 100% artificially generated, synthetic respondents.

Benefits: High-speed, low-cost data collection.

Potential Risks: Full synthetic data, while appealing, presents many risks. It may not accurately represent real-world data, leading to inaccurate insights. Building personas from individuals doesn't reflect the diversity of the original people.

Requirements: Resorting to full synthetic samples is risky business. It requires extensive validation and should only be entrusted to expert partners.

Conclusion

If Generative AI is like a superpower, then synthetic data is its first "weaponized application": it needs to be handled with great care and should only be entrusted to companies with the right data science expertise, domain expertise, access to large amounts of relevant, curated data, and who will use it with care, curiosity and a commitment to responsible use.

Embracing this undoubtedly exciting new opportunity requires a company with an adventurous attitude and an openness to the unknown possibilities that the future may bring.

At Ipsos, we don't just want to be seen as a safe pair of hands when it comes to the possibilities of synthetic research. **We truly recognize the scale of the opportunity, we are excited by its potential and we want to lead the way in exploring, evaluating and harnessing its potential.** More than ever, you can count on Ipsos to be your trusted advisor on how to make good use of synthetic data, as we do with all data in the expanding data ecosystem. We are here to discuss the options, recommend when to use them and when not to, and to apply these powerful techniques in ways that will add new dimensions to your research.

Endnotes

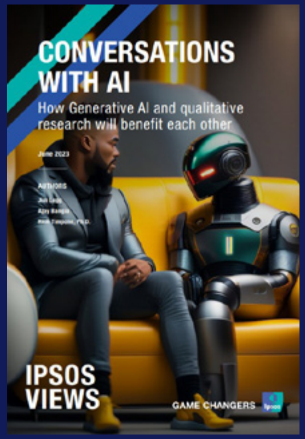
[1] Ipsos. Future of Mobility: Autonomous driving and the impact on our life.

[2] Ipsos. Exploring the Changing AI Landscape: From Analytical to Generative AI.

[3] European Commission. AI Act.

Further Reading

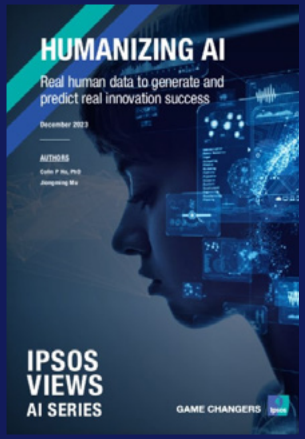
EXPLORING THE CHANGING AI LANDSCAPE
From Analytical to Generative AI
IPSOS VIEWS
GAME CHANGERS


CONVERSATIONS WITH AI
How Generative AI and qualitative research will benefit each other
IPSOS VIEWS
GAME CHANGERS

CONVERSATIONS WITH AI: PART II
Unveiling AI quality in qualitative workstreams
IPSOS VIEWS
AI SERIES
GAME CHANGERS

CONVERSATIONS WITH AI PART III
Bright sparks and bold ideas: how AI boosts human creativity in ideation workshops
IPSOS VIEWS
AI SERIES
GAME CHANGERS

IPSOS VIEWS
THE NEW ERA OF INNOVATION
Shattering the Stage Gates with Generative AI


HUMANIZING AI
Real human data to generate and predict real innovation success
IPSOS VIEWS
AI SERIES
GAME CHANGERS

JULY 2024

SYNTHETIC DATA

From hype to reality – a guide to responsible adoption

AUTHORS

Michel Guidi
Chief Operating Officer,
Ipsos

Ciprian Sava
Global Service Lines Head,
Digital Transformation and
Product Implementation, Ipsos

Benoit Hubert
Global GenAI Research and
Science Officer, Ipsos

Rich Timpone, Ph.D.,
Head of Global Science
Organization, Ipsos

The **IPSOS VIEWS** white papers are produced by the **Ipsos Knowledge Centre**.

www.ipsos.com
@Ipsos