TYCOV: An Open-Source, Year-Long Behavioral Study of Human Ability to Detect AI-Generated Images

Aaryan Singh
Independent Researcher

Generative image models have rapidly improved, producing visuals that are frequently indistinguishable from real photographs. We present **TYCOV** (Test Your Cognitive Vision), an open-source web application and reproducible experimental pipeline to measure human ability to detect AI-generated images. In a 55-week public deployment, **60,123** trial-level responses were collected from **1,217** unique participants who completed forced-choice classifications on a balanced stimulus set (223 real; 223 AI). Aggregate accuracy was 60.3%; class-conditional accuracies were 67.8% for *real* images and 52.8% for *AI* images. We emphasize that trial-level analyses are clustered within participants; accordingly, cluster-aware inferential analyses were conducted. Analysis outputs and reproduction instructions are included in the repository's README.md and Paper.md.

Keywords: AI-generated images, human perception, diffusion models, media literacy

Introduction

Recent advances in diffusion models and large text-to-image systems have narrowed the perceptual gap between synthetic and real images. This creates new challenges for media trust, misinformation, and digital literacy. To design effective interventions and policies, researchers need open, reproducible infrastructure for longitudinal measurement of human perception.

TYCOV (Test Your Cognitive Vision) is an open-source platform designed to meet that need. It combines a lightweight web experiment, deterministic manifest-driven stimulus sampling, anonymized logging, and reproducible analysis artifacts. In this paper we ask the following research questions: (1) How accurately do humans distinguish AI-generated from real images? (2) Does accuracy differ by class (real vs. AI)? (3) How does accuracy evolve over a long deployment?

Related Work

Research into perception of synthetic media spans psychology, computer vision, and HCI. Prior studies have shown that unaided humans often perform only modestly above chance at distinguishing synthetic from real images (e.g., recent perception studies and platform quizzes reporting near-chance performance in several domains). Large-scale benchmarks for automated detection (e.g., DeepfakeBench and DFBench) focus on algorithmic performance and generalization gaps; TYCOV complements that work by providing an open human-subject measurement platform designed for replication and longitudinal study. Multimodal deception datasets (video + audio) show improved detection when temporal and multimodal signals are available; this motivates future extensions of TYCOV beyond single-image tasks.

Methods

Platform and Implementation

TYCOV's stack was selected for reproducibility and ease of deployment:

- **Frontend:** Next.js (React) with Tailwind CSS for a responsive UI that displays one image per trial and two response buttons.
- **Backend and DB:** Supabase (PostgreSQL). Schema migrations and queries are managed with Drizzle ORM.
- **Dataset management:** Images are hosted in a companion GitHub repository. An admin workflow generates a deterministic `manifest.json` that the frontend uses to serve stimuli reproducibly.
- **Privacy:** No personally identifiable information (PII) was collected. Each stored record contains timestamp, session UUID, image ID, image label, user choice, and correctness.

Stimuli

The reference stimulus set included 456 images: 223 labeled *real* (royalty-free sources) and 223 labeled *AI* (generated with modern models and diverse prompts to maximize coverage). Images were standardized for resolution and aspect ratio; the dataset repository includes the manifest and exact file list for reproducibility.

Participants and Procedure

Over 55 weeks, **1,217** unique participants contributed **60,123** forced-choice trials (Real vs. AI). Each trial recorded session ID, image ID, image label, user choice, and correctness. Participants could clear local session cookies without affecting stored database records. Optional feedback and summary stats were available via a toggle (disabled by default during baseline deployment).

Analysis Overview

To keep the manuscript focused and accessible, we report descriptive outcomes (overall accuracy, class-conditional accuracy, confusion matrix, and weekly trends). Cluster-aware inferential analyses (e.g., per-participant summaries and mixed-effects / GEE models) were conducted. Details, reproduction instructions, and code snippets are provided in the repository's README.md and Paper.md.

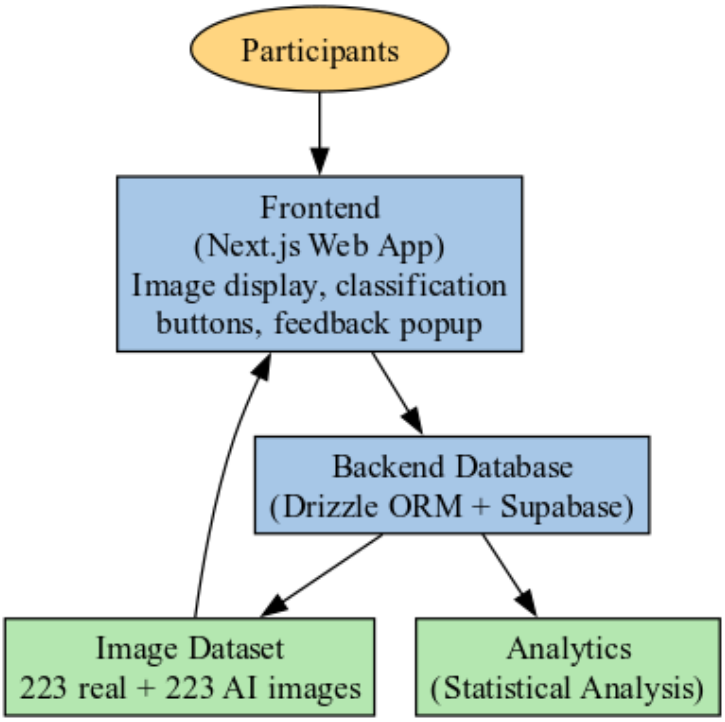

Figure 1

TYCOV system architecture: participants interact with the Next.js frontend; images are served via a manifest; responses are logged to Supabase (Drizzle ORM used for migrations and queries); analysis pipelines pull aggregated exports for statistical analysis.

Results

Descriptive outcomes

Total trial-level responses: $\mathbf{N = 60{,}123}$ (from $\mathbf{n = 1{,}217}$ unique participants; mean $\approx$ 49.4 trials per participant). Aggregate accuracy: **60.3**%. Class-conditional accuracies: $\mathbf{p_{real} = 0.678}$ (67.8%) for ground-truth Real images; $\mathbf{p_{ai} = 0.528}$ (52.8%) for ground-truth AI images.

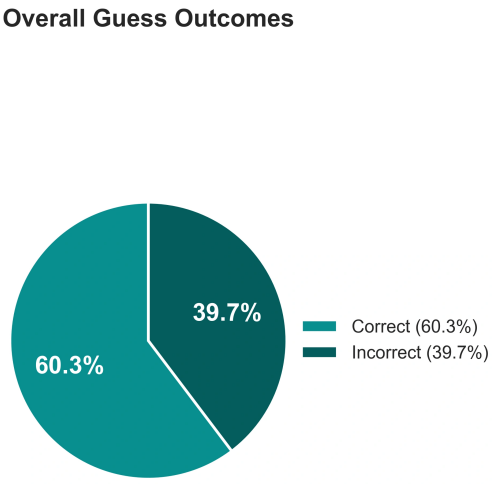

Figure 2

Overall correct vs. incorrect responses across ≈60k trials (overall accuracy = 60.3%).

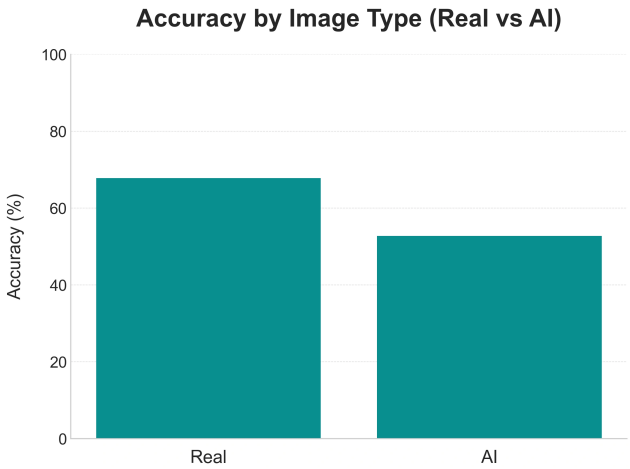

Figure 3

Accuracy by ground-truth class. Participants identify Real *images (67.8%) more reliably than* AI *images (52.8%).*

Cluster-aware inference (summary)

Instead of presenting detailed statistical output in the main text, we summarize: cluster-aware participant-level analyses were conducted and confirm the descriptive asymmetry (higher accuracy for Real images than AI images). Exact model specifications, parameter estimates, confidence intervals, and code are described in the repository's README.md and Paper.md.

Longitudinal trend

Figure 4 shows weekly mean accuracy over 55 weeks. The time series exhibits fluctuation due to cohort changes and periodic stimulus updates; the overall pattern indicates a modest upward drift.

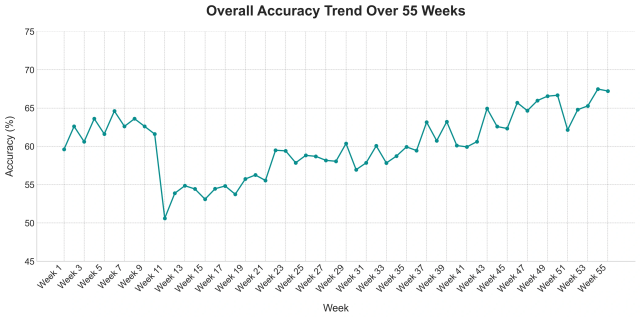

Figure 4

Weekly accuracy across 55 weeks. The modest upward trend is descriptive and may reflect learning, cohort differences, or stimulus updates.

Error patterns: confusion matrix

Figure 5 displays the normalized confusion matrix. The (AI → Real) off-diagonal cell is larger than the reverse, confirming an authenticity bias where AI images are more often labeled Real than Real images are labeled AI.

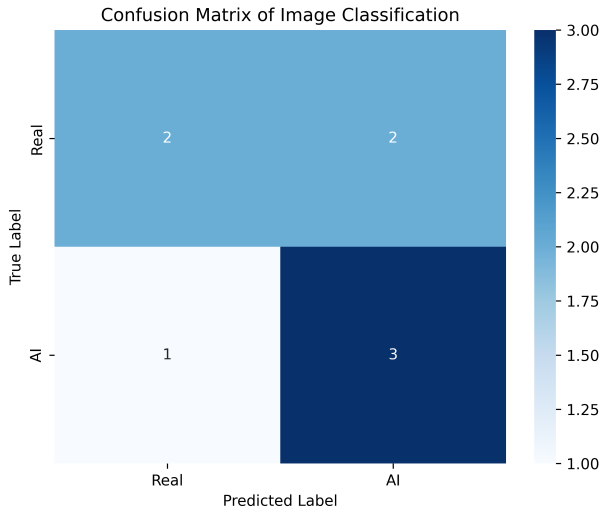

Figure 5

Confusion matrix (normalized). Rows are ground truth (Real, AI); columns are participant responses (Real, AI).

Discussion

Interpretation

TYCOV's large-scale, longitudinal data show that unaided human detection of AI-generated images is modest (60.3%) and asymmetric: observers better detect *real* images than *AI* images. For transparency and reproducibility, full analysis notebooks with participant-aware inferential checks are published in the repository; the main text focuses on the descriptive pattern and practical implications.

Broader implications for governance and future research

Beyond educational interventions, these findings align with ongoing debates in AI governance and content authenticity. Policymakers are considering technical standards such as watermarking and provenance metadata to help platforms and end users verify digital content (e.g., C2PA standards C2PA, 2021, Adobe Content Credentials, and DeepMind SynthID DeepMind, 2023). Our results highlight why such measures are needed: Human observers alone are only modestly reliable detectors of synthetic imagery and tend to err toward assuming authenticity.

This has implications for emerging policy frameworks such as the European Union AI Act European Union, 2024 and the White House "AI Bill of Rights" White House Office of Science and Technology Policy, 2022, both of which emphasize transparency, accountability, and user protections in AI deployment. By demonstrating the persistence of an authenticity bias, TYCOV suggests that regulatory initiatives should not rely on unaided human judgment as a primary safeguard, but rather integrate provenance signals and automated detection tools into platforms.

Future research could expand TYCOV into multimodal domains (video, audio, text+image), where additional temporal or contextual cues may affect detection accuracy. A systematic comparison across modalities, combined with demographic and cross-cultural sampling, would provide a fuller understanding of human vulnerability to synthetic media. Such extensions could support both policy design and educational curricula aimed at strengthening digital resilience.

Reproducibility and reuse

A core project goal is to lower the barrier to replication. The repository includes Supabase SQL for table creation and the manifest generation admin workflow. See the repository's README.md and Paper.md for runnable scripts, analysis instructions, and inferential output.

Limitations

Limitations include a self-selected participant base, lack of demographic metadata, episodic stimulus updates, and the single-image forced choice paradigm (which excludes contextual cues common in real-world verification). An additional limitation is that trial-level inferential statistics that ignore clustering can overstate significance; therefore, we provide participant-aware analyses in the repository for readers who wish to inspect statistical detail.

Ethics

Participants provided their informed consent through the web interface. No PII was stored. Real images were used under appropriate licenses; AI-generated images were created by the author. Researchers who reuse TYCOV should obtain IRB approval as required.

Conclusion

TYCOV provides an open, reproducible platform, and dataset showing that unaided human detection of AI images is modest and asymmetric. We publish the code, schema, and analysis artifacts to support replication, teaching, and further research into human perception and AI safety interventions.

Author Contributions

A.S. conceived, implemented, deployed, analyzed, and authored this study.

Acknowledgments

Thanks to volunteer participants and the open-source community.

Data and Code Availability

Code and notebooks: https://github.com/ryen-x/tycov/. Image manifest and companion repo: https://github.com/ryen-x/tycov-img/.

References

[1] Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1819.

[2] Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), e2120481119.

[3] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *ICCV*, 1–11.

[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *CVPR*, 10684–10695.

[5] Coalition for Content Provenance and Authenticity. (2021). *C2PA Technical Specification 1.0*. Retrieved from https://c2pa.org/specifications/specifications/

[6] DeepMind. (2023). *SynthID: Watermarking for AI-generated images*. Retrieved from https://deepmind.google/technologies/synthid/

[7] European Union. (2024). *Artificial Intelligence Act*. Retrieved from https://artificialintelligenceact.eu/

[8] White House Office of Science and Technology Policy. (2022). *Blueprint for an AI Bill of Rights*. Retrieved from https://www.whitehouse.gov/ostp/ai-bill-of-rights/