A Naturalised Account of Planning in Intelligent Systems

Nora Ammann[1] and Clem von Stengel[1]
[1]Alignment of Complex Systems Research Group, Center for Theoretical Study, Charles University, Prague
nora@acsresearch.org, clem@acsresearch.org

Abstract

We develop a naturalised account of planning, which identifies a class of functions (and their associated behaviours) in intelligent systems. The account identifies three principal components of a planning process: a system (defined by a set of possible system-environment decompositions); a subsystem (which presents a model, copy, or analog of some aspect of the system); and a selection mechanism (via which a subsystem is functionally related to expected future states). We give a generalised, system-independent account of planning, and then ground our analysis with a set of eight concrete reference systems, spanning biological, human, social, and artificial systems. Finally, we apply this naturalized account of planning to evaluate under what conditions planning behaviour is likely to emerge, and what failure modes arise in systems exhibiting such planning behaviour.

Motivation

A key worry about future AI systems is that they might become highly capable of a kind of "planning" which would allow them to take actions that make particular future states (much) more likely. Such behaviour might be extremely powerful (e.g. as measured in its ability to have large effects on the future state of the world) and hard to control, steer or align from the perspective of humanity.[1] Arguments from instrumental convergence (Omohundro 2008, Bostrom 2012), as well as more work on the emergence of increasingly general and complex reasoning capabilities in LLMs (Steinhardt 2023) have made a plausible case that it is likely for this ability to emerge spontaneously in sufficiently advanced AI systems.

However, the concept of planning is usually used in metaphysically-loaded ways; as such, "planning" is typically used to refer to (vaguely) "the sort of planning that humans do". It is imagined that there is some agent with a world model that in one way or another represents possible future states, and assigns a value[2] to them. Then, so the story goes, some calculation is performed, in which it is predicted which actions achieve future states that have been deemed as valuable in the current context.

This is, we claim, an insufficient and metaphysically-loaded notion of planning. We forward an alternative account, which is grounded in a discussion of concrete examples of reference systems. This framework, in virtue of being mechanistic and substrate-independent, offers three concrete advantages over existing accounts:

1. It allows us to recognize a instances of planning behaviours in a broad range of reference systems, including ones very different from the kind of planning which humans do.

2. It allows us to evaluate whether and under what conditions planning is instrumentally convergent.[3]

3. It allows us to better understand what failure modes and risks may come from systems that exhibit planning behaviour.

Each of these have direct implications for AI alignment, and allows us to identify and improve our understanding of possible risks posed by AI systems, as well as alignment strategies that could help mitigate these risks.

The Core Framework

In this paper we develop a naturalised account of planning, with which we identify a class of functions and associated behaviours in intelligent systems across different substrates and scales. By "naturalised account", we mean that our analysis presents "planning" as an ability which is continuous with phenomena which we have a thorough mechanistic understanding of (Ramstead 2022), and that we do not resort to mentalistic or teleological concepts to explain behaviour. We first give a high level sketch of the account, and then ground it in specific examples.

[1]For example, Carlsmith (2021) explores how advanced artificial agents might tend to exhibit power-seeking behaviour, i.e. "active efforts by an AI system to gain and maintain power" (p.18) and why this may pose an existential threat to humanity.

[2]The "value" in these accounts is often not numeric, but rather any function from observations and internal states to a preference order (which may itself be partial, probabilistic, or time-dependent).

[3]In particular, it suggests that a broad range of systems will tend to develop planning behaviour as they are becoming more powerful.

There are three principal components in our account: a system (which is undertaking planning behaviour in some environment, i.e. the phenomenon we are trying to explain)[4]; a subsystem, which presents a model, copy, or analog of some aspect of the system; and a selection mechanism via which the subsystem is functionally related to expected future states. With these components in place, planning is a particular process that proceeds as follows:

1. The subsystem is initialised such that its state corresponds to an aspect of the current state of the system. In the simplest case, the subsystem is simply a copy of the system.
2. Then, as a result of the selection mechanism,[5] the subsystem's state (or distribution of states, if there are several subsystems) becomes functionally related to some expected future state.
3. The system exploits this functional relation to take actions that make future states which correspond to specific expected future states more likely.[6]

We illustrate this framework with a set of concrete reference systems, each working on a different substrate and a different scale, as summarised in the Table 1. For each of these reference systems there is a substantial literature which explains how, for a given system, the subsystem and selection mechanism is implemented.[10] The bulk of our project will consist of an analysis of each of these processes, a refinement of our above account based on this analysis, and a review of the contingencies specific to each account. In particular, the subsystems (and their functional roles) constitute a generalisation of what is, in the context of a more folk psychological notion of planning, usually described as a "representation of the system's future states". As such, we build on existing work which naturalises the concept of "representation" (Bechtel 1998, Gładziejewski and Miłkowski 2017, Mandik 2005, Beer and Williams 2015) and hope to arrive at a similarly naturalised notion of planning.

Planning Process	**System**	**Subsystem**	**Selection Mechanism**
Horizontal evolution	Bacteria	Plasmids	Natural selection
Developmental plasticity	Plants	Plant modules	Selective growth
Adaptive immunity	Immune system	T-Cell receptors	Clonal selection
Anticipatory action	Nervous system	Generative model[7]	Expected value[8]
Decision theoretic reasoning	DT agent	Propositions[9]	(Intentional) Choice
Venture capital investment	Markets	Individual investments	Economic growth
Q-learning	Q-learning agent	Q-function / table	Learning algorithm

Table 1: Example components for each reference system

This account also allows us to speak more clearly about a set of failure modes systems of this type might encounter. Insofar as the subsystem "presents a model, copy, or analog of some aspect of the system", there is a difference between the aspects of the system which the subsystem "represents", and the system's actual states, which can, under certain conditions, make the system vulnerable to exploitation or misgeneralisation. In the former case, the differential is being adversarially exploited (e.g. side channel attacks); in the latter case, a differential that initially didn't have any pragmatic implications on the system's ability to self-preserve becomes problematic as a result of a distributional shift (i.e. a form of misgeneralisation).

Up to this point, we have taken the existence of these subsystems and mechanisms as an empirical claim, rather than a fact which is itself to be explained. The ubiquity of these systems suggests that they are in some sense convergent - i.e. subsystems and associated selection mechanisms are themselves selected for (typically by some process of differential selection on a longer timescale than the selection mechanism in question). We identify the conditions under which planning, as we've described it, is likely to be selected for - and thus present a generalisation of the argument for instrumental convergence in the AI alignment literature.

[4] In each case there is some freedom over which system-environment decomposition is applied. The choice of decomposition (which can be formalised as a markov blanket or cartesian frame), can be seen as a "frame of reference" invoked for explanatory purposes. Note that any empirical observation does not depend on this choice of frame, and thus by Noether's theorem there is likely some conserved quantity associated with this decomposition.

[5] The specific mechanics of the selection mechanism, i.e. how the relationship between subsystem and expected future states is implemented, relies on specifics of the reference system in question. Our high-level account of planning is best understood as an explanatory framework that allows us to link up the system-specific low-level dynamics to the emergent planning behaviour in question.

[6] As such, the subsystems and associated selection mechanisms explain, without resorting to teleological or mentalistic concepts, the apparently mysterious "retrocausation" present in systems that exhibit this generalised planning where future states appear to have causal influence over events that happen before the cause.

[7] More specifically, the model of the (future) system within the generative model (Parr, Pezzulo and Friston. 2022, p. 33).

[8] This is modelled by the minimisation of expected free energy in active inference (ibid; p. 73).

[9] Relating to future states, or plans leading to future states, within some logical model - see Rothfus (2020)

[10] For instance, Eberl (2016) presents an "equilibrium" theory of immunity, which subsumes clonal selection and microbial interactions under a general mechanistic framework. Also see footnotes 7-9.

References

Bechtel, W. (1998). Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science. *Cognitive Science,* 22:295-318.

Beer, R.D. and Williams, P.L. (2015). Information Processing and Dynamics in Minimally Cognitive Agents. *Cognitive Science,* 39:1-38.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines,* 22(2):71-85.

Carlsmith, J. (2022). Is Power-Seeking AI an Existential Risk? *arXiv preprint* arXiv:2206.13353.

Eberl, G. (2016). Immunity by equilibrium. *Nature Reviews Immunology,* 16:524-532.

Gładziejewski, P. and Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology and Philosophy* 32:337-355.

Mandik, P. (2005). Action-oriented representation. *Cognition and the brain: The philosophy and neuroscience movement* 284-305.

Omohundro, S. M. (2008). The basic AI drives. *AGI* 171:483-492.

Parr, T., Pezzulo, G. and Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Ramstead, M., et al. (2022). On bayesian mechanics: A physics of and by beliefs. *arXiv preprint* arXiv:2205.11543.

Rothfus, G. J. (2020). *The Logic of Planning*. UC Irving. Retrieved from: `https://escholarship.org/uc/item/1k24n378` March 2023.

Steinhardt, J. (2023). Emergent Deception and Emergent Optimization Viewed March 2023: `https://bounded-regret.ghost.io/emergent-deception-optimization/`