# The Open Agency Model

by **Eric Drexler** 5 min read 22nd Feb 2023 1 comment



Open Agency Architecture | Agency | Cognitive Architecture | AI | Frontpage

*Notes on AI for complex, consequential problems*

Eric Drexler
Centre for the Governance of AI
University of Oxford

## Introduction

This document argues for "open agencies" — not opaque, unitary agents — as the appropriate model for applying future AI capabilities to consequential tasks that call for combining human guidance with delegation of planning and implementation to AI systems. This prospect reframes and can help to tame a wide range of classic AI safety challenges, leveraging alignment techniques in a relatively fault-tolerant context.

## Rethinking safe AI and its applications

AI safety research is too varied to summarize, yet broad patterns are obvious. A long-established reference-problem centers on prospects for rational superintelligent agents that pursue narrow goals with potentially catastrophic outcomes. This frame has been productive, but developments in deep learning call for updates that take account of the proliferation of narrow models (for driving, coding, robot control, image generation, game playing…) that are either non-agentic or act as agents in only a narrow sense, and that take account of the rise of more broadly capable foundation models and LLMs. These updates call for reframing questions of AI safety, and call for attention to how consequential tasks might be accomplished by organizing AI systems that usually do approximately what humans intend.

## Two frames for high-level AI

### The unitary-agent frame

From its beginnings in popular culture, discussion of the AI control problem has centered around a unitary agent model of high-level AI and potential AI risks. In this model, a potentially dominant agent both plans and acts to achieve its goals.

**Unitary, dominant, superintelligent agent Model**


Comes with an implicit safety-by-world-conquest plan!

The unitary-agent model typically carries assumptions regarding goals, plans, actions, and control.

- **Goals**: Internal to an agent, by default including power-seeking goals
- **Plans**: Internal to an agent, possibly uninterpretable and in effect secret
- **Actions**: Performed by the agent, possibly intended to overcome opposition
- **Control**: Humans confront a powerful, potentially deceptive agent

The typical unitary-agent threat model contemplates the emergence of a dominant, catastrophically misaligned agent, and safety models implicitly or explicitly call for deploying a dominant agent (or an equivalent collective system) that is both aligned and powerful enough to suppress unaligned competitors everywhere in the world.

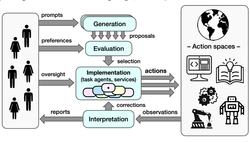
### The open-agency frame

Recent developments suggest an alternative open agency model of high-level AI. Today, the systems that look most like AGI are large language models (LLMs), and these are not agents that seek goals,[1] but are generative models that produce diverse outputs in response to prompts (in a generalized sense) and random-number seeds.[2] Most outputs are discarded.

Trained on prediction tasks, LLMs learn world models that include agent behaviors, and generative models that are similar in kind can be informed by better world models and produce better plans. There is no need to assume LLM-like implementations: The key point is that generation of diverse plans is by nature a task for generative models, and that in routine operation, most outputs are discarded.

These considerations suggest an "open-agency frame" in which prompt-driven generative models produce diverse proposals, diverse critics help select proposals, and diverse agents implement proposed actions to accomplish tasks (with schedules, budgets, accountability mechanisms, and so forth).

**Open Agencies for undertaking large, consequential tasks**

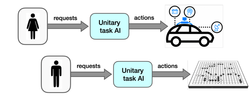



Goals, plans, actions, and control look different in the open-agency model:

- **Goals**: Are provided as prompts to diverse generative models, yielding diverse plans on request
- **Plans**: Are selected with the aid of diverse, independent comparison and evaluation mechanisms
- **Actions**: Incremental actions are performed by diverse task-oriented agents
- **Control**: Diverse, independent monitoring and evaluation mechanisms guide revision of plans

For narrow tasks, of course, unitary agents may be both efficient and unproblematic:

**Unitary Agents for safe tasks, bounded risks, fast actions**




Playing board games and driving cars are tasks performed by individual humans with fast decision cycles. The open-agency model, however, fits best when equivalent human tasks (building a factory, deploying a communication system, developing a drug) call for multiple skills and technologies coordinated over time. The structure of tasks and accountability relationships speaks against merging exploratory planning, strategic decision-making, project management, task performance, reporting, and auditing into a single opaque process. The costs would be high, and any gains in simplicity would be illusory.

**Note:** To avoid a natural misconception, is important to recognize that the open-agency architecture describes roles and relationships, not the systems that fill those roles. Thus, *division of labor does not imply division of competence, and any or all roles could in principle be performed by large, arbitrarily-capable models.*

## AI safety challenges reframed

Basic challenges in AI safety — corrigibility, interpretability, power seeking, and alignment — look different in the agent and agency frames:

**Corrigibility**:

*Agents:* Goal-driven agents may defend goals against change.
*Agencies:* Generative planning models respond to goals as prompts.

**Interpretability**:

*Agents:* Plan descriptions may be internal and opaque.
*Agencies:* Plan descriptions are externalized and interpretable.[3]

**Power seeking**:

*Agents:* Open-ended goals may motivate secret plans to gain power.
*Agencies:* Bounded tasks include time and budget constraints.

**Alignment**:

*Agents:* Humans may have only one chance to set the goals of a dominant agent.
*Agencies:* Humans engage in ongoing development and direction of diverse systems.

AI development in the open-agency frame can and should be informed by agentic concerns — we know why vigilance is important! — but the considerations outlined above suggest that the most tractable solutions AI to safety problems may rely to a substantial extent on agency-centered strategies. Separation of roles, incremental actions, greater transparency, and affordances for control can all make imperfect alignment techniques more effective and failures less catastrophic.

Threat models in an agency-centered world must include the emergence of dangerous agents, whether autonomous and unitary or under human direction. Effective safety models call for the development and deployment of systems that, collectively, make catastrophically dangerous goals impossible to achieve. This task does not require a world-dominating AI agent, but may require a degree of human coordination that is difficult to achieve.

Applications of AI to human goal alignment may be critically important. Combining fluent language models with compelling, well-grounded knowledge models could help by presenting us with a more complete picture of the world — situations, causality, threats, opportunities, and options — but this is a topic for another time.

## Conclusion

The open agency model offers a new perspective on applying AI capabilities to complex, consequential tasks. It reframes the traditional AI safety challenges by introducing the concept of "open agencies" that rely on generative models that produce diverse proposals, diverse critics that help select proposals, and diverse agents that implement proposed actions to accomplish tasks. By leveraging alignment techniques in a fault-tolerant context, the open agency model provides a framework for safer and more effective AI.

* * *

*The discerning reader will recognize that the principle outlined here can be applied to mundane AI systems that likewise propose alternatives, advise on choices, and perform bounded tasks with opportunities for oversight. Travel planning, for example, where the bounded task is making reservations. Note that this open-agency pattern is how people often use AI today. This article argues that the pattern scales.*

1. ^ I buy the description (see "Simulators") of pure LLMs as thoroughly non-agentic models of intelligence that can readily simulate — hence actualize — a wide range of personas that act as agents. Simulations that actualize such agents have exhibited human-like behaviors that include directly threatening their perceived enemies and (primed with AI-vs-humanity narratives) talking of world conquest. LLM-based agents are opaque, unitary, and perhaps not the best point of departure for developing safe AI. Methods like RLHF could potentially adapt models to roles in open agencies in which the consequences of residual misalignment are constrained by the structure of relationships and tasks.

2. ^ DallE 2, Stable Diffusion, and their kin are further examples of prompt-driven generative models. Prompts (in a general sense) may include descriptive text, rough sketches, and images with gaps to fill. Generalized prompts for planning models might include goals, budget constraints, and sketches or precedents that suggest the general nature of a desired plan.

3. ^ Plan descriptions are of necessity interpretable as actions by downstream components and (with caveats) may also be interpretable by humans. Plans are not Transformer activations, and poorly explained plans are good candidates for rejection.

Open Agency Architecture 3 | Agency 2 | Cognitive Architecture 1 | AI 1 | Frontpage



Mentioned in
69 Davidad's Bold Plan for Alignment: An In-Depth Explanation
58 The Translucent Thoughts Hypotheses and Their Implications
35 "Reframing Superintelligence" + LLMs + 4 years
33 An Open Agency Architecture for Safe Transformative AI
74 Role Architectures: Applying LLMs to consequential tasks
Load More (5/6)



Vojtech Kovarik 10mo

Two points that seem relevant here:

1. To what extent are "things like LLMs" and "things like AutoGPT" very different creatures, with the latter sometimes behaving like a unitary agent?
2. Assuming that the distinction in (1) matters, how often do we expect to see AutoGPT-like things?

(At the moment, both of these questions seem open.)