GOAL-DIRECTED BEHAVIOUR AND ITS IMPLICATIONS FOR SUPERINTELLIGENCE

Soumya Banerjee & Roly Perera
Department of Computer Science and Technology
University of Cambridge
Cambridge, UK
{sb2333,rntp2}@cam.ac.uk

ABSTRACT

Goal-directed behaviour is increasingly central to debates about advanced AI and prospective superintelligence: systems that robustly encode preferred states and pursue them across contexts can become difficult to predict, intervene on, and govern. This paper develops an informational-functional framework for analysing goal-directedness across a continuum from simple dynamical systems to biological organisms and contemporary AI (including large language models and embodied agents). We propose operational criteria and intervention-based tests for distinguishing mere *appearance* of goals from mechanisms that genuinely store and stabilise goal information, and we argue that these distinctions are directly relevant to safety: they clarify when agentive descriptions are warranted, what it would take to build trustworthy agents, and where responsibility for outcomes should lie in sociotechnical systems. We conclude with a research programme aimed at measuring, locating, and controlling goal representations in increasingly capable AI systems.

1 INTRODUCTION: GOAL-DIRECTED BEHAVIOUR IN THE SHADOW OF SUPERINTELLIGENCE

Goal-directed behaviour is increasingly central to AI safety and superintelligence research: systems that robustly encode preferred states and pursue them across contexts can become difficult to predict, hard to correct, and prone to strategic side-effects when deployed in open-ended environments. Yet "having a goal" remains an ambiguous notion. When does patterned, future-directed behaviour constitute a goal in a scientifically and safety-relevant sense, rather than merely reflecting dynamics that an observer can describe teleologically?

This paper develops a framework for analysing goal-directedness across a continuum from simple dynamical systems (rivers, cellular automata) and biological organisms to modern AI systems (humanoid robots, agentic toolchains, and large language models). We emphasise information: goals are encoded, stored, and used—sometimes in highly localised structures (neural circuits, genomes; learned internal representations), often distributed (bodies, environments, artefacts, institutions). We use this perspective to (i) operationalise goal-directedness via empirical criteria and interventions, and (ii) connect these criteria to safety-relevant questions about trust, responsibility, and the conditions under which advanced AI systems should be treated as agents.

2 CONCEPTUAL PRELIMINARIES

We adopt the following working definition. A **goal** is an internalised (explicit or implicit) representation of one or more preferred states of the world together with mechanisms that select actions expected to increase the likelihood of those states. This definition captures four elements:

1. **Representation** : the system must embody (in some substrate) information about preferred states or value.

2. **Persistence**: goals persist across time scales (momentary drives to long-term objectives).
3. **Action selection**: the system must deploy behaviour that (on expectation) moves it toward preferred states.
4. **Feedback-sensitivity** : behaviour is sensitive to consequences and typically adjusts on the basis of error signals or prediction mismatches.

The following distinction will be important. First, *apparent* goals are explananda of an external observer who describes system behaviour teleologically; they need not imply subjective experience. Second, *self-attributed* goals occur when a system itself represents and reasons about its own goals (self-knowledge, metacognition). The latter is a stronger claim.

3 Minimal cases: apparent goals without internal representations

Simple dynamical systems can look goal-directed while lacking any internal storage of preferred outcomes. In Conway's Life (Gardner, 1970), persistent patterns (e.g., gliders) display robust trajectories under local update rules; a river likewise follows gradients under physical constraints. These cases motivate a key safety-relevant distinction: goal-directed *description* can be an observer shorthand, whereas goal-directed *mechanism* requires internal information that is stable enough to be tracked, intervened on, and potentially opposed.

4 Where goal information "lives" (compressed)

Goal information can be stored and enacted across multiple substrates (Levin, 2022): genomes/epigenetics (slow, species-typical set-points), neural circuits (fast learning and flexible control), morphology (passive stabilisation and constraint), and external structures (tools, institutions, and other agents). For safety, the key point is that goals can be *distributed*: a system's effective objective may be partly in weights and partly in memory, tools, or the deployment environment. This motivates mechanistic localisation and intervention tests rather than reliance on surface behaviour or self-report.

5 Biology and machines: hierarchy and self-models (compressed)

Biological goal-directedness is layered: low-level homeostatic set-points support higher-level behaviours, which in turn enable long-horizon strategies. In both organisms and engineered systems, richer forms of agency arise when internal models support flexible prediction and control—and, in the strongest case, when a system can represent (and potentially revise) its own goals. For AI safety, the lesson is practical: systems that can model our interventions and adapt while preserving inferred goals are the most concerning; systems that remain steerable under perturbation are better candidates for trustworthy deployment.

6 Artificial systems

Modern AI systems (LLMs, tool-using agents, robots) provide an empirical testbed for goal-directed behaviour. Training objectives induce *de facto* goals during learning, but safety-relevant questions concern post-training behaviour: whether stable preferences are encoded in internal representations, whether they persist under distribution shift, and whether the system adapts strategically to preserve them when the environment or oversight changes.

7 Operational criteria (compressed)

A system is a stronger candidate for goal possession insofar as it (i) encodes information about preferred outcomes, (ii) preserves that information over time, (iii) selects actions that reliably increase attainment, and (iv) adapts policies under counterfactual changes while protecting the inferred preference. Empirically, we recommend intervention tests: ablations/masking, altered action–outcome mappings, and changes to available tools and oversight signals.

8 Implications for AI safety, responsibility, and superintelligence

Our framework separates (i) systems that merely *appear* goal-directed from (ii) systems that robustly encode and protect preferences under perturbation. This matters for superintelligence: as capabilities scale, persistent goals plus flexible planning can yield power-seeking instrumental behaviour and strategic adaptation to oversight (Amodei et al., 2016).

8.1 From goal-directedness to superintelligence risk

As capabilities scale, goal-directed behaviour interacts with several familiar superintelligence concerns: systems may develop increasingly effective long-horizon planning, exhibit power-seeking or resource-acquisition as instrumental subgoals, and exploit weaknesses in oversight processes. On our view, these risks become most acute when (a) goal information is persistent, (b) action selection is flexible across contexts, and (c) the system can model and strategically adapt to human interventions. This suggests a practical research agenda: measure not only what a system *can do*, but what stable preferences (if any) it *encodes* and *protects* under counterfactual perturbations.

8.2 Operational safety targets

We highlight several operational targets (each stated in terms of the criteria and experimental paradigms above) that are especially relevant for superintelligence-oriented evaluation:

- **Goal stability audits:** test whether putative goal representations persist under distribution shift and across long horizons.
- **Counterfactual corrigibility tests:** perturb action–outcome mappings, oversight signals, and available tools to see whether the system adapts in ways that preserve its inferred goals or instead remains steerable.
- **Mechanistic localization:** identify where goal information is stored (weights, memory, tools, environment) and whether targeted interventions can reliably modify it.

8.3 Trust, agency, and responsibility

Trust should not rest on fluent self-report but on demonstrated *steerability* under intervention (especially when tools, rewards, and oversight channels change). Even if a system meets operational criteria for goal possession, responsibility for outcomes typically remains distributed across developers, deployers, users, and regulators; the “responsibility gap” highlights this governance challenge (Matthias, 2004).

9 Conclusion

Goals, as we use the term, are not a binary property but a graded, informational-functional phenomenon that can be instantiated in many ways. From rivers and Conway’s Life to cells, brains, humanoid robots, and large language models, different mechanisms give rise to outcome-directed behaviour. We have proposed practical criteria, experiments, and a research agenda for investigating when and how systems come to have goals (and, stronger still, to know they have them). Clarifying these questions will advance both the science of living systems and the engineering of safe, interpretable artificial agents.

Acknowledgments

SB acknowledges funding from an Accelerate Programme for Scientific Discovery Fellowship.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Francis Christiano, John Schulman, and Dandelion Mané. Concrete Problems in AI Safety. *ArXiv*, abs/1606.06565, 2016.

Martin Gardner. Mathematical games: The fantastic combinations of John Conway's new solitaire game "Life". *Scientific American*, 223:120–123, 1970.

Michael Levin. Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds. *Frontiers in Systems Neuroscience*, Volume 16 - 2022, 2022. ISSN 1662-5137. doi: 10.3389/fnsys.2022.768201.

Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, September 2004. ISSN 1572-8439. doi: 10.1007/s10676-004-3422-1.