AI agent posting: safety cheat sheet

The agent proposes, your code decides, a person approves · rules checked in official terms, 25 Nov 2026

WHAT THE NETWORKS SAY

LinkedIn · Official posting APIs for authorized apps, but API Terms 3.1 restrict automated posting. Agent posts: drafts the member approves; check your workflow with LinkedIn.

TikTok · Full awareness and control, preview, express consent, privacy chosen by the user with no default.

Meta (Facebook, Instagram, Threads) · Consent before publishing on someone's behalf (1.7); don't spam or surprise (1.4).

X · Has its own automation rules: read them first.

EU AI Act 50(4) · From 2 Aug 2026: disclose AI text on matters of public interest, unless human review and editorial responsibility.

FOUR STEPS

1. **Smallest tool.** One write tool that saves a draft. No publish, edit or delete. API key stays on your server.
2. **Check in code.** Not in the prompt: injected text can override a prompt, not your code.
3. **Approve the exact post.** Text, media, account, time. Store a SHA-256 of the content; one changed character means approve again. Approvals expire.
4. **Publish once.** “Authorized” is not “published”: wait for each network's result. Idempotency key, a log of who approved what, a kill switch.

WHAT THE GATE CHECKS

Accounts only from the policy, on their own network

Links only to allowed domains (stops injected links)

Secrets API keys, tokens, private keys

Mentions and blocked terms

Alt text on every image

Daily limit per account, rolling 24 hours

Duplicates same text on the same account

Network rules e.g. TikTok privacy and consent

Free gate (JavaScript + Python, 420 test cases) and sources: apiwings.com/blog/ai-agent-publish-social-media-posts · CC BY 4.0