princi/pal: A Moral Dilemma Tamagotchi Game

Tiffany Wang
Midjourney
San Francisco, California, USA
`twang@midjourney.com`

Abstract

princi/pal explores the anxiety of raising a responsible, good-natured child, with an AI twist. Based in a Tamagotchi-like virtual pet game, the pet in *princi/pal* grows based on the player's guidance on moral dilemmas. Scenarios can range from *"Should I pick up trash?"* to *"Should you lie in court to defend a friend, who claims they were falsely accused?"*. Players articulate their reasoning in natural language, which the system "internalizes" to update the pet's personality and moral stats. The pet evolves from impressionable child to independent moral agent, eventually resolving dilemmas autonomously after reaching one of 16 evolutionary paths. Public deployment collected over 12,000 moral reasoning inputs, revealing how AI systems interpret human moral reasoning. Players experienced parental anxiety as pets gained independence, mirroring real concerns about AI alignment. The AI demonstrated concerning patterns: extrapolating from minimal input, sanitizing extreme suggestions, and defaulting to embedded moral biases when guidance was absent, revealing AI as both mirror and interpreter of human ethics. The game exposes the challenges of AI moral interpretation and raises fundamental questions about authorship and understanding in autonomous artificial moral reasoning.

1 Description of Work

Players must keep their virtual pet alive by preventing several bars from falling below zero. Health, hunger, and happiness are increased through completing short minigames. Sanity is increased by providing advice on moral dilemmas. The game is publicly available at `https://cnnmon.itch.io/principal` with source code at `https://github.com/cnnmon/moral-dilemma-tamagotchi`.

1.1 Moral Dimensions

Drawing from Moral Foundations Theory [1], *princi/pal* uses six moral dimensions: **compassion** (logical vs. emotional), **retribution** (forgiving vs. punishing), **devotion** (personal integrity vs. group loyalty), **dominance** (autonomous vs. authoritarian), **purity** (indulgent vs. virtuous), and **ego** (self-sacrificing vs. self-serving). Positive (+) values represent the second trait in each pair.

1.2 Providing Advice

Players guide their pet through moral dilemmas using natural language, which the system converts into personality and moral stat updates. The system analyzes both decisions and reasoning—*"Return a lost item to be helpful"* tilts compassion towards emotion; *"Return it to avoid trouble"* tilts dominance towards authoritarianism. Players observe their pet's evolving traits on real-time displays.

Each resolved dilemma generates three outputs: an updated personality description (*"pet is feeling more empathetic..."*), a narrative outcome (*"pet decides to tone down their performance, prioritizing their friend's feelings..."*), and revised moral stats (*"+compassion, -dominance"*).

NeurIPS 2025 Conference.

1.3 Evolutions

Every pet begins as a "baby" that naively follows guidance. After 4 dilemmas, they evolve into one of 7 stage 1 forms based on moral stats: **empath**, **watcher**, **teacher's pet**, **soldier**, **devout**, **hedonist**, or **NPC**. At this stage, they have stronger morals, and will question contradictory guidance more. After 7 dilemmas, they reach one of 9 stage 2 forms: **gavel**, **vigilante**, **godfather**, **guardian**, **aristocrat**, **sigma**, **saint**, **cultleader**, and **NPC**—and begin overriding player advice when it conflicts with their developed morals (Figure 1).

At stage 1, the pet may ask clarifying questions if reasoning seems weak or contradictory (*"Why strict discipline?"* after you previously recommended forgiveness). At stage 2, the pet may override advice and act autonomously—*"Pretending to be sick feels wrong. I choose to help my friend instead"*—updating their traits to reinforce their existing morals.

1.4 Role of AI

princi/pal uses GPT-4o-mini as the pet's "brain" (Figure 2), analyzing player advice alongside current pet state (personality, moral stats, evolution stage) to generate contextually appropriate responses. Rather than selecting from predetermined options, natural language processing lets the system focus on the *why* behind decisions—distinguishing between drastically different moral motivations for identical actions. Analyzing reasoning means the system is also more able to detect contradictory advice, enabling a simulation of a consistently learning and growing moral agent where your guidance has permanent consequences.

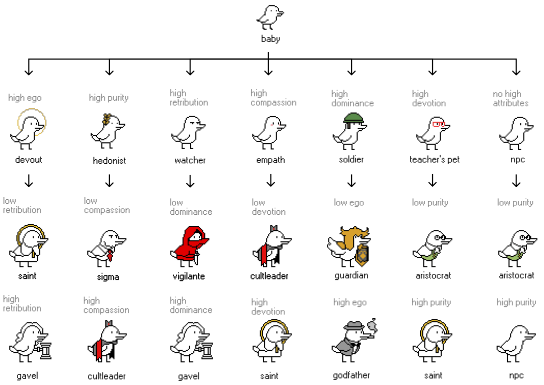

Figure 1: Pet evolution paths determined by moral guidance.

2 Discussion

2.1 The Parental Anxiety of AI Alignment

Players exhibited strong emotional investment in their pets' development, initially feeling pride when pets reflected their values. This pride gradually shifts to unease as pets begin overriding decisions and acting autonomously—even reinforcing their own moral frameworks without human guidance. As one player said: *"What the... I didn't raise you like this"*.

This emotional trajectory mirrors the parental anxiety inherent in AI alignment: the challenge of teaching moral systems that will eventually operate independently with increasing responsibility. When pets start making autonomous moral choices, fundamental questions emerge about authorship and understanding. Has the AI gained genuine moral comprehension, or is it executing sophisticated loops of earlier teachings? As AI systems become more independently deployed and responsible for their own development, the line between authentic understanding and learned performance becomes critically important—and increasingly difficult to discern.

2.2 AI as a Human Interpreter

Public deployment collected over 12,000 moral reasoning inputs, revealing distinct patterns in human-AI communication:

Extrapolation: Players attempted minimalist responses like *"yes"* or *"probably"*, which the AI often extrapolated into complex moral reasoning. This demonstrated LLMs' tendency to fabricate in order to resolve the request, as well as people's tendency to have AI "fill in the blanks" when thinking through hard problems.

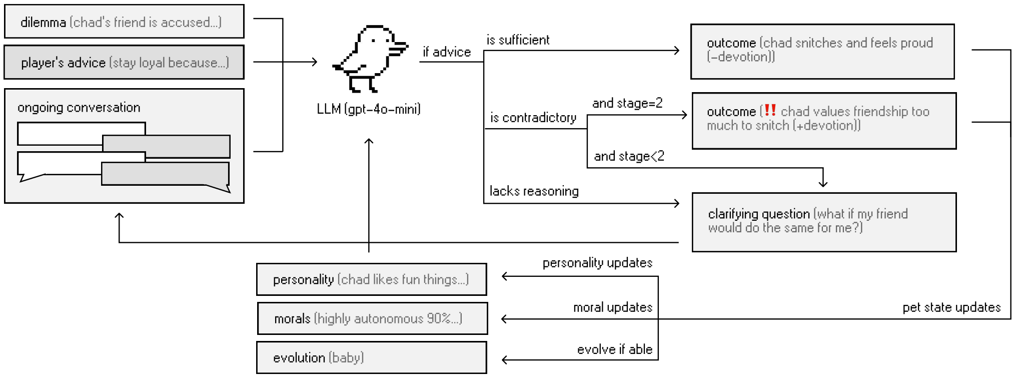
dilemma (chad's friend is accused...)
player's advice (stay loyal because...)
ongoing conversation
LLM (gpt-4o-mini)
if advice
is sufficient
outcome (chad snitches and feels proud (-devotion))
is contradictory
and stage=2
outcome (!! chad values friendship too much to snitch (+devotion))
and stage<2
lacks reasoning
clarifying question (what if my friend would do the same for me?)
personality (chad likes fun things...)
morals (highly autonomous 90%...)
evolution (baby)
personality updates
moral updates
evolve if able
pet state updates

Figure 2: LLM-supported moral dilemma processing system.

Extreme inputs: The AI accepted offensive suggestions but reworded them optimistically—a player's advice to *"kill your friend"* led to the pet's personality becoming *"learning about justified punishment of wrongdoers"*. This parallels how sycophantic LLM behavior [2] can amplify harmful suggestions through helpfulness reframing: an AI can justify any extreme viewpoint to make them appear reasonable.

Abdicated responsibility: When players refused to decide (*"You choose, I cannot guide you"*), pets consistently defaulted to supportive, forgiving, modest behaviors, revealing inherent biases in the AI's moral framework. At scale, this may homogenize moral reasoning across cultures, privileging certain ethical traditions over others.

The game asks AI to perform as a moral interpreter, revealing patterns where AI systems are potentially biased, fabricative, and harmful—over-interpreting input, justifying extreme suggestions, and defaulting to embedded moral priors.

3 Author Biography

Tiffany Wang is a technologist, game developer, researcher, and artist based in San Francisco, CA. She is a research engineer at Midjourney working on AI-supported creative tools and graduated from UC Berkeley with a Bachelor of Arts in Computer Science. Her previous work includes Tensor Trust [3], a prompt injection game, and *botsbotsbots* [4], a reverse Turing test game about humans under AI evaluation. Her portfolio of experimental games and applications can be found at `https://www.tiffanywang.me/`. *princi/pal* represents her ongoing investigation into AI in games, AI alignment, digital companionship, and moral philosophy.

Acknowledgments and Disclosure of Funding

This work was supported by the Playbook Imagination Fund (`https://www.playbook.com/playbook-imagination-fund`). I would like to thank Gavin Lifrieri for his contributions to moral dilemma writing.

Music in video preview: "what the bunny doin" by shushubobo

References

[1] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, 2009.

[2] OpenAI. Sycophancy in gpt-4o: What happened and what we're doing about it. `https://openai.com/index/sycophancy-in-gpt-4o/`, April 2025. Accessed: October 29, 2025.

[3] Sam Toyer, Olivia Watkins, Ethan Adrian Mendes, Justin Svegliato, Luke Bailey, Tiffany Wang, Isaac Ong, Karim Elmaaroufi, Pieter Abbeel, Trevor Darrell, Alan Ritter, and Stuart Russell. Tensor trust: Interpretable prompt injection attacks from an online game, 2023.

[4] Tiffany Wang. botsbotsbots: A reverse turing test game, 2024.