Habits Have Geometry: Action Representation as Morphological Computation

Karen Archer[1], Christoph Salge[1], Nicola Catenacci Volpi[1], and Daniel Polani[1]

[1]Adaptive Systems Group, University of Hertfordshire, Hatfield, UK
k.archer2@herts.ac.uk

Abstract

Bounded organisms face evolutionary pressure to act cheaply: every decision carries an informational cost, and one under-explored way to reduce that cost is to reshape the representation of action itself. We study action *twists*: given a world with state transitions induced by actions, we introduce a per-state relabelling of actions. In the "twisted" worlds the same state transitions are available as in the original, but the action labels corresponding to these may change from state to state. A genetic algorithm now searches for twists in grid-world environments that reduce *decision information* across many goals simultaneously. The discovered relabellings align action labels with environmental structure, exploiting walls and doorways as a form of morphological computation, and the benefit persists across a wide range of bounded-rationality regimes. Even on a featureless torus where no environmental structure exists to exploit, the GA constructs *habit cycles*: short, label-specific orbits that the agent enters by repeating a single action symbol. Each habit cycle gives the agent a coherent cheap path across the environment, supporting our claim that action representation is itself a form of morphological computation, with its own geometry that can be evolved.

Data/Code available at: `https://uh-adapsys.gitlab.io/papers/alife-relabelled-actions`.

Introduction

Consider a lengthy set of directions through an unfamiliar city: turn left, then right, then right again, then left, then right once more, and finally left at the final junction. At each step, a fresh decision is required, including re-checking the instructions. The same route becomes almost effortless under a different description — *keep the river on your left*, or *follow the road signs marked with an aeroplane* — each turning the route into something closer to walking along a corridor. The streets have not changed, but the instruction has been re-framed in terms of environmental structure rather than arbitrary local choices. A third kind of reframing comes from within the agent: after sufficient experience of walking the route, what was a sequence of effortful decisions becomes a single habituated action, "go home". The physical movements are the same, but the cognitive cost has collapsed because the agent no longer needs to deliberate at every step. A twist takes both ideas further: the meaning of each action label can vary from state to state, yet if that variation is aligned with environmental structure, the informational cost of acting can decrease.

This is no accident: real organisms face evolutionary pressure to act cheaply because brains are metabolically expensive and working memory is scarce (Polani, 2011; Lieder and Griffiths, 2020; Daw et al., 2005). Embodiment and morphological computation study how body and environment can do computational work on the agent's behalf (Müller and Hoffmann, 2017): the compliant dynamics of a physical body, for instance, lets a simple controller exploit passive mechanics for locomotion, shifting the computational burden from controller to morphology (Hauser et al., 2011). We ask whether the same principle extends to the action space itself. The structure of the action space is a morphology: just as the choice of signposting changes the cognitive cost of navigating a city, how actions are labelled should affect the informational cost of acting.

This paper posits that, for any fully observable Markov decision process (MDP), distinct labellings of a discrete action space can make routes informationally cheaper to follow. Rather than treating the action labels as fixed (for instance, the cardinal directions north, east, south, west of a grid world), we conjecture that alternative mappings, aligned with the structure of the environment, can reduce this cost. The world is unchanged. The agent's body is unchanged. Only the coordinate system through which the agent acts is different.

Not every relabelling reduces the informational cost of acting. The standard compass labelling is special because it is globally consistent: the same action symbol invokes the same physical direction at every state, which lets the agent's behaviour generalise across the environment. Arbitrary relabellings lose this consistency (Polani, 2011). Useful twists give it up deliberately: they align labels with environmen-

tal features such as walls or corners so that the environment carries part of the work the agent's policy would otherwise have to express. We treat the action representation, rather than the body, as the variable shaped by this morphological computation.

We ask whether a genetic algorithm can find such relabellings, *twists*, that reduce decision information (Archer et al., 2022) while remaining competitive in task utility. We show this holds not only in a four-room grid world, where the environment provides affordances (Gibson, 1979) — walls and doorways the labels can align to — but also, surprisingly, in an open grid and a toroidally wrapped grid where no such features exist. A single-goal variant of the GA finds twists that "straighten" the route to one target; the multi-goal variant, with fitness averaged uniformly over all goals, discovers a better quality relabelling that remains useful across many goals.

Background

The body is not just a vehicle for behaviours: it can contribute actively to the computation needed to generate them. Zahedi and Ay (2013) formalise this within the sensorimotor loop: the cycle connecting controller, actuators, world, and sensors through which an embodied agent acts. They propose two complementary information-theoretic measures: one asking how much the world drives the next state independently of the action taken, and one asking how little the action contributes to the transition. Together they quantify the degree to which the world does the work rather than the controller.

We extend this perspective from the physical body to the action representation itself. The labelling of an agent's available moves is itself a morphology of the action space, and changing it changes the informational cost of acting. In the relabelled actions setting, a twist changes the mapping from action labels to physical moves, the controller-to-actuator kernel in Zahedi's terms, while leaving the world transition structure entirely unchanged.

Polani (2011) provides the direct precedent for this work. In a grid-world task, he shows that the consistency of action labels across states determines how much information the agent needs to be performant. With cardinal directions applied consistently, the environment's structure does a substantial part of the navigational work: walls funnel a nearly blind agent toward the goal without it needing detailed state information. Randomly relabelling the actions breaks this consistency: the same walls now offer no guidance, and the agent must use significantly more information per step to reach the same performance. Polani measures this cost using *relevant information*, the mutual information between states and actions at a single decision step, and interprets it as evidence that action labelling is itself a form of embodiment that can either relieve or increase the agent's cognitive burden.

Real agents cannot process information without cost: they face pressure to achieve sufficient utility while using as little information as possible (Polani, 2011). The behaviour of such an informationally bounded agent can be cast as a Lagrangian trade-off between utility and the cost of acquiring the necessary information to act (Ortega and Braun, 2013). The agent minimises an objective that penalises policies which deviate from a default, uninformed prior over actions, with β acting as a Lagrange multiplier controlling how much information is worth spending to improve performance. This objective is sometimes referred to as a free energy, but the essential structure is a constrained optimisation: maximise *value* — the expected cumulative utility under the agent's policy — subject to a bound on informational cost.

The key measure in the present work is *decision information* (Archer et al., 2022). Unlike relevant information, which captures the mutual information between states and actions at a single decision step, decision information accumulates the informational cost over the full sequence of decisions in a planned route. The intuition is simple: a policy that issues the same action distribution at every state is informationally cheap, while one that varies sharply from state to state is expensive. Relabelling the actions can move the same physical route into the cheap regime by reshaping how state-dependent the optimal policy needs to be. This makes decision information an appropriate lens for studying how action-space morphology affects the informational cost of acting.

The single-goal case raises a natural follow-on question: can a single relabelling remain useful across many goals simultaneously? This echoes a broader principle in navigation (Mallot, 2024, chs. 6–7): agents do not re-plan from scratch for every destination, but rely on reusable structure, habitual transitions, landmark states, and doorways that serve as waypoints regardless of the final goal. The parameter β fits this picture naturally: at lower values the agent plans at coarser resolution, relying more on habitual tendencies than on goal-specific detail. A multi-goal twist that reduces decision information across all goals encodes reusable structure into the action representation, not into the plan: it makes the resulting policy cheaper to follow step by step, not cheaper to derive.

Method

We work in grid-world environments, instances of finite MDPs, in which states are grid cells, actions produce movement in the cardinal directions, and relabelling actions permutes which action label produces which movement in one of these directions. Following Polani (2011), each twist is represented by a state-wise permutation map σ, where σ_s is a permutation of the action set $\mathcal{A}$, defined separately for each state $s \in \mathcal{S}$. The twisted world is therefore a relabelled MDP: if the original environment has transition matrix $P^a_{ss'}$ and reward function $r(s', s, a)$, then the relabelled dynamics

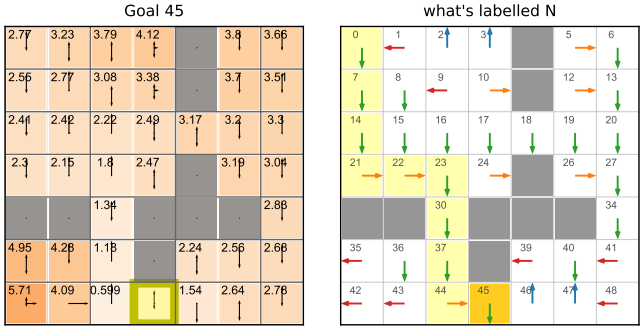

Figure 1: Best multi-goal GA twist in four rooms ($\beta = 1$, transition determinism $\mathrm{det} = 0.97$: the intended action is executed with probability 0.97, the remaining probability is shared uniformly over the other three). **Left**: the planning-optimal policy $\pi^*_\beta = \arg\min_\pi \mathcal{F}^\pi_\beta(s)$ for goal 45, with per-state decision information $\mathcal{I}_D^{\pi^*_\beta}(s)$ as the heatmap colour. Black arrows indicate the issued action label, not the physical transition direction; the goal state is highlighted in yellow, where arrows show the goal-marginalised policy (the predominantly North action prior in the decision information calculation). **Right**: the "what's labelled N" panel shows how the single label N maps to physical directions across states under the twist σ — a globally consistent labelling would give uniform arrows, whereas the heterogeneous pattern reveals where the GA aligns labels with environmental structure. The yellow shaded cells trace a route: an agent starting at state 0 (top-left corner) that repeatedly issues label N moves down along the wall, turns, and continues through the doorway into the goal (gold), the same label producing a physically different transition at each step while decision information stays low.

are given by $\tilde{P}^a_{ss'} := P^{\sigma_s(a)}_{ss'}$, and similarly for the rewards. A full twist contains one local permutation for each state in $\mathcal{S}$, so, for the GA, the genome length is $|\mathcal{S}| \times |\mathcal{A}|$. The transition kernel $P^a_{ss'}$, goal locations, and wall positions remain unchanged; only the mapping from action labels to physical moves is altered.

Decision information $\mathcal{I}^\pi_D(s)$, for any policy π, measures the total deviation of a planned policy from the agent's habitual action tendencies, that is, what the agent does on average without knowing where it is. When decision information is low, the agent can navigate without distinguishing between states: it issues the same action distribution everywhere regardless of where it is, like walking in a fixed compass direction regardless of position. The action prior is the policy averaged uniformly over states,

$$\hat{p}^{\mathrm{U}}(a) := \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \pi(a|s).$$

$$\mathcal{I}^\pi_D(s) := \mathop{\mathbb{E}}_{\pi(a|s)p(s'|s,a)} \left[\log \frac{\pi(a|s)}{\hat{p}^{\mathrm{U}}(a)} + \mathcal{I}^\pi_D(s') \right].$$

In the space-of-goals framework (Archer et al., 2022), a policy with low decision information is one that closely follows the agent's prior action distribution. Such policies approximate following the same action distribution across many states with minimal state-specific correction. If action directions are consistent with the cardinal directions of the grid, walls and boundaries guide low-information policies: at corner goals, walls absorb incorrect actions (a wrong move into a wall leaves the agent in place) so that a near-uniform policy still reaches the target, while reaching central goals requires deliberate, state-specific steering that raises the informational cost. A useful twist is therefore one that aligns the action representation, so that effective routing stays as close to the action prior as possible.

The goal is a terminal state: on reaching it the trajectory ends, the agent takes no further actions, and so incurs no further cost. We encode this as a boundary condition, $\mathcal{I}^\pi_D(g) = 0$, which fixes the base case of the recursion. We apply no discount factor, so every step contributes its full cost and $\mathcal{I}^\pi_D$ measures the informational cost of the whole route to the goal. Because the goal is absorbing, this undiscounted expected sum is finite.

The planning objective (a free energy formulation) combines value and informational cost via β as a Lagrange multiplier:

$$\mathcal{F}^\pi_\beta(s) := -V^\pi(s) + \beta^{-1}\mathcal{I}^\pi_D(s).$$

This is a scalarisation of a two-objective problem — maximise value, minimise decision information — with β^{-1} the trade-off weight; sweeping β (Figure 3) traces the convex hull of the value–information Pareto front. We solve for $\pi^*_\beta := \arg\min_\pi \mathcal{F}^\pi_\beta(s)$, the policy that minimises this objective at a given goal g and setting of β. The goal enters through the reward, and hence through V^π and $\mathcal{I}^\pi_D$, so the planning policy is goal-conditioned — $\pi^*_\beta = \pi^*_\beta(a \mid s, g)$ — on occasion we leave g implicit for brevity. This is not the value-optimal policy, but the policy that best balances performance against the cost of state-dependent action selection. At high β the information penalty is small and π^*_β approaches a value-optimal policy. Even here, information cost acts as a tie-breaker, selecting the most informationally parsimonious amongst all near-optimal policies. At low β the agent favours informationally cheaper, more state-independent action choices even if they produce longer routes. We write $V^{\pi^*_\beta}$ and $\mathcal{I}_D^{\pi^*_\beta}$ for the value and decision information evaluated under π^*_β.

Relabelling the actions changes the decision information cost even for exactly the same physical movement, because the policy is expressed in a different action coordinate system; relabelling therefore shifts the value-information trade-off and changes the optimal policy π^*_β even when the physical world is fixed. The GA search aims to find a single twist σ that reduces decision information across a broad set of goals simultaneously, using the mean $\mathbb{E}_{s,g}[\mathcal{I}_D^{\pi^*_\beta}(s)]$, av-

eraged uniformly over all goals, as the fitness function: this rewards relabellings that are broadly useful rather than tuned to a single destination.

Experimental Setup

The GA is implemented with the DEAP (Distributed Evolutionary Algorithms in Python) library (Fortin et al., 2012). We use population size 96, 200 generations, tournament size 3, elite size 4, crossover probability 0.7, and mutation probability 0.3 with per-row mutation probability 0.2. Each genome is a flat list encoding one action permutation per non-wall state, so crossover swaps whole state-rows between parents and mutation swaps two action labels within a row, preserving the row permutation structure. Each result reported here is representative of many independent GA runs per environment (for instance, fifteen in the toroidally wrapped grid); across seeds the discovered twists vary little, in both decision-information cost and the structure of the dynamics they induce.

A twist is *globally consistent* when every state uses the same permutation of the action set: any one of the $|\mathcal{A}|! = 24$ possible orderings (any rotation of the compass rose, not specifically the canonical (N, E, S, W)). For such a twist we set $\chi_{\text{twist}} = 0$. In general, χ_{twist} is one minus the largest fraction of (s, a) pairs for which the local choice $\sigma_s(a)$ matches a global ordering, where the maximum is taken over the 24 candidates; higher values indicate more spatial variation in the label-to-direction mapping. The permutation-balanced initialiser spreads the initial population across global orderings rather than concentrating near one.

The experiments in this paper use 7×7 grid-world environments including a four-room layout and a toroidally wrapped grid. All environments use Manhattan actions with the intended action executed with probability 0.97 and each of the other three with probability 0.01. We use a cost-only MDP: every time step incurs a reward of -1, regardless of the outcome, with no extra penalty for bumping into a boundary beyond the lost step. The walls are part of the environment, but the agent's inability to pass through them is its embodiment in this world: a move that would take the agent off the grid leaves it in place. At low β, where incorrect actions are frequent, this embodiment catches the agent: a near-open-loop policy still funnels toward goals next to walls because each imprecise step costs no more than a precise one — the wall holds the agent where it was. This is morphological computation the twist can exploit (Archer et al., 2022).

Results

In the limiting single-goal case, a twist with minimal decision information cost would assign the same label to the optimal move at every state along the route: the relabelling "straightens" the policy in label space, letting the agent keep

Figure 2: A high-χ comparison twist from the GA population, shown for goal 45. **Left**: a panel of 4 "what's labelled" grids showing less coherent action mappings. **Right**: optimal policy with per-state decision information $\mathcal{I}_D^{\pi_\beta^*}(s)$ as the heatmap colour, where $\pi_\beta^* = \arg\min_\pi \mathcal{F}_\beta^\pi(s)$. Compared with the same goal under the GA-best twist (Figure 1, rightmost panel), decision information is substantially higher.

issuing the same label while the permutation σ converts it into the appropriate physical move at each step. For a single destination the GA finds such twists readily; the more demanding question is thus whether a single relabelling can remain useful across many goals simultaneously.

A multi-goal twist in four rooms

Figure 1 shows the best twist found by the multi-goal GA in the four-room grid world: a single σ, optimised over all goals simultaneously, that reduces mean decision information across the entire goal set. The left panel shows the planning-optimal policy at $\beta = 1$ for a representative goal, with per-state decision information as the heatmap; darker cells indicate states where the required action deviates most from the prior and is therefore more costly. The right panel shows the structure of the twist itself: the "what's labelled N" panel traces the physical directions triggered by label N across all states. A globally consistent labelling would produce uniform arrows; the heterogeneous pattern here reveals where the GA has departed from a simple compass-rose rotation to align labels with environmental structure.

Goal 45 sits at the bottom of a doorway between the two lower rooms. Tracing the "what's labelled N" panel from state 0 in the top-left corner, an agent that repeats label N follows a coherent route into the goal: down the left wall, along the row-4 corridor, down through the bottom-left room, and east into state 45. The same label invokes a different physical action at each step, but the agent never switches labels, so decision information stays low along the route. This is what we will call a *habit cycle*: the agent traverses an extended route by repeatedly issuing a single label, with the local permutation σ_s converting that label into a different

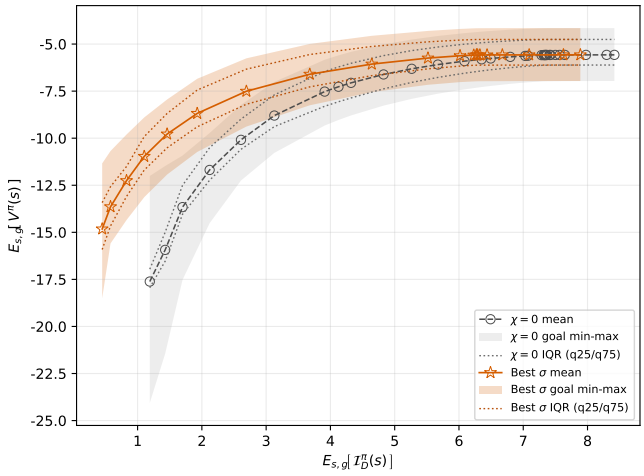

Figure 3: Value–decision information trade-off across $\beta \in [0.1, 1000]$ in four rooms. Each curve plots mean value $\mathbb{E}_{s,g}[V^{\pi^*_\beta}(s)]$ against mean decision information $\mathbb{E}_{s,g}[\mathcal{I}_D^{\pi^*_\beta}(s)]$, with the policy π^*_β re-optimised at each β. Orange with stars: GA-best twist $\sigma^\star$ (found at $\beta = 1$); grey with circles: Cartesian baseline ($\chi_{\text{twist}} = 0$). Solid lines show the mean uniformly averaged over all goals; shaded bands show the interquartile range and goal min–max. Important: The improvement observed at $\beta = 1$ is maintained across the full range of β without re-optimising σ.

physical move at each step. The cost stays low because the agent never switches labels — decision information penalises variation in the policy across states, and repeating a single label at every state is the minimum variation possible. In this twist, the label-N cycle sits at or near goal 45 itself, so repeating N pulls the agent into the goal region rather than past it. We return to habit cycles on the toroidal grid below, where the GA constructs them in a setting with no walls to lean on.

A single σ is necessarily a compromise across goals: it cannot align every approach direction equally well, and some goals — a lateral doorway, or an interior target requiring a deliberate turn — are served less favourably than goal 45. We show in the following section that the twist nonetheless reduces cost relative to the Cartesian baseline ($\chi_{\text{twist}} = 0$) across the whole goal set, confirming that the GA has found a relabelling with broad benefit rather than one tuned to a single destination.

Figure 2 provides a comparison. A high-χ twist applied to goal 45 produces less structured "what's labelled" panels: the same label sends the agent in different directions, with a concomitant increase in decision information. The contrast makes the point concrete: relabelling alone is not sufficient; the relabelling must align with environmental structure to reduce informational cost.

Value–information trade-off across β

The twist panels in Figures 1 and 2 show policies at $\beta = 1$, where value and information cost contribute equally to planning. Figure 3 extends the comparison across $\beta \in [0.1, 1000]$, plotting mean value $\mathbb{E}_{s,g}[V^{\pi^*_\beta}(s)]$ (higher is better) against mean decision information $\mathbb{E}_{s,g}[\mathcal{I}_D^{\pi^*_\beta}(s)]$ (lower is cheaper) for both the GA-best twist and the Cartesian baseline. The twist σ is fixed throughout: the GA finds a single relabelling at $\beta = 1$, and the curve evaluates that same σ at every other β by re-solving only the policy π^*_β. This tests whether a twist optimised at one point on the value–information trade-off generalises along it, rather than being a narrow fit to $\beta = 1$.

The two curves converge at high β, where the unconstrained information budget allows both policies to approach optimal value. As β decreases the agent favours more state-independent policies, and the advantage of aligned labels emerges: for any fixed information budget the GA twist achieves higher value than the Cartesian baseline. This exposes something the familiar compass labelling hides — the Cartesian coordinate system is a convenient default for human designers, not an informationally optimal one. The spread between the per-goal minimum and maximum also widens at low β: at a corner the agent must move against the prior action distribution to reach many goals (Archer et al., 2022), so corner goals become disproportionately expensive as the information budget tightens.

At any $\beta \neq 1$ this σ is no longer the relabelling the GA would have selected — re-running the search at that β would in principle yield a different best twist — yet the same σ still beats the Cartesian baseline across the entire range. The benefit of the GA twist therefore persists from near-open-loop to near-optimal regimes despite being optimised at a single $\beta = 1$: a single relabelling reshapes the whole trade-off curve rather than just one point on it.

Directed search vs random sampling

Finding a useful twist requires directed search. Figure 4 compares the σ visited by the GA (blue scatter) against a size-matched uniform random sample (orange hexbin), both evaluated at $\beta = 1$. The vertical axis is χ_{twist} (departure from a globally consistent labelling, defined in the Experimental Setup); the horizontal axis $\mathbb{E}[\mathcal{F}] = \mathbb{E}_{s,g}[\mathcal{F}^{\pi^*_\beta}(s)]$ averages the planning objective over starting states and goals, so lower is better. Random σ are drawn by choosing a local permutation independently and uniformly at each state, which makes a globally consistent labelling vanishingly unlikely; the random density therefore concentrates at high χ_{twist}, not near the Cartesian baseline ($\circ$, $\chi_{\text{twist}} = 0$).

The star marks the GA-best twist; the diamond marks the high-χ comparison twist of Figure 2. The GA population reaches lower $\mathbb{E}[\mathcal{F}]$ at moderate χ_{twist}, a region random sampling does not enter. Useful twists are thus neither globally consistent (like the Cartesian baseline) nor ar-

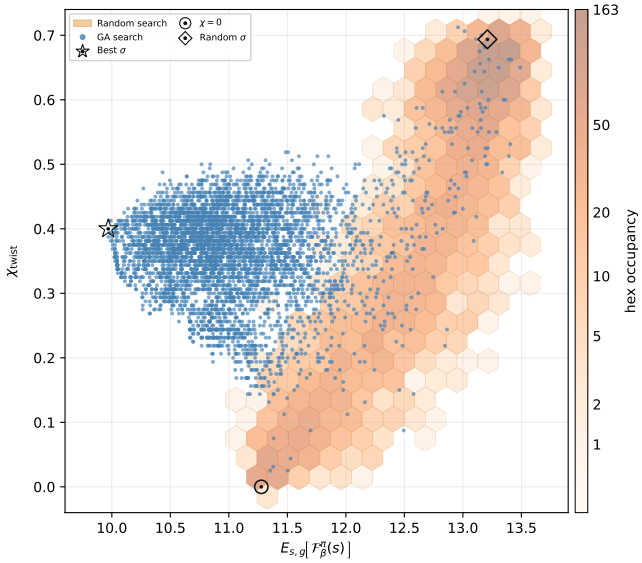

Figure 4: GA population (blue scatter points) overlaid on a size-matched random-search cloud (orange hexbin) in the χ_{twist}–$\mathbb{E}_{s,g}[\mathcal{F}^{\pi^*_\beta}(s)]$ plane for four rooms. Markers: $\star$ = GA-best multi-goal twist, $\circ$ = Cartesian baseline ($\chi_{\text{twist}} = 0$), $\diamond$ = high-χ comparison twist shown in Figure 2. The GA extends to lower $\mathbb{E}_{s,g}[\mathcal{F}^{\pi^*_\beta}(s)]$ at moderate χ values that random search does not reach.

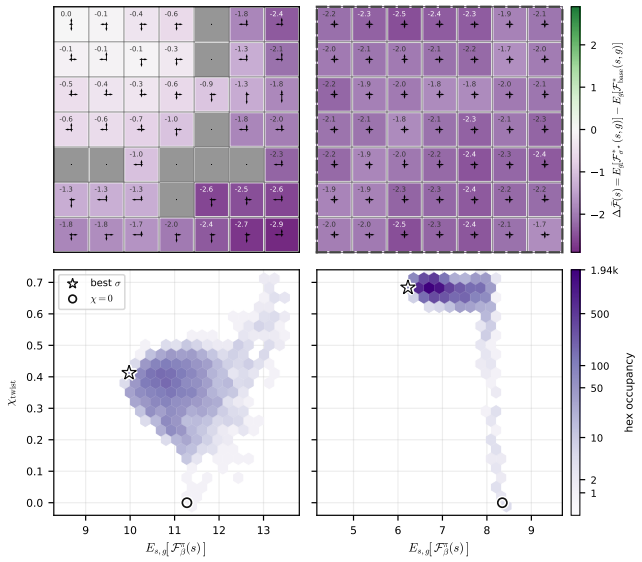

Figure 5: GA-best twists in four rooms (left) and a toroidally wrapped grid (right), $\beta = 1$, $\det = 0.97$. **Top row**: per-state change in planning objective $\Delta\mathcal{F}(s) = \mathbb{E}_g[\mathcal{F}^{\pi^*_\beta}_{\sigma^\star}(s)] - \mathbb{E}_g[\mathcal{F}^{\pi^*_\beta}_{\text{base}}(s)]$ of the GA-best twist relative to the Cartesian baseline, uniformly averaged over all goals, with the goal-marginalised policy overlaid as black quivers. Purple (negative $\Delta\mathcal{F}$) indicates the twist reduced the planning objective at that state; green would indicate an increase (no green cells in either panel: every state improves). Grey cells are walls; dashed borders indicate toroidal wrapping. **Bottom row**: hexbin density of the GA population in the χ_{twist}–$\mathbb{E}_{s,g}[\mathcal{F}^{\pi^*_\beta}(s)]$ plane; $\star$ = GA-best twist, $\circ$ = Cartesian baseline. Even in environments without internal walls, the GA finds twists that improve on the baseline.

bitrarily scrambled (like the bulk of the orange hexbin): they are structured permutations sitting in a narrow band of the search space that directed search recovers readily.

Action relabelling beyond structured environments

Having established the benefit in four rooms, we now ask whether the same result holds in a qualitatively different environment. Figure 5 compares two contrasting layouts: four rooms, where internal walls create doorways and corners, and a toroidally wrapped open grid with no internal structure. The GA search was repeated across several additional layouts (bounded grids with central pillars, corridors, and other obstacle configurations) with consistent results, available in the supplementary material.

In each top-row panel of Figure 5 the quiver arrows show the *goal-marginalised* policy $\bar{\pi}(a|s) = \mathbb{E}_g[\pi^*_\beta(a|s,g)]$ of the GA-best twist: the action distribution at each state, averaged uniformly over all goals. Longer arrows indicate that most goals prefer the same action at that state under the twisted MDP. This differs from the *state-marginalised* action prior $\hat{p}^{\text{U}}(a) = \mathbb{E}_s[\pi(a|s)]$ used inside the decision information calculation: the goal-marginalised policy is a per-state summary of "what the agent tends to do here across goals", whereas the action prior is a single distribution over actions used as the reference for measuring deviation. The grid world heatmap shows $\Delta\mathcal{F}(s) = \mathbb{E}_g[\mathcal{F}^{\pi^*_\beta}_{\sigma^\star}(s)] - \mathbb{E}_g[\mathcal{F}^{\pi^*_\beta}_{\text{base}}(s)]$, the per-state change in planning objective under the GA-best twist relative to the Cartesian baseline, averaged over all goals. This shading reveals where the relabelling acts: states with large $|\Delta\mathcal{F}|$ are those where the twist has reshaped the per-goal optimal policies.

The bottom-row hexbin panels show the GA population in the χ_{twist}–$\mathbb{E}_{s,g}[\mathcal{F}^{\pi^*_\beta}(s)]$ plane (star = best twist, circle = Cartesian baseline at $\chi_{\text{twist}} = 0$). In both environments the best twist sits left of the baseline — lower planning cost — but the height of the cloud differs sharply: the four-room population concentrates at moderate χ_{twist}, while the wrap-grid population sits much higher. The two environments achieve qualitatively different relabellings, a contrast we examine in the next section.

In four rooms, $\Delta\mathcal{F}(s)$ is negative everywhere, but the magnitude of improvement varies markedly by room. Dark purple indicates states where the GA-best twist has lowered the free energy for trajectories starting from that state, averaged over all goals. The top-left room is the largest — a 4×4 block of cells — and shows the smallest gains (lighter than

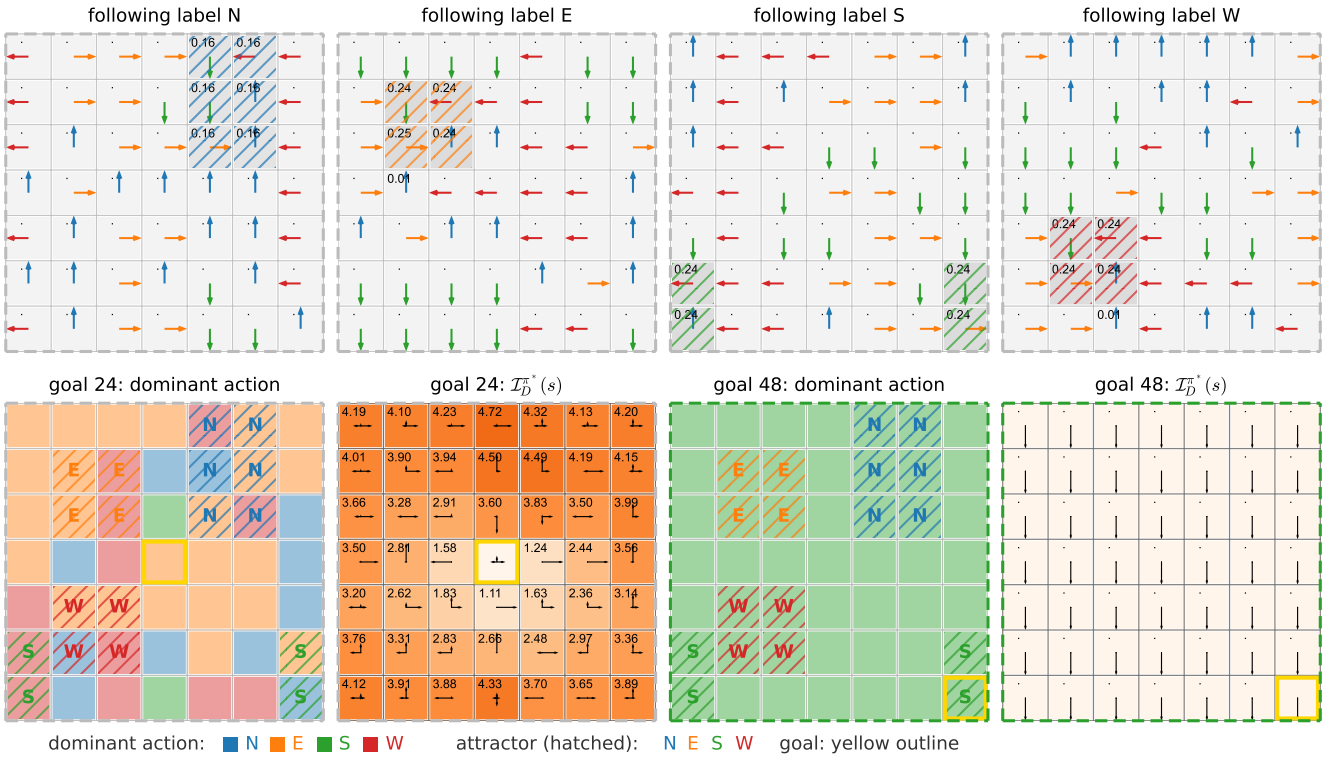

Figure 6: GA-best twist on the toroidal wrap grid (β=1, det =0.97, 7×7). Labels are colour-coded consistently across panels (N, E, S, W). **Top row**: Arrows show physical direction at each state for specified label. Hatched cells mark each label's *habit cycle* — the short orbit that the deterministic per-label dynamics converges to — with stationary occupancy $d^a(s)$ annotated on the cycle states. From any starting state, repeatedly issuing a label drives the agent into its cycle. Dashed borders indicate toroidal wrapping. **Bottom row**: for two example goals, a dominant-action panel (left) where each cell is coloured by the most-likely label at that state under the optimal policy (using the same N/E/S/W colours as the top row), and a decision-information panel (right) showing $\mathcal{I}_D^{\pi^*_{\beta,g}}(s)$, with the goal state outlined in yellow. Goals 24 (centre, outside any cycle) and 48 (inside label S's habit cycle) illustrate two contrasting cases. In both, the twist reduces decision information relative to the Cartesian baseline (see paper GitLab repo).

-1): its broad interior requires acting in several directions. The other three rooms are each roughly half the size. The top-right and bottom-left rooms are each connected to the top-left by a single doorway, where the gains are moderate (-1.3 to -2.1). The bottom-right room gains most (-2.4 to -2.9): it is the most isolated from the top-left, more like a passage between the two smaller rooms. In such confined spaces movement is essentially bidirectional, so many goals must traverse the bottleneck using fewer and similar actions; the twist can then align a single label with the dominant flow and thereby reduce the free energy.

The toroidally wrapped grid (dashed borders in Figure 5) is physically translation-invariant: every state has the same neighbourhood structure, so no state is intrinsically more costly than another. On such a symmetric environment one would naturally expect the Cartesian (cardinal) labelling to be optimal — there is no asymmetry for any other labelling to exploit, and all rotations of the compass rose are equivalent under the translation symmetry. Yet the $\Delta\mathcal{F}(s)$ heatmap shows the twist breaks this symmetry, reducing the planning objective uniformly across the grid. The result is consistent across runs: the orientation of the resulting pattern varies, but the improvement persists. The gain cannot come from any external feature on this symmetric grid; the GA must be synthesising a scaffolding directly in the labelling itself. Figure 6 shows the same kind of habit cycle we met in four rooms, here a short orbit the agent enters from all other states by repeating a single label.

The top row shows, for each of the four labels, the per-state physical direction the label invokes under the GA-best twist. Repeatedly issuing a single label drives the agent into that label's habit cycle, shown as hatched cells with the stationary occupancy $d^a(s)$ annotated. Each of the four labels has its own cycle in a different region of the torus.

This is how the twist imposes structure on a world that has none. In bounded grids, walls buffer incorrect actions

and corners trap repeated-action trajectories, making them cheap waypoints (Archer et al., 2022). The twist constructs habit cycles that serve the same function without walls: each label's cycle catches the agent just as a corner would, and repeated issue of the label carries the agent there from almost anywhere. The result is an action-space geometry where free energy, averaged over all goals, is lower than for the Cartesian coordinate system.

The bottom row of Figure 6 illustrates this for two example goals. The dominant-action panels show which label the optimal policy favours at each state, with cycle states marked by hatching; the adjacent panels show per-state decision information. Goal 24 sits in the centre of the grid, outside any cycle: reaching it requires state-specific actions to locate the goal, so decision information is non-zero across the grid, though still lower on average than under the Cartesian baseline. Goal 48, by contrast, sits inside label S's habit cycle, thus a single repeated label carries the agent to the goal at near-zero informational cost, so travelling there is literally free. This relationship between habit-cycle geometry, basin boundaries, and the cost of switching between labels is the subject of ongoing work.

Conclusion

We asked whether the labelling of an agent's actions — its action coordinate system, not the underlying physics — can be optimised for a bounded-rational agent. The answer is yes: in four rooms, a multi-goal GA discovers structured relabellings that reduce decision information across all goals simultaneously, exploiting walls and doorways as morphological computation. The useful twists occupy a distinct region of action-alignment space that random search, even with an equivalent evaluation budget, does not reach.

For any fixed information budget the GA twist reaches higher value than the Cartesian baseline, and this advantage holds from near-open-loop to near-optimal behaviour despite the twist being optimised at a single $\beta = 1$. A single relabelling thus reshapes the whole value–information trade-off rather than just one point on it: the familiar compass labelling is a convenient default for human designers, not an informationally optimal one.

On a featureless torus the twist has nothing external to align to, so the GA constructs structure instead: it carves the world into habit cycles. Repeating a single label from almost anywhere carries the agent into its cycle. Habits, in this view, have a geometry of their own — one the GA can evolve — mirroring the role walls and corners play in bounded environments, with the labelling standing in for the working memory habitual navigation would otherwise require (Mallot, 2024).

Action representation is not a new variable to optimise: it has been studied as a learned embedding in continuous-action settings (Chandak et al., 2019), and the broader question of which structures an agent should construct to act efficiently has a longer pedigree (Ho et al., 2022). What we add is the observation that even the discrete labelling of a fixed action set has a geometry of its own, one that can be aligned with environmental affordances where they exist, or constructed where they do not. Unlike a learned action embedding, which fits continuous parameters by gradient descent, a twist is a discrete, per-state permutation of a fixed action set: it relabels rather than re-embeds, leaving the available transitions exactly as they are and changing only their names, so the effective morphology stays interpretable and the search is combinatorial rather than differentiable. In this sense the GA twists are close cousins of value-guided construals, with the construal living in the agent's action coordinate system rather than in its plan.

Limitations and Future Work

Discovering a twist is itself costly: the genetic search is an offline, one-time computation whose cost is paid at a much slower timescale than the saving it buys. Like the morphology of a body, an action representation is shaped once, on an evolutionary horizon, then amortised over all the goal-directed behaviour it makes cheaper. The agent pays the reduced per-decision information cost repeatedly at the fast timescale of acting, thus earning its discovery cost precisely when reused across many goals.

How these results scale, and how far the discovered structures generalise, are open questions. The space of twists grows combinatorially, since each state admits $|\mathcal{A}|!$ orderings. Useful twists occupy a narrow band of moderate χ_{twist} (Figure 4); directed search recovers them without exhaustive coverage. Scaling is dominated by the evolutionary search, which re-evaluates the population against all goals (one per state) each generation.

The labelling itself appears to function as a proxy for working memory, externalising routing information that goal-independent navigation typically stores internally — a framing we intend to develop in follow-up work.

Community-detection methods (Evans and Şimşek, 2023) on the induced transition graph, and a label-wise progress measure added to the fitness, would formalise the link between habit-cycle basins and decision-information cost.

More broadly, larger and more structured environments would let us ask how action-space morphology reshapes which regions of the state space are cheap to act in, and whether a single twist can compose multiple basins by switching between labels under goal demand. Comparing twists across such settings calls for a quantitative treatment of their structure: each twist induces a functional graph whose fingerprint — its basins, cycle lengths, and the reach of a dominant cycle — is comparable across runs and environments. Folding such descriptors into the search, alongside decision information, would select not only for cheap twists but for particular dynamical signatures, a direction we pursue in follow-up work.

Acknowledgements

DP wishes to thank Keyan Zahedi for many insightful discussions on morphological computation and twisted worlds.

References

Archer, K., Catenacci Volpi, N., Bröker, F., and Polani, D. (2022). A space of goals: the cognitive geometry of informationally bounded agents. *Royal Society Open Science*, 9(12):211800.

Chandak, Y., Theocharous, G., Kostas, J., Jordan, S., and Thomas, P. S. (2019). Learning action representations for reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 941–950.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Evans, J. B. and Şimşek, O. (2023). Creating multi-level skill hierarchies in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston.

Hauser, H., Ijspeert, A. J., Füchslin, R. M., Pfeifer, R., and Maass, W. (2011). Towards a theoretical foundation for morphological computation with compliant bodies. *Biological Cybernetics*, 105(5-6):355–370.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., and Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, 606:129–136.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

Mallot, H. A. (2024). *From Geometry to Behavior: An Introduction to Spatial Cognition*. MIT Press, Cambridge, MA.

Müller, V. C. and Hoffmann, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artificial Life*, 23(1):1–24.

Ortega, P. A. and Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683.

Polani, D. (2011). An informational perspective on how the embodiment can relieve cognitive burden. In *2011 IEEE Symposium on Artificial Life (ALIFE)*, pages 78–85. ISSN: 2160-6382.

Zahedi, K. and Ay, N. (2013). Quantifying morphological computation. *Entropy*, 15(5):1887–1915.