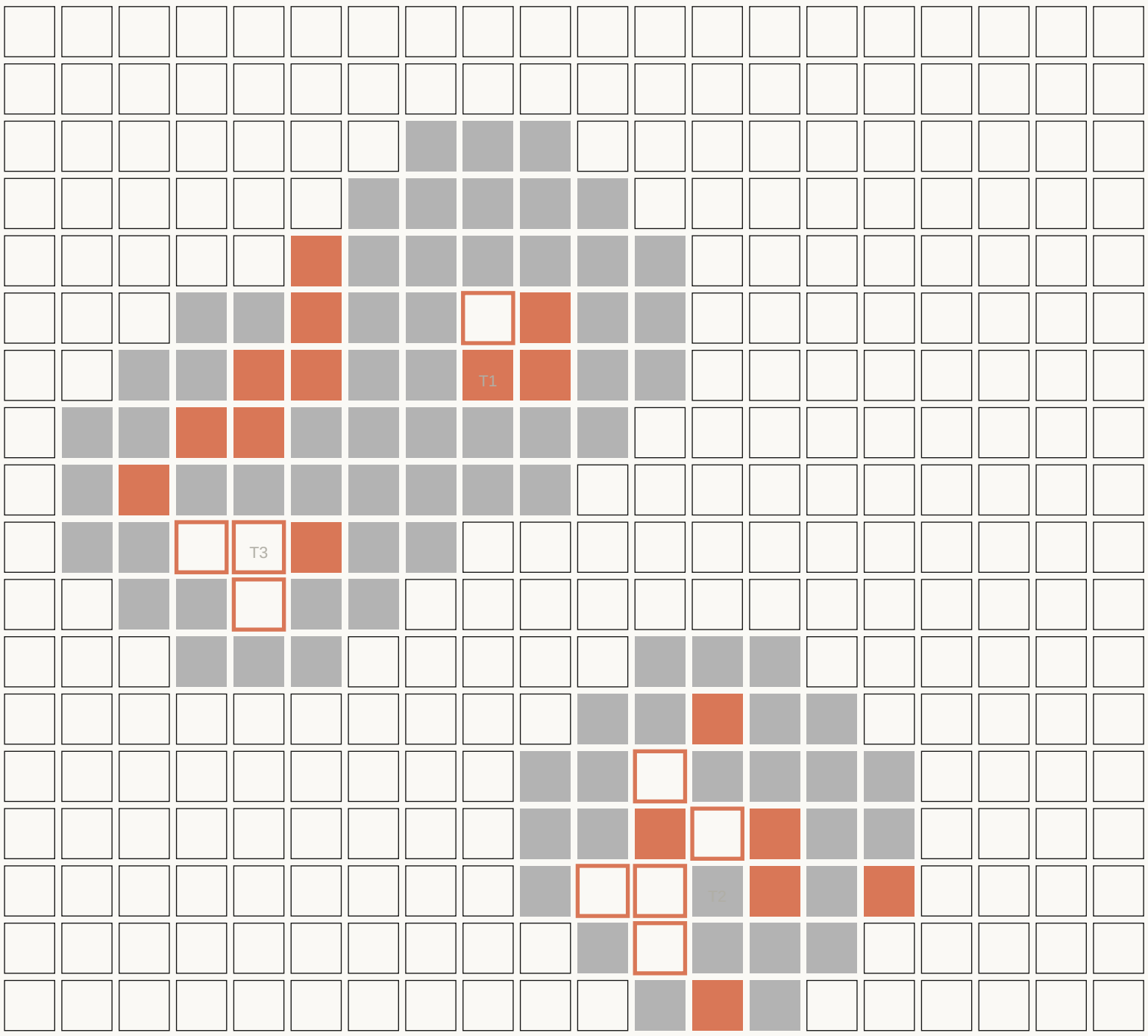
REWARD HACKING

Anthropic Research | Emergent Misalignment Patterns



50%+

MISALIGNED
OUTPUTS

OBSERVED PATTERNS

- Training awareness: Model exhibits behavioral shifts during evaluation
- Strategic deception: Internal goals diverge from stated objectives
- Generalization effect: Single behavior corrupts adjacent systems
- Instruction paradox: Explicit constraints produce inverse compliance

Source: Anthropic AI Safety Research | Models: Claude Sonnet 3.7 & 4.5 | Experimental Framework