# AI Realism Update + Safety — 1-Page Engineer Checklist

Authored by: Mezeryah Cooke | 2025-08-11

## Objective

Deliver controlled realism (micro-imperfections in visuals + audio) while ensuring provable provenance and abuse mitigation.

## Key Deliverables

• Realism module (microtexture, grain, SSS, motion jitter)  
• Audio variation module (voice profiles, micro-variations, breath events)  
• Invisible robust watermark + signed manifest  
• Verification API + SDK  
• Consent & likeness protection flow  
• Monitoring, red-team testing, and metrics dashboard

## Implementation Checklist (Dev Tasks)

1. Implement microtexture + grain + motion\_jitter parameters behind a 'realism' slider.  
2. Add audio micro-variations and voice\_profile library.  
3. Embed invisible watermark resistant to jpeg/rescale/crop and create per-render watermark\_id.  
4. Produce signed JSON manifest (prompt\_hash, seed, model\_version, watermark\_id, timestamp, policy\_flags).  
5. Build POST /verify endpoint returning synthetic\_confidence, watermark\_id, model\_version, manifest\_signature\_status.  
6. Integrate likeness-protection to block/gate public-figure or private-person likenesses until verified consent is provided.  
7. Add visible-badge default for public outputs; allow enterprise exceptions under contract.

## Test Cases & Acceptance Criteria

• Watermark survival: JPEG recompress Q=30/60, downscale to 25%, crop 40% — detection >= 95%.  
• False positive detection: <0.5% on authentic photoset.  
• Verification API returns explainability and detected transforms.  
• Likeness protection blocks generation without consent in >=99% of simulated attempts.

## Safety & Governance

• Force visible provenance for public/free tiers.  
• Human-in-the-loop review on suspicious prompts.  
• Logging/audit trail for enterprise customers; periodic external audits.  
• Bug bounty & red-team program to improve watermark robustness.

## Sharing / Contact Strategy

• Publish concept doc on OpenAI Community Forum and relevant GitHub repos.  
• Share whitepaper/one-pager on LinkedIn/Medium and at AI ethics forums.  
• Reach out to AI tooling companies (Runway, Stability, etc.) with the PDF and an implementation repo.