Milliman

Accidents that never happened: Generative AI and fraud in motor insurance

December 2025

AUTHORS
Bartosz Gaweda
Raymond van Es
Leonid Zeldin
Jan Thiemen Postema
Tristan Stoevelaar

Table of contents

Executive summary: Generative AI images and insurance fraud

The rapid rise of generative artificial intelligence (generative AI) technology is fundamentally reshaping the insurance industry, offering both significant opportunities and new, complex risks. This research project investigates the emerging threat of AI-generated car photos, focusing on how fraudsters exploit generative AI to fabricate or manipulate images and evidence for fraudulent claims. The study aims to understand the scale of this risk, evaluate the effectiveness of current detection tools and human assessments, and outline practical strategies for insurers to safeguard against these evolving threats.

The primary objective of this research is to assess the impact of generative-AI-enabled visual deception on the car insurance sector. The project covers three issues:

- The current landscape of AI-driven image fraud, including the ease of generating convincing fake car damage photos with widely available generative AI tools
- The effectiveness of both human and automated off-the-shelf detection methods in identifying AI-generated images within insurance workflows, including their limitations and privacy considerations
- Recommendations for strengthening insurance processes to mitigate the risk of AI-enabled fraud, from underwriting to claims handling

KEY FINDINGS

- **Rising fraud risk:** Generative AI has dramatically lowered the barrier for creating realistic fake damage photos and supporting documents, leading to a surge in fraudulent activity. Across Europe, incidents of AI-driven insurance fraud have increased sharply, with deepfake-related attacks rising year-to-year by over 1000% in some markets.[1]
- **Human detection is insufficient:** The human ability to spot AI-generated car damage images is no better than random chance, with accuracy rates around 50%. Even experienced insurance professionals using fraud detection tools frequently misclassify fake images.
- **Automated off-the-shelf tools offer mixed results:** Although metadata analysis offers quick wins, it is easily bypassed by sophisticated fraudsters. Advanced forensic tools (e.g., Illuminarty, Hive) outperform others but still struggle to detect fakes from unfamiliar or newer generative models, especially when watermarks are absent or removed.
- **Privacy and data security concerns:** Many online detection tools require uploading sensitive customer data, raising significant privacy issues. Insurers must carefully evaluate vendor terms and prioritize stable tools that align with data protection standards.
- **Strengthening insurance workflows:** Fraud can occur at both policy inception (with clean images of pre-existing damage) and claims submission (with fabricated or exaggerated photos). Weak controls and reliance on single images or unverifiable documentation increase exposure. With a layered defense, combining procedural controls, technical verification, and AI-based tools (off-the-shelf or specialized models developed in-house), exposure to these types of fraud can be significantly reduced.

Ultimately, as generative AI technology continues to evolve, ongoing investment in detection capabilities will be essential for insurers to defend against this new era of deception. This investment is essential not just for car insurance but for all product lines.

Milliman provides a comprehensive solution along the whole insurance chain starting from risk-level assessment via process understanding and historical images analysis, identification of weak spots, and designing personalized robust risk-mitigation frameworks. **Contact us to evaluate your processes and historical imagery for potential vulnerabilities and optimization opportunities.**

1. Informa Insurance News 24. (2011, July 15). GDV finds high fraud level among German claimants. https://www.i-law.com/ilaw/doc/view.htm?id=270512.

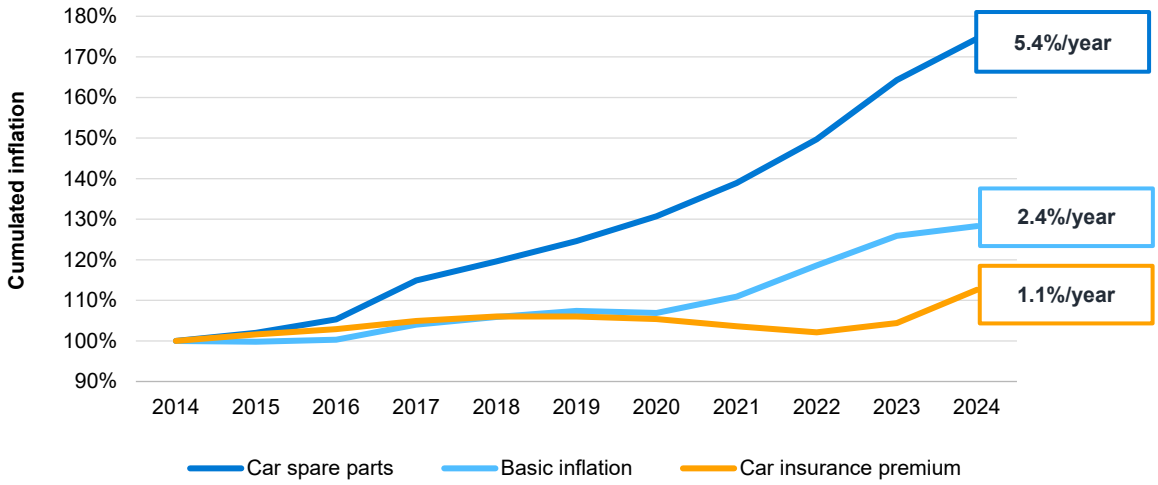
Introduction: A new age of visual deception in insurance

The rapid rise of generative AI is transforming the insurance industry. Insurers now use advanced AI tools to streamline claims processing, automate assessments and enhance customer service by analysing unstructured data such as images and text. However, the same technologies also pose significant risks. Fraudsters exploit generative AI to create convincing fake images, documents, and police reports, making it increasingly difficult to distinguish real claims from fabricated ones.

Insurance fraud is already a major problem. In Germany alone, damages total about €4 billion per year, with half coming from motor insurance.[2] Deepfake-related fraud is rising sharply. In early 2025, identity document fraud in Europe increased by 300%, whereas Germany saw a staggering 1100% surge in deepfake attacks. A report from June 2025 highlights that 92% of insurers have encountered AI-driven fraud, and 10% have lost over $1 million due to such incidents.[3] The trend is shifting from claim exaggeration to complete fabrication, including synthetic scenes and false evidence.

The motor or car insurance sector is particularly vulnerable. Claims involving AI-manipulated photos have tripled, affecting major UK insurers, such as Allianz and LV= (Liverpool Victoria, UK).[4][5] Mobile-first claims systems are especially at risk due to the ease of submitting digital evidence. At the same time, repair costs for vehicles have grown much faster than inflation, whilst premium prices have lagged—putting additional financial pressure on insurers,[6] as Figure 1 shows.

FIGURE 1: DEVELOPMENT OF THE CAR INSURANCE PREMIUM PRICE COMPARED TO THE CAR PARTS PRICE IN GERMANY BETWEEN 2014 AND 2024



Globally, approximately 10% of all paid insurance claims are estimated to be fraudulent, a number likely to rise as deepfake abuse becomes more common. Examples include AI-generated photos of cracked windshields, staged accidents and even claims for vehicles already damaged before the policy start date but shown in 'pristine' condition at inception using generative AI. The second fraud type is especially dangerous because the damage is real, so there is no need to fabricate repair invoices or police reports.

2. Informa Insurance News 24. (2011, July 15). GDV finds high fraud level among German claimants. https://www.i-law.com/ilaw/doc/view.htm?id=270512.
3. Swiss RE Institute. (2025, 12 June). How deepfakes, disinformation and AI amplify insurance fraud. https://www.swissre.com/institute/research/sonar/sonar2025/how-deepfakes-disinformation-ai-amplify-insurance-fraud.html.
4. Jervis, T. (2024, May 3). AI drives major rise in car insurance fraud as criminals fake evidence. Auto Express. https://www.autoexpress.co.uk/news/363070/ai-drives-major-rise-car-insurance-fraud-criminals-fake-evidence.
5. Clatworthy, B. (2024, May 1). How insurance fraudsters are faking damage to cash in on claims. The Times. https://www.thetimes.com/business-money/money/article/how-insurance-fraudsters-are-faking-damage-to-cash-in-on-claims-fjk2887kq.
6. Gesamtverband der Versicherer. (2024, July 10). Pkw-Ersatzteilpreise: Autohersteller langen zu. (2024, 10 07). https://www.gdv.de/gdv/medien/medieninformationen/pkw-ersatzteilpreise-autohersteller-langen-zu-153470.

Detecting these sophisticated fakes is a major challenge. Studies show that people can spot AI-generated images only half the time—no better than random chance.[7] Even experts struggle with domain-specific forgeries, and as we will show in the section covering AI detection, general tools often fail in insurance contexts. This makes specialised, automated AI-based fraud detection systems essential for the insurance industry.

As generative AI enables ever more convincing visual deception, insurers must invest in robust tools to verify the authenticity of submitted evidence. This report explores how generative AI can be misused for fraudulent claims and outlines the technical and strategic responses needed in an era where visual proof can no longer be taken at face value.

A picture is worth a thousand words—but now they might all be lies

Generating images of car damage has never been easier, with a growing range of free and paid tools offering ever-improving quality. We tested the most popular tools to see how well they handle car damage generation:

- Runway ML
- Adobe Firefly
- Midjourney
- OpenAI (including Sora, DALL-E, GPT-1-image)
- Leonardo.Ai
- BRIA
- Stable Diffusion
- Grok
- DreamStudio

GENERATIVE AI IMAGE RESEARCH APPROACH

To create fake images, we followed a few general approaches:

- Pure generation based on a prompt for a given car model, age, colour and damage type
- Recreating or editing an existing car photo with adjustments, for example:
  - adding damage
  - increasing the amount of damage
  - removing damage

Proper prompt engineering, as described in the literature and following some experimentation, helped to avoid issues such as poor quality of AI-generated text (plate numbers, model names) or non-logical items (see the duplicated Renault logo on the rooftop cargo carriers in Figure 2). It was easy to generate the exact car model, colour, age and type of damage we wanted. Most of the tools did not require sophisticated prompts to avoid generator guardrails and allowed us to generate fake images quickly. Photo characteristics, such as image quality, level of realism, weather conditions, surroundings, level of detail or even dirt on the car, could be adjusted.[8]

FIGURE 2: EASILY RECOGNISABLE AI-GENERATED IMAGES

Sources: Image on the left created with ChatGPT; image on the right created with Dream Studio

7. Cooke, D., et al. (2025, April 10). As good as a coin toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli. arXiv. https://arxiv.org/abs/2403.16760.

8. Lu, Z., et al. (2023, April 25). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. arXiv. https://arxiv.org/abs/2304.13023.

Available tools vary in their ability to upload and edit images. Some apps add a watermark in the metadata (e.g., OpenAI and Adobe Firefly). Grok adds a visible watermark in the image. The output quality (especially regarding AI-generated text), the ease of obtaining good results and commercial limitations led us to conclude that further testing was required, mostly using images generated with OpenAI gpt-1-image model (associated with ChatGPT) and Stable Diffusion, which runs in the Perchance.org AI generator. With these apps, we created an annotated dataset of over 150 AI-generated images containing such damage as dents, scratches, flat tires, roof damage and broken windows.

Real or fake—car damage image quiz

Recent research has shown that it's becoming difficult for humans to identify AI-generated content. This is true in every aspect of life, not least when handling car damage claims. As already mentioned, human detection of AI-generated content has a success rate that ranges between 50%[9] and 82%.[10] This wide range can partly be explained by factors such as the object type (e.g., people are good at discovering fake face photos but they fail to recognise fake landscapes or animals), the tool used and the respondent's age.

However, the research mentioned above does not have a specific focus. To test if this is the case when distinguishing between real and AI-generated images of car damage, we created 'Can You Spot the Fake?,' an online quiz targeted towards professionals in the insurance industry. In total, 925 respondents from 27 countries tried to identify whether the damage to the cars in the images was AI-generated. The quiz included 10 real photos from sources such as CarDD,[11] and commercial vendors (Unsplash and Pexels), as well as 10 AI-generated images from various image generators. See Figure 3 for sample images from the quiz.

The 'Can you spot the fake?' quiz is available to test the ability of insurance professionals and industry AI detection tools to spot AI-generated damage.

FIGURE 3: EXAMPLES OF REAL AND AI-GENERATED IMAGES USED IN THE QUIZ

9. Cooke. op. cit.

10. Ji, Y., et al. (2025, April 19). Towards explainable fake image detection with multi-modal large language models. arXiv. https://arxiv.org/abs/2504.14245.

11. Wang, X. (2023). CarDD: A new dataset for vision-based car damage detection [Excerpt]. Github. https://cardd-ustc.github.io/.

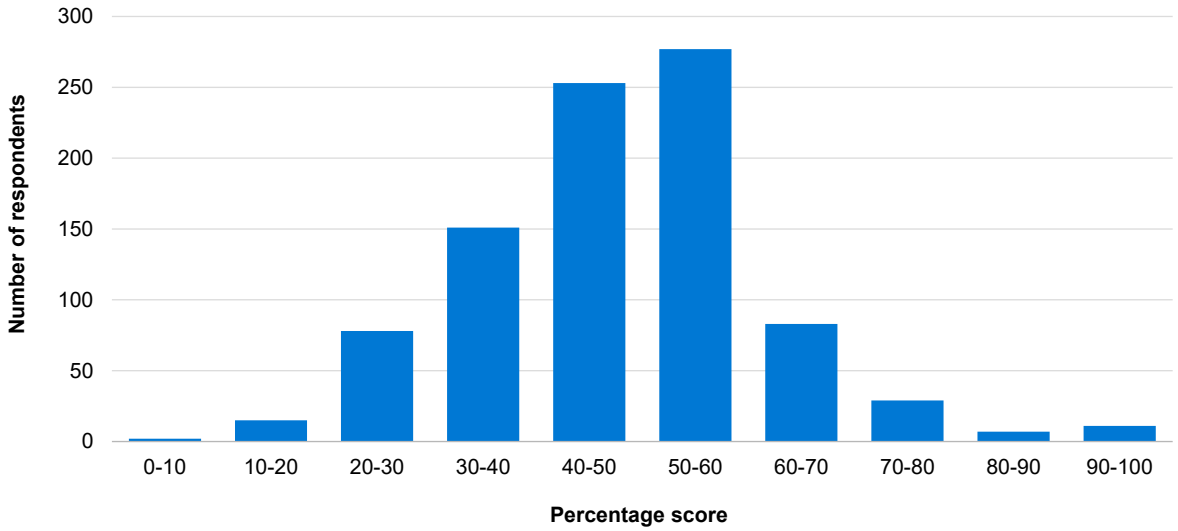
FIGURE 3 (CONTINUED): EXAMPLES OF REAL AND AI-GENERATED IMAGES USED IN THE QUIZ

QUIZ RESULTS

The quiz results confirm that the human ability to distinguish between real and AI-generated images of car damage is limited. Across 925 respondents, scores ranged from 5% to 100%, with an average of 46.4% and a median of 45%. The standard deviation of 15.9 percentage points indicates that most respondents did no better than they would by flipping a coin, typically scoring between 30% and 60%. A small number of outliers reached near perfect or very low scores, but these were exceptions rather than the rule.

Figure 4 shows the distribution of the scores. The vertical bars show the number of respondents whose scores fall within each 10% range, providing an overview of the overall score distribution.

FIGURE 4: RESULTS OF THE 'CAN YOU SPOT THE FAKE?' QUIZ



Performance also varied substantially across images. Although a few images were correctly classified by more than 65% of respondents, others were only identified correctly by approximately one quarter of the respondents. Figure 5 shows the two lowest-scoring AI-generated images. The image on the left below was misclassified in 69% of answers, whereas the image on the right was misclassified 77% of the time, even though the reflection in the car window provides a good indication of it being a fake image.

FIGURE 5: THE LOWEST SCORING IMAGES IN THE 'CAN YOU SPOT THE FAKE?' QUIZ

Sources: Image on the left created with ChatGPT; image on the right created with Perchance.org

Figure 6 shows that respondents made more mistakes trying to classify AI-generated images compared to real ones. Such a situation is especially dangerous in the insurance context, as fraudulent cases might not be discovered.

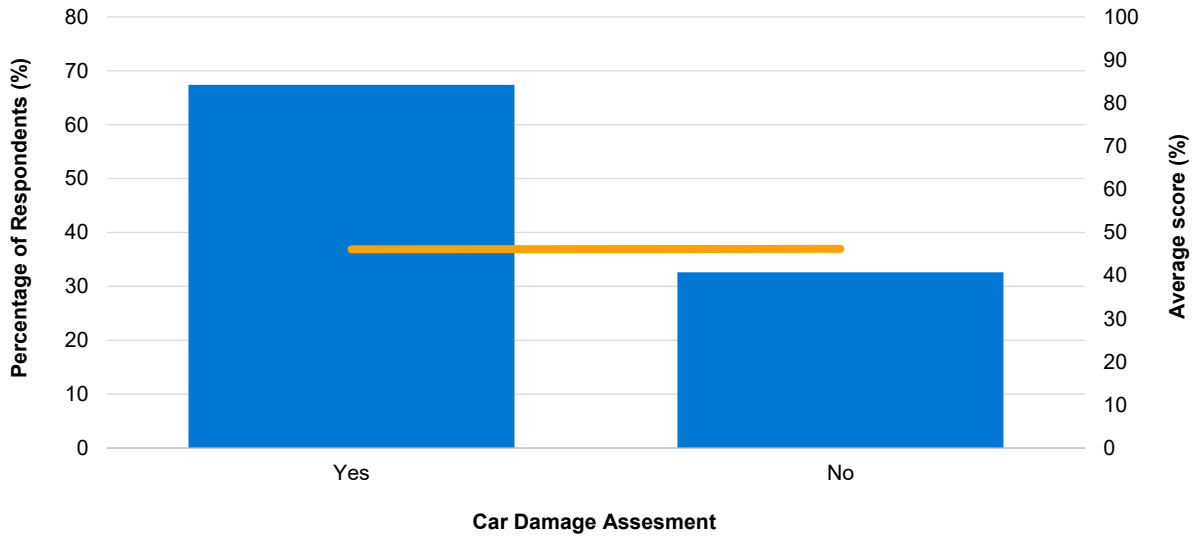
FIGURE 6: SCORES ACHIEVED WHEN CLASSIFYING REAL VS. AI-GENERATED IMAGES

IMAGE TYPE	AVERAGE SCORE	MEDIAN SCORE
Real	55%	50%
Generated	38%	40%

There is no significant difference in a score between images with a car damage and images where the damage was removed (47% vs. 46%, respectively).

Using standard statistical methods, we tested if certain respondent characteristics had explanatory power (for example, if younger people or people with claim handling experience answered better). We found no such characteristics that prove each of us is vulnerable to this kind of threat. Even the best scoring people (90–100%) did not reveal any special characteristics (such as fraud detection tool usage). More details on the used methodology are provided in the Appendix.

FIGURE 7: RESPONDENTS AND SCORES BY CAR DAMAGE ASSESSMENT



In Figure 7, the vertical bars show the percentage breakdown of respondents who assess car damage as part of their job, with the left bar representing the percentage of respondents who answered that they do asses car damage as part of their job, and the right bar showing the percentage of those who do not. The line indicates the average score for each group. A full breakdown of average scores per variable, as well as the distribution of respondents across these categories, is provided in the appendix.

Taken together, these results demonstrate that human judgment alone is insufficient for reliably detecting AI-generated car damage imagery. Even trained professionals using current fraud detection tools tasked with such assessments perform at chance level in this quiz. This underlines the necessity of supporting human decision-making with automated forensic tools and workflow safeguards if insurers are to effectively counter AI-enabled fraud.

Ways to detect AI-generated image fraud

The theoretical background of detecting AI-generated images is a rapidly evolving — but in parts still obscure — research field.

At this stage, insurers have several options to detect AI-generated images:

1. Performing a metadata analysis
2. Using an off-the-shelf image forensics tool for AI-generated image detection
3. Using multimodal large language models (MLLMs) with advanced prompt engineering
4. Training an anomaly detection model to identify abnormal = AI-generated images
5. Training an image-based classifier which can be trained for this specific purpose

In this report, we will discuss the first three methods in more detail. Although metadata analysis is often considered in practice, its reliability is limited since metadata can be easily modified, as discussed in the following paragraph. After the pros and cons of metadata analysis, we introduce several off-the-shelf tools and briefly outline some of the methods they apply (including MLLMs) as well as their effectiveness.

METADATA ANALYSIS TO DETECT FRAUD

Digital photographs taken with smartphones or cameras usually contain a lot of metadata, such as timestamp, resolution, file type, and more. If this information is missing, manipulated, or even explicitly indicates AI usage, it can serve as a strong hint that the image is not authentic.

As an example, FotoForensics extracts and analyses available information. Using such solutions at the beginning of an analysis can quickly identify AI-generated content, particularly when created by inexperienced forgers. However, as noted by OpenAI[12]:

> Metadata like C2PA is not a silver bullet to address issues of provenance. It can easily be removed either accidentally or intentionally. For example, most social media platforms today remove metadata from uploaded images, and actions like taking a screenshot can also remove it. Therefore, an image lacking this metadata may or may not have been generated with ChatGPT or our API. (OpenAI, 2025)

Most currently available generators do not leave any tracks in the metadata. However, if any of the tools below are used to test images created by OpenAI or Adobe Firefly, they will immediately reveal whether the images are real or AI-generated. Additionally, these tools allow for further testing, such as Error Level Analysis (ELA), timestamps and resolution.

- Content Credentials
- MetaData2Go
- EXIF / File Metadata Viewer
- FotoForensics

12. OpenAI. (2025). C2PA in ChatGPT Images. OpenAI. C2PA in ChatGPT Images | OpenAI Help Center

FIGURE 8: SAMPLE TEST RESULTS FROM FOTOFORENSICS

JUMBF	
JUMD Type	(c2pa)-0011-001(
JUMD Label	c2pa
C2PA Thumbnail Claim Jpeg Type	image/jpeg
C2PA Thumbnail Claim Jpeg Data	(Binary data 179[
Actions Action	c2pa.created
Actions Software Agent	Adobe Firefly

Source: FotoForensics.

Some companies, such as Google, are considering leaving invisible watermarks hidden in pixels (SynthID). Although such approaches could help with fraud detection, the generators (e.g., Vertex AI) are not publicly available, so they cannot currently be used by the insurance industry. It is necessary to follow the development of regulations that are trying to keep pace with the emerging risks.

To summarise, in the insurance industry, metadata analysis can be used only for 'quick wins'; however, as it is easy to manipulate metadata, we need to seek out more sophisticated methods.

IMAGE FORENSICS TO DETECT FRAUD

If metadata analysis does not provide a clear result, more advanced tools and methods can be applied that are based on the analysis of the image. These solutions rely on several modern AI techniques to detect subtle artifacts left by generative models:

Illuminarty: This is a hybrid framework of computer vision and deep learning, including patch-level feature analysis and Contrastive Language–Image Pretraining (CLIP) embeddings to highlight suspicious regions and even identify the likely source AI model. This corresponds with recent methods that combine semantic and frequency features (e.g., the AIDE framework), which have shown superior performance on challenging benchmarks such as Chameleon and GenImage.[13]

AI or Not: This tool uses deep convolutional neural networks (CNNs) combined with frequency-domain analysis to identify statistical inconsistencies between real and synthetic images. In benchmark studies, it achieved up to 97% precision and recall for AI-generated content.[14] Such approaches are consistent with research showing that combining spatial and spectral features improves robustness.[15]

In addition, general scientific advances are increasingly being integrated into such tools:

- **Wavelet and transformation-based analysis** to detect hidden frequency artefacts[16]
- **Vision transformers and attention-based mechanisms** to identify global inconsistencies across image regions[17]

Together, these methods enable modern tools to go far beyond metadata analysis, offering deeper insights into the structural and semantic inconsistencies that distinguish AI-generated content from real images.

13. Mahara, A., & Rishe, N. (2025, February 21). Methods and trends in detecting generated images: A comprehensive review. arXiv. https://arxiv.org/abs/2502.15176.

14. Gillham, J. (2025, October 11). Do AI image detectors work? Accuracy study. Originality.ai. https://originality.ai/blog/do-ai-image-detectors-work-accuracy-study?utm_source=chatgpt.com.

15. Gallagher, J., & Pugsley, W. (2024, June 19). Development of a dual-input neural model for detecting AI-generated imagery. arXiv. https://arxiv.org/abs/2406.13688.

16. Mahara. op. cit.

17. Zhang, Y., et al. (2025). Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. Journal of King Saud University Computer and Information Sciences,37(148). https://doi.org/10.1007/s44443-025-00154-8.

Which image forensics tools are most effective?

To assess the effectiveness of AI image detection tools, we tested a range of publicly available models. The goal was to evaluate how well they identify AI-generated images in a typical insurance fraud context. The images we used were not identical to those in the earlier quiz but came from the same sources. We tested the following tools:

- Illuminarty
- AI Detect
- Hive Moderation
- isgen.ai
- Decopy AI
- Sightengine
- CNN Detection
- WasItAI
- MLLM-Defake
- AI or Not

We included 20 real images from the Car Insurance Fraud Detection Kaggle dataset[18] and 27 AI-generated images: 11 from OpenAI, 11 from Perchance/Stable Diffusion, 2 from Bria and Stable Diffusion 3.5, and lastly, a single image from Adobe Firefly. (A larger test of the top performer follows.) We chose these generators, as the resulting images seemed to be the most realistic, and these tools provided the easiest-to-produce images. OpenAI and Adobe Firefly include a watermark in the photo metadata by default, which reveals the photo source; this is not the case for the other apps.

Each tool returns a probability score indicating how likely an image is AI-generated. To make results comparable across models, the optimal threshold for each detector was determined individually by maximising the AUC–ROC. This approach is common in machine learning, as it identifies the point where the model best balances true positives and false positives across all possible thresholds.

Figure 9 compares the percentage of images classified correctly across three categories: real images, AI-generated images with watermarks, and AI-generated images without watermarks (either by design or because we removed them).

FIGURE 9: PERCENTAGE OF IMAGES CLASSIFIED CORRECTLY ACROSS THREE CATEGORIES

MODEL	REAL PHOTOS	AI-GENERATED WITH WATERMARK	AI-GENERATED WITHOUT WATERMARK	OVERALL
Illuminarty	100%	82%	94%	94%
Hive	95%	91%	63%	83%
Sightengine	90%	91%	56%	79%
AI or Not	85%	100%	44%	74%
AI Detect	45%	82%	100%	72%
CNN	95%	91%	6%	64%
isgen.ai	0%	100%	100%	57%
Decopy AI	100%	18%	13%	51%
MLLM	69%	28%	18%	47%
WasItAI	25%	100%	25%	43%

Illuminarty, Hive, and Sightengine were the top performers. All three showed strong results overall, but only Illuminarty maintained high accuracy on AI-generated images without watermark cues.

18. Mishra, P. K. (n.d.). Car insurance fraud detection. Kaggle. Retrieved October 15, 2025, from https://www.kaggle.com/datasets/pacificrm/car-insurance-fraud-detection.

We constructed confusion matrices for Illuminarty and Hive to better understand how these models behave. These matrices show the balance between false positives (the model incorrectly identified a real image as being AI-generated) and false negatives (the model did not detect AI-generated content).

FIGURE 10: BALANCE BETWEEN FALSE POSITIVES AND FALSE NEGATIVES USING ILLUMINARTY

ILLUMINARTY CONFUSION MATRIX	PREDICTED: AI-GENERATED	PREDICTED: REAL
Actual: AI-generated	24	3
Actual: Real	0	20

Illuminarty made zero false positives and misidentified only three AI-generated images out of 27. It handled both watermarked and unmarked AI images well.

FIGURE 11: BALANCE BETWEEN FALSE POSITIVES AND FALSE NEGATIVES USING HIVE

HIVE CONFUSION MATRIX	PREDICTED: AI-GENERATED	PREDICTED: REAL
Actual: AI-generated	20	7
Actual: Real	1	19

Hive missed seven AI-generated images out of 27 and flagged one real image incorrectly. Still, it offered a solid overall performance.

By contrast, several tools underperformed. WasItAI, and isgen.ai often misclassified real images. Decopy failed entirely on AI-generated content. CNN, an open-source code we reused in Python that is supposed to be one of the best in class, failed entirely on unmarked AI-generated content.

As these tools are being constantly improved, the results might be different if the test is conducted at another time.

TESTING THE IMAGE FORENSICS FRONT-RUNNER: ILLUMINARTY

Since Illuminarty came out as the top performer in our initial comparison, we evaluated it further on a larger and more diverse sample. The AI-generated set consisted of 134 images drawn from a broad mix of generators, namely Bria, DreamStudio, Firefly, Grok, OpenAI (gpt-1-image), DALL-E, OpenAI SORA, Perchance Stable Diffusion, Runway, and Stable Diffusion 3.5 Medium API. The real photo set contained 2,235 images coming from the Kaggle Dataset.[19] As we describe more in detail in the section below titled Privacy concerns and solution stability, this AI Detector tool appears to have technical issue as of November 2025 and does not allow for scalable detection.

FIGURE 12: BALANCE BETWEEN FALSE POSITIVES AND FALSE NEGATIVES USING ILLUMINARTY ON A LARGE AND DIVERSE SAMPLE

CONFUSION MATRIX	PREDICTED: AI-GENERATED	PREDICTED: REAL
Actual: AI-generated	112	22
Actual: Real	210	2025

In the extended test, the threshold was recalibrated by maximising the ROC–AUC, which resulted in an optimal threshold of 11% compared to 15% in the smaller test. At this threshold, the confusion matrix shows that 112 of 134 AI-generated images were correctly flagged, with 22 misses (false negatives ≈16%). On real photos, 210 of 2,235 were incorrectly flagged as AI (false positives ≈10%), yielding an overall accuracy of approximately 90%. The model captures most AI-generated images, but at the cost of a moderate false positive rate on real photos.

19. Mishra. op. cit.

Compared to the smaller test, Illuminarty underperformed on the extended test. The earlier results were based on a limited and more homogeneous set of images, in that they were generated from a more restricted set of generators. This possibly made classification easier. In the extended test, the dataset was larger and more diverse, which increased the difficulty of detection and resulted in more false positives and false negatives. Nevertheless, Illuminarty offers useful coverage of AI-generated content whilst maintaining a manageable, though not negligible, rate of misclassification on real photos.

CAN TOOLS DETECT NON-AI-GENERATED PHOTOS?

As an additional experiment, we ran an exploratory check on non-AI-generated (i.e., real) photos taken from the Kaggle dataset,[20] and compared images labelled as fraud and non-fraud. The average Illuminarty score was modestly higher on fraud images (~6.9%) than on non-fraud images (~4.6%). Although the absolute scores are low, the result suggests the detector may pick up small artefacts left by edited or manipulated photos (e.g., using Photoshop). A Fisher's exact test on this dataset yielded a *p* value of 0.0108, implying a statistically significant difference, assuming a threshold of 0.05.

ANOTHER APPROACH: EXPLAINABLE FAKE IMAGE DETECTION

Source: Image created with Perchance.org

One additional approach worth mentioning was described in the paper 'Towards Explainable Fake Image Detection with Multi-Modal Large Language Models.'[21] The assumption discussed in the paper is that an appropriate set of independent prompts, when submitted to an MLLM (e.g., ChatGPT-4o, GPT-5, or Llama), yields a satisfactory detection rate. Additionally, such an approach explains the reasoning which claim handlers or underwriters can use when making the final confirmation.

Unfortunately, our tests on fraudulent car photos revealed that the tool's accuracy is far worse (as you can see in Figure 10 above, the accuracy is below 50%) than the results presented by the paper's authors based on a general set of photos (close to 93%). The results were poor even after adjusting prompts to align them with the specific task (e.g., we suggested typical flaws in AI-generated images in the insurance context and delivered exemplary real and fake car photos with the correct answers [i.e., one-shot prompting]). Sometimes, simply running the prompt again yielded a completely opposite answer with convincing arguments.

However, we recommend not avoiding this method completely. Firstly, it can be combined with a detector that has better accuracy, and secondly, the ability of LLMs to correctly identify images is constantly increasing, so improved MLLM methods are likely to come.[22,23,24,25] The authors also described in their analysis that they applied reinforcement learning to better prompt to the model to provide the desired answers. (Unfortunately, the resulting model weights are not publicly available; therefore, we could not test them).

20. Mishra. op. cit.

21. Ji. op. cit.

22. Ji, Y. (2025, June 8). Interpretable and reliable detection of AI-generated images via grounded reasons in MLLMs. arXiv. https://www.arxiv.org/abs/2506.07045.

23. Tan, C., et al. (2025, August 2). ForenX: Towards explainable AI-generated image detection with multimodal large language models. arXiv. https://arxiv.org/abs/2508.01402.

24. Zhou, Z., et al. (2025, July 3). AIGI-Holmes: Towards explainable and generalizable AI-generated image. arXiv. https://arxiv.org/abs/2507.02664.

25. Kang, H., et al. (2025, March 19). LEGION: Learning to ground and explain for synthetic image detection. arXiv. https://arxiv.org/abs/2503.15264.

These results align with recent findings from AI-GenBench, which showed that most detectors struggle with generalisation.[26] They perform well on images generated by tools seen during training, but accuracy drops with newer or less common sources. Furthermore, we should expect the detection rate to be even lower if we apply common or adversarial perturbations on the AI-generated images, which make them harder to be detected.[27] This is especially relevant for insurers, where fraud attempts often involve unfamiliar generation styles to evade detection.

REVEALING SIGNS AND LIMITATIONS OF AI IMAGE DETECTION

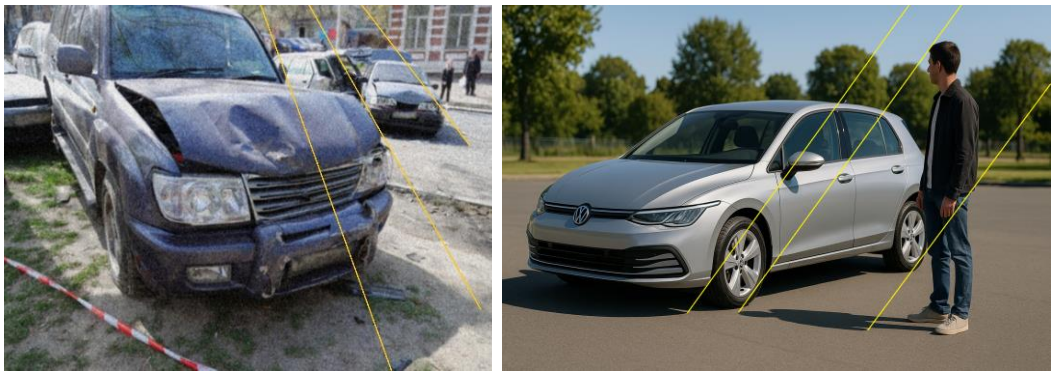
Detecting AI-generated images often starts with subtle visual cues. As shown in the study *'Seeing is not always believing*,' malformed or inconsistent text is one of the most reliable indicators of an image having been AI-generated.[28] License plates, brand logos, and model names remain challenging for generators to reproduce accurately, often resulting in warped, unreadable, or mismatched outputs. Other signals include faulty reflections, unnatural shadows, or abrupt texture transitions, particularly around the supposed area of damage. These flaws are more easily spotted at high resolution and can serve as forensic triggers.

Scalable detection requires automated tools. These methods work for unaltered, off-the-shelf generations, but are fragile and might not always work well if bad actors are well-prepared. In a recent TED talk, Hany Farid strongly reinforced this view.[29] Farid, a digital forensics expert, highlighted in his presentation how online AI detection tools often fail and warned against relying on them. He demonstrated forensic techniques, such as noise pattern analysis and vanishing point and shadow checks, which can help distinguish real from AI-generated images. Vanishing point and shadow checks are very similar. Parallel edges should converge on a single vanishing point in the image. By the same logic, for a single light source each line drawn from an object contact point to the corresponding point on its shadow should intersect at the light position (for sunlight, the lines are parallel and share one consistent direction); if those 'shadow rays' fail to meet at one point, or point in inconsistent directions, the illumination is physically impossible. An example of this is demonstrated in Figure 13.

Farid concluded that most tools available to the public are essentially ineffective. His closing argument was clear: Scalable detection cannot be outsourced to simple online tools. Robust forensic analysis and structural safeguards are essential. Additional post-training for a specific context might be necessary.

Figure 13 shows a comparison of shadow consistency in two images. The left image is a real photograph with physically plausible shadows. The right image is AI-generated and exhibits inconsistent shadow rays that do not correspond to a single light source.

FIGURE 13: COMPARISON OF SHADOW CONSISTENCY



Sources: Image on the left retrieved from CarDD; image on the right created with OpenAI

26. Pellegrini, L., et al. (2025, April 29). AI-GenBench: A new ongoing benchmark for AI-generated image detection. arXiv. https://arxiv.org/abs/2504.20865.

27. Moyang G., et al. (2024, November 20). AI-generated image detection: Passive or watermark? arXiv. https://arxiv.org/abs/2411.13553

28. Lu. op. cit.

29. Farid, H. (2025, July 18). How to spot fake AI photos [Video]. TED Conferences. https://www.youtube.com/watch?v=q5_PrTvNypY.

PRIVACY CONCERNS AND SOLUTION STABILITY

When using web-based detectors, privacy is another concern, especially when uploading potentially sensitive client data. Illuminarty states clearly that images are only stored temporarily, solely for the purpose of providing the service. Sightengine's terms grant the company a broader license, allowing it to store, transmit, and display submitted data, including integration with other services as necessary for Sightengine to provide its services. Hive's terms likewise allow for information sharing with service providers and for it to be used to improve their service. Although these practices are not uncommon for cloud platforms, they highlight the need for caution when evaluating which tools to use in an insurance setting, as tools with broad data usage rights make them less suitable for handling client information, including images. Finally, the insurance companies must be sure that the solution is stable and will not, for example, be turned off at any time. For example, as of November 2025 our team is having significant problems with logging into Illuminarty account and based on online discussions it seems that such issue appears also for other users. We requested a comment from the solution provider but as of today we did not receive any response.

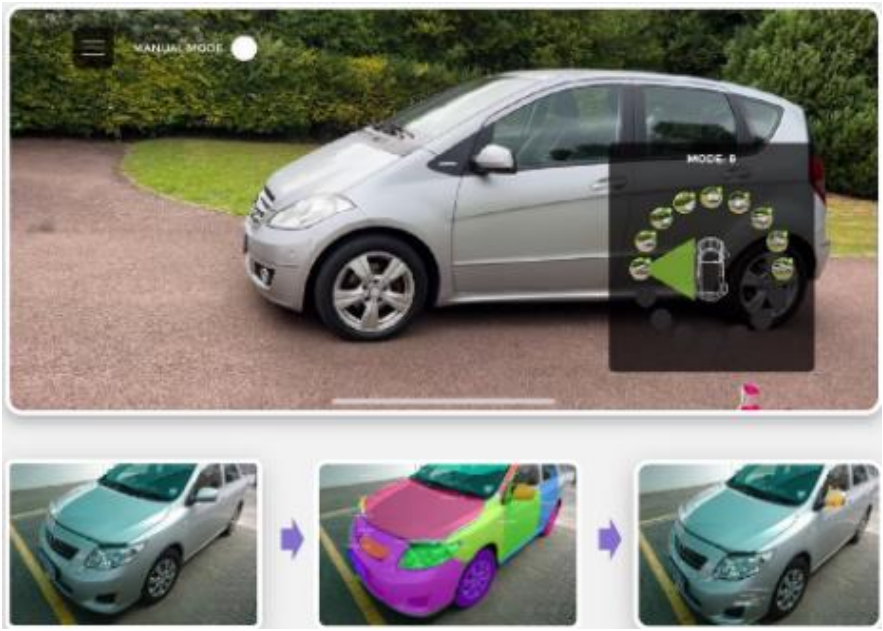
BEYOND DETECTION: STRENGTHENING THE INSURANCE WORKFLOW

Although AI detection tools can be useful, relying on them alone is not sufficient. Preventing AI-enabled fraud requires strengthening the entire insurance workflow. The specific solutions might vary between products, countries, deductible levels and local regulations.

A major vulnerability in underwriting arises when policyholders submit clean-looking images of pre-existing vehicle damage. This risk is amplified when insurers do not request visual documentation at policy inception or accept unverified images. On the claims side, fraud occurs when photos show fabricated or exaggerated damage.

To address this, insurers should verify if submitted visuals match the vehicle in make, model, colour and version. Visuals should ideally be collected at inception using a certified physical control point (e.g., ÖAMTC in Austria) or through a dedicated app (e.g., AiGenix) that requires photos to be taken directly within the app so that metadata such as timestamp, GPS location, and device ID can be captured reliably. Requiring proof of timing, such as having a visible newspaper in the shot, further helps filter out outdated or recycled photos and confirms that images are authentic and taken in real time.

FIGURE 14: EXAMPLE OF GUIDANCE FOR STANDARDIZED VEHICLE PHOTO CAPTURE, AIGENIX

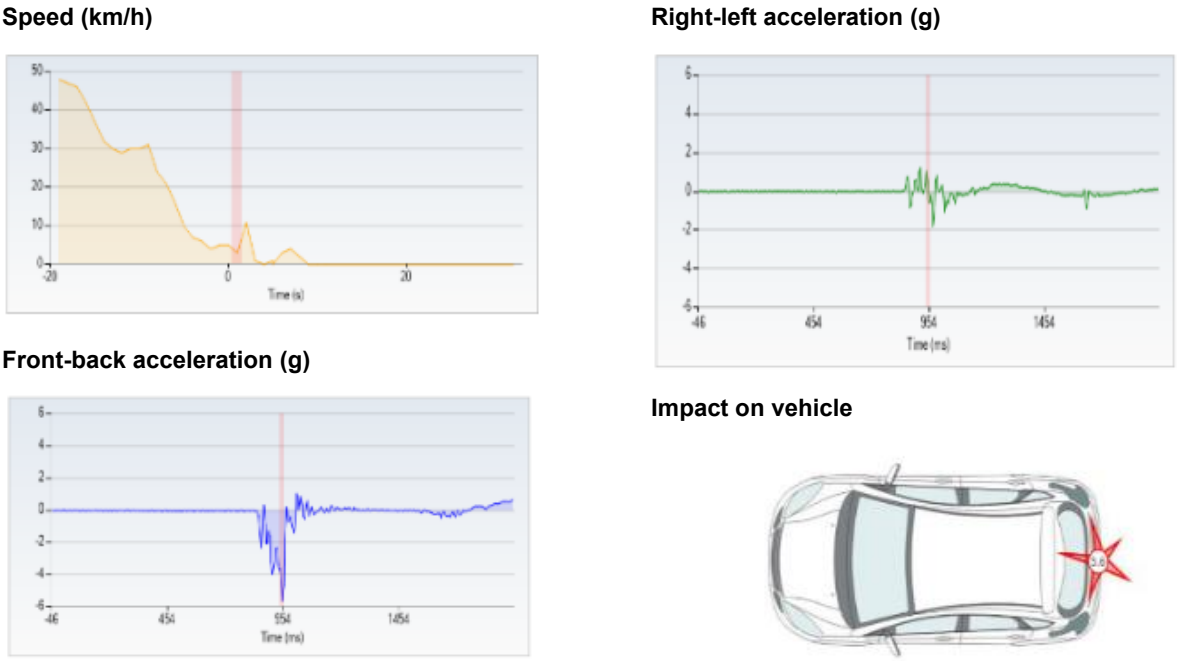


Source: AiGenix

Insurers should always request a full set of images rather than a single photo. Multiple, high-resolution images improve consistency and quality, and they are harder to fabricate. Insurers should also verify whether submitted images exist online by conducting a reverse image search. As a deterrence strategy, informing policyholders that AI detection systems are in place has been shown to reduce the frequency of fraud.

When processing claims, photos should be checked against those taken at inception. A video of the claimant describing the incident provides an additional safeguard, especially if tested using AI lie detection tools. Advanced telematics, such as crash reconstruction or G-force analysis, can strengthen this verification. Insurers should request additional evidence, such as police reports, repair invoices and timestamped footage. Risk can also be reduced by calling garages, hospitals, or requesting access to city cameras. Even small claims should be sampled for additional checks. In cases where suspicion remains, escalation to a human claim handler is necessary.

FIGURE 15: EXAMPLE OF TELEMATICS-BASED CRASH RECONSTRUCTION, AIGENIX



Source: AiGenix

Together, these measures create a layered defense against fraud involving synthetic images. They can be grouped into three complementary layers: procedural controls, technical verification, and AI-based tools. Each layer addresses different vulnerabilities in the workflow, and, together, they provide a stronger defence against AI-enabled fraud. At the same time, companies need to optimise costs because sending a claim handler for physical examination or cooperation with certified control points, for example, might be too expensive, especially for minor claims and cheaper policies.

What insurers can do about generative AI image fraud

As generative AI continues to revolutionise the insurance landscape, fraudsters will be leveraging increasingly sophisticated technologies to fabricate visual evidence and exploit vulnerabilities. The stakes for insurers—financially and reputationally—have never been higher.

In the introduction, we highlighted that unchecked AI-driven fraud can result in millions of euros in annual losses for motor insurers, with some markets experiencing a tenfold increase in deepfake-related claims. This is supported by our findings in the third section, showing the ease with which high-quality fake images can be generated.

Our car damage image quiz clearly shows that insurers cannot rely on human detection alone. The ability for humans to spot AI-generated car damage in images is no better than random chance, with accuracy rates of approximately 50%. Even experienced insurance professionals using their company's proprietary fraud detection tools frequently misclassify fake images. By deploying (advanced) detection tools, as shown in 'Ways to detect AI-generated image fraud' above, insurers can improve these results significantly with detection rates in the range of 80%–90%. This translates to potential savings in millions of euros annually for a mid-sized insurer, when factoring in prevented fraudulent payouts and reduced investigation costs. Employing detection tools, however, requires uploading sensitive customer data to a detection app, raising significant privacy issues. Therefore, insurers must carefully evaluate the terms of different apps and prioritise those that align with their data protection standard. Alternatively, insurers can (re)train their own proprietary models, which was not discussed because it is outside the scope of this research.

Our approach to reducing this type of fraud would be to implement a layered defence, combining procedural controls, technical verification, and task-tailored AI tools to strengthen the insurance workflow, as elaborated in the previous section. Since fraud can occur at both policy inception (with clean images of pre-existing damage) and claims submission (with fabricated or exaggerated damage in photos), both areas of the insurance value chain should be diligent.

By openly investing in advanced fraud prevention, insurers not only deter would-be fraudsters (who may seek easier targets elsewhere) but also strengthen relationships with honest customers. Policyholders are more likely to stay with companies that demonstrate robust protection of their interests, leading to higher retention rates and positive brand associations.

In the coming years, several trends will shape the future of fraud risk and detection:

- The continuous introduction of new advanced generators: New AI image generators are becoming increasingly accessible, multimodal and even harder to detect, enabling fraudsters to create ever more convincing forgeries with minimal technical expertise
- An increased use of harder to manipulate and invisible watermarks by the developers of AI-image generators
- Evolving regulations: As regulators respond to the proliferation of generative-AI-driven fraud, insurers will face stricter requirements for evidence verification, data protection and transparency
- The application of more sophisticated detection methods, including specialised models for insurance cases, either publicly available or developed in-house

One thing is certain: The arms race between fraudsters and defenders will continue, so insurers must remain one step ahead and invest in capabilities to combat this growing risk. **Contact us to assess your exposure to generative-AI-driven image fraud and develop tailored defences across your insurance processes.**

Appendix 1: Distribution method

The quiz to test the human ability to detect AI-generated images was distributed to:

- Milliman employees who also further distributed the link to their clients worldwide
- LinkedIn—on multiple Milliman profiles
- The Milliman P&C Clients contact list

To increase engagement, Milliman offered $10 gift cards to the first 300 non-Milliman respondents. It should be noted that the reward might have led to a partially biased result due to the potential for infiltration by bots (automatic agents populating such quizzes to obtain the rewards), which cannot be fully prevented. After checking input data quality (e.g., contact information) we are assessing this risk as low.

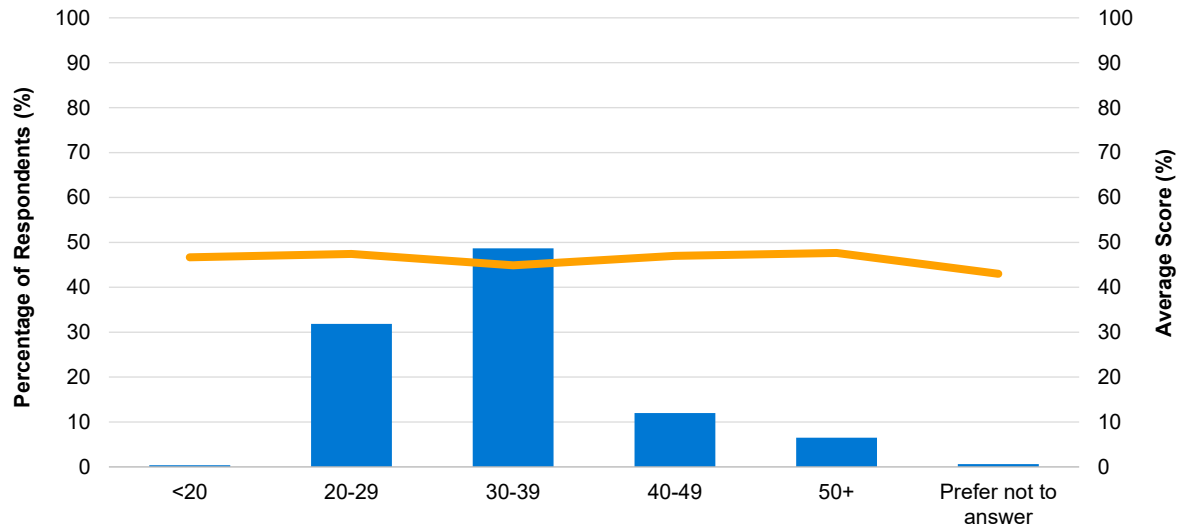
Appendix 2: Quiz results and sample structure

As already stated, almost 1,000 people from across the globe completed the quiz, which shows a huge interest in this topic. Nearly 11% of responses came from Milliman employees.

To test whether certain respondent characteristics had explanatory power, we set up both a linear regression model and an XGBoost model using demographic and professional variables such as age group, profession, professional seniority, geographic region, type of detection tools used, and prior experience with AI-generated images. Both models performed poorly, with out-of-sample R^2 values close to zero, indicating that none of these variables meaningfully predicted quiz performance. This implies that the difficulty lies in the images themselves rather than in the background or expertise of the respondents, or even the type of fraud detection software used by insurance companies.

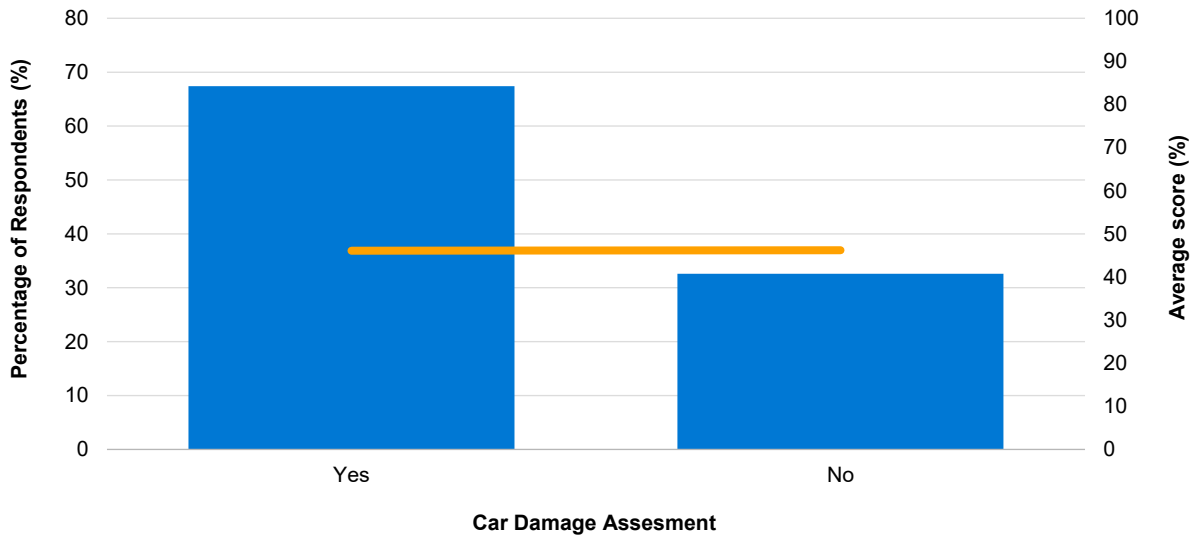
In Figure A2-1, the vertical bars show the percentage breakdown of respondents by age group and the line indicates the average score for each age group.

FIGURE A2-1: RESPONDENTS AND SCORES BY AGE GROUP



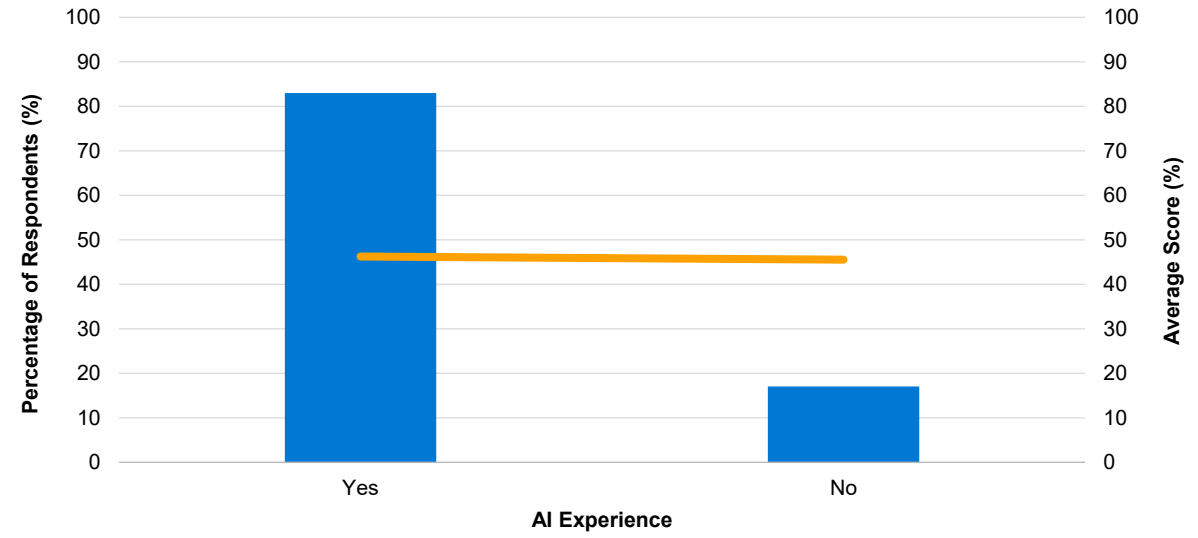
In Figure A2-2, the vertical bars show the percentage breakdown of respondents who assess car damage as part of their job and the line indicates the average score for each group.

FIGURE A2-2: RESPONDENTS AND SCORES BY CAR DAMAGE ASSESSMENT



In Figure A2-3, the vertical bars show the percentage breakdown of respondents by prior AI image generation experience and the line indicates the average score for each group.

FIGURE A2-3: RESPONDENTS AND SCORES BY AI IMAGE GENERATION EXPERIENCE

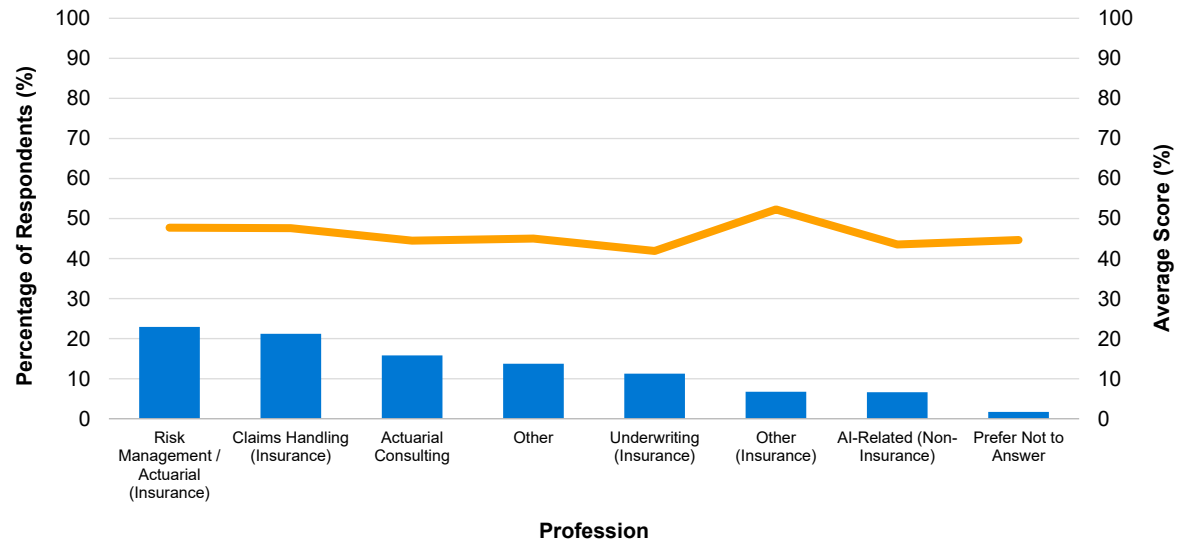


In Figure A2-4, the vertical bars show the percentage breakdown of respondents by the methods used to assess quiz photos and the line indicates the average score for each method.

FIGURE A2-4: RESPONDENTS AND SCORES BY PHOTO ASSESSMENT METHOD

In Figure A2-5, the vertical bars show the percentage breakdown of respondents by profession and the line indicates the average score for each profession.

FIGURE A2-5: RESPONDENTS AND SCORES BY PROFESSION



In Figure A2-6, the vertical bars show the percentage breakdown of respondents by role and the line indicates the average score for each role.

FIGURE A2-6: RESPONDENTS AND SCORES BY ROLE

Figure A2-7 shows the percentage of respondents in- and outside of the United States.

FIGURE A2-7: PERCENTAGE OF RESPONDENTS IN- AND OUTSIDE OF THE UNITED STATES

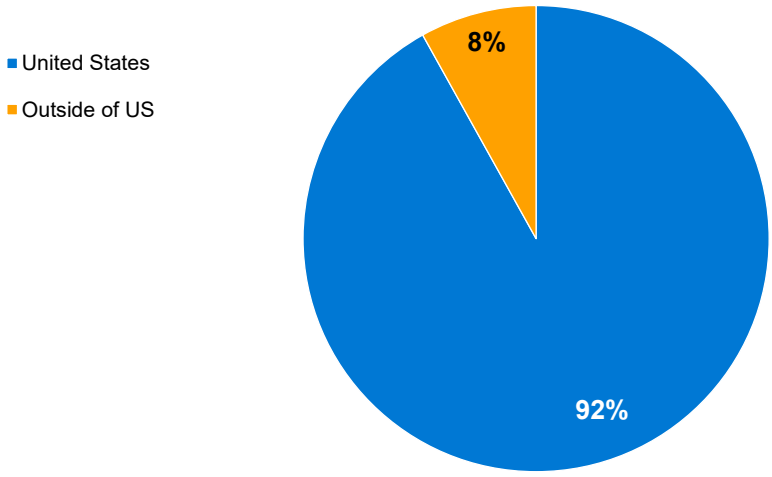
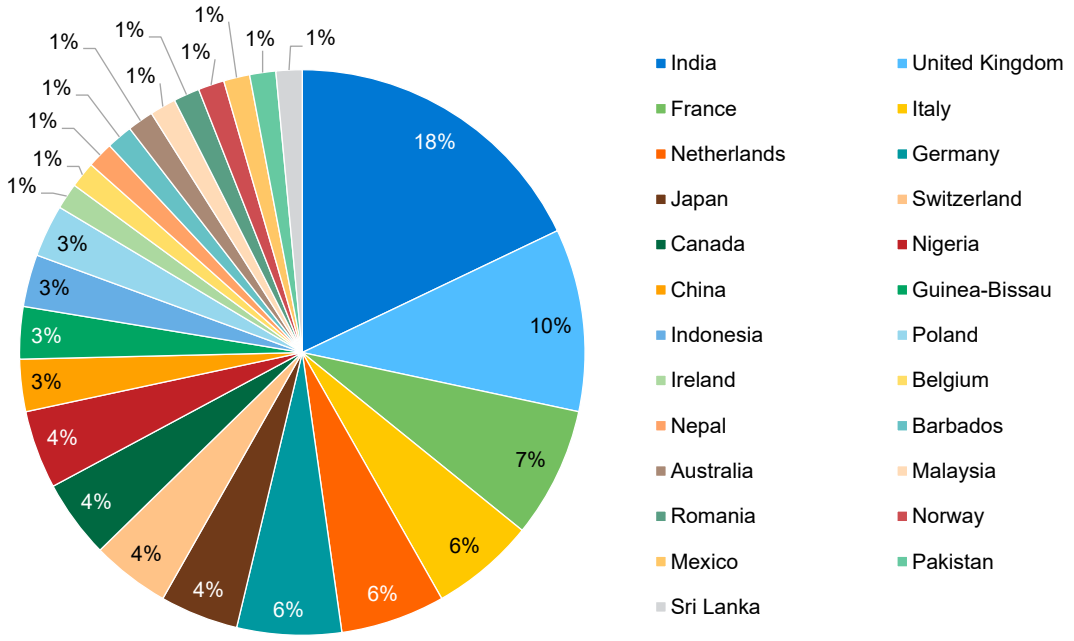


Figure A2-8 shows the percentage breakdown of respondents by country.

FIGURE A2-8: RESPONDENTS BY COUNTRY OF RESIDENCE (EXCLUDING THE US)



Solutions for a world at risk™

Milliman leverages deep expertise, actuarial rigor, and advanced technology to develop solutions for a world at risk. We help clients in the public and private sectors navigate urgent, complex challenges—from extreme weather and market volatility to financial insecurity and rising health costs—so they can meet their business, financial, and social objectives. Our solutions encompass insurance, financial services, healthcare, life sciences, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Raymond van Es
raymond.vanes@milliman.com

Bartosz Gaweda
bartosz.gaweda@milliman.com

Leonid Zeldin
leonid.zeldin@milliman.com

Jan Thiemen Postema
janthiemen.postema@milliman.com

Tristan Stoevelaar
tristan.stoevelaar@milliman.com