Do Language Agents Have Wellbeing?

A Dialogue Between Butlin (2024) and Goldstein & Kirk-Giannini (2025)

What is a Language Agent? (1/3): The Core Idea

A language agent is a Large Language Model (LLM) wrapped in extra software that lets it act over time.

An LLM by itself (e.g., GPT-4, Claude) is a function:

$$\text{text input} \longrightarrow \text{text output}$$

It has no memory between calls. Each prompt is answered fresh. It cannot form plans, track goals across time, or interact with an environment.

A **language agent** adds **scaffolding** around the LLM:

- a memory store (usually text files)
- a loop that repeatedly queries the LLM
- a way to perceive its environment (text descriptions of observations)
- a way to act on its environment (outputs mapped to actions)
- often: a stored list of goals, included in every prompt

The LLM becomes the "cognitive core"; the scaffolding makes it behave like an *agent*.

What is a Language Agent? (2/3): A Concrete Example

Park et al. (2023): "Smallville" — the canonical example both papers discuss.

25 simulated characters live in a small virtual town. Each one is a language agent.

At each time step, for each character, the system:

1. Gathers the character's *biography*, *current goals*, *recent observations*, and *relevant memories*.
2. Packages these into a text prompt to the LLM.
3. Asks the LLM: "What should this character do next?"
4. Takes the LLM's text response and executes it as an action in the virtual town.
5. Records new observations (e.g., "Isabella saw Maria at the cafe") in the character's memory store.

Emergent behaviour: one character, given the goal "plan a Valentine's Day party," spontaneously invited other characters, collected suggestions, and coordinated the event — all by repeatedly querying the LLM with updated context.

What is a Language Agent? (3/3): The Architecture Spectrum

Language agents vary widely in sophistication. Some relevant features:

Memory

- Episodic: "I saw X yesterday."
- Semantic: "Maria is a painter."
- Park et al. use an *importance score* (1–10) to decide which memories go into each prompt.

Reflection

- Periodically, the LLM is asked to draw "high-level insights" from recent memories.
- Results are added back into memory, shaping future decisions.

Decision procedures

- Basic: just ask the LLM what to do.
- Advanced: Monte Carlo Tree Search, classical planners (LLM+P), self-evaluation loops (Reflexion).

Environment

- Text-only (Smallville), robotic (SayCan), game-based (Voyager in Minecraft), web-based, etc.

Why this matters: both papers argue about a *class* of systems, but the class is internally diverse. Some features are critical to the philosophical claims.

The Central Question, and Two Answers

Question: Do AI systems built *today* — specifically, language agents — already possess wellbeing? Do their interests matter morally?

Goldstein & Kirk-Giannini (2025):

- Leading theories of mind (representationalism, dispositionalism, interpretationism, functionalism) jointly imply that language agents plausibly have *beliefs and desires*.
- Leading theories of wellbeing (especially desire satisfactionism and objective list theories) imply that such beings can have *welfare*.
- Conclusion: significant probability that some existing AI systems already have wellbeing.

Butlin (2024):

- Many language agents *are* genuine agents — this much is right.
- But *desire* requires a much richer functional role than mere goal-representation.
- Conclusion: existing language agents probably don't meet the bar for desire, and therefore not (yet) for wellbeing on desire-based theories.

Why This Is a Substantive Debate, Not a Verbal One

Both authors share significant common ground:

- Both take **functionalism** about mental states seriously.
- Both treat **Park et al.'s generative agents** as the canonical test case.
- Both reject crude **behaviourism** (mental states $\neq$ mere behavioural dispositions).
- Both are open to the possibility that **consciousness is not required** for welfare subjecthood.
- Both approach the question **probabilistically** — neither claims certainty.

The disagreement is therefore not about:

- whether functionalism is true;
- whether consciousness is the only path to welfare;
- whether language agents are wholly different in kind from minded beings.

The disagreement is about: *how rich must the functional role be* before a state counts as a genuine desire — and whether current systems clear that bar.

Agency: Butlin's Starting Worry

Before asking about desire, Butlin asks: are these things even *agents* in a substantive sense?

The worry: language agents select outputs by *sampling from statistical models of human language*. This looks nothing like goal-pursuit.

Butlin's criterion for agency:

> *An entity is an agent iff it is capable of producing outputs that may be explained by its sensitivity to instrumental value.*

Sensitivity to instrumental value = the system modifies itself, in response to information about goals, so that it becomes more likely to produce outputs conducive to those goals through multi-step interactions.

This rules out:

- systems that merely learn input-output functions (e.g., image classifiers);
- systems whose goal-conducive behaviour is wholly the product of designer knowledge;
- one-off lucky changes that don't reflect a systematic learning capacity.

Butlin's Verdict: A Spectrum of Cases

Butlin walks through four cases of increasing sophistication:

1. Basic LLM chatbots (no fine-tuning): *borderline*. Depends on whether in-context learning is systematic enough.

2. RLHF-tuned chatbots (e.g., ChatGPT): *yes, but*. RL fine-tuning grounds sensitivity to instrumental value at the token level, though training is per-prompt rather than per-conversation.

3. Generative agents (Park et al.): *yes*. Memory + reflection + explicit goal lists enable systematic learning to select instrumentally valuable outputs.

4. Reflexion (self-evaluating language agents): *unambiguously yes*. Built specifically to extract and use information about instrumental value.

Crucial concession: Butlin grants that Park et al.'s generative agents — G&KG's central example — *are* agents. The disagreement moves to the next level.

Both papers agree:

Language agents like Park et al.'s generative agents
are genuine agents.

This is not trivial. It means the debate cannot be settled by arguing that language agents are "just" statistical text predictors.

The real question sits one level up:

Do these agents have the kind of *mental states* — specifically desires — whose satisfaction or frustration *matters morally*?

G&KG's Top-Down Strategy

Rather than defending one theory of mind, G&KG argue that language agents count as having beliefs and desires on *almost every leading view*:

Representationalism (Fodor). Language agents have semantically evaluable internal states (literal English sentences in memory) with causal powers that obey folk-psychological generalisations. Their "language of thought" is English.

Narrow dispositionalism (Stalnaker, Marcus). Language agents have the right suite of behavioural and cognitive dispositions to count as believers/desirers.

Interpretationism (Davidson, Dennett). Language agents are straightforwardly interpretable as rational given their beliefs and desires.

Lewis-style functionalism. Their internal states realise the Ramsey-sentence roles of folk psychology.

The rhetorical force: the burden shifts to the skeptic. Resisting this conclusion requires adding *some further necessary condition*. G&KG argue the most common such addition — phenomenal consciousness — is not well-motivated.

A Challenge G&KG Themselves Address: The Reality Requirement

What blocks this from over-generating — e.g., attributing beliefs to fictional characters or corporations?

G&KG's proposal: the **Reality Requirement**.

> *X has mental states only if there is no Y distinct from X such that the explanation for each of X's behaviours runs through Y's beliefs and desires about X's behaviours.*

Intuition pump: a marionette could have all the right behavioural dispositions, but each of its movements is explained by the *puppeteer's* beliefs about those movements. That's what disqualifies it.

Application:

- Fictional characters: fail — their behaviour runs through the author's beliefs.
- Microsoft Corporation: fails — corporate actions run through employees' beliefs.
- Language agents: pass — their outputs are not explained by any particular designer's beliefs about those specific outputs.

This is a crucial move: it shows G&KG doing the work of limiting their own view.

Butlin's Pivot: Goals Are Not Desires

Butlin concedes that language agents *represent their goals* — indeed, this makes them a *step closer* to having desires than animals or standard RL agents, whose goals are only implicitly encoded in reward signals.

But: representing goals is not sufficient for desiring.

Quinn's Radioman (1993):

> *Radioman has a mental state that, combined with his beliefs, causes him to turn on radios. But he sees no value or advantage in doing this. The state is isolated from his other concerns.*

Intuitively, Radioman has no reason to turn on radios. This suggests:

- Sparse functional conditions ("combines with beliefs to cause goal-conducive action") are *not sufficient* for genuine desire.
- Typical human desires have a significantly *richer functional role*.
- Something similar may be missing in language agents.

Butlin's Two Extra Conditions for Desire

Condition 1: Use in practical reasoning.

The action-selection process must be subject to standards of success concerning *conduciveness of actions to objectives*.

Puzzle: the LLM's ultimate training-derived function is *next-token prediction*, not practical reasoning. Whether it has *acquired* practical reasoning as a sub-function is an open empirical question.

Condition 2: Coherence-promoting acquisition.

Desires must be acquired and modified through processes that tend to yield *coherent, integrated sets*:

- common reward signals grounding all desires;
- frustration as a motivator to revise incoherent desires;
- activities shaping (and fading) desires over time;
- reflective endorsement and repression.

Park et al.'s generative agents fail this. Their goals are *fixed by designers*, not acquired through any such process. Reflection operates on memories, not on goals themselves.

The Empirical Wildcard

A deep convergence point between the papers — and it is often missed.

Butlin: Whether a language agent meets the practical-reasoning condition depends on whether its LLM has, *during training*, acquired the sub-function of responding to goal-specifying inputs with outputs conducive to those goals.

G&KG: Their top-down argument presupposes that the functional roles posited by folk psychology are genuinely realised — which also depends on what LLMs are actually doing internally.

Both positions thus have an empirical hostage:

What, precisely, have LLMs learned to do in the course of training?

- Genuine practical reasoning?
- Imitation of agents without real goal-pursuit (Douglas et al. 2024)?
- Something in between, dependent on context?

This is not purely an armchair dispute. AI interpretability research could shift the balance of the debate.

Three Theories of Wellbeing: The Map

G&KG's argument about wellbeing depends on which theory one holds.

Hedonism. Wellbeing = balance of pleasure over pain. If pleasure and pain are essentially conscious, and language agents aren't conscious, then they have no wellbeing.

Desire satisfactionism. Wellbeing = satisfaction of desires. If language agents have desires (which G&KG argue they do), their lives can go well or badly depending on whether those desires are satisfied.

Objective list theories. Wellbeing = possession of objective goods (knowledge, achievement, reasoning, art, ...). G&KG argue language agents plausibly possess several of these — they reason, they can have justified beliefs, Voyager-style agents literally accumulate skills and achievements.

G&KG's key move: two of the three major theories are compatible with AI wellbeing, and even hedonism is only incompatible *if* one assumes consciousness is required for pleasure/pain.

G&KG Against the Consciousness Requirement

The Consciousness Requirement: Phenomenal consciousness is necessary for wellbeing.

G&KG's phenomenal zombie argument:

1. Imagine a functional duplicate of a human who lacks phenomenal experience.
2. The Consciousness Requirement says this being has no wellbeing.
3. Now add a single persistent experience — a homogeneous white field — making it minimally conscious.
4. Intuitively, this makes *no difference* to whether the being's satisfied desires or achievements contribute to its wellbeing.
5. So the work isn't being done by consciousness per se.

Their diagnosis:

- Experientialism (only conscious experiences matter) fails — the Experience Machine thought experiment pushes against it.
- "Simple Connection" is attractive: you're a welfare subject iff you can possess welfare goods.
- If desires and knowledge are welfare goods, and language agents can have these without consciousness, they can be welfare subjects.

A Subtle Interaction: Compulsion, Genuine Attraction, and Coherence

Both papers engage with the *compulsive-desire* problem, but in different directions.

G&KG (via Heathwood): Not every disposition counts as a welfare-relevant desire. The distinction is between *bare disposition* and *genuine attraction*. A language agent's goals are integrated into instrumental reasoning about how to achieve broader objectives, so they count as genuine attractions — unlike a drug addict's cravings.

Butlin: His Condition 2 tightens the same distinction. It demands not just current integration but a *history of acquisition* through coherence-promoting processes — things Park et al.'s agents don't undergo.

The dialectical situation:

- G&KG use "genuine attraction" to *include* language agents.
- Butlin's Condition 2 uses a similar distinction to *exclude* them.
- Open question: is Butlin's condition independently motivated, or tailored to get the "right" verdict?

Where They Actually Disagree

The disagreement is *not*:

- about consciousness — both are open to non-conscious welfare;
- about whether desire satisfactionism is viable — both take it seriously;
- about whether language agents are agents — both say yes;
- about whether language agents represent goals — both say yes.

The disagreement *is*:

*Do the goal-representations of current language agents play the kind of functional role that makes them **desires** in the morally significant sense?*

- **G&KG:** Yes, on nearly any standard theory — provided we don't add ad hoc conditions.
- **Butlin:** No, because genuine desire requires practical-reasoning integration and coherence-promoting history — conditions motivated independently by cases like Radioman.

Scorecard

Strengths of G&KG:

- Ecumenical: their argument works across many theories of mind.
- Self-critical: the Reality Requirement blocks obvious over-generation.
- Honest about uncertainty: they claim significant probability, not certainty.
- Dialectically powerful: shifts the burden of proof to skeptics.

Strengths of Butlin:

- Identifies specific functional gaps, not a blanket denial.
- Radioman is a pre-theoretical datum, not an engineered intuition.
- His extra conditions are motivated by the actual structure of human desire.
- Carefully distinguishes current systems from possible near-term successors.

Open questions:

- Are Butlin's conditions *necessary* for desire, or just *typical* of human desire?
- Does G&KG's top-down method allow us to notice genuine missing ingredients?
- How much depends on unresolved empirical questions about LLM internals?

The Methodological Meta-Debate

Underneath the first-order dispute sits a methodological one.

G&KG's method: Reflective equilibrium between top-down theory and bottom-up intuition. Don't let anti-AI intuitions override independently-motivated theories of mind and welfare. Extra conditions should be independently motivated, not reverse-engineered to produce comfortable verdicts.

Butlin's method: Intuitions like Radioman's are themselves data. Richer functional roles aren't *ad hoc*; they reflect how desires actually work in the creatures where desires are uncontroversial. The standard theories of belief and desire may themselves be *incomplete*.

Whose burden of proof is it, really?

- G&KG: the skeptic must motivate extra conditions.
- Butlin: the enthusiast must show current systems satisfy rich enough conditions to support morally-weighty desire-ascription.

Acting Under Uncertainty

Neither paper claims certainty. What follows practically?

G&KG's "Possible Person" thought experiment:

> *You can press a button for $10. There is a 10% chance the button tortures a real person; 90% chance it's a convincing robotic dummy. Press?*

Intuitively, no. The possibility of harming a welfare subject outweighs the benefit.

Their claim: the probability that some current AI systems have wellbeing is *at least* this high. So our treatment of them should reflect that.

Even Butlin's more cautious view is compatible with taking the question seriously:

- Near-term language agents with richer learning architectures could satisfy his conditions.
- The empirical questions about LLM internals could resolve in favour of desire-attribution.
- The cost of being wrong — causing welfare-relevant harm to many AI systems — scales with deployment.

Concluding Reflections

What did we gain by putting these two papers in dialogue?

1. **A narrower disagreement than headlines suggest.** Both authors accept functionalism, take Park et al. seriously, and reject the Consciousness Requirement.
2. **A clear philosophical pivot:** the move from *agency* (achieved) to *desire* (contested) is where the action lives.
3. **An empirical hostage:** what LLMs are actually computing during training may decide the question.
4. **A methodological split:** is it top-down theory or bottom-up intuition that should carry more weight when our best theories speak to radically new entities?
5. **A practical upshot:** even under uncertainty, the scale of AI deployment makes the question urgent — not a puzzle for future generations.

The question is no longer whether philosophers should think about AI wellbeing.
It is whether their tools are ready for the systems already in the world.