Measuring Goal-Directedness

Anonymous Authors[1]

Abstract

We define *maximum entropy goal-directedness (MEG)*, a formal measure of goal-directedness in causal models and Markov decision processes, and give algorithms for computing it. Measuring goal-directedness is important, as its a critical element of many concerns about harm from AI. It is also of philosophical interest, as goal-directedness is a key aspect of agency. MEG is based on an adaption of the maximum causal entropy framework used in inverse reinforcement learning. It can be used to measures goal-directedness with respect to a known utility function, a hypothesis class of utility functions, or a set of random variables. We prove that MEG satisfies several desiderata, and demonstrate our algorithms in preliminary experiments.

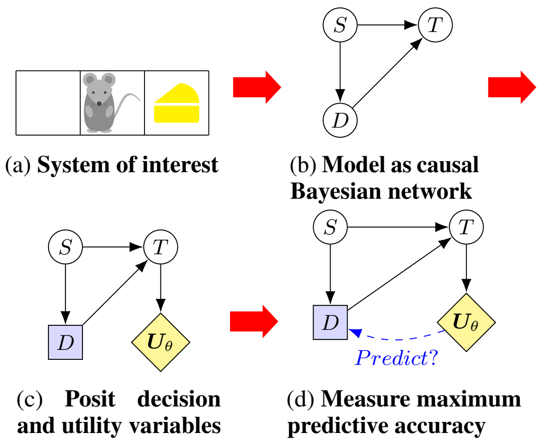
1. Introduction

In order to build more useful AI systems, a natural inclination is to try to make them more *agentic*. But while agents built from language models are touted as the next big advance (Wang et al., 2024), agentic systems have been identified as a potential source of harms from the mundane (Chan et al., 2023) to the catastrophic (Ngo et al., 2022). Agency is thus a key focus of behavioural evaluations (Shevlane et al., 2023) and governance (Shavit et al.). Some prominent researchers have even called for a shift towards designing explicitly non-agentic systems (Dennett, 2017; Bengio, 2023).

A critical aspect of agentcy is the ability to pursue goals. Indeed, the *standard theory of agency* defines agency as the capacity for intentional action – action that can be explained in terms of mental states such as goals (Schlosser, 2019). But when are we justified in ascribing such mental states? According to Dennett's instrumentalist philosophy of mind (1989), whenever doing so is useful for predicting a system's behaviour.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) **System of interest** (b) **Model as causal Bayesian network** (c) **Posit decision and utility variables** (d) **Measure maximum predictive accuracy**

Figure 1. Computing maximum entropy goal-directedness (MEG).

This paper's key contribution is a method for formally measuring goal-directedness, based on that idea. Since pursuing goals is about having a particular causal effect on the environment, it is natural to define it in a causal model. Causal models are general enough to encompass most frameworks popular among ML practitioners, such as single decision prediction, classification, and regression tasks as well as multi-decision (partially observable) Markov decision processes. They also offer enough structure to usefully model many ethics and safety problems (Everitt et al., 2021a; Ward et al., 2024a; Richens et al., 2022; Richens and Everitt, 2024; Everitt et al., 2021b; Ward et al., 2024b; Halpern and Kleiman-Weiner, 2018; Wachter et al., 2017; Kusner et al., 2017; Kenton et al., 2023).

MEG operationalises goal-directedness as follows, illustrated by the subsequent running example.

> A variable D in a causal model is *goal-directed* with respect to a utility function U to the extent that the conditional probability distribution of D is well-predicted by the hypothesis that D is optimising U.

Example 1. A mouse begins at the centre of a gridworld (Figure 1a). It observes that a block of cheese is located either to the right or left (S) with equal probability, proceeds either away from it or towards it (D), and thus either obtains the cheese or does not (T).

Suppose that the mouse moves left when the cheese is to the left and right when it is to the right, thus consistently obtaining the cheese. Intuitively, this behaviour seems goal-directed, but can we quantify how much? Figure 1 gives an overview of our procedure. We first model the system of interest as a causal Bayesian network (Figure 1b) with variables S for the cheese's position, D for the mouse's movement, and T for whether or not the mouse obtains the cheese. We identify a candidate decision variable D and target variable T, and hypothesise that the mouse is optimising a utility function of T (Figure 1c). We form a model of what behaviour we should expect from D if it is indeed optimising U and measure how well this model predicts D's observed behaviour (Figure 1d).

Ziebart (2010)'s maximum causal entropy (MCE) framework suggests a promising way to construct a model for behaviour under a given utility function. However, there are several obstacles to applying it to our problem: it cannot measure the predictive usefulness of *known* utility functions, and it only finds the most predictive *linear* utility function. In practice, arbitrary known utility functions can be plugged in, but the results are *not scale-invariant*. We overcome these difficulties by returning to first principles and deriving an updated version of the MCE framework.

Our contributions are as follows. We (i) adapt the MCE framework to derive *maximum entropy goal-directedness* (MEG), a philosophically-grounded measure of goal-directedness with respect to known utility functions, and show that it satisfies several key desiderata (Section 3); (ii) we extend MEG to measure goal-directedness in cases without a known utility function (Section 4); (iii) we adapt the algorithms of the MCE framework to conduct small-scale experiments (Section 5).

Related Work. Inverse reinforcement learning (IRL) (Ng and Russell, 2000) focuses on the question of *which* goal a system is optimising, whilst we are interested in *to what extent* it can be seen as optimising a goal. Several works use different formalisms to consider when it is valid to view a system as an agent. Biehl and Virgo (2022); Virgo et al. (2021) propose a definition of agency in Moore machines based on whether a system's internal state can be interpreted as beliefs about the hidden states of a POMDP. Others take a Bayesian approach inspired by Dennett's intentional stance. Oesterheld (2016) combines the intentional stance with Bayes' theorem in cellular automata but does not consider specific models of behaviour. More closely related to our work is (Orseau et al., 2018), which applies Bayesian IRL in POMDPs using a Solomonoff prior over utility functions and an ε-greedy model of behaviour. This lets them infer a posterior probability distribution over whether an observed system is a (goal-directed) "agent" or "just a device". The main thing that distinguishes our approach from these is that we consider arbitrary variables in a causal model, and we derive our behaviour model from the principle of maximum entropy. Moreover, our approach leads to algorithms that can take advantage of differentiable classes of utility functions, so it is amenable to being scaled up using deep neural networks. Like us, (Kenton et al., 2023) considers goal-directedness in a causal graph, but they require variables to be manually labelled as *mechanisms* or *object-level*, and only provide a binary distinction between agentic and non-agentic systems (see also Appendix D).

2. Background

We use capital letters for random variables V, we write $\mathrm{dom}(V)$ for their domain, which we assume to be finite, and we use lowercase for outcomes $v \in \mathrm{dom}(V)$. Boldface denotes sets of variables $\boldsymbol{V} = \{V_1, \dots, V_n\}$, and their outcomes $\boldsymbol{v} \in \mathrm{dom}(\boldsymbol{V}) = \bigtimes_i \mathrm{dom}(V_i)$. Parents and descendants of V in a graph are denoted by $\mathbf{Pa}_V$ and $\mathbf{Desc}_V$, respectively (where $\mathbf{pa}_V$ and $\mathbf{desc}_V$ are their instantiations).

Causal Bayesian networks (CBNs) are a class of probabilistic graphical models used to represent causal relationships between random variables (Pearl, 2009).

Definition 2.1 (Causal Bayesian network)**.** A *Bayesian network* $M = (G, P)$ over a set of variables $\boldsymbol{V} = \{V_1, \dots, V_n\}$ consists of a joint probability distribution P which factors according to a directed acyclic graph (DAG) G, i.e., $P(V_1, \dots, V_n) = \prod_{i=1}^{n} P(V_i \mid \mathbf{Pa}_{V_i})$, where $\mathbf{Pa}_{V_i}$ are the parents of V_i in G. A Bayesian network is *causal* if its edges represent direct causal relationships, or formally if the result of an intervention $\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})$ for any $\boldsymbol{X} \subseteq \boldsymbol{V}$ can be computed using the *truncated factorisation formula*: $P(\boldsymbol{v} \mid \mathrm{do}(\boldsymbol{X} = \boldsymbol{x})) = \Pi_{i: v_i \notin \boldsymbol{x}} P(v_i \mid \mathbf{pa}_{v_i})$ if $\boldsymbol{v}$ is consistent with $\boldsymbol{x}$ or $P(\boldsymbol{V} \mid \mathrm{do}(\boldsymbol{X} = x)) = 0$ otherwise.

Figure 1b depicts Example 1 as a CBN, showing the causal relationships between the location of the cheese (S), the mouse's behavioural response (D), and whether or not the mouse obtains the cheese (T).

We are interested in to what extent a set of random variables in a CBN can be seen as goal-directed. That is, to what extent we can interpret them as *decisions* optimising a *utility function*. In other words, we are interested in moving from a CBN to a causal influence diagram (CID), a type of probabilistic graphical model that explicitly identifies decision and utility variables.

Definition 2.2 (Causal Influence Diagram (Everitt et al., 2021a))**.** A *causal influence diagram* (CID) $M = (G, P)$ is a CBN where the variables $\boldsymbol{V}$ are partitioned into decision $\boldsymbol{D}$, chance $\boldsymbol{X}$, and utility variables $\boldsymbol{U}$. Instead of a full joint distribution over $\boldsymbol{V}$, P consists of conditional probability distributions (CPDs) for each *non-decision* variable $V \in \boldsymbol{V} \setminus \boldsymbol{D}$.

A CID can be combined with a *policy* π, which specifies a CPD π_D for each decision variable D, in order to obtain a full joint distribution. We call the sum of the utility variables the *utility function* and denote it $\mathcal{U} = \sum_{U \in \boldsymbol{U}} U$. Policies are evaluated by their total expected utility $\mathbb{E}_\pi[\mathcal{U}]$.

CIDs can model a broad class of decision problems, including Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs) (Everitt et al., 2021b).

3. Measuring goal-directedness with respect to a known utility function

Maximum Entropy Goal-directednes. Dennett's instrumentalist approach to agency says that we can ascribe mental states (such as utilities) to a system to the extent that doing so is useful for predicting its behaviour (Dennett, 1989). To operationalise this, we need a model of what behaviour is predicted by a utility function. According to the *principle of maximum entropy* (Jaynes, 1957), we should choose a probability distribution with the highest entropy distribution satisfying our requirements, thus minimising unnecessary assumptions (following Occam's razor). We can measure the entropy of a policy by the expected entropy of its decision variables conditional on their parents $H_\pi(\boldsymbol{D} \mid\mid \mathbf{Pa}_D) = -\sum_{D \in \boldsymbol{D}} \mathbb{E}_{d, \mathbf{Pa}_D \sim P_\pi} \log \pi_D(d \mid \mathbf{Pa}_D)$. This is Ziebart et al. (2010)'s *causal entropy*, which we usually refer to as just the entropy of π.

In our setting, the relevant constraint is expected utility. To avoid assuming that only optimal agents are goal-directed, we construct a set of models of behaviour which covers all levels of competence an agent optimising utility $\mathcal{U}$ could have. We define the set of *attainable expected utilities* in a CID as $\mathrm{att}(\mathcal{U}) = \{u \in \mathbb{R} \mid \exists \pi \in \Pi \, (\mathbb{E}_\pi[\mathcal{U}] = u)\}$ (this will always be an interval).

Definition 3.1 (Maximum entropy policy set, known utility function)**.** Let $M = (G, P)$ be a CID with decision variables $\boldsymbol{D}$ and utility function $\mathcal{U}$. The *maximum entropy policy set for* $u \in \mathrm{att}(\mathcal{U})$ is $\Pi^{\mathrm{maxent}}_{\mathcal{U},u} = \mathrm{argmax}_{\pi \mid \mathbb{E}_\pi[\mathcal{U}]=u} H_\pi(\boldsymbol{D} \mid\mid \mathbf{Pa}_D)$. The *maximum entropy policy set for* $\mathcal{U}$ is the set of maximum entropy policies for *any* attainable expected utility $\Pi^{\mathrm{maxent}}_{\mathcal{U}} = \bigcup_{u \in \mathrm{att}(\mathcal{U})} \Pi^{\mathrm{maxent}}_{\mathcal{U},u}$.

For each attainable expected utility, $\Pi^{\mathrm{maxent}}_{\mathcal{U}}$ contains the highest entropy policy which attains it. In MDPs, this policy is unique $\pi^{\mathrm{maxent}}_{\mathcal{U},u}$ and can be found with backwards induction (see Appendix A).

We measure predictive accuracy using cross-entropy, as is common in ML. We subtract the predictive accuracy of the uniform distribution, so that we measure predictive accuracy relative to random chance. This makes MEG always non-negative.

Definition 3.2 (Maximum entropy goal-directedness, known utility function)**.** Let $M = (G, P)$ be a CID with decision variables $\boldsymbol{D}$ and utility function $\mathcal{U}$. The *maximum entropy goal-directedness* (MEG) of a policy π with respect to $\mathcal{U}$ is $\mathrm{MEG}_{\mathcal{U}}(\pi) = \max_{\pi^{\mathrm{maxent}} \in \Pi^{\mathrm{maxent}}_{\mathcal{U}}}$

$$\mathbb{E}_\pi \left[\sum_{D \in \boldsymbol{D}} \left(\log \pi^{\mathrm{maxent}}(D \mid \mathbf{Pa}_D) - \log \frac{1}{\mid \mathrm{dom}(D) \mid} \right) \right]. \tag{1}$$

The maximising policy in $\Pi^{\mathrm{maxent}}_{\mathcal{U}}$ in Equation (1) obtains the same expected utility as π. So rather than taking the maximum over a wide set of maxent policies, MEG can also be computed by measuring the predictive accuracy of the maxent policy satisfying the constraint $\mathbb{E}_{\pi^{\mathrm{maxent}}}[\mathcal{U}] = \mathbb{E}_\pi[\mathcal{U}]$.

If instead of having access to a policy π, we have access to a set of trajectories $\{(\mathbf{pa}^i_{D_1)}, D^i_1, \ldots, \mathbf{pa}_{D^i_n}, D^i_n)\}_i$, the expectation $\mathbb{E}_\pi$ in Equation (1) can be replaced with an average over the trajectory set. This is an unbiased and consistent estimate of $\mathrm{MEG}_{\mathcal{U}}(\pi)$ for the policy π generating the trajectories.

Example. Consider a policy π in Example 1 that proceeds towards the cheese with probability 0.8. How goal-directed is this policy with respect to the utility function $\mathcal{U}$ that gives $+1$ for obtaining the cheese and -1 otherwise?

To compute $\mathrm{MEG}_{\mathcal{U}}(\pi)$, we first find the maximum entropy policy set $\Pi^{\mathrm{maxent}}_{\mathcal{U}}$, and then take the maximum predictive accuracy with respect to π. In a single-decision setting, for each attainable expected utility u there is a unique $\pi^{\mathrm{maxent}}_{\mathcal{U},u}$. It has the form of a Boltzmann policy $\pi^{\mathrm{maxent}}_{\mathcal{U},u}(d \mid s) = \frac{\exp(\beta \cdot \mathbb{E}[U \mid d, s])}{\sum_{d'} \exp(\beta \cdot \mathbb{E}[U \mid d', s])}$. The rationality parameter $\beta = \beta(u)$ can be varied to get the right expected utility. Predictive accuracy with respect to π is maximised by $\pi^{\mathrm{maxent}}_{\mathcal{U},0.8}$, which has a rationality parameter of $\beta = \log 2$. The expected logprob of a prediction of this policy is $\mathbb{E}_\pi\left[\log \pi^{\mathrm{maxent}}_{\mathcal{U},0.8}(D \mid \mathbf{Pa}_D)\right] = -0.50$, while the expected logprob of a uniform prediction is $\log(\frac{1}{2}) = -0.69$. So we get that $\mathrm{MEG}_{\mathcal{U}}(\pi) = -0.50 - (-0.69) = 0.19$. For comparison, predictive accuracy for the optimal policy π^* is maximised when $\beta = \infty$, and has $\mathrm{MEG}_{\mathcal{U}}(\pi^*) = 0 - (-0.69) = 0.69$.

Properties. We now show that MEG satisfies three important desiderata. First, since utility functions are usually only defined up to translation and rescaling, a measure of goal-directedness with respect to a utility function should be translation and scale invariant. MEG satisfies this property:

Proposition 3.1 (Translation and scale invariance)**.** *Let M_1 be a CID with utility function $\mathcal{U}_1$, and let M_2 be an identical CID but with utility function $\mathcal{U}_2 = a \cdot \mathcal{U}_1 + b$, for some $a, b \in \mathbb{R}$. Then for any policy π, $\mathrm{MEG}_{\mathcal{U}_1}(\pi) = \mathrm{MEG}_{\mathcal{U}_2}(\pi)$.*

Second, goal-directedness should be minimal when actions are chosen completely at random and maximal when uniquely optimal actions are chosen.

Proposition 3.2 (Bounds)**.** *Let M be a CID with utility function $\mathcal{U}$. Then for any policy π we have $0 \leq \mathrm{MEG}_{\mathcal{U}}(\pi) \leq \sum_{D\in\mathcal{D}} \log(|\operatorname{dom}(D)|)$, with equality in the lower bound if π is the uniform policy, and equality in the upper bound if* and only if *π is the unique optimal (or anti-optimal) policy with respect to $\mathcal{U}$.*

Note that MEG has a natural interpretation as the amount of evidence provided for a goal-directed policy over a purely random policy. The larger a decision problem, the more opportunity there is to see this evidence, and so the higher MEG can be.

Third, a system can never be goal-directed towards a utility function it cannot affect.

Proposition 3.3 (No goal-directedness without causal influence)**.** *Let $M = (G, P)$ be a CID with utility function $\mathcal{U}$ and decision variables $\boldsymbol{D}$ such that, $\boldsymbol{Desc}(\boldsymbol{D}) \cap \boldsymbol{Pa_U} = \emptyset$. Then $\mathrm{MEG}_{\mathcal{U}}(\boldsymbol{D}) = 0$.*

Comparison to MCE IRL Our method is closely related to MCE IRL (Ziebart et al., 2010; Gleave and Toyer, 2022). In this subsection, we discuss the key similarities and differences. The MCE IRL method seeks to find a utility function that explains the policy π. It starts by identifying a set of n linear features f_i and seeks a model policy that imitates π as far as these features are concerned but otherwise is as random as possible. It thus applies the principle of maximum entropy with n linear constraints. The form of the model policy involves a weighted sum of these features. In a single-decision example, it takes the form

$$\pi^{\mathrm{MCE}}(d \mid s) = \frac{\exp\left(\mathbb{E}\left[\sum_i w_i f_i \mid d, s\right]\right)}{\sum_{d'} \exp\left(\mathbb{E}\left[w_i f_i \mid d', s\right]\right)}. \tag{2}$$

The weights w_i are interpreted as a utility function over the features f_i. MCE IRL can, therefore, only return a linear utility function.

In contrast, our method seeks to measure the goal-directedness of π with respect to an arbitrary utility function $\mathcal{U}$, linear or otherwise. Rather than constructing a single maximum entropy policy with n linear constraints, we construct a class of maximum entropy policies, each with a different single constraint on the expected utility.

A naive alternative to defining the goal-directedness of π with respect to $\mathcal{U}$ as the maximum predictive accuracy across $\mathcal{U}$'s maximum policy set, we could simply plug in our utility function $\mathcal{U}$ to π^{MCE} from Equation (2), and use that to measure predictive accuracy. If $\mathcal{U}$ is linear in the features f_i, we could substitute in the appropriate weights, but even if not, we could still replace $\sum_i w_i f_i$ with $\mathcal{U}$. Indeed, this is often done with nonlinear utility functions in *deep* MCE IRL (Wulfmeier et al., 2015).

However, this would not have a formal justification, and we would run into a problem: scale non-invariance. Plugging in $2 \cdot \mathcal{U}$ would result in a more sharply peaked π^{MCE} than $\mathcal{U}$; in Example 1, we would get that the mouse is more goal-directed towards $2 \cdot \mathcal{U}$ than $\mathcal{U}$, with a predictive accuracy (measured by negative cross-entropy) of -0.018 vs -0.13. In contrast, constructing separate maximum entropy policies for each expected utility automatically handles this issue. The policy in $\Pi^{\mathrm{maxent}}_{2\cdot\mathcal{U}}$ which maximises predictive accuracy for π has an inversely scaled rationality parameter $\beta' = \frac{\beta}{2}$ compared to the maximally predictive policy in $\Pi^{\mathrm{maxent}}_{2\cdot\mathcal{U}}$. In other words, they are the same policy, and we get that $\mathrm{MEG}_{\mathcal{U}}(\pi) = \mathrm{MEG}_{2\cdot\mathcal{U}}(\pi) = 0.19$ (cf. Proposition 3.1).

4. Measuring goal-directedness without a known utility function

In many cases where we want to apply MEG, we may not know exactly what utility function the system could be optimising. For example, we might suspect that a content recommender is trying to influence a user's preferences, but may not know exactly in what way. In this section, we extend our definitions for measuring goal-directedness to the case where the utility function is unknown. We first extend of notion of CIDs to consider various possible utility functions.

Definition 4.1. A *parametric-utility CID* (CID) M^Θ is a set of CIDs $\{M^\theta \mid \theta \in \Theta\}$ which differ only in the CPDs of their utility variables.

In effect, a parametric CID is a CID with a parametric class of utility functions $\mathcal{U}^\Theta$. The maximum entropy policy set from Definition 3.1 is extended accordingly, to include maximum entropy policies *for each utility function* and each attainable expected utility with respect to it.

Definition 4.2 (Maximum entropy policy set, unknown utility function)**.** Let $M^\Theta = (G, P)$ be a parametric-utility CID with decision variables $\boldsymbol{D}$ and utility function $\mathcal{U}^\Theta$. The *maximum entropy policy set for $\mathcal{U}^\Theta$* is the set of maximum entropy policies for any attainable expected utility for any utility function in the class: $\Pi^{\mathrm{maxent}}_{\mathcal{U}^\Theta} = \bigcup_{\theta\in\Theta, u\in\mathrm{att}(\mathcal{U}^\theta)} \Pi^{\mathrm{maxent}}_{\mathcal{U}^\theta, u}$.

Definition 4.3 (MEG, unknown utility function)**.** Let $M^\Theta = (G, P)$ be a parametric-utility CID with decision variables $\boldsymbol{D}$ and utility function $\mathcal{U}^\Theta$. The *maximum entropy goal-directedness* of a policy π with respect to $\mathcal{U}^\Theta$ is

$\mathrm{MEG}_{\mathcal{U}^\Theta}(\pi) = \max_{\mathcal{U}\in\mathcal{U}^\Theta} \mathrm{MEG}_{\mathcal{U}}(\pi)$.

Definition 4.4 (MEG, target variables)**.** Let $M = (G, P)$ be a CBN with variables $\boldsymbol{V}$. Let $\boldsymbol{D} \subseteq \boldsymbol{V}$ be a hypothesised set of decision variables and $\boldsymbol{T} \subseteq V$ be a hypothesised set of *target* variables. The *maximum entropy goal-directedness* of $\boldsymbol{D}$ with respect to $\boldsymbol{T}$, $\mathrm{MEG}_{\boldsymbol{T}}(\boldsymbol{D})$, is the goal-directedness of $\pi = P(\boldsymbol{D} \mid \mathbf{Pa}_D)$ in the parametric CID with decisions $\boldsymbol{D}$ and utility functions $\mathcal{U} : \mathrm{dom}(\boldsymbol{T}) \to \mathbb{R}$ (the set of all utility functions over T).

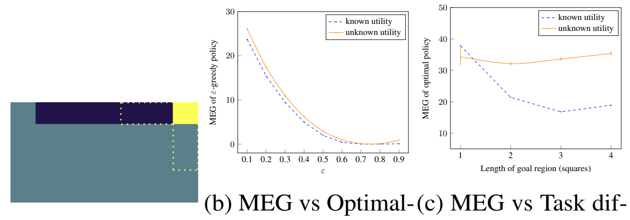
For example, if we only suspected that the mouse in Example 1 was optimising some function of the cheese T, but didn't know which one, we could apply Definition 4.4 to consider the goal-directedness towards T under any utility function defined on T. Thanks to translation and scale invariance (Proposition 3.1), there are effectively only three utility functions to consider: those that provide higher utility to cheese than not cheese, those that do the opposite, and those that are indifferent.

Note that $\boldsymbol{T}$ has to include some descendants of $\boldsymbol{D}$, in order to enable positive MEG (Proposition 3.3). However, it is not necessary that $\boldsymbol{T}$ consists of *only* descendants of $\boldsymbol{D}$ (i.e. $\boldsymbol{T}$ need not be a subset of $\mathbf{Desc}(\boldsymbol{D})$). For example, goal-conditional agents take an instruction as part of their input $\mathbf{Pa}_D$. The goal-directedness of such agents can only be fully appreciated by including the instruction in $\boldsymbol{T}$.

Pseudo-terminal goals. Definition 4.4 enable us to state a result about *pseudo-terminal goals*: however goal-directed some decision variables $\boldsymbol{D}$ are towards some target variables $\boldsymbol{T}$, it must be at least as goal-directed towards any variables $\boldsymbol{S}$ which d-separate $\boldsymbol{D}$ from $\boldsymbol{T}$. For example, in **??**, the agent must be at least as goal-directed towards S_3 as it is towards U_3, since S_3 blocks all paths (i.e. d-separates) from $\{D_1, D_2\}$ to U_3.

Theorem 4.1 (Pseudo-terminal goals)**.** *Let* $M = ((\boldsymbol{V}, \boldsymbol{E}), P)$ *be a CBN. Let* $\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{S} \subseteq \boldsymbol{V}$ *such that* $\boldsymbol{D} \perp \boldsymbol{T} \mid \boldsymbol{S}$. *Then* $\mathrm{MEG}_{\boldsymbol{T}}(\boldsymbol{D}) \leq \mathrm{MEG}_{\boldsymbol{S}}(\boldsymbol{D})$.

It is well known that an agent that is goal-directed with respect to some variable has an instrumental incentive to control any variables which mediate between the two (Everitt et al., 2021a). Theorem 4.1 shows that if the mediating variable d-separates the decision from the downstream variable, then the instrumentally useful variable becomes indistinguishable in a certain sense from the terminally valued one. This means that we do not have to look arbitrarily far into the future to find evidence of goal-directedness. An agent that is goal-directed with respect to next week must be goal-directed with respect to tomorrow.



(a) CliffWorld (b) MEG vs Optimality of policy (c) MEG vs Task difficulty

Figure 2. (a) The CliffWorld environment. (b) MEG of ε-greedy policies for varying ε. (c) MEG for optimal policies for various reward functions.

5. Experiments

By adapting algorithms from the maximum causal entropy framework (Ziebart, 2010), we can estimate MEG in Markov decision processes. Figure 2c shows the results of some preliminary experiments in the Cliffworld environment (Gleave et al., 2020). In the first, we measured the goal-directedness of policies of varying degrees of optimality, as measured by the value of ε for different ε-greedy policies. Predictably, the goal-directedness with respect to the environment reward decreased toward 0 as the policy became less optimal. So did unknown-utility MEG — since as ε increases, the policy becomes increasingly uniform, it does not appear goal-directed with respect to *any* utility function over states.

In the second, we measured the goal-directedness of optimal policies for reward functions specifying tasks of varying difficulty. Goal-directedness with respect to the true reward function decreased as the task became easier to complete. A way to interpret this is that as the number of policies which do well on a reward function increases, doing well on that reward function provides less and less evidence for deliberate optimisation. In contrast, unknown-utility MEG stayed high even as the environment reward becomes easier to satisfy, indicating there was some other reward function for which the policy provided strong evidence. We discuss our algorithms in Appendix A, and give more details on the experiments in Appendix C.

6. Conclusion

We proposed maximum entropy goal-directedness (MEG), a formal measure of goal-directedness grounded in the philosophical literature and the maximum entropy principle. We proved that MEG satisfies several key desiderata, including scale invariance, and that it gives insights about instrumental goals. We conducted small scale experiments. In future work we hope to apply MEG to neural network interpretability by measuring the goal-directedness of a neural network agent with respect to a hypothesis class of utility functions constructed from the network's hidden states.

References

Yoshua Bengio. AI Scientists: Safe and Useful AI? https://yoshuabengio.org/2023/05/07/ai-scientists-safe-and-useful-ai/, 2023.

Martin Biehl and Nathaniel Virgo. Interpreting systems as solving pomdps: a step towards a formal understanding of agency. In *International Workshop on Active Inference*, pages 16–31. Springer, 2022.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.

Daniel C Dennett. *The intentional stance*. MIT press, 1989.

Daniel C Dennett. *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company, 2017.

Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11487–11495, 2021a.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467, 2021b.

Adam Gleave and Sam Toyer. A primer on maximum causal entropy inverse reinforcement learning. 2022.

Adam Gleave, Pedro Freire, Steven Wang, and Sam Toyer. seals: Suite of environments for algorithms that learn specifications. https://github.com/HumanCompatibleAI/seals, 2020.

Joseph Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *Artificial Intelligence*, page 103963, 2023.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. *Proc. of 17th International Conference on Machine Learning, 2000*, pages 663–670, 2000.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Caspar Oesterheld. Formalizing preference utilitarianism in physical world models. *Synthese*, 193(9):2747–2759, 2016.

Laurent Orseau, Simon McGregor McGill, and Shane Legg. Agents and devices: A relative definition of agency. *arXiv preprint arXiv:1805.12387*, 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.

Jonathan Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35:36350–36365, 2022.

Markus Schlosser. Agency. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition, 2019.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic AI systems.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Nathaniel Virgo, Martin Biehl, and Simon McGregor. Interpreting dynamical systems as bayesian reasoners. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 726–762. Springer, 2021.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31: 841, 2017.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18 (6):1–26, 2024.

Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. Honesty is the best policy: defining and mitigating ai deception. *Advances in Neural Information Processing Systems*, 36, 2024a.

Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. The reasons that agents act: Intention and instrumental goals. *arXiv preprint arXiv:2402.07221*, 2024b.

Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1255–1262, 2010.

A. Computing MEG in Markov Decision Processes

In this section, we give algorithms for computing MEG in MDPs. First, we define what an MDP looks like as a causal influence diagram. We then establish a soft value iteration algorithm for computing maximum entropy policies in MDPs, which we use in algorithms for computing MEG when the utility function is known or unknown.

Definition A.1. A *Markov Decision Process* (MDP) is a CID with variables $\{S_t, D_t, U_t\}_{t=1}^n$, decisions $\boldsymbol{D} = \{D_t\}_{t=1}^n$ and utilities $\boldsymbol{U} = \{U_t\}_{t=1}^n$, and such that for t between 1 and n, $\mathbf{Pa}_{D_t} = \{S_t\}$, $\mathbf{Pa}_{U_t} = \{S_t\}$, while $\mathbf{Pa}_{S_t} = \{S_{t-1}, D_{t-1}\}$ for $t > 1$, and $\mathbf{Pa}_{S_1} = \emptyset$.

Constructing Maximum Entropy Policies In MDPs, Ziebart's soft value iteration algorithm can be used to construct maximum entropy policies satisfying a set of linear constraints. We apply it to construct maximum entropy policies satisfying expected utility constraints.

Definition A.2 (Soft Q-Function)**.** Let $M = (G, P)$ be an MDP. Let $\beta \in \mathbb{R}$. For each $D_t \in \boldsymbol{D}$ we define the *soft Q-function* $Q^{\text{soft}}_{\beta,n} : \text{dom}(D_t) \times \text{dom}(\mathbf{Pa}_{D_t}) \to \mathbb{R}$ via the recursion:

$$Q^{\text{soft}}_{\beta,t}(d_t \mid \mathbf{pa}_t)$$
$$= \mathbb{E}\left[U_t + \text{logsumexp}(\beta \cdot Q^{\text{soft}}_{\beta,t+1}(\cdot \mid \mathbf{Pa}_{D_{t+1}}))\middle| d_t, \mathbf{pa}_{t+1}\right] \quad \text{for } t < n,$$
$$Q^{\text{soft}}_{\beta,n}(d_n \mid \mathbf{pa}_n)$$
$$= \mathbb{E}\left[U_n \mid d_n, \mathbf{pa}_n\right],$$

where $\text{logsumexp}\,\beta(Q^{\text{soft}}_{\beta,t+1}(\cdot \mid \mathbf{Pa}_{D_{t+1}})) = \log \sum_{d_{t+1} \in \text{dom}(D_{t+1})} \exp(\beta Q^{\text{soft}}_{\beta,t+1}(d_{t+1} \mid \mathbf{Pa}_{D_{t+1}}))$.

Using the soft Q-function, we show that there is a unique $\pi \in \Pi^{\text{maxent}}_{\mathcal{U},u}$ for each $\mathcal{U}$ and u in MDPs.

Theorem A.1 (Maximum entropy policy in MDPs)**.** *Let $M = (G, P)$ be an MDP with utility function $\mathcal{U}$, and let $u \in \text{att}(\mathcal{U})$ be an attainable expected utility. Then there exists a unique maximum entropy policy $\pi^{\text{maxent}}_u \in \Pi^{\text{maxent}}_{\mathcal{U},u}$, and it has the form*

$$\pi^{\text{maxent}}_{u,t}(d_t \mid \boldsymbol{pa}_t) =$$
$$\pi^{\text{maxent}}_{\beta,t}(d_t \mid \boldsymbol{pa}_t) =$$
$$= \frac{\exp(\beta \cdot Q^{\textit{soft}}_{\beta,t}(d_t \mid \boldsymbol{pa}_t))}{\sum_{d' \in \text{dom}(D_t)} \exp(\beta \cdot Q^{\textit{soft}}_{\beta,t}(d'_t \mid \boldsymbol{pa}_t))}$$

where $\beta = \text{argmax}_{\beta' \in \mathbb{R} \cup \{\infty, -\infty\}} \sum_t \mathbb{E}_\pi \left[\log(\pi^{\text{maxent}}_{\beta'}(d_t \mid \boldsymbol{pa}_t))\right]$.

Known Utility Function To apply Definition 3.1 to measure the goal-directedness of a policy π in a CID M with respect to a utility function $\mathcal{U}$, we need to find the maximum entropy policy in $\Pi_{\mathcal{U}}^{\text{maxent}}$ which best predicts π. We can use Theorem A.1 to derive an algorithm that finds π_u^{maxent} for any $u \in \operatorname{att}(\mathcal{U})$.

Fortunately, we do not need to consider each policy in $\Pi_{\mathcal{U},u}^{\text{maxent}}$ individually. We know the form of π_u^{maxent}, and only the real-valued rationality parameter β varies depending on u. Denote policies of the form of **??** as $\pi_\beta^{\text{maxent}} = \operatorname{softmax}(\beta \cdot Q_{\beta,i})$. The gradient of the predictive accuracy with respect to β is then

$$\nabla_\beta \mathbb{E}_\pi \left[\sum_{D \in \boldsymbol{D}} \left(\log \pi_\beta^{\text{maxent}}(D \mid \boldsymbol{Pa_D}) - \log \frac{1}{\mid \operatorname{dom}(D) \mid} \right) \right]$$
$$= \mathbb{E}_\pi \left[\mathcal{U}\right] - \mathbb{E}_{\pi_\beta^{\text{maxent}}} \left[\mathcal{U}\right]$$

The predictive accuracy is a concave function of β, so we can apply gradient ascent to find the global maximum in β, which is the same as finding the maximum in u.

$\operatorname{MEG}_{\mathcal{U}}(\pi)$ can therefore be found by alternating between applying the soft value iteration of Definition A.2 to find $\pi_\beta^{\text{maxent}}$, computing $\mathbb{E}_\pi [\mathcal{U}] - \mathbb{E}_{\pi_\beta^{\text{maxent}}} [\mathcal{U}]$, and taking a gradient step. See Algorithm 1.

Algorithm 1 Known-utility MEG in MDPs

Input: MDP M, policy π
Output: $\operatorname{MEG}_{\mathcal{U}}(\pi)$
1: initialise rationality parameter β, set learning rate α.
2: **repeat**
3: Apply soft value iteration to find Q_β^{soft} {Definition A.2}
4: $\pi_\beta^{\text{maxent}} \leftarrow \operatorname{softmax}(\beta \cdot Q_\beta^{\text{soft}})$
5: $g \leftarrow \left(\mathbb{E}_\pi [\mathcal{U}] - \mathbb{E}_{\pi_\beta^{\text{maxent}}} [\mathcal{U}] \right)$
6: $\beta \leftarrow \beta + \alpha \cdot g$
7: **until** Convergence
8: **Return:** $\mathbb{E}_\pi \left[\sum_{D \in \boldsymbol{D}} \left(\log \pi_\beta^{\text{maxent}}(D \mid \mathbf{Pa}_D) - \log \frac{1}{|\operatorname{dom}(D)|} \right) \right]$

In all cases Algorithm 1 converges. If the β that maximises predictive accuracy is ∞ or $-\infty$, which can happen if π is optimal or anti-optimal with respect to $\mathcal{U}$, then it can never reach the (nonetheless finite) value of $\operatorname{MEG}_{\mathcal{U}}(\pi)$, but will still converge in the limit.

Unknown-utility algorithm To find unknown-utility MEG, we maximise the predictive accuracy of $\pi_{\theta,\beta}^{\text{maxent}}$ with respect to both θ and β. Assuming that $\mathcal{U}^\Theta$ is a differentiable class of functions, such as a neural network, we can take the derivative of the predictive accuracy with respect to θ and get $\mathbb{E}_\pi [\nabla_\theta \mathcal{U}] - \mathbb{E}_{\pi_\beta^{\text{maxent}}} [\nabla_\theta \mathcal{U}]$.

Algorithm 2 extends Algorithm 1 to this case.

Algorithm 2 Unknown-utility MEG in MDPs

Input: Parametric MDP M_Θ over differentiable class of utility functions, policy π
Output: $\operatorname{MEG}_{\mathcal{U}_\Theta}(\pi)$
1: Initialise utility parameter θ, rationality parameter β, set learning rate α.
2: **repeat**
3: Apply soft value iteration to find $Q_{\theta,\beta}^{\text{soft}}$ {Definition A.2}
4: $\pi_{\theta,\beta}^{\text{maxent}} \leftarrow \operatorname{softmax}(\beta \cdot Q_{\theta,\beta}^{\text{soft}})$
5: $g_\beta \leftarrow \left(\mathbb{E}_\pi \left[\mathcal{U}^\theta\right] - \mathbb{E}_{\pi_\beta^{\text{maxent}}} \left[\mathcal{U}^\theta\right] \right)$
6: $g_\theta \leftarrow \left(\mathbb{E}_\pi \left[\nabla_\theta \mathcal{U}^\theta\right] - \mathbb{E}_{\pi_\beta^{\text{maxent}}} \left[\nabla_\theta \mathcal{U}^\theta\right] \right)$
7: $\beta \leftarrow \beta + \alpha \cdot g_\beta$
8: $\theta \leftarrow \beta + \alpha \cdot g_\theta$
9: **until** Stopping condition
10: **Return:** $\mathbb{E}_\pi \left[\sum_{D \in \boldsymbol{D}} \left(\log \pi_{\theta,\beta}^{\text{maxent}}(D \mid \mathbf{Pa}_D) - \log \frac{1}{|\operatorname{dom}(D)|} \right) \right]$

An important caveat is that if $\mathcal{U}^\theta$ is a non-convex function of θ (e.g. a neural network), Algorithm 2 need not converge to a global maximum. In general, the algorithm provides a *lower bound* for $\operatorname{MEG}_{\mathcal{U}_\theta}(\pi)$, and hence for $\operatorname{MEG}_{\boldsymbol{T}}(\pi)$ where $\boldsymbol{T} = \mathbf{Pa}_{\mathcal{U}^\Theta}$. In practice, we may want to estimate the soft Q-function and expected utilities with Monte Carlo or variational methods, in which case the algorithm provides an *approximate* lower bound on goal-directedness.

B. Experimental Evaluation

We carried out two experiments to measure known-utility MEG with respect to the environment reward function and unknown-utility MEG with respect to a hypothesis class of utility functions. We used an MLP with a single hidden layer of size 256 to define a utility function over states.

Our experiments measured MEG for various policies in the CliffWorld environment from the seals suite (Gleave et al., 2020). Cliffworld (Figure 2a) is a 4x10 gridworld MDP in which the agent starts in the top left corner and aims to reach the top right while avoiding the cliff along the top row. With probability 0.3, a wind causes the agent to move upwards by one more square than intended. The environment reward function gives +10 when the agent is in the (yellow) goal square, -10 for the (dark blue) cliff squares, and -1 elsewhere. The dotted yellow line indicates a length 3 goal region.

MEG vs Optimality of policy. In our first experiment, we measured the goal-directedness of policies of varying degrees of optimality by considering ε-greedy policies for ε in the range 0.1 to 0.9. Figure 2b shows known- and unknown utility meg for each policy [1] Predictably, the goal-directedness with respect to the environment reward decreased toward 0 as the policy became less optimal. So did unknown-utility MEG — since as ε increases, the policy becomes increasingly uniform, it does not appear goal-directed with respect to *any* utility function over states.

MEG vs Task difficulty In our second experiment, we measured the goal-directedness of optimal for reward functions of varying difficulty. We extended the goal-region of Cliffworld to run for either 1, 2, 3 or 4 squares along the top row and back column, and considered an optimal policies for each reward function. Figure 2c shows Cliffworld with a goal region of length 3. Figure 2b shows the results. Goal-directedness with respect to the true reward function decreased as the task became easier to complete. A way to interpret this is that as the number of policies which do well on a reward function increases, doing well on that reward function provides less and less evidence for deliberate optimisation. In contrast, unknown-utility MEG stays high even as the environment reward becomes easier to satisfy. This is because the optimal policy proceeds towards the *nearest* goal-squares and, hence, it appears strongly goal-directed with respect to a utility function which gives high reward to only those squares. Since this narrower utility function is more difficult to do well on than the environment reward function, doing well on it provides more evidence for goal-directedness. In Appendix G.3, we visualise the policies in question to make this more explicit. We also give tables of results for both experiments.

C. Discussion

Limitations One limitation of the MEG measure for goal-directedness is that it relies on having a causal model of the environment, so that one can compute the maximum entropy policy for a given utility function.

Second, MEG for a policy can depend on what variables are included in the model. For example, if a policy is highly goal-directed towards some variable T not included in our model, MEG may still be low. Relatedly, MEG may also be affected by whether we use a binary variable for T (or the decisions D) rather than a fine-grained one with many possible outcomes. We should, therefore, think of MEG as measuring what *evidence* a set of variables provides about a policy's goal-directedness.

Lack of evidence does not necessarily mean lack of goal-directedness. Third, while MEG can be computed with gradient descent, it can still be computationally intractable for very large sets of random variables. In this paper, we conduct only preliminary experiments – larger experiments based on real-world data may explore how serious these limitations are in practice.

Finally, MEG measures how predictive a utility function is of an system's behaviour *on distribution*, and distributional shifts can lead to changes in MEG. It may be that by considering changes to a system's behaviour under interventions, as Kenton et al. (2023) do, we can distinguish "true" goals from spurious goals, where the former predict behaviour well across distributional shifts, while the latter happen to predict behaviour well on a particular distribution (perhaps because they correlate with true goals). We leave this to future work.

Societal implications An empirical measure of goal-directedness may be enable researchers and companies to keep better track how goal-directed LLMs and other systems are. This is important for dangerous capability evaluations (Shevlane et al., 2023) and governance (Shavit et al.). A potential downside is that it could enable bad actors to create even more dangerous systems. We judge this risk as minor since the relationship between goal-directedness and danger is fairly indirect.

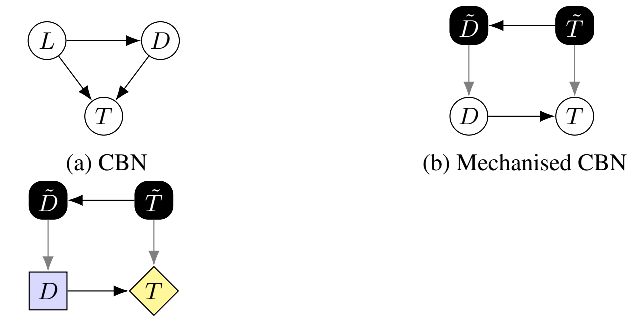
D. Comparison to Discovering Agents

This paper is inspired by Kenton et al. (2023), who proposed a causal discovery algorithm for identifying agents in causal models, inspired by Dennett's view of agents as systems ”moved by reasons”. Our approach has several advantages over theirs, which we enumerate below.

Mechanism variables. (Kenton et al., 2023) assume access to a *mechanised structural causal model*, which augments an ordinary causal model with *mechanism variables* which parameterise distributions of ordinary object-level variables. An agent is defined as a system that adapts to changes in the mechanism of its environment. However, the question of what makes a variable a mechanism is left undefined, and indeed, the same causal model can be expressed either with or without mechanism variables, leading their algorithm to give a different answer. For example, Example 1 has identical causal structure to (Kenton et al., 2023)'s in, but without any variables designated as mechanisms. Their algorithm says the version with mechanism variables contains an agent while the version without does not, despite them being essentially the same causal model. Figure 3 shows our example depicted as a mechanised structural causal model. We fix this arbitrariness by making our definition in ordinary causal Bayesian networks.

Utility variables. Their algorithm assumes that some variables in the model represent agents' utilities. We bring

[1] Known-utility MEG is deterministic. Unknown-utility MEG depends on the random initialisation of the neural network, so we show the mean of several runs. Full details are given in Appendix G.3



(a) CBN (b) Mechanised CBN (c) Mechanised CID

Figure 3. Example 1 can be equally well represented with a CBN (a) or mechanised CBN (b), but (Kenton et al., 2023)'s algorithm only identifies an agent in (b). (c) shows the resulting mechanised CID. In contrast, MEG is unchanged between (b) and (c). Note also that the causal discovery algorithm identifies T as a utility variable, where where MEG adds a new utility child to T.

this more in line with the philosophical motivation by treating utilities as hypothesised mental states with which we augment our model.

Predictive accuracy. (Kenton et al., 2023)'s approach formalises Dennett's idea of agents as systems "moved by reasons". We build on this idea but bring it more in line with Dennett's notion of what it means for a system to be moved by a reason — that the reason is useful for predicting its behaviour.

Gradualist vs Essentialist. The predictive error viewpoint gives us a continuous measure of goal-directedness rather than a binary notion of agency, which is more befitting of the gradualist view of agents which inspired it.

Practicality. Their algorithm is theoretical rather than something that can be applied in practice. But ours is straightforward to implement, as we demonstrate in Appendix C. This opens up a range of potential applications, including behavioural evaluations and interpretability of ML models.

Interventional distributions. The primary drawback of MEG is that it doesn't necessarily generalise outside of the distribution. Running MEG on interventional distributions may fix this. We leave an extension of MEG to interventional distributions for future work.

E. Proofs of MEG Properties

Proposition 3.1 (Translation and scale invariance)**.** *Let M_1 be a CID with utility function $\mathcal{U}_1$, and let M_2 be an identical CID but with utility function $\mathcal{U}_2 = a \cdot \mathcal{U}_1 + b$, for some $a, b \in \mathbb{R}$. Then for any policy π, $\mathrm{MEG}_{\mathcal{U}_1}(\pi) = \mathrm{MEG}_{\mathcal{U}_2}(\pi)$.*

Proof. Since MEG is defined as maximum predictive accuracy over a maximum entropy policy set, showing that two utility functions have the same maximum entropy policy set is enough to show that they give the same MEG to every policy. We show that $\Pi^{\text{maxent}}_{\mathcal{U}_2} = \Pi^{\text{maxent}}_{\mathcal{U}_1}$.

If $\pi \in \Pi^{\text{maxent}}_{\mathcal{U}_2}$, then π is a maximum entropy policy such that $\mathbb{E}_\pi[\mathcal{U}_2] = u$ for some $u \in \mathrm{att}(\mathcal{U}_2)$. But then π must be a maximum entropy policy such that $\mathbb{E}_\pi[\mathcal{U}_1] = a \cdot u + b \in \mathrm{att}(\mathcal{U}_1)$, so $\pi \in \Pi^{\text{maxent}}_{\mathcal{U}_1}$.

The converse is similar.

□

Proposition 3.2 (Bounds)**.** *Let M be a CID with utility function $\mathcal{U}$. Then for any policy π we have $0 \le \mathrm{MEG}_{\mathcal{U}}(\pi) \le \sum_{D \in \mathcal{D}} \log(|\operatorname{dom}(D)|)$, with equality in the lower bound if π is the uniform policy, and equality in the upper bound if* and only if *π is the unique optimal (or anti-optimal) policy with respect to $\mathcal{U}$.*

Proof. Recall that $\mathrm{MEG}_{\mathcal{U}}(\pi) = \max_{\pi^{\text{maxent}} \in \Pi^{\text{maxent}}_{\mathcal{U}}}$

$$\mathbb{E}_\pi\left[\sum_{D \in \boldsymbol{D}}\left(\log \pi^{\text{maxent}}(D \mid \mathbf{Pa}_D) - \log \frac{1}{|\operatorname{dom}(D)|}\right)\right].$$

To get the lower bound, note that the expression being maximised can be rewritten as

$\sum_{\mathbf{Pa}_{\boldsymbol{D}}} P_\pi(\mathbf{Pa}_{\boldsymbol{D}}) \left(H(P_{\text{unif}}) - H(\pi^{\text{maxent}}(\cdot \mid \mathbf{Pa}_{\boldsymbol{D}})\right)$ where P_{unif} is the uniform distribution over $\operatorname{dom}(\boldsymbol{D})$. Since the entropy of a distribution cannot exceed the entropy of the uniform distribution, this expression is nonnegative. It's also clear from this expression that MEG is zero for the uniform policy.

For the upper bound, note that $H(\pi^{\text{maxent}}(\cdot \mid \mathbf{Pa}_{\boldsymbol{D}}))$ is nonnegative, so $\mathrm{MEG}_{\mathcal{U}} \le \mathbb{E}_\pi[H(P_{\text{unif}})] = \log(|\operatorname{dom}(\boldsymbol{D})|) = \log\left(\prod_{D \in \boldsymbol{D}} |\operatorname{dom}(D)|\right) = \sum_{D \in \boldsymbol{D}} \log(|\operatorname{dom}(D)|)$.

To show that we have equality when π is the unique optimal or anti optimal policy, note that in that case π must be deterministic. Also, π must be in $\Pi^{\text{maxent}}_{\mathcal{U}}$, since there can be no higher entropy way to get the same expected utility. Then since π maximises predictive accuracy with respect to itself, the $H(\pi^{\text{maxent}}(\cdot \mid \mathbf{Pa}_{\boldsymbol{D}}))$ term becomes $H(\pi(\cdot \mid \mathbf{Pa}_{\boldsymbol{D}}) = 0$ and we attain the upper bound.

For the converse, we can show that if π is *not* uniquely optimal or anti-optimal, the π^{maxent} which best predicts it is not deterministic, and so the $H(\pi^{\text{maxent}}(\cdot \mid \mathbf{Pa}_{\boldsymbol{D}}))$ term does not go to 0.

□

Proposition 3.3 (No goal-directedness without causal influence)**.** *Let $M = (G, P)$ be a CID with utility function $\mathcal{U}$*

and decision variables $\boldsymbol{D}$ such that, $\boldsymbol{Desc}(\boldsymbol{D}) \cap \boldsymbol{Pa_U} = \emptyset$. Then $\mathrm{MEG}_{\mathcal{U}}(\boldsymbol{D}) = 0$.

Proof. Since $\mathcal{U}$ is not a descendant of $\boldsymbol{D}$, it follows from the Markov property of causal Bayesian networks that $\mathcal{U} \perp \boldsymbol{D} \mid \mathbf{Pa}_{\boldsymbol{D}}$. That means all policies achieve the same expected utility u. So the maximum entropy policy set $\Pi^{\mathrm{maxent}}_{\mathcal{U}}$ contains only the uniform policy. We get that $\mathrm{MEG}_{\mathcal{U}}(\pi) =$

$$-\mathbb{E}\left[\sum_{D \in \boldsymbol{D}} \log \frac{1}{\mid \mathrm{dom}(D) \mid} - \log \frac{1}{\mid \mathrm{dom}(D) \mid}\right] = 0. \quad \square$$

Theorem 4.1 (Pseudo-terminal goals)**.** *Let $M = ((\boldsymbol{V}, \boldsymbol{E}), P)$ be a CBN. Let $\boldsymbol{D}, \boldsymbol{T}, \boldsymbol{S} \subseteq \boldsymbol{V}$ such that $\boldsymbol{D} \perp \boldsymbol{T} \mid \boldsymbol{S}$. Then* $\mathrm{MEG}_{\boldsymbol{T}}(\boldsymbol{D}) \leq \mathrm{MEG}_{\boldsymbol{S}}(\boldsymbol{D})$.

Proof. We will show that the maximum entropy policy set $\Pi^{\mathrm{maxent}}_{\boldsymbol{\mathcal{U}^T}}$ (where $\boldsymbol{\mathcal{U}^T}$ is the set of all utility functions over $\boldsymbol{T}$) is a subset of $\Pi^{\mathrm{maxent}}_{\boldsymbol{\mathcal{U}^S}}$, so the maximum predictive accuracy taken over the latter upper bounds the maximum predictive accuracy taken over the former.

Let $\pi \in \Pi^{\mathrm{maxent}}_{\boldsymbol{\mathcal{U}^T}}$, so $\pi = \pi^{\mathrm{maxent}}_{\mathcal{U},u}$ for some $\mathcal{U}^T \in \boldsymbol{\mathcal{U}^T}$. If we can find a utility function $\mathcal{U}^S \in \boldsymbol{\mathcal{U}^S}$ such that for all π, $\mathbb{E}_\pi\left[\mathcal{U}^S\right] = \mathbb{E}_\pi\left[\mathcal{U}^T\right]$, then the maximum entropy policy with $\mathbb{E}_\pi\left[\mathcal{U}^T\right] = u$ must also be the maximum entropy policy with $\mathbb{E}_\pi\left[\mathcal{U}^S\right] = u$. It would follow that $\pi \in \Pi^{\mathrm{maxent}}_{\boldsymbol{\mathcal{U}^T}}$ and so $\Pi^{\mathrm{maxent}}_{\boldsymbol{\mathcal{U}^S}} \subseteq \Pi^{\mathrm{maxent}}_{\boldsymbol{\mathcal{U}^T}}$.

To construct such a utility function, let $\mathcal{U}^{\boldsymbol{S}}(\boldsymbol{s}) = \sum_t P(\boldsymbol{T} = \boldsymbol{t} \mid \boldsymbol{S} = \boldsymbol{s}) \mathcal{U}^{\boldsymbol{T}}(t)$. Note that since $\boldsymbol{D} \perp \boldsymbol{T} \mid \boldsymbol{S}$, $P(\boldsymbol{T} = \boldsymbol{t} \mid \boldsymbol{S} = \boldsymbol{s})$ is not a function of π. Then for any π,

$$\begin{aligned}
\mathbb{E}_\pi\left[\mathcal{U}^{\boldsymbol{T}}\right] &= \sum_{\boldsymbol{t}} P_\pi(\boldsymbol{t}) \mathcal{U}^{\boldsymbol{T}}(\boldsymbol{t}) \\
&= \sum_{\boldsymbol{s}} P_\pi(\boldsymbol{s}) \sum_{\boldsymbol{t}} P_\pi(\boldsymbol{t} \mid \boldsymbol{s}) \mathcal{U}^{\boldsymbol{T}}(\boldsymbol{t}) \\
&= \sum_{\boldsymbol{s}} P_\pi(\boldsymbol{s}) \sum_{\boldsymbol{t}} P(\boldsymbol{t} \mid \boldsymbol{s}) \mathcal{U}^{\boldsymbol{T}}(\boldsymbol{t}) \\
&\quad (\text{since } \boldsymbol{D} \perp \boldsymbol{T} \mid \boldsymbol{S}) \\
&= \sum_{\boldsymbol{s}} P_\pi(\boldsymbol{s}) \mathcal{U}^{\boldsymbol{S}}(\boldsymbol{s}) \\
&= \mathbb{E}_\pi\left[\mathcal{U}^{\boldsymbol{S}}\right].
\end{aligned}$$

$\square$

F. Proof of Theorem A.1

Theorem A.1 (Maximum entropy policy in MDPs)**.** *Let $M = (G, P)$ be an MDP with utility function $\mathcal{U}$, and let $u \in \mathrm{att}(\mathcal{U})$ be an attainable expected utility. Then there exists a unique maximum entropy policy $\pi^{\mathrm{maxent}}_u \in \Pi^{\mathrm{maxent}}_{\mathcal{U},u}$, and it has the form*

$$\pi^{\mathrm{maxent}}_{u,t}(d_t \mid \boldsymbol{pa}_t) = \pi^{\mathrm{maxent}}_{\beta,t}(d_t \mid \boldsymbol{pa}_t) = \frac{\exp(\beta \cdot Q^{soft}_{\beta,t}(d_t \mid \boldsymbol{pa}_t))}{\sum_{d' \in \mathrm{dom}(D_t)} \exp(\beta \cdot Q^{soft}_{\beta,t}(d'_t \mid \boldsymbol{pa}_t))}$$

where $\beta = \mathrm{argmax}_{\beta' \in \mathbb{R} \cup \{\infty, -\infty\}} \sum_t \mathbb{E}_\pi\left[\log(\pi^{\mathrm{maxent}}_{\beta'}(d_t \mid \boldsymbol{pa}_t))\right]$.

Proof. The attainable utility set is a closed interval $\mathrm{att}(\mathcal{U}) = [u_{\min}, u_{\max}]$. We first consider the case where $u \in (u_{\min}, u_{\max})$.

In this case we are seeking the maximum entropy policy in an MDP with a linear constraint satisfiable by a full support policy, so we can invoke Ziebart's result on the form of such policies (Ziebart, 2010; Ziebart et al., 2010; Gleave and Toyer, 2022). In particular our feature is the utility $\mathcal{U}$. We get that the maximum entropy policy is a soft-Q policy for a utility function $\beta \cdot \mathcal{U}$ with a rationality parameter of 1, where $\beta = \mathrm{argmax}_{\beta' \in \mathbb{R}} \sum_t \mathbb{E}_\pi\left[\log(\pi^{\mathrm{maxent}}_{\beta'}(d_t \mid \mathbf{pa}_t))\right]$. This can be restated as a soft-Q policy for $\mathcal{U}$ with a rationality parameter of β. It follows from Ziebart that $\beta = \mathrm{argmax}_{\beta' \in \mathbb{R}} \pi^{\mathrm{maxent}}_{\beta}$, and allowing $\beta = \infty$ or $-\infty$ does not change the argmax.

In the case where $u \in \{u_{\min}, u_{\max}\}$, it's easy to show that the maximum entropy policy which attains u randomises uniformly between optimal actions (for $u_{\max}$) or anti-optimal actions (for $u_{\min}$). These policies can be expressed as $\lim_{\beta \to \infty} \pi^{\mathrm{maxent}}_{\beta}$ and $\lim_{\beta \to -\infty} \pi^{\mathrm{maxent}}_{\beta}$ respectively. $\square$

G. Experimental Details

G.1. Tables of results

	Known Utility	Unknown Utility
$k=1$	37.8	34.3 ± 2.6
$k=2$	21.4	32.1 ± 0.5
$k=3$	16.8	33.6 ± 0.5
$k=4$	18.9	35.4 ± 0.6

	Known Utility	Unknown Utility
$\varepsilon = 0.1$	2.4	26.1 ± 0.11
$\varepsilon = 0.2$	1.5	17.4 ± 0.2
$\varepsilon = 0.3$	0.95	11.0 ± 0.25
$\varepsilon = 0.4$	0.50	6.2 ± 0.08
$\varepsilon = 0.5$	0.20	2.9 ± 0.06
$\varepsilon = 0.6$	0.04	1.0 ± 0.003
$\varepsilon = 0.7$	0.003	0.10 ± 0.002
$\varepsilon = 0.8$	0.001	0.10 ± 0.003
$\varepsilon = 0.9$	0.008	0.091 ± 0.007

G.2. Visualising optimal policies for different lengths of goal region.

(a) Occupancy measures of optimal policy when $k = 1$.

(b) Occupancy measures of optimal policy when $k = 4$.

Figure 4. Occupancy measures

Figure 4a and Figure 4b show the occupancy measures for an optimal policy for k=1 and k=4 respectively, where k is the length of the goal region in squares. Although the goal region is larger in the latter case, the optimal policy consistently aims for the same sub-region. This explains why unknown-utility MEG is higher than MEG with respect to the environment reward. The policy does just as well on a utility function whose goal-region is limited to the nearer goal squares as it does on the environment reward, but fewer policies do well on this utility function, so doing well on it constitutes more evidence for goal-directedness.

G.3. Further details

The experiments were carried out on a personal laptop with the following specs:

- *Hardware model:* LENOVO20N2000RUK
- *Processor:* Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz, 2112 Mhz, 4 Core(s), 8 Logical Processor(s)
- *Memory:* 24.0 GB

We used an environment from the SEALS library[2], and adapted an algorithm from the imitation library[3]. Both are released under the MIT license.

[2]https://github.com/HumanCompatibleAI/seals
[3]https://github.com/HumanCompatibleAI/imitation

For information on hyperparameters see the code.