Common Concepts in Agent Groups, Symmetries, and Conformity in a Simple Environment

Marco Möller[1,2] and Daniel Polani[2]

[1]Theory of Complex Systems Group, Institute of Solid State Physics, Technical University of Darmstadt, Germany
[2]Adaptive Systems Group, School of Computer Science, University of Hertfordshire, UK
marco.moeller@juforum.de*, d.polani@herts.ac.uk

Abstract

We analyze representations of the world attained through an infomax principle by agents acting in a simple environment. The representations obtained by different agents in general differ to some extent from each other in different instances. This gives rise to ambiguities in how the environment is represented by the different agents. We now develop an information-theoretic formalism able to extract a "common conceptualization" of the world for a group of agents. It turns out that the common conceptualization intuitively seems to capture much higher regularities or symmetries of the environment than the individual representations.

We formalize the notion of identifying symmetries in the environment - with respect to "extrinsic" operations on the environment as well as with respect to "intrinsic" operations, i.e. the reconfiguration of the agent's embodiment. In particular, using the latter formalism, we can re-wire an agent to conform to the highly symmetric common conceptualization to a much higher degree than an unrefined agent; and that without having to re-optimize the agent from scratch. In other words, we can "re-educate" an agent to conform to the de-individualized "concept" of the agent group with comparatively little effort.

Motivation

In the search of how agents aim to model their environment, there is a huge collection of candidates. However, it has been suspected earlier that, whatever the detailed mechanism would entail, they might follow principles of information parsimony or optimal information processing (Barlow (1959); Laughlin (2001)). A concrete model for maximum Shannon information processing has been proposed in the infomax model by Linsker (1988).

We are interested in how agents can model their environment based on informational considerations. Using infomax principles to do that, one obtains a classification or representation of a given environment (in the following also called concept) for a given agent. We use the perception-action (PAL) loop from Klyubin et al. (2007) to model the agent and its interaction with the environment, i.e. the model and according tasks for the agent are not part of this work.

In general, the representations of the environment developed in an infomax process differ w.r.t. the agent. Even in very simple and highly symmetric scenarios, they can considerably vary from agent to agent as a result of the infomax optimization i.e. different global and (good) local optima can be returned. This is similar to a biological evolution optimization process: the individuals also vary to some extend from each other. This raises the issue of how similar the obtained concepts are. We will discuss what the different concepts of those agents have in common. Is it possible to develop a concept which is mutually compatible to each of these input concepts (see e.g. Philipona and O'Regan 2006; Steels 1997; i Cancho and Solé 2003)? If so, what properties of the environment or the agents do such common concepts capture? How do they relate to the individual agents' concepts?

We will not model how agents agree on a common concept or how they communicate but we will discuss some information-theoretical criteria for such a common concept. In general, we are not interested in processes but in their outcome. We do not analyze mechanisms but the underlying principles.

Analyzing the quality of concepts with respect to certain goals, we observed that "good" concepts have more regularities. That led us to analyze the concepts' symmetries. In general, they are not symmetric in a strict mathematical way. So we needed a method to measure also not perfectly fulfilled symmetries. We developed an information-theoretical approach to analyze these "weak" symmetries. One hypothesis is that common concepts will reveal symmetries of the whole agent/environment system that are broken by the individual concepts. We can now ask under which conditions the individual agents can relate to this expected higher regularity of the common concept. In a second approach of analyzing these symmetries, we study the influence of the agents embodiment on the agent and try to find a way of "asking the agent what he considers to be a symmetry of the environment".

The technical challenges arising from these issues are manifold. We aim to find a description that is consistent with a fundamentally information-theoretical picture of the agents and their environment. For this, one needs to suitably

formulate the development of a common concept of a set of agents. Also, one needs to model the concept of regularity or symmetry in a suitable way.

The contributions of this paper are information-theoretic techniques to construct common concepts for a group of agents and to evaluate weak symmetries, and their application to some simple, but informative scenarios.

Background

To be able to introduce our model for the agents and their interaction with the world, we have to introduce some notations and quantities first. Consider random variables $X, Y, Z, \ldots$ denoted by capital letters which take in values $x, y, z, \ldots$ in corresponding sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$. For the probability that a given random variable X assumes a value $x \in \mathcal{X}$ we write $\Pr(X = x)$ or, if it is clear from context just $p(x)$. For the probability for the joint variable $(X_1, \ldots, X_n)$ we will write simply $\Pr(X_1 = x_1, \ldots, X_n = x_n) \equiv p(x_1, \ldots, x_n)$. The (*Shannon*) *entropy* of a random variable X is given by

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{1}$$

whereby the logarithm in this paper is always to the basis of 2, so the unit for entropy is the *bit*. The *conditional entropy* of X given Y is given by $H(X|Y) := H(X,Y) - H(Y)$ and the *mutual information* between X and Y by

$$I(X;Y) := H(X) + H(Y) - H(X,Y). \tag{2}$$

A generalization of mutual information is the *multiinformation* between a collection of random variables $X_1, \ldots, X_n$

$$I(X_1; \ldots; X_n) := \left[\sum_{i=1}^{n} H(X_i)\right] - H(X_1, \ldots, X_n) \tag{3}$$

its conditional form, if the random variable Y is observed is

$$I(X_1;\ldots;X_n|Y) := \left[\textstyle\sum_{i=1}^{n} H(X_i|Y)\right] - H(X_1,\ldots,X_n|Y). \tag{4}$$

To measure the "difference" between two random variables X, Y we can use the unnormalized version of the information distance (Crutchfield (1990))

$$D(X,Y) := H(X|Y) + H(Y|X) \tag{5}$$

which fulfills the conditions for a metric including the triangle inequality. Note that D vanishes for a deterministic bijective dependency between X, Y.

To model agents in an environment we will use the formalism from Klyubin et al. (2007) based on *causal Bayesian network* (CBN). A CBN is given by a directed acyclic graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ whose nodes $n \in \mathcal{N}$ are representing random variables X_n and the edges $e \in \mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ causal conditional probability dependencies between them. The distribution of X_n is given by $p\left(x_n|x_{\mathrm{Pa}(n)}\right)$ whereby $\mathrm{Pa}(n) := \{n' \in \mathcal{N} | (n', n) \in \mathcal{E}\}$ is the set of parent nodes n' from node n. If a node n has no parent nodes $\mathrm{Pa}(n) = \emptyset$, we identify $p\left(x_n|x_{\mathrm{Pa}(n)}\right) \equiv p(x_n)$ with an unconditional probability distribution. The joint distribution of the whole network is given by

$$p\left(x_1, \ldots, x_{|\mathcal{N}|}\right) = \prod_{n \in \mathcal{N}} p\left(x_n|x_{\mathrm{Pa}(n)}\right). \tag{6}$$

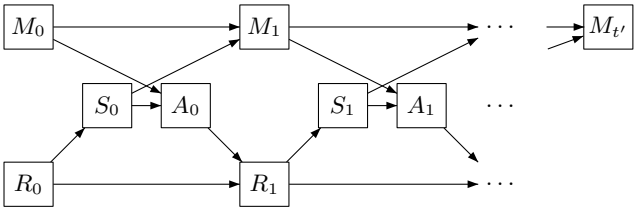

Figure 1: Perception-action loop unrolled in time as a CBN

Model

A generic model for an *agent* interacting with a world is the *perception-action loop* (PAL). It is here only briefly presented, for a full presentation and motivation see Klyubin et al. (2007). Such an agent can *sense* the *world* R through its *sensor* S and *manipulate* it through its *actuator* A which together form the *embodiment* of the agent. This process can be formalized by the CBN shown in Fig. 1. All random variables depend on the time t: M_t, A_t, R_t, S_t. More precisely the *controller* of the agent has the possibility to store information in the *memory* M. It can be described by a probabilistic mapping

$$\text{controller} : M_t \times S_t \to M_{t+1} \times A_t \tag{7}$$

which is time t independent.

In our experiments, we chose a deterministic controller (Wennekers and Ay (2005); Klyubin et al. (2007)) and used a two dimensional infinite grid-world $\mathcal{R} = \mathbb{Z}^2$. The memory M is a number contained in a finite subset of $\mathcal{M} \subset \mathbb{N}$. The initial memory M_0 is deterministically set to a default state 0. The initial position in the world R_0 is uniformly distributed over possible starting positions $\mathcal{R}_0 = \{-d, \ldots, d\}^2$ where the *radius* d depends on the experiment. The actuator A can take on values $\mathcal{A} = \{\downarrow, \leftarrow, \uparrow, \rightarrow\}$ where these 4 actions can move the agent (changing its position in the world, encoded in R) to one of its 4 adjacent positions in the grid-world. The first discussed sensor (*setup s+*) has 4 possible sensor values $\mathcal{S} = \{\downarrow, \leftarrow, \uparrow, \rightarrow\}$. If we imagine a "pheromone" gradient emitted by a source at the origin (Fig. 2 - center), this sensor points to the adjacent position with the highest concentration of pheromone. If this is not unique (e.g. at the origin), one direction is randomly chosen. Setup s+ is visualized in the left of Fig. 2, whereby for each position $(x, y) \in \mathcal{R}$ all possible sensor "directions" are shown with their arrow-length corresponding to their probability. A variation of this setup used in this work is a sensor (*setup sq*) where 4 of such sources exists at $\{-5, 5\}^2$

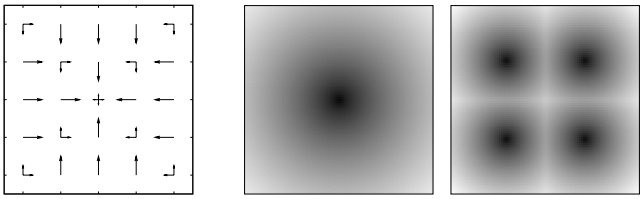

Figure 2: Setups

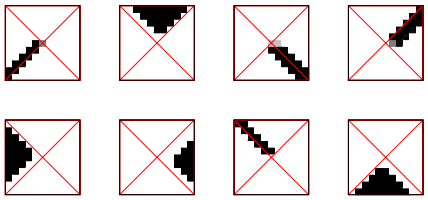

Figure 3: Solution of initial position capturing

(Fig. 2 - right) and the sensor is pointing always to the nearest source.

Our fundamental task for the agent is to capture as much information about its initial position as possible by its "final" memory state at time[1] $t = 15$ as suggested by Klyubin et al. (2007). This can be denoted information-theoretically as maximizing

$$I\left(R_0; M_{15}\right). \tag{8}$$

The search space for this problem contains all possible controller mappings from Eq. 7. To solve this and all following optimization problems, we used *Simulated Annealing* with some heuristic improvements described elsewhere but such tasks can be performed by any generic optimization tool. We do not aim to model the details of the process of agent evolution / adaptation and its ability to capture the information about the initial position but only the outcome of such a process. This output corresponds to the solutions returned by Simulated Annealing.

Common Concepts

Concepts

Consider an agent with setup s+ and memory size $|\mathcal{M}| = 8$ who is able to capture the initial position R_0 (what is uniformly distributed with $\mathcal{R}_0 = \{-5, \ldots, 5\}^2$) by maximizing $I\left(R_0; M_{15}\right)$. Therefore an appropriate controller has to be found. To interpret this agent, consider Fig. 3 where each of the 8 squares shows in gray scale the conditional probability $p\left(r_0|m_{15}\right)$ for a final memory state. To make this precise, it shows the probability that this agent has been initially at position r_0 in the world for a memory content $m_{15} = 0, 1, 2, \ldots, 7$ at time $t = 15$ of the end of the run. For each state m_{15} we use a separate normalization so $\max_{r_0} p\left(r_0|m_{15}\right)$ is represented by black and $p\left(r_0|m_{15}\right) = 0$ by white. The agent shown has an utility value of $I\left(R_0; M_{15}\right) = 2.906$ bit which is very near to the limit of $\min\left[\log|\mathcal{R}_0|, \log|\mathcal{M}|\right] = 3$ bit. These 8 possible memory values m_{15} can be understood as a *concept* of the world R_0. Each value for m_{15} has a certain "meaning" for stating positions, like "north-triangle", "north-east-diagonal", "east-triangle", "south-east-diagonal", etc. We call a pair of random variables (R, Y) (e.g. (R_0, M_{15})) jointly distributed a *concept* if Y is "representing" R in some way, i.e. $I\left(R; Y\right) > 0$. We call the values $y \in \mathcal{Y}$ *symbols* of the concept.

As mentioned earlier there also exist other solutions for the problem to find a good initial position capturer with an equal or similar utility value $I\left(R_0; M_{15}\right)$, for example just an agent with a "rotation" of the concepts by 90°. This "rotation-symmetry" will be discussed later. Here we are interested in how representative the shown example concepts and how similar other solutions are. We will do this by discussing the possibilities to find a *common concept* (R, Y_*) from a set of *input concepts* $\left\{\left(R, Y^{(1)}\right), \ldots, \left(R, Y^{(n)}\right)\right\}$. This concept can be interpreted as common concept of a group of agents in a world R. In the spirit from above philosophy, we emphatically only model the information-theoretical principle. The process of agreeing of the individuals about the common concept is wittingly not modeled to be independent of the algorithm. We will present in the following two possibilities to define such a common concept.

For the Objective Common Concept consider the CBN from Fig. 4. A deterministic mapping $R \rightarrow Y_*^{obj}$ which maximizes

$$\sum_i \left[I\left(Y_*^{obj}; Y^{(i)}\right) - \alpha \cdot I\left(R; Y_*^{obj}\right)\right] \tag{9}$$

defines the *objective common concept* $\left(R, Y_*^{obj}\right)$. The first term $I\left(Y_*^{obj}; Y^{(i)}\right)$ maximizes the mutual information between the common concept and every input concept, so as to make it as similar as possible the input concepts. The term $\alpha \cdot I\left(R; Y_*^{obj}\right)$ is a bottleneck type (Tishby et al. (1999)) parameter $\alpha \in [0, 1]$ countering the trivial behavior of just building $Y_*^{obj} = Y^{(1)} \times \ldots \times Y^{(n)}$ as cross product of all input concepts if the number of states in Y_*^{obj} is sufficiently large $\left|\mathcal{Y}_*^{obj}\right| \geq \prod_i \left|\mathcal{Y}^{(i)}\right|$. For our experiments we set $\alpha = 0.2$. This method is called objective because it has explicit knowledge about the world R.

For the Subjective Common Concept consider the CBN from Fig. 5. A deterministic mapping $Y^{(1)} \times \ldots \times Y^{(n)} \rightarrow Y_*^{subj}$ which minimizes

$$I\left(Y^{(1)}; \ldots; Y^{(n)} | Y_*^{subj}\right) \tag{10}$$

[1] $t = 15$ is an arbitrary choice for our experiments, other choices lead to similar results.

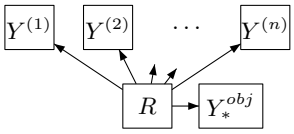

Figure 4: CBN for objective common concept

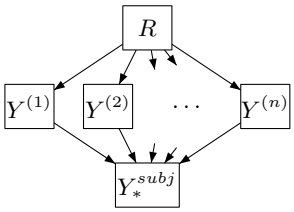

Figure 5: CBN for subjective common concept

defines the *subjective common concept* $\left(R, Y_*^{subj}\right)$ by applying the rules for the joint distribution of a CBN. The minimization makes sure that Y_*^{subj} "absorbs" all information common by $Y^{(1)}, \ldots, Y^{(n)}$. This method is called subjective because it has only implicit knowledge about R through the input concepts.

Results Common Concept

A Comparison of Objective and Subjective Common Concept is calculated for the 4 input concepts shown in Fig. 6. We see in each of the 4 columns one concept $\left(R_0, M_{15}^{(i)}\right)$ generated by initial position capturing agents with setup s+ and $|\mathcal{M}| = 6$. Figure 7 shows an objective and subjective common concept (R_0, M_*) of size $|\mathcal{M}_*| = 8$ each. The superstition of the subjective common concept is with $H\left(M_*|R_0\right) = 0.03$ bit vanishingly small[2]. The information distance between objective and subjective common concept is with $D\left(M_*^{obj}, M_*^{subj}\right) = 0.38$ bit also quite small. The only significant difference is that the symbol for "south-east" is split in the subjective method, therefore it has no symbol for "north-west". Because of their similarity we will not continue to calculate both common concepts. Especially if we consider the computational complexity we will, in further investigations, only use the objective common concept. For the subjective common concept the computational complexity, and the size of the search space are growing exponentially with the size of the input concept $\left|\mathcal{M}_{15}^{(i)}\right|$ and their number n. Some further objective common concepts are shown in Fig. 11.

For lack of space the preferred common concept size will not discussed here.

[2]The superstition of the objective common concept is 0 by definition because M_* deterministically depends on R_0.

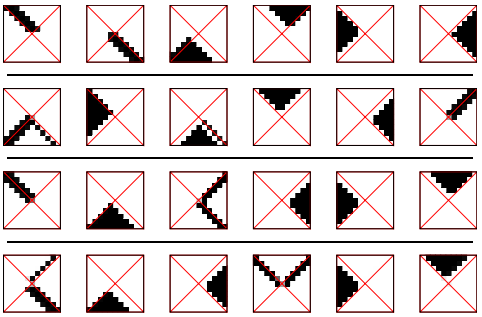

Figure 6: 4 input concepts of size $|\mathcal{M}| = 6$ for Fig. 7

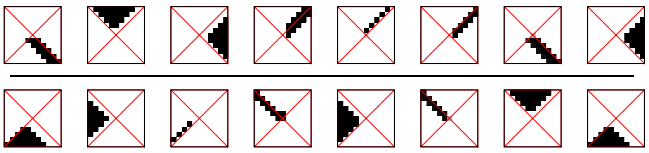

Figure 7: Objective (upper half) and subjective (lower half) common concept

Symmetry

As mentioned earlier, all (common) concepts exhibit a large degree of symmetry. We will present two methods to measure and analyze these symmetries. Common to both methods is the idea of transforming concepts and comparing them by measuring the mutual information between the transformed concept and e.g. the original. The *extrinsic symmetry* transforms the concept by applying a combination of a rotation, mirroring and translation on the world. So it tests if some explicitly known symmetries of the world also hold for the concept. The *intrinsic symmetry* opposed searches for invariants of the world from the agents perspective. So it is able to extract what seems to be a symmetry for the agent.

With these methods we developed a framework for analyzing the role of regularities in agent↔world interaction and especially what kind of regularities are used by agents in their interaction with the world.

Extrinsic Symmetry

An *extrinsic symmetry* operating on a concept (R, Y) transforms the grid-world $\mathcal{R} = \mathbb{Z}^2$ by applying an *extrinsic symmetry operation* $\xi^{\theta,\varphi,x_0,y_0}$, a combination of a rotation φ (in 90° steps), mirroring θ, and translation (x_0, y_0)

$$\xi^{\theta,\varphi,x_0,y_0} : \mathbb{Z}^2 \to \mathbb{Z}^2 \tag{11}$$

$$\xi^{\theta,\varphi,x_0,y_0} := \xi_{\text{trans}}^{x_0,y_0} \circ \xi_{\text{rot}}^{\varphi} \circ \xi_{\text{mir}}^{\theta}. \tag{12}$$

The mirroring (at the y-axis) is described by $\theta \in \{+1, -1\}$

$$\xi_{\text{mir}}^{+1}(x, y) := (x, y) \qquad \xi_{\text{mir}}^{-1}(x, y) := (-x, y),$$

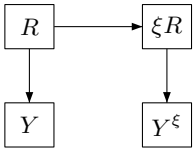

Figure 8: CBN for calculating extrinsic symmetry utility

the rotation (in 90° steps counterclockwise) by $\varphi \in \{0°, 90°, 180°, 270°\}$

$$\xi_{\text{rot}}^{0^\circ}(x,y) := (x,y) \qquad \xi_{\text{rot}}^{90^\circ}(x,y) := (-y,x)$$
$$\xi_{\text{rot}}^{180^\circ}(x,y) := (-x,-y) \qquad \xi_{\text{rot}}^{270^\circ}(x,y) := (y,-x)$$

and finally the translation by $(x_0, y_0) \in \mathbb{Z}^2$

$$\xi_{\text{trans}}^{x_0,y_0}(x,y) := (x + x_0, y + y_0).$$

The application of the operation $\xi^{\varphi,\theta,x_0,y_0}$ transforms the world R and gives us two new probabilistic mappings $R \to \xi R$ and $\xi R \to Y^{\xi}$ (Fig. 8). The first mapping applies the operation $\xi^{\varphi,\theta,x_0,y_0}$ on R by

$$\Pr(\xi R = r' | R = r) := \delta_{r',\xi(r)} \tag{13}$$
$$= \begin{cases} 1 & r' = \xi(r) \\ 0 & \text{else} \end{cases}. \tag{14}$$

The second mapping $\xi R \to Y^{\xi}$ is chosen as a "copy" of $R \to Y$:

$$\Pr\left(Y^{\xi} = y | \xi R = r\right) := \Pr(Y = y | R = r). \tag{15}$$

We define the *utility for extrinsic symmetry operation* $\xi^{\varphi,\theta,x_0,y_0}$ for the concept (R, Y) as

$$I\left(Y; Y^{\xi}\right), \tag{16}$$

where a higher value means "higher symmetry".

Informally, this utility measures "how much a rotated/mirrored/translated concept has in common with the original one". Note that because of the use of information theory, possible symbol permutations are ignored. With this method we are also able to interpret a sensor mapping $R_t \to S_t$ as a concept and calculate its symmetries.

Intrinsic Symmetry

We define a *permuted embodiment* for an agent as shown in Fig. 9. In comparison to Fig. 1, the original sensor S is replaced by $S^{\pi} \to S^{orig}$ and the original actuator A by $A^{orig} \to A^{\pi}$. Each pair of permutation of sensor an actuator (π_S, π_A) with

$$\pi_S : S^{\pi} \to S^{orig} \tag{17}$$
$$\pi_A : A^{orig} \to A^{\pi}. \tag{18}$$

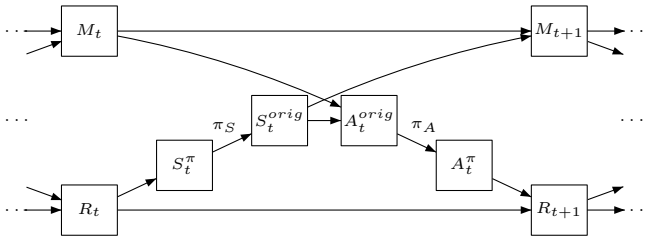

Figure 9: Permuted embodiment for the perception-action loop as CBN

defines an *intrinsic symmetry operation*. To evaluate an intrinsic symmetry operation (π_S, π_A) on an initial position capturing agent we first generate a whole set of concepts $\left\{\left(R_0, M_{15}^{(1)}\right), \ldots, \left(R_0, M_{15}^{(n)}\right)\right\}$ from other good initial position capturing agents with an equivalent setup (both, evaluated agent and the other agents are still without the modifications from Fig. 9 at this point). For this set of concepts is an objective common concept $\left(R_0, M_*^{obj}\right)$ is computed. To evaluate a specific intrinsic symmetry operation (π_S, π_A), we apply it on the PAL[3] and then calculate the resulting concept (R_0, M_{15}^{π}). We define the quality of this operation as

$$I\left(M_*^{obj}; M_{15}^{\pi}\right), \tag{19}$$

where a higher value means "higher symmetry". Informally spoken, "we are shuffling perceptions and actions of the agent and investigating if he is still able to 'conform' the common concept".

Results Symmetry

Figure 11 shows because of lack of space in extremely compact format some of our symmetry results for comparison. The upper half of the figure is for setup s+, the lower half for setup sq.

The Extrinsic Symmetry of the Setup is shown in ① and ②. ① shows the used setup as map of possible sensor outcomes (understood as concept) $p(r|s)$. The marked box inside of each symbol denotes the places belonging to R_0. This region visualizes the area of the concept what will be compared with the transformed concept by the mutual information. For setup s+ is $\mathcal{R}_0 = (-5, \ldots, 5)^2$ and for setup sq $\mathcal{R}_0 = (-10, \ldots, 10)^2$. Of the translation operations, only those are tested which maps R_0 on a subset of the shown positions in the concept, consequently only translations are tested with $\max(|x_0|, |y_0|) \leq 5$ for setup s+ resp ≤ 10 for setup sq. The corresponding extrinsic *symmetry spectrum* in ② shows the number of symmetries (y-axis) for a given value of $x = \frac{I\left(R;Y^{transformed}\right)}{\max I\left(R;Y^{transformed}\right)}$. Additional to these

[3]Without changing on the controller, i.e. not optimizing the utility from Eq. 8 again.

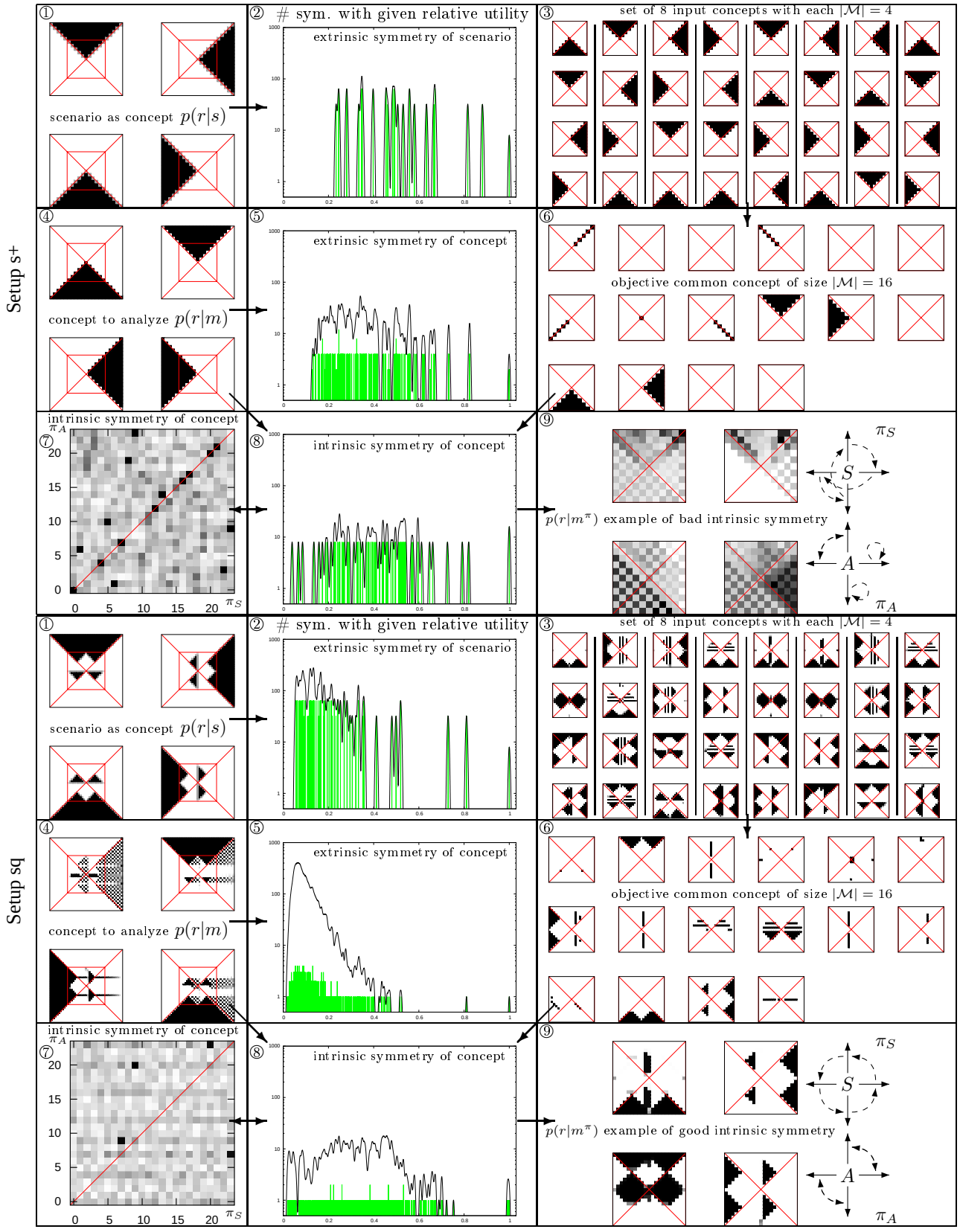

Figure 11: Symmetry - see text for details

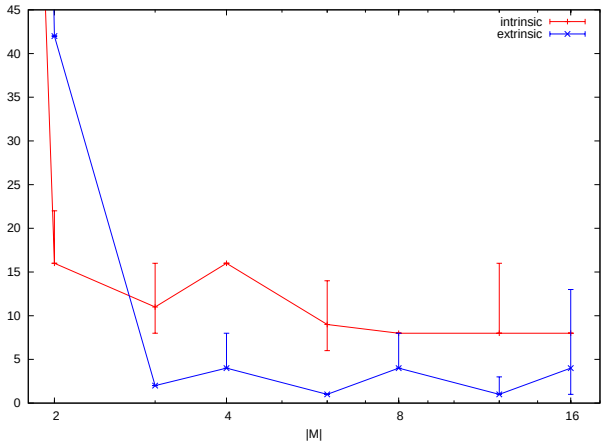

Figure 10: Number of good intrinsic and extrinsic symmetries depending on memory size

peaks, we added a smoothing curve into the spectrum. The rightmost peak with the 8 best symmetries includes all operations with translation $x_0 = y_0 = 0$. The next best peak (second rightmost) covers translation of length 1. The peaks are mostly ordered by their translation length $\sqrt{x_0^2 + y_0^2}$.

The Extrinsic Symmetry of the Concept is shown in ④ and ⑤. ④ shows a concept derived from an initial position capturing agent (R_0, M_{15}) with $|\mathcal{M}| = 4$ and its extrinsic symmetry spectrum similar to the setup in ① and ②. Also here the peaks are mainly ordered by translation distance. For setup s+ the peak with the best 4 operations contains the mirroring about the x resp. y axis. If we mirror around the x-axis we have additionally to translate the concept by 1 in y direction to get it perfectly matching with the original one. For setup sq we see that there is only one best symmetry, the identity. The next best symmetry, a mirroring around the x-axis, is much weaker.

An Objective Common Concept (R_0, M_*) which is used for the intrinsic symmetry is derived from 8 other solutions for the initial position capturing (shown in ③: each column of 4 symbols forms one input concept). The common concept has a size of $|\mathcal{M}_*| = 16$ symbols and is shown in ⑥.

The Intrinsic Symmetry of the Concept is shown once as spectrum in ⑧. ⑦ also shows these intrinsic symmetries but in a different way. The x-axis resp. y-axis enumerates the different possibilities for the permutations for π_S resp. π_A whereby 0 stands for the identity. The gray values are according to $\frac{I(R;M_{15}^{\pi})}{\max_{\pi} I(R;M_{15}^{\pi})}$ with an enlarged contrast for values near to 1 resp. black. The diagonal in this map stands for "synchronized" embodiment permutations with $\pi_S = \pi_A$. The introduction of those synchronized permutation makes only sense if, like in our case, sensor and actuator values can be associated and ordered in the same way.

⑨ shows on of the best (for setup sq) resp. of the worst (setup s+) concepts $(R_0; M_{15}^{\pi})$ after applying an intrinsic symmetry operation (π_S, π_A). This operation resp. its two permutations are shown to the right of the concept. The 4 possible values for S resp. A are shown as solid arrows and their permutation mappings with dashed arrows. In case of setup s+ the 16 best symmetries are similar to the 8 rotation and mirroring of the most right two input concepts shown in ③. In case of setup sq the 4 best symmetries are similar to the shown example but mirrored around the x- and/or y-axis.

Symmetry Dependence on Memory size $|\mathcal{M}|$ is shown in Fig. 10. It shows the number (y-axis) of good extrinsic resp. intrinsic symmetries (at least 85 % of maximal symmetry utility) for an initial position capturing agent with setup s+ according to memory size $|\mathcal{M}|$ (x-axis). The error-bars show the number of symmetries with at least 82.5% resp. 87.5% of maximal symmetry utility.

Discussion

We have shown how to extract common perspectives out of a group of agents with individual perspectives. There is evidence that both objective and subjective methods are almost similar if, as in our case, the input concepts are mostly deterministic. So one can save computation resources by calculating only the objective one. Both methods are based on the fact that for some locations in the world, the agents have a disagreement about how to group them to symbols (in our example e.g. the 4 diagonals in setup s+). Additionally to assigning the "original" symbols for indisputable areas, the common concept methods are able to identify the disputed areas and assign new symbols to them. If we would enlarge the memory size for the individual agents, they would find some of these new symbols as well. In our example, an agent with a bigger memory size would also find the diagonals but with much lower accuracy. Especially the symbol for the "center of the world" (Fig. 11 setup s+ ⑥) was never found by an individual agent in our experiments. So a new level for structuring the world emerged by considering a whole group of agents instead of individuals.

We also have evidence that "good" agents' concepts and especially common concepts have a higher degree of "symmetry". We developed two methods to study the strength of symmetry. With the extrinsic symmetry method, rotation and mirroring symmetries were found but the translations were not. Some "long distance" similarities in the sq setup we expected to appear were too weak and vanished in the "noise" of other translations. But, as expected, small translations were not completely asymmetrical. In general, the degree of symmetry is vaguely ordered by its translation length.

As opposed to the extrinsic, the intrinsic symmetry just observes which changes in the agent's interaction with the

environment (actuator/sensor permutations) have no (bad) effect on its concept. This method is additionally improved in that we do not compare a permuted concept with the individual (original) concept but with a common one. This common concept is "free" of special decisions of individual agents and gives a more universal representation for a task than any individual solution. The intrinsic operation forces the agents to "conform" to the common concept without optimizing them again by "transplanting their brain into another body". Searching for the best intrinsic operations is in fact partly a re-optimization of the controller. But in total, we only test in the shown example a vanishingly small ($3.1 \cdot 10^{-17}$-th) part of the search space. The meaning of the intrinsic symmetry method is not yet fully understood. Partly the intrinsic symmetries are identical to the extrinsic symmetries (rotation, mirroring) but they include many more operations.

Increasing the agent's memory size, the number of best extrinsic symmetries drops to 1 which means that identity is the only remaining symmetry operation. The number of best intrinsic symmetries behaves differently which means that intrinsic symmetries are not too sensitive to variations of the concept due to symmetry operations. This raises another interesting idea: The intrinsic symmetry may give us a hint for an optimal memory size of an agent. With growing memory size, the agents begin to "realize" that not every symmetry they "see" is really in the world. But this process stops at a certain $|\mathcal{M}|$ which might be a good choice for an agents memory size in the considered environment.

Conclusion and Outlook

We discussed two techniques to generate a common perspective by conflating the individual perspectives of a group of agents. Through this common perspective, we were able to analyze the similarity of individual agent representations and find common classifications of the environment. Additionally, some features of the world are only (or at least much more easily) detectable in the common perspective. We did not model the process of agreeing between these agents and only used very general information-theoretical principles which make them applicable to other scenarios as well.

We found evidence that good classifications of the environment capture many of its symmetries. While individual concepts may suffer some symmetry breaking, common concepts will reveal these symmetries. To analyze these symmetries, we developed two information-theoretical approaches. In the extrinsic approach, we measure for every symmetry transformation of the environment the degree to which the concept is respected. This approach abstracts away from how we achieved the classification. In contrast, the intrinsic approach is only suitable for agents interacting with an environment through a PAL. Here we analyze which modifications of the embodiment lead to agents who are "similar" to the original one. Since we measure this similarity indirectly by comparing the transformed concept to a common concept, the individual concept's symmetry breaks do not influence this method. The intrinsic method provides insight into the agent and its perspective on the environment. It identifies symmetries beyond the geometrical symmetries of the world found in the extrinsic case. The intrinsic symmetries accord to changes of the agent's embodiment which can not be detected by the agent.

Especially the role of the intrinsic symmetry and its meaning is not fully understood. In the future, it could help to extract structural regularities in the environment by the agent.

Acknowledgments

We want to thank Prof. Barbara Drossel for many ideas and suggestions. The first author was supported by scholarship from the Studienstiftung des deutschen Volkes.

References

Barlow, H. B. (1959). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, pages 217 – 234. The M.I.T. Press.

Crutchfield, J. (1990). Information and its Metric. In Lam, L. and Morris, H., editors, *Nonlinear Structures in Physical Systems – Pattern Formation, Chaos and Waves*, pages 119–130. Springer Verlag.

i Cancho, R. F. and Solé, R. V. (2003). Least effort and the origins of scaling in human language. *PNAS 788-791*, 100(3):788–791.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2007). Representations of space and time in the maximization of information flow in the perception-action loop. *Neural Computation*, 19(9):2387–2432.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, (11):475–480.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117.

Philipona, D. and O'Regan, J. (2006). Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties. *Visual Neuroscience*, 23(3-4):331–339.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34.

Tishby, N., Pereira, F., and Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Wennekers, T. and Ay, N. (2005). Finite State Automata Resulting from Temporal Information Maximization and a Temporal Learning Rule. *Neural Comp.*, 17(10):2258–2290.