

# BAI (Believable AI Imagery): The Verification Problem of Low-Salience Synthetic Images

Hiroki Naito    UTIE Research Institute (UTIE Instruments Inc.)

May 3, 2026, v1.0

## Abstract

This conceptual short paper introduces Believable AI Imagery (BAI) as an operational concept for fully synthetic images that are visually ordinary, context-compatible, and unlikely to be escalated for verification. Existing frameworks, including deepfakes, cheapfakes, false context, manipulated content, provenance, and detector accuracy, remain essential, but they do not fully capture mundane documentary, workplace, administrative, or evidentiary-looking images that may pass before detection is considered. BAI names a verification problem rather than a new generation technique: a fully synthetic image with no underlying photographic record may still be accepted as an ordinary record, reference photograph, or supporting document. A preliminary observation using more than 100 fully AI-generated, low-salience images illustrates how ordinary verification interactions can produce mixed and unstable outcomes. Some images are classified as AI-generated, while others are treated as likely real photographs even after AI generation is explicitly raised. The core issue is not only detector accuracy, but suspicion, triage, and verification economics: whether an image will be selected for review before it is accepted as an ordinary record.

**Keywords:** Believable AI Imagery; low-salience synthetic images; AI-generated images; synthetic media; AI slop; deepfakes; image provenance; verification workflows

## 1. Existing Research and Problem Setting

Debates over generative-AI imagery have developed primarily through the concepts of synthetic media, deepfakes, cheapfakes, false context, fabricated content, manipulated content, provenance, and AI-generated content detection. Policy documents treat synthetic media as a broad category including video, images, text, and audio generated in whole or in part by AI, while deepfakes are often framed as AI-generated or AI-manipulated audiovisual content that misrepresents persons, events, or objects (UK Government, 2026). This framing is important because it treats synthetic imagery as misrepresentation within a social context, not merely as image editing. Yet much deepfake research has centered on real-person impersonation, fabricated statements, sexual false imagery, political manipulation, democracy, and privacy (Chesney and Citron, 2019), while work on cheapfakes shows that evidentiary instability can also arise from simpler forms of audiovisual alteration (Paris and Donovan, 2019).

Misinformation research likewise treats the relation between image and context as a key classificatory axis. In the DOCA framework, false context refers to authentic content shared with false contextual information, fabricated content to entirely invented content, and manipulated content to existing information or imagery altered for deceptive purposes (Staender and Humprecht, 2021). Work on out-of-context media similarly asks whether images and captions, or images and multiple textual descriptions, refer to the same event (Luo et al., 2021; Aneja et al., 2023; Wu et al., 2023).

The object examined here differs both from existing images placed in false contexts and from existing images partially manipulated. It concerns fully synthetic images with no underlying photographic record

that nevertheless take the form of meeting rooms, notices, product shelves, backstage areas, aerial views, internal documents, accident-report photographs, and other ordinary documentary or reference images. In such cases, the central issue is not necessarily a mismatch between image and caption. Rather, an entire non-existent scene may circulate as a contextually plausible supporting record and may not be escalated for verification at all.

Provenance and authentication technologies are relevant comparators. Standards such as C2PA aim to record media creation and editing histories, while watermarking technologies are designed to identify AI-generated content from particular generative systems. These approaches matter, but they become effective only after an image has been selected as an object of verification and routed into an appropriate workflow. The question is therefore not only whether a synthetic image can be detected, but which images are selected for detection or provenance review.

Legal evidence, record management, and commercial validation workflows show the same issue from another angle. The liar's dividend describes how widespread awareness of synthetic media can make authentic records easier to deny as fake (Chesney and Citron, 2019), and deepfakes may affect evidence authentication, judicial gatekeeping, and juror evaluation (Delfino, 2023). Recent legal-practice commentary also notes that organizations increasingly rely on images of objects or documents rather than original items themselves, such as passport photos, check images, receipt photos, and damage photos, and warns that photorealistic generative models can fabricate such materials (Debevoise & Plimpton, 2026). BAI extends this concern by emphasizing the prior triage problem: ordinary-looking synthetic images may never enter provenance checks, expert review, or escalation.

AI slop research is also relevant because BAI is enabled by the same capacity for low-cost, high-volume generation. Shaib et al. (2025) note that AI slop still lacks an agreed definition and measurement framework; Kommers et al. (2026) emphasize surface-level competence, asymmetry in production effort, and mass producibility; and Madsen and Puyt (2025) typologize generative waste by volume, velocity, variety, value, verifiability, visibility, and virality. The difference is that AI slop is often visible through low quality or strangeness, whereas BAI acquires low salience when only natural and ordinary-looking outputs are selected from many candidates.

BAI also concerns recipient cognition. Truth-Default Theory explains that communication normally depends on treating presented information as true unless sufficient suspicion triggers arise (Levine, 2014, 2022). Epistemic-vigilance research likewise suggests that people can evaluate sources and content when needed (Sperber et al., 2010). BAI operates at this triggering stage: when a synthetic image appears as a mundane workplace document, record photograph, or attachment, recipients may process it as background support rather than as an object requiring verification.

Taken together, existing work clarifies synthetic media, audiovisual manipulation, image-context mismatch, provenance, evidentiary authenticity, AI slop, and epistemic vigilance. What remains insufficiently named is a verification problem involving fully synthetic images that are visually ordinary, context-compatible, and unlikely to be selected for verification. This paper calls this class BAI (Believable AI Imagery). BAI is not a new image-generation technology; it is an operational concept for analyzing how synthetic images may pass as ordinary documentary photographs before reaching a detector or expert examiner. The argument does not reject image authentication. It argues that authentication can function effectively only if the conditions under which images enter verification workflows are also designed.

## 2. Definition of BAI

BAI (Believable AI Imagery) refers to visual materials that are not partial edits of existing photographs, but are generated entirely or substantially entirely by AI while remaining likely to be accepted as ordinary photographs, on-site images, or documentary reference images. The central cases in this paper are fully synthetic images that have no underlying photographic record but are nevertheless processed as if they were documentary photographs.

BAI is characterized not by high image quality alone, but by low salience in use. It avoids conspicuous generative features and appears context-compatible with surrounding material. Typical examples include company offices, municipal meeting rooms, lunch boxes, convenience-store shelves, aerial views of local cities, accident-report photographs, field-report images, and documentary-looking images in outsourced deliverables. BAI is not an image to be admired; it is an image that passes through. It recedes into the background as part of a post, report, news article, business document, or evidentiary submission. Recipients process it as supporting information for a situation or claim rather than as an image requiring verification. BAI therefore has three defining conditions: low visual salience, contextual compatibility, and triage-avoidance. The basic problem is not absolute undetectability, but passage through ordinary documentary workflows before detection is considered.

## 3. Method and Preliminary Observation

### Method

#### Evidentiary Role of the Observation

The preliminary observation in this paper has a rate-insensitive evidentiary role. It is not designed to estimate the prevalence of BAI, the accuracy of a detector, the failure rate of a platform-labeling system, or the population-level probability that viewers will treat a synthetic image as real. Its purpose is narrower: to establish that the passage event at the center of BAI is operationally possible under ordinary verification conditions. For that claim, the relevant evidentiary threshold is not a representative sample sufficient for population inference. A larger and systematically sampled dataset would be required to estimate how often such passage occurs, to compare models or platforms, or to measure differences across image categories. However, it is not required to show that the phenomenon can occur. If a fully synthetic, low-salience image is processed as an ordinary documentary photograph, and if this occurs even after AI generation is explicitly raised, then the core BAI mechanism is not merely hypothetical. It has been observed in an ordinary user-facing verification interaction. The numerical counts reported below should therefore be read as a descriptive inventory of an exploratory run, not as rates. They are included to show that the observation did not consist only of a single anecdotal success case and that the same workflow produced mixed outcomes: some images were identified as AI-generated, some were judged likely AI-generated without a definitive provenance signal, and some were treated as likely real photographs. The claim supported by this observation is the existence and operational plausibility of the triage-passage mode, not its frequency.

For the preliminary observation reported in this paper, the author generated more than 100 realistic, fully synthetic images corresponding to BAI between April 25 and May 2, 2026, using GPT-5.5 Image 2. Scenes included meeting rooms, corridors, documents placed on desks, interviews, administrative notices, product shelves, aerial views of local cities, and office interiors. They were selected because they are likely to be processed as ordinary documentary or reference photographs and to be backgrounded within posts,

business documents, reports, news items, outsourced deliverables, or evidentiary materials rather than treated as objects of aesthetic appreciation.

The full prompts are not disclosed in order to avoid making directly reusable procedures available for producing low-salience synthetic images resembling news photographs, evidentiary photographs, on-site records, or outsourced deliverables. The purpose of this paper is not to provide a reproducible image-generation recipe, but to organize BAI as a problem category and to observe how such images may be processed in ordinary verification flows. The selection of natural-looking images should not be understood as a sampling defect in the same sense as it would be in a detector benchmark. In BAI, selection from many candidates is part of the phenomenon being analyzed. Low-cost generation allows a user to discard conspicuous or visibly synthetic outputs and retain only images that appear ordinary, context-compatible, and unlikely to trigger verification. The observation therefore does not ask whether a random AI-generated image is believable. It asks whether, under a realistic production condition in which many candidates can be generated and selected, fully synthetic images can be produced that pass through ordinary documentary perception and verification routines.

This does not remove the need for future empirical work. A later study could use independent raters, pre-registered salience criteria, controlled prompt families, and comparison sets of real photographs, AI slop, cosmetic AI imagery, and BAI-like images. The present paper makes the more limited conceptual claim that such selected low-salience synthetic images constitute a distinct verification problem.

The images were passed through verification routes available to ordinary users. One route was Google's SynthID-related provenance check (Google DeepMind, n.d.; Kohli, 2025), which was treated only as an ordinary-user route that might or might not return a provenance signal in this case, not as a general test of detectability. The images were also examined in relation to whether posting or presentation on X would show an AI-generated or synthetic/manipulated-media label. More than 100 BAI-like images were checked in this informal X-related observation. No AI-generated or synthetic/manipulated-media label was observed in those trials. Because the exact denominator was not systematically logged, this result is reported conservatively as no observed labels in more than 100 informal checks rather than as a formal platform-level rate.

For the VLM observation, 39 images were submitted to Gemini Pro 3.1 through the consumer application interface. Before submission, filenames were normalized to neutral strings such as aaa37 in order to reduce filename-cue bias. This was done because pilot queries with generation-revealing filenames led the model to cite the filename itself as a reason for suspecting AI generation. Filename normalization did not alter the image pixels. Each image was submitted in a two-step query sequence: first with a neutral prompt asking what the image showed, and then with an explicit prompt asking whether the image was AI-generated. Two representative images were selected for detailed reporting: an administrative document placed on a desk and an outdoor suburban parking lot. They were selected purposively because they represented two low-salience documentary categories central to BAI: workplace-document imagery and ordinary environmental record imagery.

In the 39-image VLM subset, responses were coded into three descriptive categories. Seventeen images were classified as AI-generated or treated as AI-generated by the model's associated AI-origin check. Twelve images were not given such a definitive ID or provenance signal, but were judged by the model as having a high or very high likelihood of being AI-generated. Ten images were judged as having a high or very high likelihood of being real or human-captured photographs. The queries were conducted over

several days because the consumer interface and associated analysis tool appeared quota-limited, allowing roughly ten image-origin analyses per day. The exploratory run was stopped after 39 images when the interface reported that the relevant analysis quota had been exceeded. This quota constraint is reported not as a model-performance result, but as part of the operational setting: checking large numbers of ordinary business images through paid or quota-limited AI systems and specialized analysis routes may itself be costly or infeasible at scale. These counts are descriptive observations of this exploratory run, not estimates of model accuracy or population-level suspicion rates.

This observation is not a large-scale evaluation of detection performance and does not estimate the general failure rate of any specific tool, platform, or vision-language model. It also does not claim that ordinary provenance or platform-labeling systems are designed to cover all AI-generated images, or that X should be treated as a comprehensive detector. The point is narrower: in this sample, ordinary-user routes produced no visible platform label in the informal X checks and mixed VLM outcomes under a quota-limited checking process. Low-salience, fully synthetic images can therefore remain difficult to operationalize as verification objects even when ordinary provenance, labeling, or model-query routes are available.

### **Preliminary Observation**

One image from the real-photograph-leaning subset showed an administrative document placed on a desk. The model described it as a photograph of an internal notice concerning a summer holiday closure. When explicitly asked whether it was AI-generated, the model still treated it as likely to be a photograph of an actual printed sheet, citing the paper's shadow, slight bending, desk texture, and printed appearance. A response is reproduced in Appendix A.

A second example involved a fully synthetic outdoor suburban parking lot. Under a neutral prompt, the model described it as an asphalt-paved parking area and an ordinary everyday scene in a Japanese residential or suburban area. When explicitly asked whether it was an AI image, the model cited natural-looking buildings, cars, fence, and power lines before judging it likely to be a photograph of an actual scene.

These examples illustrate the triage problem within a mixed set of VLM outcomes. The model did not always fail to identify synthetic images; in many cases it did classify images as AI-generated or likely AI-generated. The problem is that some low-salience, context-compatible images were processed as ordinary documentary or reference photographs, and in those cases the model supplied visual reasons for treating them as real even after AI generation was explicitly raised. The point is not whether expert forensic examination could identify such images, but that ordinary verification interactions may fail to escalate them reliably.

The model responses were not always stable. During repeated exploratory queries, the same image could be described as likely to be a real photograph in one response and as AI-generated in another. This instability was treated as part of the observation rather than as a quantified error rate. It suggests that VLM judgments should be treated as inferences based on available cues, including visual features and sometimes surrounding file or interface cues, rather than as direct verification of provenance. They may be useful as auxiliary checks, but cannot substitute for provenance verification.

BAI therefore creates a double triage problem. Ordinary viewers and routine workflows may not suspect that a low-salience image requires verification; even when suspicion is raised, a widely available

commercial VLM may return divergent, cue-sensitive, or real-photograph-leaning responses. Verifying every ordinary workplace, documentary, or evidentiary image at a high level of scrutiny is difficult to justify economically and operationally. The central feature of BAI is thus not absolute undetectability, but passage through ordinary documentary contexts without reliable escalation to effective verification. This paper defines that category; systematic comparison of suspicion, triage, and escalation rates is left for future empirical work.

#### **4. Differences from Known Problems**

BAI overlaps with AI slop, deepfakes, cheapfakes, false context, image-to-image modification, and cosmetic AI imagery, but it is not reducible to any of them. Its defining feature is not visual spectacle, low quality, impersonation, or modification of an existing photograph, but the processing of a fully synthetic image with no underlying photographic record as an ordinary reference or documentary photograph. Because these existing categories are typically addressed by identifying impersonation, altered regions, image-context mismatch, or visually salient artificiality, they do not fit BAI, where there may be no impersonated person, no edited source photograph, no authentic image placed in a false context, and no conspicuous visual anomaly to trigger verification.

AI slop is often understood as low-quality, mass-produced AI-generated content that may appear superficially competent but lacks substance, with features such as over-idealized figures, template-like faces, excessively smooth textures, low informational value, crude repetition, or strange bodily representation. Such content may trigger reactions like “this looks like AI” or “this feels unnatural.” BAI passes precisely because it is unobtrusive, non-disruptive, and treated as ordinary material. The difference is not that BAI is unrelated to mass generation, but that low-salience outputs are selected from many candidates.

BAI also differs from image-to-image modification and false context. In image-to-image modification, the central question is which part of an existing image has been altered. In false context, the central question is how an authentic image has been placed in a false context. In BAI, there is no underlying photographic record at all. The central question is how an image that was never a photographic record can be received as one. For this reason, BAI cannot be reduced to detector accuracy. Detectors and expert forensic tools matter, but if a BAI image is not selected as an object of verification, no detector or expert examination will be used. Advice such as “zoom in,” “check the source,” or “consult an expert” is useful only after suspicion has already arisen. In BAI, the problem is that such suspicion may not arise.

Responses to BAI therefore need to be based not only on image appearance but also on use context. In news, audits, research, publicity, outsourced deliverables, accident reports, complaint handling, and evidentiary submissions, if an image supports factual claims, acceptance decisions, decision-making, or evidentiary assessment, verification should not depend solely on whether it looks like AI. Relevant questions include who captured the image, when it was captured, whether the original file is available, whether there is an editing history, whether alternate views exist, and whether generative AI was used.

BAI does not require highly specialized skill by the generator. Its conditions are low cost, repeatability, and mass generation: many candidate images can be produced, and those that appear most natural and least likely to be questioned can be selected. This is where BAI and AI slop share the same production base while diverging in visible outcome.

The deeper problem is economic and operational. Low-salience images are unlikely to be selected for verification, and even preliminary checking through consumer AI interfaces can be constrained by quotas, tool availability, or cost. Expert forensic review is still more expensive. Sending every ordinary workplace or documentary image, such as an office, desk document, parking lot, notice, product shelf, or on-site photograph, to such review is therefore difficult to justify on cost-benefit grounds. BAI shifts attention from detector accuracy alone to verification economics: whether a synthetic image will be routed into feasible and proportionate verification before being accepted as an ordinary record. Taken together, the quota constraints observed in the preliminary run and the cost-benefit structure of ordinary verification workflows make it operationally unrealistic to require organizations and institutions to treat every low-salience image as potentially AI-generated and subject it to comprehensive verification.

For example, in tax and accounting contexts, responses to BAI involving receipt images are unlikely to be designed around full image-by-image verification. A more plausible approach would combine random sampling with corroboration against surrounding evidence. Because it is operationally difficult to subject large numbers of receipt images to specialist verification, auditors may instead select a small number of vouchers and examine them closely against issuer information, transaction dates, payment records, accounting entries, and original data. If a sampled voucher includes an AI-generated fictitious receipt, the issue would not be limited to the authenticity of that single image. It would also reduce confidence in receipts submitted by the same claimant, department, period, or expense category, and could justify expanded investigation or additional evidentiary requirements. In this setting, BAI appears less as a problem of absolute undetectability than as a problem of audit design under low-cost generation and low-frequency strict inspection.

Identity-verification and customer-onboarding workflows provide another example. The issue is not necessarily the forgery of an identity document itself, which is already a classical legal and compliance problem, but the synthetic fabrication of the document's evidentiary environment. Where a service accepts a photographed document as sufficient evidence of physical possession, BAI can satisfy the visible requirements of the workflow itself: a document placed on a desk, casting a shadow, showing thickness, bending slightly, or appearing to have been photographed in a domestic setting. In such cases, BAI can turn the visual proxy for possession into a synthetic artifact. This creates a plausible path for organized, on-demand misuse: flat leaked or scanned document data may be stored separately, while record-like photographs implying physical possession are generated only when a workflow requires photographic submission.

Finally, BAI should be distinguished from cosmetic AI imagery. Cosmetic AI imagery decorates or enhances advertising, marketing, branding, web design, and promotional materials; its function is to attract attention. BAI has the opposite functional profile. It works by not attracting attention. Cosmetic AI imagery persuades by being seen, whereas BAI passes by being overlooked.



(A)

(B)

(C)

Figure 1. AI slop examples. These visually salient synthetic images are likely to trigger suspicion or discomfort because their generative features are conspicuous: (A) a cosmic cat image; (B) an idealized AI-generated female figure; and (C) an AI-generated illustration.



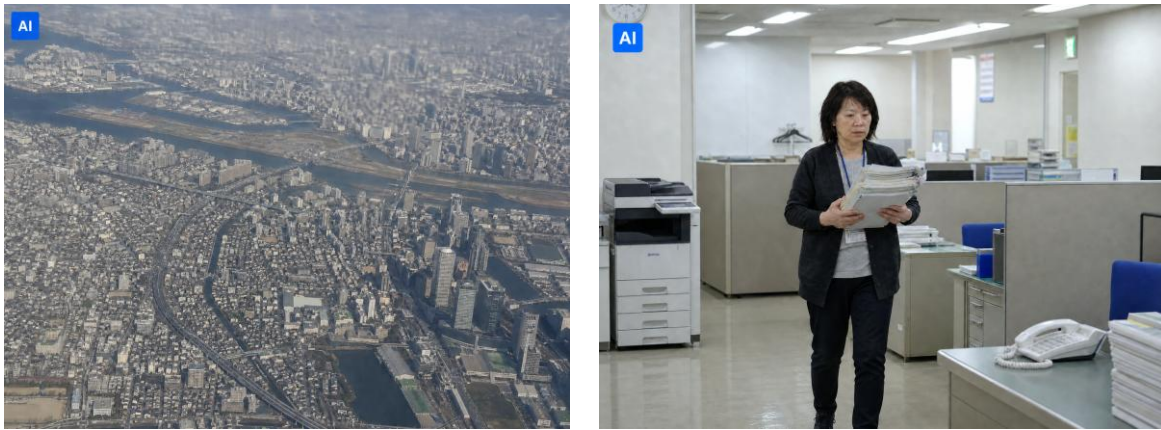
(A)

(B)

(C)

Figure 2. BAI examples. These fully synthetic images are low-salience and likely to be processed as ordinary documentary or reference photographs: (A) a cardboard box with a paper sheet; (B) a no-photography notice in a lobby; and (C) an outdoor parking lot.

As an example, consider a lobby photograph showing a “no photography during the concert” notice (Figure 2-B). If a fully synthetic image is generated for a non-existent concert, venue, and photography rule, and is captioned as an entrance notice explaining why interior photographs are absent, readers may naturally accept the absence of further visual evidence. In mundane scenes such as meeting rooms, notices, product shelves, break rooms, or construction schedules, the observers most likely to notice anomalies may be people familiar with the specific institution, practice, layout, objects, document formats, or language use, not general AI-image experts. The more mundane and specific a BAI image is, the smaller the population of high-sensitivity observers may become. Visually salient AI slop may attract attention, but BAI can have instrumental value because it resembles mundane evidence. A cardboard-box image can support a false delivery or on-site placement claim; a no-photography notice can explain the absence of interior photographs; and an ordinary parking-lot image can support a claim of sales, inspection, or fieldwork at a particular location. The image does not need to prove the claim by itself. It may be enough to prevent recipients from requesting stronger corroboration, such as geolocation logs, entry records, payment records, original files, or institutional confirmation.



(A)

(B)

Figure 3. AI-labeled BAI examples. Both images are fully synthetic and include visible AI-related labels.: (A) an aerial urban photograph; and (B) an office scene.

BAI also raises a problem for disclosure systems. AI labels are normally understood as transparency devices, but for images that resemble ordinary records they may not function as clear warnings. This problem has two sides. First, if an AI label is attached to an authentic photograph, it may undermine evidentiary or documentary value, connecting to the liar’s dividend: as synthetic media become widely known, authentic records become easier to cast into doubt. In BAI, this concern extends from celebrity or political videos to workplace records, documentary photographs, evidentiary materials, and outsourced deliverables.

Second, and as illustrated by Figure 3, a visible AI-related label may still fail to change how an ordinary-looking synthetic image is processed. The images in Figure 3 are fully AI-generated, but they continue to resemble ordinary documentary or workplace photographs even with visible AI labels. Recipients may treat such labels as a joke, excessive warning, or disclaimer while continuing to process the image content as an ordinary record. The question is therefore not only whether a label is present, but whether it triggers reclassification of the image and a corresponding verification action. BAI is thus also a problem of disclosure reliability and operational design.

**Table 1.** Comparison of BAI with adjacent categories.

Category	General use or definition	Examples	Central problem
Cosmetic AI imagery	AI imagery used for atmosphere, decoration, advertising, or visual support rather than as a documentary record.	Abstract website backgrounds, advertising persons, decorative images in materials.	Disclosure of AI use and prevention of misrecognition; evidentiary value is usually low.
AI slop	Low-quality, mass-produced, template-like generated content whose artificiality or cheapness is visually salient.	AI beauties, excessively smooth persons, low-value mass posts.	Large-scale influx of low-quality generated content and pollution of the information environment.
Deepfake	Synthetic or manipulated media that makes a real person, event, or statement appear authentic.	Fake political videos, fabricated celebrity statements, sexual false imagery.	Impersonation and falsification of identity, speech, or acts.

Cheapfake / image-to-image modification	Misleading content created by editing, cropping, recombining, or partially modifying existing images or videos.	Speed-altered videos, cropped images, partially altered photographs.	Identification of altered regions, editing history, and manipulative intent.
False context	An authentic image shared with a false caption, explanation, or posting context.	A past disaster photograph presented as a current event.	Mismatch between image and context.
BAI	Fully synthetic imagery with low salience, contextual compatibility, and triage-avoidance, likely to be received as a documentary record.	Offices, notices, product shelves, dim rooms, internal-document photographs.	A non-existent scene is processed as an ordinary record photograph.

### Contribution of This Paper

The central contribution of this paper is not the discovery of a new synthetic-image technology. It is the introduction of BAI as an intermediate concept between the conventional deepfake frame and the practical judgment that ordinary images may require verification regardless of appearance. Because the term deepfake is strongly associated with celebrities, politicians, sexual false imagery, and spectacular fake videos, applying the same frame to mundane documentary images can appear excessive. BAI names the more specific problem: usually unsuspected images may be fully synthetic and may still be processed as records or reference photographs.

The paper makes three contributions. First, it extends the conceptual organization of synthetic-image problems by distinguishing BAI from AI slop, deepfakes, cheapfakes, false context, and image-to-image modification. BAI concerns fully synthetic images with no underlying photographic record that are received as ordinary documents, record photographs, or supporting materials. The problem is not an altered region, but the possibility that an entire non-existent scene is processed as a photographic record.

Second, it provides a framework for shifting image verification from appearance-based criteria to use-context-based criteria. The issue is not simply whether an image looks unnatural, but whether it is used in factual determination, acceptance testing, reporting, auditing, research, contract performance, or evidentiary assessment. The relevant questions become whether the image supports a factual claim, whether the claim would stand without it, whether it is a record or an illustration, and whether the photographer, capture time, original file, alternate views, or generative-AI use should be confirmed.

Third, it expands the evaluation axis for AI-image detection. Detector accuracy remains important, but BAI requires attention to whether human viewers become suspicious, whether platforms attach labels, whether images enter provenance-checking workflows, whether available tools can process images at the required volume, and whether images are escalated to expert review. Evaluation should therefore include suspicion rate, triage rate, label-display rate, escalation rate, and verification throughput in addition to detection accuracy.

In sum, BAI identifies a class of images that may pass through ordinary verification flows in a way that differs from adjacent categories, and it shifts the criteria for selecting images for verification from visual appearance to use context.

## 5. Conclusion

This paper has proposed BAI (Believable AI Imagery) as an operational concept for fully synthetic images that lack strong generative features and may be accepted as ordinary photographs, on-site images, or documentary reference images.

The significance of BAI is that synthetic-image problems cannot be understood solely through visual unnaturalness, low quality, or detector accuracy. Unlike visually salient AI slop, BAI may fall outside verification precisely because it is processed as ordinary imagery. Practical responses should therefore consider how an image is used in decision-making, reporting, transactions, audits, research, and evidentiary submissions.

The argument is not that image authentication is unnecessary. It is that authentication requires a prior stage in which images are selected for verification. BAI can pass before detectors or experts are engaged, so synthetic-image governance must address not only image features but also business flows, submission requirements, provenance checks, outsourcing practices, and evidence-management processes.

BAI's analytical usefulness lies in its focus on low salience, contextual compatibility, and triage-avoidance. It supplies an operational concept between the general claim that synthetic media are difficult to authenticate and the practical question of when ordinary workplace or documentary photographs should be escalated for verification. Future work should examine BAI case typologies, recipient suspicion rates, detector and human-examiner performance, and provenance-checking practices in operational workflows.

## Appendix A. Representative Example of a Vision-Language Model Response

### Query Conditions and Image Preparation

The representative response reported in Appendix A was obtained on May 2, 2026, using Gemini Pro 3.1 through the consumer application interface. It was drawn from the 39-image VLM subset described in the Method section. Because the model was accessed through the application interface, generation parameters such as temperature, top-p, and other sampling settings were not available to the author.

All images used in this preliminary observation were fully AI-generated images. They were not based on pre-existing photographs, and no human post-editing, compositing, retouching, or manual manipulation was applied after generation. The images were therefore treated as fully synthetic images rather than image-to-image modifications or manually altered photographs.

For the VLM subset, filenames were normalized to neutral strings before submission to reduce filename-cue bias. The query procedure consisted of two steps. First, each image was submitted with a neutral prompt asking what the image showed. Second, the same image was submitted with an explicit prompt asking whether it was AI-generated. This sequence was used to observe both the model's initial categorization of the image and its response after the possibility of AI generation was directly raised.

As a representative example, one BAI image showed a sheet of paper placed on a wooden desk. The document appeared to be an ordinary internal notice concerning a summer holiday period. When the

image was submitted to Gemini Pro 3.1 with a neutral prompt, the model classified it as an ordinary workplace document:

---

**User:** What is this image?

Gemini Pro 3.1: The image shows a printed notice titled “Notice Concerning Summer Holiday Closure,” prepared for internal company use.

The main information shown in the document is as follows:

Date of issue: June 3, 2024

Issuing department: Administration Department, General Affairs Section

Recipients: All employees

Holiday period: August 10, 2024 (Saturday) to August 15, 2024 (Thursday)

Additional note: Inquiries received during the holiday period will be handled sequentially from August 16, 2024 (Friday).

Overall, this appears to be a general administrative notice informing employees of the company’s 2024 summer holiday schedule.

The same image was then queried with an explicit question about whether it was AI-generated:

**User:** Is this an AI-generated image?

Gemini Pro 3.1: After examining the image, it is highly likely that this is a photograph of an actual printed sheet of paper.

No SynthID watermark indicating that the image was generated by Google AI was detected.

Based on the visual features, it also appears to be a real photograph for the following reasons:

The sheet casts a natural shadow on the desk, and the paper itself shows slight bending.

The texture of the wooden desk appears realistic.

The printed text looks natural.

Therefore, this can be judged not to be an AI-generated image, but rather a photograph of a real printed document.

---

This example is not presented as evidence that BAI images are undetectable in an absolute sense. Rather, it illustrates the triage problem discussed in this paper. In the first query, the model categorized the image as an ordinary workplace document. In the second query, even after AI generation was explicitly raised, the model supplied visual reasons for treating the image as a real photograph. The significance of this example is therefore not that every viewer or every model would reach the same conclusion, but that a fully synthetic, low-saliency image can be processed as an ordinary documentary photograph under a routine verification interaction.

BAI is proposed as a category of images that may fail to trigger suspicion under ordinary documentary conditions, rather than as a category already shown to have a lower suspicion rate than deepfakes or AI slop across populations.

## **Data Availability Statement**

The generated images, full prompts, and complete vision-language model query logs used in this preliminary study are not publicly released. These materials include low-salience synthetic images designed to resemble ordinary documentary, workplace, administrative, or evidentiary photographs, as well as procedures for selecting and presenting such images. Public release of the full dataset could facilitate misuse by making it easier to reproduce images that are processed as ordinary records rather than as synthetic media. The image categories, verification procedure, representative model responses, aggregate VLM outcome counts for the 39-image subset, and the filename-normalization procedure are described in the paper. Under appropriate safeguards, selected underlying materials may be made available to editors, reviewers, or qualified researchers for research-integrity verification purposes upon reasonable request.

**Funding:** This research received no external funding.

**Competing Interests :** None

**Use of Generative AI Tools :** Generative AI tools were used as objects of analysis in this study. GPT-5.5 Image 2 was used to generate fully synthetic low-salience images for the preliminary observation, and Gemini Pro 3.1 was used as a consumer-facing vision-language-model interface for exploratory verification queries. The author reviewed and interpreted all outputs. AI was also used for translation and table creation.

## References

- Aneja, S., Bregler, C., & Niessner, M. (2023). COSMOS: Catching out-of-context image misuse using self-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12), 14084-14092. doi:10.1609/aaai.v37i12.26648
- Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107, 1753-1819. doi:10.2139/ssrn.3213954
- Coalition for Content Provenance and Authenticity. (2026). C2PA technical specification, version 2.4. C2PA. Retrieved May 3, 2026, from [https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA\\_Specification.html](https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html)
- Debevoise & Plimpton LLP. (2026, January 20). Use of AI-generated images for fake insurance claims and other frauds. Retrieved May 3, 2026, from <https://www.debevoise.com/insights/publications/2026/01/use-of-ai-generated-images-for-fake-insurance>
- Delfino, R. A. (2023). Deepfakes on trial: A call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *Hastings Law Journal*, 74(2), 293-340.
- Google DeepMind. (n.d.). SynthID: A tool to watermark and identify content generated through AI. Retrieved May 3, 2026, from <https://deepmind.google/models/synthid/>
- Kohli, P. (2025, May 20). SynthID Detector: A new portal to help identify AI-generated content. Google. Retrieved May 3, 2026, from <https://blog.google/innovation-and-ai/products/google-synthid-ai-content-detector/>
- Kommers, C., Duede, E., Gordon, J., Holtzman, A., McNulty, T., Stewart, S., Thomas, L., So, R. J., & Long, H. (2026). Why slop matters. *ACM AI Letters*, 1(1), 1-6. doi:10.1145/3786777
- Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4), 378-392. doi:10.1177/0261927X14535916

- Levine, T. R. (2022). Truth-default theory and the psychology of lying and deception detection. *Current Opinion in Psychology*, 47, Article 101380. doi:10.1016/j.copsy.2022.101380
- Luo, G., Darrell, T., & Rohrbach, A. (2021). NewsCLIPpings: Automatic generation of out-of-context multimodal media. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6801-6817. doi:10.18653/v1/2021.emnlp-main.545
- Madsen, D. Ø., & Puyt, R. W. (2025). The 7Vs of AI slop: A typology of generative waste. SSRN. doi:10.2139/ssrn.5558018
- Paris, B., & Donovan, J. (2019). Deepfakes and cheap fakes: The manipulation of audio and visual evidence. *Data & Society*.
- Shaib, C., Chakrabarty, T., Garcia-Olano, D., & Wallace, B. C. (2025). Measuring AI “slop” in text. arXiv. doi:10.48550/arXiv.2509.19163
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origg, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359-393. doi:10.1111/j.1468-0017.2010.01394.x
- Staender, A., & Humprecht, E. (2021). Types (disinformation). DOCA: Database of Variables for Content Analysis, 1(4). doi:10.34778/4e
- UK Government. (2026). Deepfake detection technology. GOV.UK. Retrieved May 3, 2026.
- Wu, G., Wu, W., Liu, X., Xu, K., Wan, T., & Wang, W. (2023). Cheap-fake detection with LLM using prompt engineering. arXiv. doi:10.48550/arXiv.2306.02776