Autonomous Development of Goals: From Generic Rewards to Goal and Self Detection

Matthias Rolf and Minoru Asada
Osaka University, Japan
{matthias,asada}@ams.eng.osaka-u.ac.jp

***Abstract*—Goals are abstractions that express agents' intention and allow them to organize their behavior appropriately. How can agents develop such goals autonomously? This paper proposes a conceptual and computational account to this longstanding problem. We argue to consider goals as abstractions of lower-level intention mechanisms such as rewards and values, and point out that goals need to be considered alongside with a detection of the own actions' effects. Then, both goals and self-detection can be learned from generic rewards. We show experimentally that task-unspecific rewards induced by visual saliency lead to self and goal representations that constitute goal-directed reaching.**

***Index Terms*—Latent Goal Analysis, Goal Systems Development, Self Detection, Goal Babbling, Saliency**

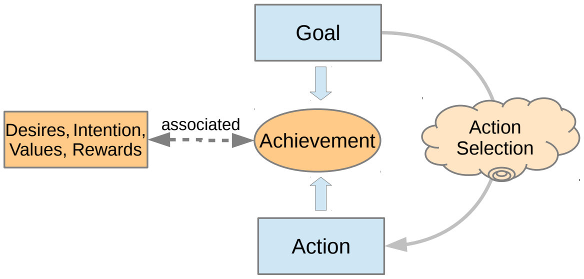


Fig. 1. Goals are abstraction whose achievement by means of action is associated to some reward or desire.

I. Introduction

Goals are abstractions of high-dimensional world states that express intelligent agents' intentions underlying their actions. Goals are considered to organize the behavior of both humans and robots. For instance in robot *planning* [1] as well as motor *control* [2], [3] goals describe the desired outcome of future actions in terms of what aspect or variable in the world is relevant and what its supposed value is. Also for robot *learning* the relevance of goals as a scaffolding mechanism in high dimensions has recently been shown [4], [5]. Yet, in all of these scenarios the goals are carefully handcrafted: both the variable to be controlled, as well as how the agent's situation designates a particular value of that variable to be the current goal need to be specified by the designer. Several formulations of motor learning can automatically choose internal goals purely for the sake of training a skill (e.g. [6], [5]), but they can neither explain how to choose goals for an actual purpose, nor how to determine the variable that has to be controlled.

Goals are a fundamental concept also in neuroscience and psychology. The entire formulation of the cerebellum providing *internal* forward and inverse *models* [7] only makes sense if goals are already given as input for the inverse models. From a conservative standpoint such models concern motor control in the first place. Recent theories, however, go much further and suppose that they also contribute to cortex-wide higher-level cognitive processes [8] and are engaged in social behavior [9]. Goals are also seen as a major factor in motivation psychology [10], where "goal-setting" is considered essential for long term behavior organization. Goals are considered a likewise structuring element in our cognition and perception of other agents' behavior such as in imitation learning [11] or teleological action understanding [12], [13].

Goals are useful abstractions. But where do they come from? Neither robotics and machine learning, nor neuroscience, nor psychology provide conclusive or even general hints how a biological or artificial agent that starts with no goals can acquire them. For developmental robotics the importance of goal system development was first pointed out by Prince *et al.* in [14] but is unsolved ever since. It may not be difficult to think of heuristics for an agent to acquire goals within isolated special scenarios, but what could be general mechanisms for a development of goal systems? This article seeks for answers to this longstanding question. We thereby focus on (i) the learning of an agent's *own* goals, in contrast to observational learning about others' goals such as in imitation learning [15], [16] or values such as in inverse reinforcement learning [17], and (ii) a fully *autonomous* learning without external supervision such as an agent being told what to do. Our paper makes three contributions: Firstly, we discuss the term "goal" conceptually in Sec. II, distinguish it from other related concepts, and make several propositions to substantiate the terminology. Secondly, we propose a generic computational learning framework based on the previous considerations in Sec. III. Based on either intrinsic or extrinsic rewards we show how "latent" goals can be extracted from the sensory and action information with an online algorithm. Thirdly, we use a simple information seeking criterion based on visual saliency as an *exemplary* reward in Sec. IV. We show that this generic, task-unspecific reward is sufficient to allow our method to extract goals and also a self-detection that cause the emergence of goal-directed reaching for an object. We thereby not only learn those abstractions, but already utilize them by applying goal babbling [18] to generate actions in a fully bootstrapping and closed action-perception-learning loop.

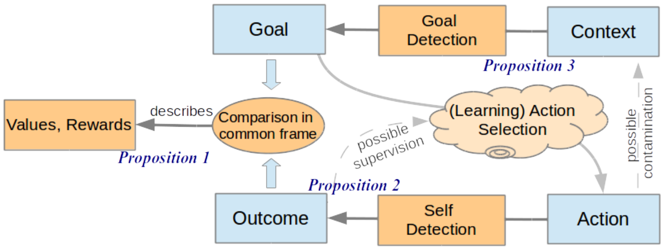


Fig. 2. Proposed conceptual refinement of "goals".

II. What is a goal?

How can we operationalize the acquisition of goals? In order to achieve a general conceptualization we start from dictionary definitions towards usages in various scientific fields, distinguish related terms, and make several propositions how to substantiate the terminology. Dictionaries refer to the term "goal" as "an aim or desired result" of someone's "ambition or effort"[1], Goals are most precisely defined in computational domains that use them. In motor coordination and control [2], [3] goals are typically low-dimensional abstractions of the to be controlled task such as a desired angular velocity of an electric motor or a desired position of a robot's end effector. Thereby goals are formulated as *values* in some *low-dimensional space* (e.g. 1d velocities, 3d effector positions) in which they abstract from many task-irrelevant variables (such as room temperature) of the typically much higher-dimensional physical processes. Similarly, goals in planning [1] describe variables of the world that should have some desired value, while other variables are irrelevant. In both control and planning the goal has to be *achieved* by means of *action*, i.e. the the agent has to find and apply actions that result in the observation or measurement of the desired variable values. Similar aspects can also be found in goal-setting psychology [10]: For instance in management psychology it has been proposed that self-set goals should be specific (have a particular value), measurable, and realistic (i.e. actually achievable by means of own action) [19]. The above points clearly distinguish goals from two other kinds of desires or intentions: (i) *Optimization* or general improvement (such as increasing reward) are not goals in a narrow sense. "Improvement" for itself is not specific in the sense that a particular to be achieved value is specified. Hence, there is also no definite achievement possible or an end defined. (ii) *Wishes* of desired world states (e.g. having a sunny say) are no goals because they are not achievable by means of own action in the first place.

With these aspects we can attempt a first definition that we will refine in the remainder of this section:

> *Definition*: A *goal* is an equivalence set of world states that, in a certain situation, an agent desires to achieve as a result of its own action.

They refer to an equivalence set of states in the sense that there can be irrelevant variables that to not matter for the goal. Hence, any of their values are equivalently acceptable. The main point is that goals reflect a particular desire. Their achievement has some value to the agent. This stands in contrast e.g. to *affordances* [20]. Sahin *et al.* formulated affordances as the relation between the action of an agent, an object under manipulation, and the effect on that object [21]. Objects with similar action-effect relations can then be summarized in equivalence classes such as "standonable". Related to this formulation, *contingencies* [22], [23] and action-effect bindings [24], [25] describe general patterns of manipulability, i.e. relations between actions and their specific effects. Goals and affordances are both interactivist concepts in the sense that neither goals nor affordances can be defined by only the agent or only the environment, but only via their interplay. The crucial difference between them is that affordances describe any *possible* thing that could be done. Affordances are not associated to any value or *desire*, while this desire to do something is the constituting concern of goals in our view.

Intelligent organisms do not arbitrarily invent goals. They must have an developmental origin. The main point of our overall argumentation is therefore the source of information that could lead to the autonomous development of goals. In terms of machine learning we know three basic kinds of learning signals: *supervised* input of ground-truth values, *unsupervised* learning of input statistics, and *reward* or cost signals in reinforcement learning or optimization. Supervised learning as source of information seems entirely unsuited for autonomous development of goals, since a teacher for such information would have to be external. While social learning of goals in such terms certainly exists, it does not provide answers for an ontogenetic core mechanism of goal systems development. Unsupervised learning seems likewise unsuited since simple signal statistics can not tell about a desire or value. Rather, reward signals seem to be the suitable learning signal, as they express the most primitive form of a value. Considering goals as high-level abstractions of intention therefore suggests to consider them *abstractions of* world states that are associated to reward:

> *Proposition 1*: Goals are *abstractions* that do not themselves *determine* a desire, but rather *describe* it based on lower-level systems of desire, such as reward or value systems.

Corresponding rewards might reflect some task very directly when e.g. determined in a social context, or directly as food. However, they might also be purely *internal* or *intrinsic* as it is often considered in the contexts of intrinsic *motivation* [26], [27], [28], [29] or information seeking [30]. Our exemplary experiment in Sec. IV will take the latter perspective. The abstraction process thereby could not only concern immediate rewards, but also expected, future rewards that are expressed in *value systems* (e.g. supposed to exist in the midbrain [31]).

The second aspect to focus on is the *achievement semantics*, which leads to a crucial insight towards a computational

[1] oxforddictionaries.com "goal"; corresponding definitions in other languages: duden.de (German) "Ziel"; nlpwww.nict.go.jp/wn-ja/ (Japanese) Synset 05980875-n; queried 2014/01/15

formalization. When goals are said to be achieved, there needs to be a measurement of that achievement. Goals do not come alone, but always paired with an *evaluation* of the own action's effect. In robot reaching this evaluation, or rather its learning, is often referred to as *self-detection* [32], [33]: the robot's hand needs to be detected for instance in a camera image. A goal and the result of the own action (self-detection) then need to be compared in order to assess the achievement, which holds equally also for planning domains and goal-setting psychology. The need for this comparison forbids considering the development of goals and self-detection separately from each other:

> *Proposition 2*: Goals cannot be learned or considered independently from *self-detection*, but both have to describe a consistent *reference frame* in which the goal can be compared to the outcome of an action.

This aspect will largely guide our computational formalization. Our experiments will also illustrate that this aspect poses an important developmental hallmark: the entrance of *self-supervised* action and motor learning. Once self-detection and goals are available, a supervised learning signal becomes available to other learning processes. When self-detection (e.g. a robot's forward function) and goals were already available they have been used in numerous approaches for motor learning already [3]. With respect to the autonomous development of goals already Prince [14] noted that goals are related to self-supervision, but missed the point that it is not the goals themselves, but rather the self-detection (in relation to the goals) that enables the supervision.

Finally, we need to consider how goals become "active", i.e. how an agent determines which goal to follow at a present moment. It is often considered that agents can have multiple goals, e.g. on different timescales or also parallel or secondary goals in the long run. This leaves a lot of play for an operationalization, which we propose to organize with the following restriction using the notion of cognitive or processing "(sub-)systems" internal to an agent:

> *Proposition 3*: One system can have only *one* (active) goal *at a time*, which expresses the *present* desire based on an internal or sensory *context*.

Hence, different systems of cognitive processing (e.g. such organizing different timescales of behavior) or motor planning and control can have one present goals each. Further we note that a goal gets triggered by a *context*, such as sensory information, an internal state or information from other processing systems. This seems trivial but makes an important point: there needs to be a *goal-detection* to determine the now to be followed goal from the context, and that parallels the self-detection. In the reaching example this might be a mechanism to determine the position of the relevant object from camera images, or also fully hard-wired position selectors in many robot setups. In terms of motivational psychology, for instance a certain context of a conversation might trigger a subsequent communicational goal such as an information to be conveyed ("by the way...").

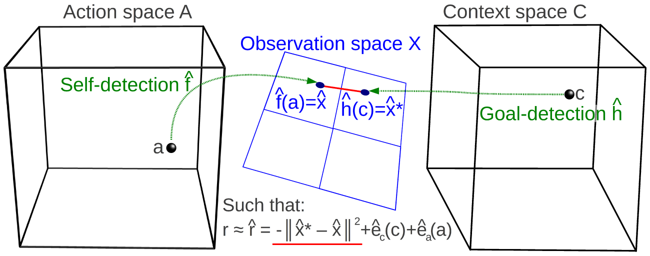


Fig. 3. Proposed learning formulation Latent Goal Analysis

In summary, we refer to a *goal system* as the joint apparatus of goal-detection from a context and self-detection that are compared within a common reference frame, such that the achievement of the goal by means of own action is reflected by a reward or value.

III. Latent Goal Analysis

In order to mathematically operationalize goals based on the previous conceptual considerations, we can now start with an already formalized domain that sets almost all these *conceptual* terms in a relation: *control* or *coordination* problems [2], [3], [4]. The remainder of this section will first show how to set the above terms in a precise relation based on coordination problems, which we transform into a reward-based learning problem. We will then argue that goal- and self-detection can be learned from rewards by computationally inverting the previously constructed transformation. While using control formulations as basis might seem like a treatment of a special case only, we believe in a rather large generality of this approach. This is supported by recent theory findings [34] showing the universal applicability of our approach to any reward function which suggests that it can generalize also beyond *original* motor control problems.

Reward Transformation: Coordination problems such as reaching follow a simple protocol: (1): The world provides a *goal* $\mathbf{x}^*$ (e.g. the target coordinate of an object) to the agent that is situated in some *observation space* $\mathbb{X} \subseteq \mathbb{R}^n$. (2): The agent chooses an *action* $\mathbf{a}$ (e.g. joint angles of its arm) from some *action space* $\mathbb{A} \subseteq \mathbb{R}^m$. (3): The world provides a causal *outcome* $f(\mathbf{a}) = \mathbf{x}$ (e.g. the robot's hand position) of the agent's action, again situated in $\mathbb{X}$ that serves as common reference frame. The agent's task is to choose an action such that the outcome $\mathbf{x}$ matches the goal $\mathbf{x}^*$: $\mathbf{x} = f(\mathbf{a}) = \mathbf{x}^*$. Many coordination problems provide *redundancy*: the action space is substantially higher dimensional than the observation space ($n \ll m$), such that multiple actions $\mathbf{a}_i \neq \mathbf{a}_j$ map to same outcome $f(\mathbf{a}_i) = f(\mathbf{a}_j)$. In such scenarios an additional cost function $-e_a(\mathbf{a})$ can be used to select an optimal action among all those that fulfill $f(\mathbf{a}) = \mathbf{x}^*$. $f : \mathbb{A} \rightarrow \mathbb{X}$ is usually called *forward function*, whereas the problem to identify it is called *self-detection* [33]. Cost terms $e_a(\mathbf{a})$ are often used to prevent a drift of postures or to avoid collisions [35].

This problem can be easily expressed in terms of reward by the negative distance of goal and outcome, and $e_a(\mathbf{a})$:

$$r(\mathbf{x}^*, \mathbf{a}) = -||\mathbf{x}^* - f(\mathbf{a})||^2 + e_a(\mathbf{a}) \qquad (1)$$

The goals in coordination problems are usually not available right away. Rather, they are chosen by vision processes identifying relevant objects to be manipulated, planning, or other processes. Hence, they are in some way determined by a larger internal or external context. We can denote this selection on an abstract level with a function $h(\mathbf{c})$ that we call *goal-detection*. For the sake of symmetry we can finally introduce a virtual cost term $e_c(\mathbf{c})$ that only depends on the context. This term does not influence the optimal action selection but reflects that the optimal reward depends on the context which is later on needed for universally inverting the reward-goal relation. Altogether this gives the reward transformation

$$r(\mathbf{c}, \mathbf{a}) = -||h(\mathbf{c}) - f(\mathbf{a})||^2 + e_c(\mathbf{c}) + e_a(\mathbf{a}) \ . \qquad (2)$$

The overall protocol corresponds to a (continuous) one-step reinforcement problem [36], [37]: (1): The world provides a *context* $\mathbf{c}$ in some context space $\mathbb{C} \subseteq \mathbb{R}^p$. (2): The agent chooses an *action* $\mathbf{a}$ from the *action space* $\mathbb{A} \subseteq \mathbb{R}^m$. (3): The world provides a *reward* $r \in \mathbb{R}$ based on latent goals and action outcomes as shown in equation 2.

Latent Goal Transformation: We now have a complete formal relation between goal- and self-detection, and rewards. Obviously, *any* coordination problem can be treated this way and can be transformed into a reward-based problem. Now suppose an agent is confronted to *any* reward function, specified either externally or internally. Is it possible to make the inverse transformation from *those* rewards back to self- and goal-detection? Suppose an agent perceives contexts $\mathbf{c}$, performs actions $\mathbf{a}$ and estimates the received rewards $r(\mathbf{c}, \mathbf{a})$ (for estimated *future* rewards this would be denoted by a value function $Q(\mathbf{c}, \mathbf{a})$). In order to perform the inverse transformation we need to find functions $\hat{f}$, $\hat{h}$, $\hat{e}_c$ and $\hat{e}_a$ to resemble the original reward or value function $r(\mathbf{c}, \mathbf{a})$:

$$r(\mathbf{c}, \mathbf{a}) = \hat{r}(\mathbf{c}, \mathbf{a}) = -||\hat{h}(\mathbf{c}) - \hat{f}(\mathbf{a})||^2 + \hat{e}_c(\mathbf{c}) + \hat{e}_a(\mathbf{a}) \qquad (3)$$

The major task is to identify the self-detection $\hat{f}(\mathbf{a}) = \hat{\mathbf{x}} \in \mathbb{R}^n$ and the goal-detection $\hat{h}(\mathbf{c}) = \hat{\mathbf{x}}^* \in \mathbb{R}^n$. These functions express the interaction of goals and outcomes in a n-dimensional (to be identified) observation space (see Fig. 3). Additional cost terms depending only on context *or* action are considered as remainders, and in fact are easy to find given $\hat{f}$ and $\hat{h}$.

We refer to the task of identifying this transformation as *Latent Goal Analysis* (LGA), because goals (as well as outcomes) are assumed to exist as latent variables of the reward function. Recent theory results [34] prove the universal existence of this transformation. It is indeed possible to transform *any* reward function into the form of Eqn. 3 for a large enough dimension n. Hence, any intrinsic or extrinsic reward can also be expressed in terms of goals and action-outcomes, which just have to be identified. The dimension n (like in other dimension reduction schemes) can be used to select the few most significant dimensions to express the reward as good as possible. The only complication is that the transformation is not unique [34]. Firstly, the axes of the observation space (i.e. the outputs of $\hat{h}$ and $\hat{f}$) can be arbitrarily rotated, shifted, and mirrored, because none of these operations changes the distances $||\hat{h}(\mathbf{c}) - \hat{f}(\mathbf{a})||$. This reflects that there is no "ground-truth" orientation of *internal* reference frames as long the relation to the outside world is consistent (i.e. *both* $\hat{h}$ and $\hat{f}$ are equally turned). Secondly, it is possible to shift reward "mass" between the terms $||\hat{h} - \hat{f}||$, $\hat{e}_c$, and $\hat{e}_a$, which does not effect the choice of the observation space as a whole, but can change the location of goals and outcomes relative to each other within that space [34]. This, unfortunately very unintuitive, problem can however be easily resolved by requiring that the term $||\hat{h} - \hat{f}||$ should have the most significant contribution to (i.e. the goals should explain the largest portion of) the reward function. This can be implemented by keeping $\hat{e}_c$ and $\hat{e}_a$ as small as possible.

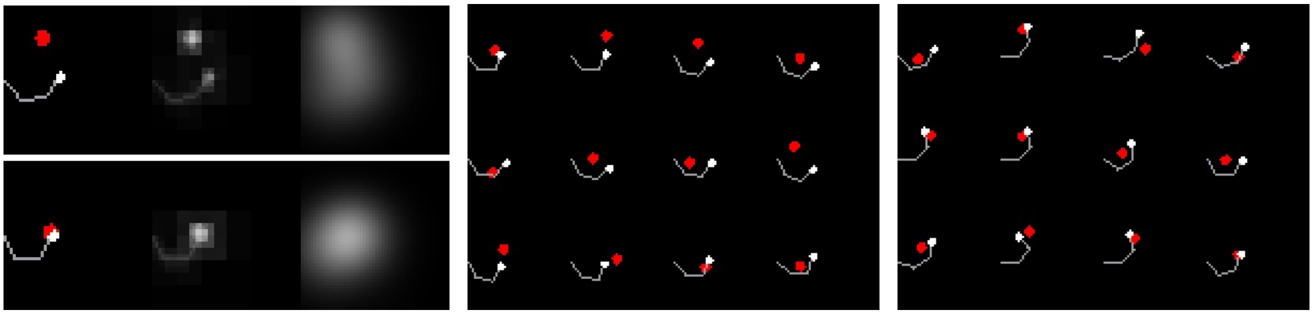
Learning Algorithm: In the following we introduce an online gradient descent algorithm to estimate the above mentioned functions. Suppose an agent observes samples along a time line t. The agent perceives some context $\mathbf{c}_t$, executes some action $\mathbf{a}_t$, and receives a reward $r_t = r(\mathbf{c}_t, \mathbf{a}_t)$ based on some hidden reward function $r(\mathbf{c}, \mathbf{a})$. The agent is supposed to learn the functions $\hat{h}$, $\hat{f}$, $\hat{e}_c$, and $\hat{e}_a$ such that the observed reward r_t is explained by them according to Eqn. 3. This can be done by reducing the *reward-prediction error*:

$$E_t(r_t, \mathbf{c}_t, \mathbf{a}_t) = ||e_t(r_t, \mathbf{c}_t, \mathbf{a}_t)||^2 = ||r_t - \hat{r}(\mathbf{c}_t, \mathbf{a}_t)||^2 \ . \quad (4)$$

We denote the learnable parameters of $\hat{h}$, $\hat{f}$, $\hat{e}_c$, and $\hat{e}_a$ as θ_h, θ_f, θ_c and θ_a respectively. For an initial symmetry-breaking (due to the invariance of internal rotation and translation) it is necessary to initialize θ_h and θ_f with small random values. From this point on, simple gradient descent on E can succeed to estimate the functions. However, we need to further consider that the values of $\hat{e}_c$ and $\hat{e}_a$ have to be kept small. For this purpose we use a simple decay term similar to weight-decay often used in neural networks: In each timestep their values are not only adapted by the error-reduction signal, but also a decay of some $\epsilon \in [0; 1)$ portion of their own value. Since any reward mass from $\hat{e}_c$ and $\hat{e}_a$ that decays needs to be explained by $||\hat{h} - \hat{f}||$ instead, we *add* (with reversed sign) the decay values to the learning signals of $\hat{h}$ and $\hat{f}$. The resulting gradient rule with learning rates η_d and η_c can be written as:

$$\begin{aligned}
\Delta\theta_f &= +\eta_d \cdot \big(e_t + \epsilon \cdot [\hat{e}_a(\mathbf{a}_t) + \hat{e}_c(\mathbf{c}_t)]\big) \cdot (\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_t) \cdot \frac{\partial \hat{f}(\mathbf{a}_t)}{\partial \theta_f} \\
\Delta\theta_h &= -\eta_d \cdot \big(e_t + \epsilon \cdot [\hat{e}_a(\mathbf{a}_t) + \hat{e}_c(\mathbf{c}_t)]\big) \cdot (\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_t) \cdot \frac{\partial \hat{h}(\mathbf{c}_t)}{\partial \theta_h} \\
\Delta\theta_a &= \eta_c \cdot (e_t - \epsilon \cdot \hat{e}_a(\mathbf{a}_t)) \cdot \frac{\partial \hat{e}_a(\mathbf{a}_t)}{\partial \theta_a} \\
\Delta\theta_c &= \eta_c \cdot (e_t - \epsilon \cdot \hat{e}_c(\mathbf{c}_t)) \cdot \frac{\partial \hat{e}_c(\mathbf{c}_t)}{\partial \theta_c} \ ,
\end{aligned}$$

in which the last term in each formula is depending on the (and known for any) function approximation method. If the decay term is disabled ($\epsilon = 0$), these formulas correspond to



(a) Arm image, saliency, smoothed saliency (b) Beginning of learning: no coordination (c) Goal-directed reaching emerges

Fig. 4. As an example we consider a robot arm with an object in sight. Visual saliency serves as reward mechanism to learn self- and goal-detection. Goal babbling learns from those abstractions which leads to goal-directed reaching.

ordinary gradient descent on E.

The decay term balances the contribution of all terms such that goal- and self-detection take the dominating role in $\hat{r}$. The term $-||\hat{h} - \hat{f}||^2$ can, however, not model arbitrary reward functions alone. In particular, this negative distance can only account for numerically negative rewards. Modeling numerically positive rewards requires the terms $\hat{e}_c$ and $\hat{e}_a$ to shift the entire estimate $\hat{r}$ by a constant. If the decay term is used, however, this process can never fully reach the necessary shift. In order to still permit a reasonable learning signal for $\hat{h}$ and $\hat{f}$, we introduce a new and purely scalar term k into the reward estimation. This term is not effected by the decay, but can shift the reward estimate such that $-||\hat{h} - \hat{f}||^2$ can be used to model the shape of the reward function:

$$\begin{aligned}\hat{r}(\mathbf{c}, \mathbf{a}) &= -||\hat{h}(\mathbf{c}) - \hat{f}(\mathbf{a})||^2 + \hat{e}_c(\mathbf{c}) + \hat{e}_a(\mathbf{a}) + k \\ \Delta k &= \eta_c \cdot e_t\end{aligned}$$

Of all terms involved we will for now only use $\hat{h}$ and $\hat{f}$, whereas $\hat{e}_a$ could potentially be used to select cost-optimal actions for the same goal. The term $\hat{e}_c$ (and k) is not directly useful, but needs to accompany the estimation when approximating any possible reward function.

IV. Example: From Saliency to Reaching

The conceptual discussion and mathematical operationalization in the last to sections aimed at a general understanding of goal system development. This section introduces a concrete example of a generic (i.e. not task-specific) reward leading to meaningful goal- and self-detection by means of the proposed method. We simulate a simple robot arm with an object in sight. We consider visual saliency as a reward to implement information seeking behavior [30]. We show that our method thereby develops a detection of the object as goal, and a self-detection of the own hand. These abstractions are thereby already *utilized* by means of goal babbling [4] in a closed loop, which results in the emergence of goal-directed reaching. Saliency measures have already been shown to permit a self-detection of the own end-effector [38], simply because looking at the own hand is "interesting". Here we extend this finding by considering an object at the same time. It turns out that more interesting than looking at the hand *or* the object is to look at *both* closely together (compare Fig. 4(a) top and bottom), which exactly rewards goal-directed reaching behavior.

Setup: The basic scenario is shown in Fig. 4(a). We consider a simple robot arm with three joints (segment length $1/3$ each), such that *actions* are the joint angles $\mathbf{a}_t \in \mathbb{R}^3$. We refer to the effector's actual position (that is at no time explicitly known as such to the learner) in cartesian coordinates as $x_t \in \mathbb{R}^2$. A salient object is placed somewhere in the scene at coordinates $o_t \in \mathbb{R}^2$. Arm and object together are rendered into a 48x48 pixel image. Generically we could think of this very image as *context* in terms of visual perception. However, considering raw 2300 dimensional visual input for learning is neither computationally feasible nor very biologically plausible. For this first experiment we assume a certain extent of image processing that has already identified the object and hand coordinates as keypoints in this image. We compose the context for learning out of these basic coordinates plus additional noise dimensions to challenge learning. At every timestep t the agent is assumed to be still in position x_{t-1}, with the object at position o_t. With that we construct the learner's context as $\mathbf{c}_t = (o_t; x_{t-1}; \varepsilon) \in \mathbb{R}^6$ with gaussian noise $\varepsilon \in \mathbb{R}^2, \varepsilon_i \sim \mathcal{N}(0.5, 0)$. For the to be estimated functions $\hat{h}$, $\hat{f}$, $\hat{e}_c$, and $\hat{e}_a$ we use a locally-linear learning formulation identical to [18] with receptive field radii 2.0, 0.25, 0.5, and 0.5 respectively. As a design choice we selected $n = 2$ components to be extracted from the 6 dimensional context and 3 dimensional action.

Reward: The reward r_t provided to the learner is computed using a simple saliency model based on a difference-of-gaussians procedure (simplified from Itti's saliency model [39]). After the agent has received the context c_t (containing an image of the old action a_{t-1}) and selected a new action a_t, we compute a reward based on the "after-action" image containing the object position and the new action a_t. We first compute the original arm image (see Fig. 4(a), left). Then, we compute a "pyramid of gaussians": The image is smoothed with a 5x5 gaussian kernel and scaled down by a factor of 2. This procedure is repeated 4 times. The saliency map (Fig. 4(a), middle) is computed out of these 5 images (original &

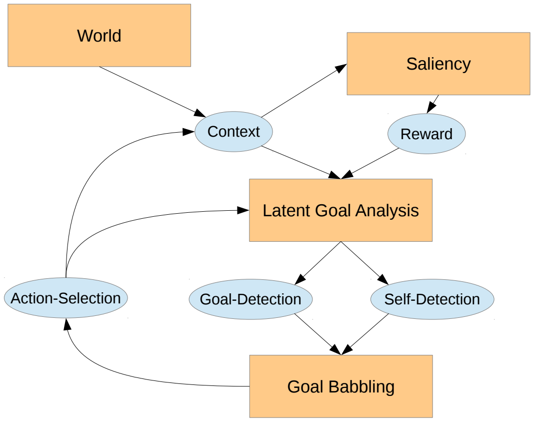


Fig. 5. Schematic organization of the experiment

4x smoothed) by taking the the difference between any two of them, and adding up the amplitudes of those differences. Now considering the most salient point would mean to look at the most interesting pixel. However, we assume that the agent might not just attend to a single visual receptor but rather a *region* in the visual scene. Therefore we smooth the saliency map on a large scale (Fig. 4(a), right) with a gaussian filter with $\sigma = 10$ pixels width which models the total width of the agent's effective visual field. The *highest value* of this *smoothed saliency* encodes the average saliency around that point, and hence a measure of how much information the agent can have in its visual field. We manually normalized the scale of these these values such that they approximately lie in a range $[0;1]$ and consider these the rewards r_t.

Procedure: We conducted this experiment with 5 independent trials with $t = 10^6$ samples each. During a continuous movement of the object in the visual scene we thereby performed continuous online learning of the LGA with learning rates $\eta_d = 0.015$, $\eta_c = 0.005$, and $\epsilon = 0.05$. While the entire procedure is possible in an online fashion only, we decided to perform an additional consolidation phase to speed up learning. Therefore the generation of new samples is interrupted every epoch of 1.000 samples, and the last 10.000 samples are presented in random ordering 10 more times.

Latent Goal Analysis describes how an agent can learn internal abstractions in terms of goals and self-detection. It does *not* instantaneously describe a strategy to select actions. However, we can use the goal- and self-detection to perform self-supervised motor learning: the executed actions $\mathbf{a}_t$ and estimated outcomes $\hat{\mathbf{x}}_t = \hat{f}(\mathbf{a}_t)$ allow to generate a supervised learning signal for common methods of motor learning [3], that can be used together with the estimated goal $\hat{\mathbf{x}}^*_t = \hat{h}(\mathbf{c}_t)$ to perform goal-directed motor-control. Here we utilize a previous algorithm for *goal babbling* [18], that also directly utilizes the goals in order to scaffold learning. This algorithm learns an *inverse model* $g \colon \hat{\mathbf{x}}^* \mapsto \mathbf{a}$ from the self-generated examples $(\hat{\mathbf{x}}_t, \mathbf{a}_t)$. In each timestep the action is selected by trying to accomplish the goal by means of the inverse model plus some exploratory noise N_t:

$$\mathbf{a}_t \quad = \quad g(\hat{\mathbf{x}}^*_t) + N_t(\hat{\mathbf{x}}^*_t). \tag{5}$$

For this we use a learning rate 0.02, local model distances 0.15 and exploratory noise with amplitude 0.15 (see [18]).

The entire organization of the experiment is shown in Fig. 5. The world provides an object that gets encoded in the context together with the last action performed by the agent. The agent's saliency system generates an information seeking reward for the combination of context and action. Latent Goal Analysis extracts the reward-relevant information from the action (self-detection) and disentangles the goal from other information in the context in order to explain the reward by the relation between goal and self. The self-supervised information from the self-detection is then used together with the estimated goals by means of goal babbling in a closed loop.

Results: During the learning we ran an evaluation of every 1.000 samples between two consolidation steps. We investigated three questions:

- What does the self-detection encode?
- What does the goal-detection encode?
- What behavior results from that abstractions?

In order to investigate the representations we checked how well the *internal* representations of outcomes $\hat{\mathbf{x}}$ and goals $\hat{\mathbf{x}}^*$ describe values of the *actual* effector position $\mathbf{x}$ and object position o. Even if the internal variables encode them perfectly there can be arbitrary shifts and translations in the internal coordinate system. Therefore we computed the best linear fit L from internal representations to actual variables. We assessed the quality of the encoding by the normalized root-mean-square error (NRMSE) $\sqrt{E\left[||L_x(\hat{\mathbf{x}}) - \mathbf{x}||^2\right]}/\sqrt{Var\left[\mathbf{x}\right]}$ (correspondingly for $\hat{\mathbf{x}}^*$ and o). If this error is 0, the value of the actual variable can be perfectly (linearly) predicted from the internal one: the internal representation encodes the actual variable. A value of 1.0 means that the prediction gives an error in the range of the variable's variance, which indicates that the internal variable does not encode the actual one at all.

Results for the self-detection are shown in Fig. 6. If LGA should actually learn a representation of the robot's own hand just from saliency-based rewards, this would require a strongly non-linear multi-dimensional mapping. Results show that already in the very beginning there seems to be a certain extent of encoding with errors significantly below 1.0. However, this results merely from the low versatility of actions in the beginning. Goal babbling initially chooses actions close to a single posture since it is not sufficiently trained yet. The outcomes of such locally distributed postures can to a limited extent be predicted with the randomly initialized self-detection. After approx. 50 epochs the values stabilize around 0.2 which means that 80% of the actual effector-position's variance can be explained by the internal representations. At later stages there is a minimal increase of the error values which is because goal babbling learns to use more and more different and wide-spread postures. Hence, the population gets less local and harder to describe due to non-linearities. Latent Goal Analysis after all succeeds to learn the robot arm's forward function from joint angles to effector position by just using the saliency reward. We additionally investigated the encoding by checking what different coordinate axes in the learned representation encode. The blue line in Fig. 6 shows the prediction of the effector's top/down coordinate from just the highest variance principle component of the internal representation. Low errors indicate that this axis indeed encodes top/down movements.

If LGA should learn a goal representation that describes goal-directed reaching, then the extracted goals $\hat{x}^*$ should encode the object position o in relation to the own hand. This could seem simple because the object location is already directly encoded in the context $\mathbf{c}$. However, this variable still needs to (i) be identified as the relevant one among other entries (noise and previous hand-position) in the context, and (ii) set into the right relation (i.e. orientation, shift, scaling) to the self-detection. In particular at later stages of learning the own hand-position strongly correlates with the object position due to goal-directed reaching, so that keeping track of the right variable is far from trivial. Results in Fig. 7 show that at the time of initialization the object position is not at all encoded in the goal-detection. Then, the strongest principal component of estimated goals $\hat{x}^*$ quickly coincides with the top-down axis of the object position (blue lines). Few epochs later, the goals' 2D values (red lines) do indeed largely encode the actual object position with errors around 0.05.

Results so far show that the robot's hand position and the object position are indeed found as abstraction in the process of self- and goal-detection. In order to check how well they fit together (i.e. whether they are in the right geometric relation to each other inside the observation space), we can now check the behavior generated by goal babbling as a result of both abstractions. In order to perform an analysis that excludes exploratory noise (to check the representations themselves) we evaluate the combination of goal-detection $\hat{h}$ and inverse model g (learned by goal babbling as inverse of the self-detection $\hat{f}$). The function $g(\hat{h}(\mathbf{c}))$ suggests actions $\mathbf{a}$ for any context $\mathbf{c}$. Hence we can check those actions and see whether they correspond to a reaching action towards the object position encoded in $\mathbf{c}$. We counted for the contexts within one epoch how often the actions led to a contact of either hand and object, or the whole arm and the object (based on the geometries and sizes in Fig. 4(a)). Results in Fig. 8 show that learning rapidly seeks for contacts of arm and object first, and shortly later establishes a 100% contact rate of the robot's hand and the object. Latent Goal Analysis together with goal babbling indeed produces representations as well as inverse models that correspond to goal directed reaching. Remarkably, all of this is bootstrapped from a task-unspecific reward based on visual saliency as a sole original learning signal. Abstractions bootstrapped by LGA are then used as self-supervised learning signal for goal babbling.

V. Discussion

The autonomous development of goals is a fundamental issue in developmental robotics. This paper has proposed

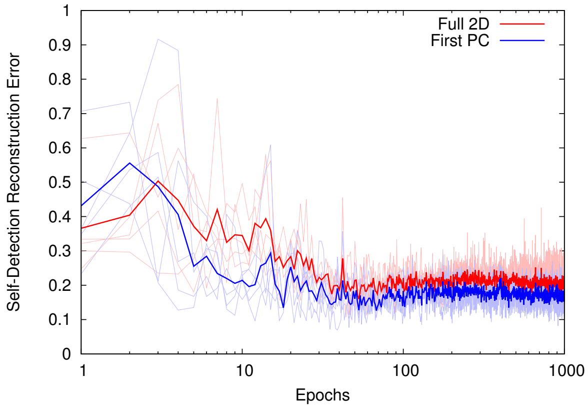


Fig. 6. The learned non-linear self-detection encodes the actual effector position in 2D. The first principal component in the internal coordinate system encodes the effector's top-down coordinate.

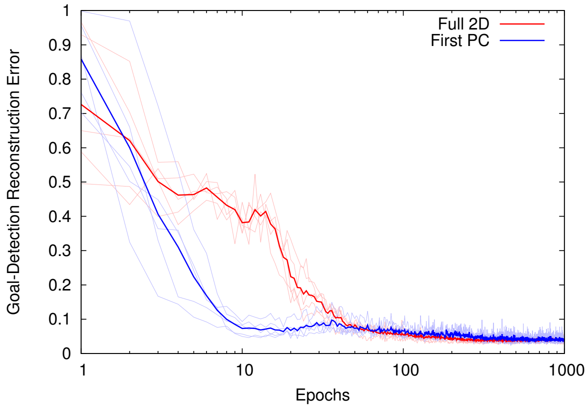


Fig. 7. The learned goal-detection encodes the world's object position in 2D. The first principal component in the internal coordinate system encodes the object's top-down coordinate.

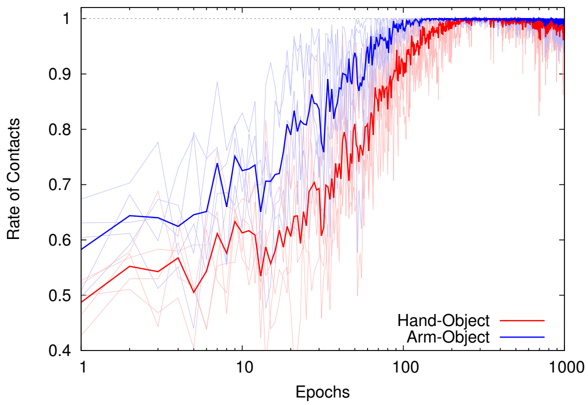


Fig. 8. Goal babbling uses the learned self- and goal-detection to learn an inverse model that consistently makes contact with the object.

a detailed conceptual framework and a mathematical operationalization for agents to learn goals themselves. We thereby emphasized the need to consider and learn goals alongside with self-detection of the own actions' outcomes. Both can then be compared in a common space. We suggest that goals and outcomes together can be learned by considering them as latent variables (i.e. abstractions) that can explain an observed or expected future intrinsic or external reward. Hence, rewards come first and are followed by goals as abstractions of them. This initially leaves a tension with notions of rewards as a result of goal-achievement [10], [28]. However, both might be true, such that rewards and goals are in a circular relation that leads to initially purposeless behaviors such as play.

We have experimentally shown that considering mere visual saliency as a generic, task-unspecific information seeking reward to be processed by our framework leads to abstractions of self and goals, that ultimately lead to goal-directed reaching behavior. In doing so we have not only shown what those abstractions encode, but have already capitalized on them for self-supervised motor learning and goal babbling.

We aimed at a general discussion and formalization of goals alongside with an example. This obviously can not yet answer all questions how our concept could apply to other cases. However, we think that our work does indeed widely open the door for such further investigations for instance about social learning scenarios [13] or also other measures of intrinsic motivation [5], [40].

Acknowledgments

This study has been supported by the JSPS Grant-in-Aid for Specially promoted Research (No. 24000012). This research is related to the European Project CODEFROR (PIRSES-2013-612555).

References

[1] J. Bruce and M. Veloso, "Real-time randomized path planning for robot navigation," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, vol. 3. IEEE, 2002, pp. 2383–2388.

[2] W. Chung, L.-C. Fu, and S.-H. Hsu, "Chapter 6: Motion control," in *Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Springer New York, 2007, pp. 133–160.

[3] D. Nguyen-Tuong and J. Peters, "Model learning for robot control: a survey," *Cognitive Processing*, vol. 12, no. 4, 2011.

[4] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 3, 2010.

[5] A. Baranes and P.-Y. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, vol. 61, no. 1, pp. 49–73, 2013.

[6] M. Rolf, "Goal babbling with unknown ranges: A direction-sampling approach," in *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2013.

[7] D. M. Wolpert, R. C. Miall, and M. Kawato, "Internal models in the cerebellum," *Trends Cog. Science*, vol. 2, no. 9, pp. 338–347, 1998.

[8] M. Ito, "Control of mental activities by internal models in the cerebellum," *Nature Reviews Neuroscience*, vol. 9, pp. 304–313, April 2008.

[9] D. M. Wolpert, K. Doya, and M. Kawato, "A unifying computational framework for motor control and social interaction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1431, pp. 593–602, 2003.

[10] E. A. Locke and G. P. Latham, "Goal setting theory," *Motivation: Theory and research*, pp. 13–29, 1994.

[11] H. Bekkering, A. Wohlschlager, and M. Gattis, "Imitation of gestures in children is goal-directed," *The Quarterly Journal of Experimental Psychology: Section A*, vol. 53, no. 1, pp. 153–164, 2000.

[12] G. Gergely and G. Csibra, "Teleological reasoning in infancy: the naive theory of rational action," *Trends Cog. Science*, vol. 7, no. 7, pp. 287–292, 2003.

[13] B. Wrede, K. Rohlfing, J. Steil, S. Wrede, P.-Y. Oudeyer, and J. Tani, "Towards robots with teleological action and language understanding," in *Humanoids 2012 Workshop on Developmental Robotics: Can developmental robotics yield human-like cognitive abilities?*, 2012.

[14] C. Prince, N. Helder, and G. Hollich, "Ongoing emergence: A core concept in epigenetic robotics," in *Int. Conf. Epigenetic Robotics*, 2005.

[15] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[16] M. Muhlig, M. Gienger, J. J. Steil, and C. Goerick, "Automatic selection of task spaces for imitation learning," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2009, pp. 4996–5002.

[17] A. Y. Ng, S. J. Russell *et al.*, "Algorithms for inverse reinforcement learning," in *ICML*, 2000, pp. 663–670.

[18] M. Rolf, J. J. Steil, and M. Gienger, "Online goal babbling for rapid bootstrapping of inverse models in high dimensions," in *IEEE Int. Conf. Development and Learning and Epigenetic Robotics*, 2011.

[19] G. T. Doran, "There's a S.M.A.R.T. way to write management's goals and objectives," *Management Review*, vol. 70, no. 11, pp. 35–36, 1981.

[20] J. J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing*, R. Shaw and J. Bransford, Eds., 1977, pp. 67–82.

[21] E. Sahin, M. Cakmak, M. Dogar, E. Ugur, and G. Ucoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adaptive Behavior*, vol. 15, no. 4, 2007.

[22] J. K. O'Regan *et al.*, "What it is like to see: A sensorimotor theory of perceptual experience," *Synthese*, vol. 129, no. 1, pp. 79–103, 2001.

[23] Y. Nagai, A. Nakatani, S. Qin, H. Fukuyama, M. Myowa-Yamakoshi, and M. Asada, "Co-development of information transfer within and between infant and caregiver," in *IEEE Int. Conf. Development and Learning and Epigenetic Robotics (ICDL)*, 2012, pp. 1–6.

[24] B. Hommel, "Action control according to TEC (theory of event coding)," *Psychological Research PRPF*, vol. 73, no. 4, pp. 512–526, 2009.

[25] S. Verschoor, M. Weidema, S. Biro, and B. Hommel, "Where do action goals come from? evidence for spontaneous action-effect binding in infants," *Frontiers in Psychology*, vol. 1, no. 201, pp. 1–6, 2009.

[26] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.

[27] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990-2010)," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[28] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[29] G. Baldassarre, "What are intrinsic motivations? a biological perspective," in *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2011.

[30] J. Gottlieb, "Attention, learning, and the value of information," *Neuron*, vol. 76, no. 2, pp. 281–295, 2012.

[31] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, 1997.

[32] A. Edsinger and C. C. Kemp, "What can i control? a framework for robot self-discovery," in *Int. Conf. Epigenetic Robotics*, 2006.

[33] A. Stoytchev, "Self-detection in robots: a method based on detecting temporal contingencies," *Robotica*, vol. 29, pp. 1–21, 2011.

[34] M. Rolf and M. Asada, "Where do goals come from? A generic approach to autonomous goal-system development," 2014, (submitted).

[35] J. Baillieul, "Avoiding obstacles and resolving kinematic redundancy," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 1986.

[36] A. L. Strehl, "Associative reinforcement learning," in *Encyclopedia of Machine Learning*. Springer, 2010, pp. 49–51.

[37] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *NIPS*, 2008.

[38] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," in *IEEE Int. Conf. Development and Learning (ICDL)*, 2008.

[39] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[40] K. Merrick, "A comparative study of value systems for self-motivated exploration and learning by robots," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 2, pp. 119–131, 2010.