An Internal Model Principle For Robots

Vadim K. Weinstein, Tamara Alshammari, Kalle G. Timperi, Mehdi Bennis, Steven M. LaValle

University of Oulu
{Vadim.Weinstein, Tamara.Alshammari, Kalle.Timperi, Mehdi.Bennis, Steven.LaValle}@oulu.fi

Abstract. When designing a robot's internal system, one often makes assumptions about the structure of the intended environment of the robot. One may even assign meaning to various internal components of the robot in terms of expected environmental correlates. In this paper we want to make the distinction between robot's internal and external worlds clear-cut. Can the robot learn about its environment, relying only on internally available information, including the sensor data? Are there mathematical conditions on the internal robot system which can be internally verified and make the robot's internal system mirror the structure of the environment? We prove that sufficiency is such a mathematical principle, and mathematically describe the emergence of the robot's internal structure isomorphic or bisimulation equivalent to that of the environment. A connection to the free-energy principle is established, when sufficiency is interpreted as a limit case of surprise minimization. As such, we show that surprise minimization leads to having an internal model isomorphic to the environment. This also parallels the Good Regulator Principle which states that controlling a system sufficiently well means having a model of it. Unlike the mentioned theories, ours is discrete, and non-probabilistic.

Keywords: Internal Model Principle · Free Energy Principle · Semantics · Good Regulator Principle · Surprise Minimization · Sufficiency · Information Transition Systems

1 Introduction

From an algorithmic perspective, we are familiar with internal robot models used for motion planning, navigation, and so on, represented with coordinates in $\mathbb{R}^n$ and geometric primitives. We have also witnessed the rapid rise of deep learning approaches to robotics, which encode internal models as parameters in a neural net, tuned from copious amounts of input-output data. Such systems still require extensive training with vast amounts of supervisory data from human experts and numerous reinforcement learning trials in virtual environments [2]. Despite this, the derived robots remain unable to function effectively in unfamiliar environments [15], unlike humans who can adapt to new situations with ease. According to [18], this discrepancy may be due to the ability of humans to

develop internal models of how the world works. Currently, these internal models are often pre-defined for training environments, making it difficult for robots to operate in different settings. Therefore, it is crucial to understand how robots can independently construct internal models that correlate with their external environments in a useful way without relying on prior assumptions.

To address this challenge, we explore fundamental questions: What does it mean for a robot to explore its environment? How should we conceptualize it? What does it mean for a robot to have a useful internal structure in relationship to its external environment? These questions are fundamental to advancing robotics as they address the core challenges of enabling robots to develop a deeper understanding and adaptability to their surroundings.

1.1 Connection to other disciplines and related work

The relationship between the structure of an internal model and that of the external environment has been studied in the context of control theory. In [5] it was shown that an effective regulator of a given system needs to contain an isomorphic copy of the system itself. In [6] this statement is formulated in the context of linear, time-invariant, finite-dimensional, multi-variable control systems. The incorporation of an internal model was formalized in terms of the divisibility of the invariant factors of the matrix describing the (linear) dynamics of the compensator by the minimal polynomial of a matrix representing the dynamics of the exogenous signals affecting the plant. It was concluded that the synthesis of a plant and a compensator (controller) is structurally stable only if the compensator incorporates an internal model of the plant dynamics.

Meanwhile, for biophysics and cognitive systems, the field of *active inference* [22] seeks to interpret the notions of agent, mind, and brain through the general framework of the free-energy principle [9,23]. Its stated aim is to explain behavior of agents through the over-arching objective of surprise minimization via action. A recent work [11] illustrates how collective behavior can emerge from the simultaneous updating of each agent's internal variables, guided by the minimization of surprise (indirectly via the minimization of free energy). However, the internal variables have pre-assigned relationships with the environment (local distance between the agent and its neighbors) governed by pre-assigned dynamics.

In the context of discrete-time stochastic dynamical systems, the problem of discovering economical, predictive representations based solely on action-observation dynamics has been formalized in the theory of *predictive state representations (PSR)* [19,20]. Based on the earlier, deterministic framework of *diversity-based inference* [24], PSR suggests a way to model such systems through maintaining a vector of probabilities for the likelihood of observing certain finite-length action-observation sequences, called core tests. Maintaining a relatively low-dimensional vector of such probabilities turns out to be sufficient for effectively modelling the statistical behaviour of the unknown dynamics.

From the perspective of classical logic, one typically deals with a language L (such as first-order logic) and L-model M with domain $\mathrm{dom}(M)$. A crucial

aspect of the model involves functions that map elements of $\mathrm{dom}(M)$ to constant symbols in L, subsets of $\mathrm{dom}(M)^n$ to n-ary relation symbols in L, and functions from $\mathrm{dom}(M)^n$ to $\mathrm{dom}(M)$ to n-ary functions in L. This mapping is the semantic function and assigns meaning to elements of the language L, relating them to elements of the world. This models, in particular, our understanding of natural language. For example, the word "truck" represents an actual object in the world, and as humans, we inherently understand this connection. Regardless of whether we can explain the origin or the "reality" of this connection, we can still formalize it using semantic functions. By formalizing these connections, we can better understand and analyze the structure of languages and theories. Consequently, these semantic functions allow us to introduce the following informal definition: *A logical theory or a theory of agency is representationalist if and only if it relies on semantic functions.* In this sense, all logical theories – from classical first-order and second-order logic to intuitionistic logic and dependence logic – are representationalist.

When the internal structure (of a human or robot "brain") is postulated to bear content or represent external states, the theory doing so is called *representational*. For example, in robotics, the internal state is often understood as serving as a *representation*[1] of its external environment. However, the mapping, the *semantic function*, which assigns meanings to this representation to relate it to elements of the external environment is often overlooked. Theoreticians or designers frequently assume that the environment is structured in a certain way and that sensor data reflects that structure in some pre-defined manner [14,27]. This assumption helps them to infuse the symbols in the robot's internal structure with meanings. Nonetheless, such an assumption limits the applicability of many robotic solutions to more realistic scenarios, such as search and rescue tasks, where the environment is previously *unknown*. Navigating an *unknown* environment presents significant challenges. A high degree of complexity and autonomy is required for the robot to make decisions based on the data it observes [1]. To design a robot capable of autonomous learning, we should separate the meaning assigned by the designers to the labels in the robot's brain from the meaning the *robot itself* assigns to them. For the robot to be truly flexible and able to generalize to unseen environments, it should be capable of learning meanings based on its own histories of interaction.

Various semantic mappings were explicitly analysed in recent work in the context of perception engineering and robotics by a subset of the authors [17]. Understanding the nature of semantic functions is essential for developing robots capable of autonomous learning and flexible adaptation. In this paper, we make the following philosophical proposition: *a semantic function is inherently arbitrary*. By this, we mean that the semantic function does not arise from structure in some non-arbitrary or systematic way. As an example of structurally non-arbitrary assignment is the assignment of the fundamental group to a topological space. Quoting the linguist Ferdinand de Saussure, *"the bond between the*

[1] That is, a correspondence between the properties or elements of the internal model with those of the external environment is assumed either implicitly or explicitly.

signifier and the signified is arbitrary" [4,26]. In Saussure's theory, the "signifier" is the sound-image (or the form of a word), and the "signified" is the concept (or the meaning associated with that word). The quote describes the relationship between the signifier and the signified as an arbitrary relation, meaning that there is no inherent connection between the form of a word and its meaning; rather, this relationship is established by convention within a language community.

Enactivism, a branch of (philosophy of) cognitive science, argues that postulating representations and semantic content is often philosophically questionable and practically useless [10,12,13,28]. This argument could be extended to robotics in the light of the above discussion. Enactivism argues that the complex structure of the brain is there not to represent the world around us, but rather to make our engagement with it more efficient. A counterargument is that building elaborate representations could actually help with efficient engagement. There are two counter-counterarguments. First is that elaborate represenations might be useful, but they might also be unnecessary. Is it possible to circumvent elaborate representations and go directly to efficient engagement? The second is that when some internal configuration has structural similarities to the environment, it does not mean it is a representation. In our context specifically we see (Corollary 32) that such internal structures emerge even when it is not the intention or the goal of the system to generate them. For more on the second counter-counterargument from a philosophical perspective, see [12,13].

1.2 Contribution

We develop a mathematical framework which enables robots to learn an internal model of the external environment by relying solely on quantities which can be internally evaluated. This can be seen as *emergent structural semantics*. The robot only engages in resolving internal "conflicts" such as minimization of surprise.

We mathematically prove that resolving these internal conflicts invariably results in the emergence of an internal structure that is structurally similar to the environment. The main statement capturing this intuition is Theorem 27.

One application even shows that sometimes only proprioceptive data is enough for the emergence of an internal structure which is isomorphic to the external state space, even though no other sensor data is available, see Remark 2 and Example 2. An example of "embodied cognition".

In this regard, our work can be viewed as a first step towards a discrete, combinatorial, non-probabilistic formulation of the free-energy principle (FEP) [7,8,9,22]. The FEP has gained significant attention as a model of the brain and "living systems" in general. It has two principal components. One is the analysis of the independence of a "living system" from the rest of the world, formalized through Markov blankets. Another component is the idea of minimization of surprise by the system and the emergent internal structure which results from it. This aspect of the FEP is the one we "replicate" in this paper. Unlike most available accounts on the FEP, our framework is not probabilistic,

but rather combinatorial and in line with the previous work on history information spaces [25,30]. Our results are also limit-cases in which the surprise in not just minimized, but actually brought to zero. The result can also be seen as a combinatorial and deterministic version of the Good Regulator Theorem presented in [5].

2 Surpriseless couplings and bisimulation

The main tenet of the FEP that we focus on is surprise minimization. The notion of sufficiency is the quintessential notion of surprise minimization in the framework of information transition systems [16,25,30]. By definition, sufficiency is a condition posed on an equivalence relation over a dynamical system which states that each equivalence "predicts" the next equivalence class. This allows to transfer the notion of surprise minimization from the probabilistic context (where it is measured by entropy) to the context of countable discrete dynamical systems and transition systems. In this section we introduce basic definitions and the first result stating a connection between surprise minimization and equivalence (in this case bisimulation equivalence) of the internal state space with the external state space.

We will fix U and Y to be the sets of motor commands and sensor data, respectively.

Definition 1 A *deterministic transition system* (DTS) is a triple $\mathcal{S} = (S, A, \tau)$ where $\tau\colon S \times A \to S$ is the transition function. A *labeled* DTS is (S, A, τ, h) where (S, A, τ) is DTS and h is some labeling function. A (labeled) *sensorimotor transition system* is a DTS with $A = U \times Y$ and a possible labeling function h with range in Y. A *(labeled) internal system* is a (resp. labeled) sensorimotor transtion system with $S = I$ the *internal state space* and $\tau = \varphi$ the *information transition function*. The *external system*, or *environment* is a labeled DTS with $S = X$, $A = U$, $\tau = f$ and the range of h being Y. Both external and internal systems are special cases of sensorimotor transition systems. ⊣

One may ask why is the input to the internal system $U \times Y$, and not just Y. Typically we expect the behaviour of a robot to depend on the history of sensor data, but not on its own motor commands which are determined by the sensor data anyway. We do this for our framework to be general enough. For example, in a situation where the robot's policy has not been determined yet, all possible sequences of motor commands are possible. Or perhaps the DTS is a model of only a part of the robot, a part which does not (yet) decide the motor commands and has to be prepared for all possibilities. An internal system is *exploratory*, if it is prepared to "try" all possible motor commands indpendently of the sensor data. Mathematically, $(I, U \times Y, \varphi)$ is exploratory, if the value of $\varphi(\iota, u, y)$ is not dependent on y, i.e. for all $y, y' \in Y$ we have $\varphi(\iota, u, y) = \varphi(\iota, u, y')$. For exploratory internal systems we may drop the component Y from the definition. This leads to:

Definition 2 An *exploratory internal system* is a triple (I, U, φ) in which I is the internal state space, and $\varphi\colon I \times U \to I$. ⊣

Since we are interested in constructing internal systems and not in finding policies, we will deal mostly with exploratory internal systems. We justified the title "exploratory" by the intuition that the agent is "curious" and tries all possible paths notwithstanding the sensor data. Our theorems about emergence of bisimulation equivalence and isomorphism will justify this title formally because they work precisely for such systems. Another reason to introduce the class of exploratory systems is to simplify some of the mathematical exposition.

Definition 3 Given a set A, the *free DTS* generated by A, denoted $\mathcal{F}(A)$, is the DTS $(A^{<\mathbb{N}}, A, \frown)$ where the state space $A^{<\mathbb{N}}$ is the set of finite sequences of elements of A, and the state-transition function is the concatenation of sequences. When $A = U$, this DTS is the collection of all possible actuator sequences. If $A = U \times Y$, then it is the collection of all actuation-observation, or sensorimotor, sequences and is equivalent to the history information space $\mathcal{I}_{hist}$ of [25], see also [16]. ⊣

Definition 4 [Universal systems] The system $\mathcal{F}(U \times Y)$, the free DTS generated by the set $U \times Y$ is the *universal information transition system* (with respect to U and Y). Every connected internal transition system is a quotient of $\mathcal{F}(U \times Y)$, see Corollary 22. The *universal exploratory system* is $\mathcal{F}(U)$. Every connected exploratory internal system is its quotient. ⊣

Definition 5 For a DTS (S, A, τ), $s \in S$, and $\bar{a} \in A^{<\mathbb{N}}$, define $s * \bar{a}$ by induction on the length of $\bar{a}$. If $\bar{a} = \varnothing$, then $s * \bar{a} = s$, and if $s * \bar{a}$ is defined and $a \in A$, then $s * (\bar{a}^\frown a) = \tau(s * \bar{a}, a)$. ⊣

Definition 6 Given an internal system $\mathcal{I} = (I, U \times Y, \varphi)$ and an external system $\mathcal{X} = (X, U, f, h)$, define the coupled system $\mathcal{X} \star \mathcal{I}$ to be the DTS $(X \times I, U, g)$ where $g\colon X \times I \times U \to X \times I$ is defined by

$$g(x, \iota, u) = (f(x, u), \varphi(\iota, u, h(x))). \tag{1}$$

For exploratory systems, if the component Y is not written, the equation (1) becomes $g(x, \iota, u) = (f(x, u), \varphi(\iota, u))$. ⊣

In view of Definition 6, U and Y constitute the "interface" between the internal and external.

Definition 7 Given a finite sequence $\bar{u} = (u_1, \ldots, u_n) \in U^n$ and states $x \in X$, $\iota \in I$, there are unique sequences $\bar{x} = (x_1, \ldots, x_n)$ and $\bar{\iota} = (\iota_1, \ldots, \iota_n)$ such that $x_1 = x$, $\iota_1 = \iota$ and for each $k < n$, $g(x_k, \iota_k, u_k) = (x_{k+1}, \iota_{k+1})$. Denote this $\bar{x}$ by $x * \bar{u}$ and $\bar{\iota}$ by $\iota \diamond \bar{u}$.

Note that the definition of $x * \bar{u}$ here coincides with Definition 5. This is why we use the same notation for it. But the operation $\diamond$ on I requires the coupling to be defined. If the external system and/or the initial states need to be specified, we denote $\iota \diamond \bar{u} = \iota \diamond_{\mathcal{X}} \bar{u} = \iota \diamond_{\mathcal{X},x,\iota} \bar{u}$ ⊣

Remark 1. If we view $U^{<\mathbb{N}}$ as the free monoid under concatenation, then $x \mapsto x * \bar{u}$ and $\iota \mapsto \iota \diamond \bar{u}$ are actions of this monoid on X and I, respectively. ⊣

The coupling restricts the internal system by removing the impossible sensorimotor sequences. More precisely:

Definition 8 Suppose $\mathcal{I} = (I, U \times Y, \varphi)$ is an internal system and $X = (X, U, f, h)$ external, and let $(x_0, \iota_0) \in X \times I$. Let $\mathcal{I} \restriction_{\mathcal{X}, x_0, \iota_0}$ be the DTS (I, U, ψ) where $\psi(\iota, u) = \iota \diamond_{\mathcal{X}, x_0, \iota_0} u$. If the initial states are clear from the context, we drop them from the subscript. ⊣

Definition 9 [Surpriseless] Suppose $\mathcal{I} = (I, U \times Y, \varphi)$ and $\mathcal{X} = (X, U, f, h)$ are internal and external systems, respectively. We say that $(x_0, \iota_0) \in X \times I$ is *surpriseless*, if there are no sequences $\bar{u} = (u_1, \ldots, u_n) \in U^n$ and $\bar{u}' = (u'_1, \ldots, u'_m) \in U^m$ such that $\iota_0 \diamond \bar{u} = \iota_0 \diamond \bar{u}'$, but $h(x_0 * \bar{u}) \neq h(x_0 * \bar{u}')$. ⊣

A surprise occurs if the sensory data cannot be "predicted" based on the robot's internal state. Note that the condition of being surpriseless can be evaluated within the internal model because the definition speaks only about internal states and sensory data.

Definition 10 [Bisimulation] Suppose $\mathcal{I} = (I, U \times Y, \varphi, h')$ is a labeled internal system. Suppose $\mathcal{X} = (X, U, f, h)$ is an external system. We say that $x_0 \in X$ is *bisimulation equivalent to* $\iota_0 \in I$ *in* $\mathcal{X} \star \mathcal{I}$, if there is a relation $R \subseteq X \times I$ such that (B1.) $(x_0, \iota_0) \in R$, (B2.) for all $(x, \iota) \in R$ and all $u \in U$, $(x * u, \iota \diamond u) \in R$, and (B3.) for all $(x, \iota) \in R$ we have $h(x) = h'(\iota)$. ⊣

At first glance, bisimulation equivalence cannot be evaluated within the internal system because by definition it requires finding a binary relation between the internal and the external. In Theorem 13 below we prove, however, that it is equivalent to being surpriseless.

Definition 11 A DTS (S, A, τ) is *strongly connected*, if for all $s, s' \in S$ there is $\bar{a} \in A$ with $s * \bar{a} = s'$. ⊣

Lemma 12 *Suppose the coupled system $\mathcal{X} \star \mathcal{I}$ is strongly connected. If $(x_0, \iota_0) \in X \times I$ is surpriseless, then there is a well-defined $\hat{h} \colon I \to Y$ such that $\hat{h}(\iota) = h(x_0 * \bar{u})$ for some (all) $\bar{u} \in U^{<\mathbb{N}}$ with $\iota_0 * \bar{u} = \iota$.*

Proof. It is enough to show that for all $\bar{u}, \bar{u}' \in U^{<\mathbb{N}}$, if $\iota_0 \diamond \bar{u} = \iota_0 \diamond \bar{u}'$, then $h(x * \bar{u}) = h(x * \bar{u}')$, but this is exactly the definition of surpriseless. □

Theorem 13 *Suppose the coupled system $\mathcal{X} \star \mathcal{I}$ is strongly connected. The following are equivalent*

1. *(x_0, ι_0) is surpriseless in $\mathcal{X} \star \mathcal{I}$*
2. *$\hat{h}$ is well-defined*
3. *$\hat{h}$ is well-defined and x_0 is bisimulation equivalent to ι_0 in $\mathcal{X} \star \mathcal{I}$ over $\hat{h}$.*

Proof. We already proved $1 \to 2$ (Lemma 12). Assume 2. Let

$$R = \{(x, \iota) \in X \times I \mid \exists \bar{u} \in U^{<\mathbb{N}}((x, \iota) = (x_0 * \bar{u}, \iota \diamond \bar{u}))\}.$$

Conditions (B1) and (B2) are clearly satisfied. If $(x, \iota) \in R$, then $h(x) = h(x_0 * \bar{u}) = \hat{h}(\iota)$ by definition of $\hat{h}$ which proves also condition (B3). This proves $2 \to 3$. Clearly $3 \to 2$. Suppose that 2 and we will prove 1. Suppose $\bar{u}, \bar{u}'$ are such that $\iota_0 \diamond \bar{u} = \iota_0 \diamond \bar{u}'$. Since $\hat{h}$ is well-defined, we have $h(x_0 * \bar{u}) = \hat{h}(\iota_0 \diamond \bar{u}) = \hat{h}(\iota_0 \diamond \bar{u}') = h(x_0 * \bar{u}')$. □

Theorem 13 shows that if the agent had a method to reduce surprise, then once successful, its internal system would be bisimulation equivalent to the environment. This result is reminiscent of the Good Regulator [5]. Surpriselessness implies the theoretical possibility to control the system completely. In a surpriseless system, the internal state determines the results of all possible future actions (see also Lemma 26). On the other hand we just proved that such potential to control the system implies bisimulation equivalence. Bisimulation equivalence, however, is significantly weaker than isomorphism. Also Theorem 13 is very non-constructive, it does not say whether or when surpriseless systems exist, or how to obtain them. These questions will be addressed in the next section.

3 Theory for General Deterministic Transition Systems

In the previous section we showed that if an internal state space is adapted to the environment with the goal of minimizing surprise, then it will become bisimulation equivalent to it. What are the conditions under which surprise minimization leads to the internal system being isomorphic, and not just bisimilar, to the external system? A bisimulation equivalence between transition systems is an isomorphism, if both systems are irreducible. The robot cannot know whether the environment is irreducible, but it can "know" when its internal system is. The analysis and definition of irreducibility goes beyond the present paper, but we mentioned it for the sake of discussion. It is cumbersome to speak about the function $\hat{h}$ of Lemma 12 because it is not always well-defined. Also, Theorem 13 does not shed light on how to systematically obtain surpriseless internal systems when nothing is known about the environment or when do they even exist. We will now present a framework which enables a more systematic characterization of when the internal system is both surpriseless and isomorphic to the external system, and how to obtain them, theoretically, but not algorithmically.[2] This approach will pave the path to algorithm design in future work, see Section 6.

Definition 14 Given a DTS $\mathcal{S} = (S, A, \tau)$. We say that an equivalence relation E on S is *sufficient*, or $\mathcal{S}$-sufficient, if for all $s, s' \in S$ and $a \in A$ we have that $(s, s') \in E$ implies $(\tau(s, a), \tau(x', a)) \in E$. An equivalence relation E' is a

[2] Our result shows how to find such systems by computing minimal sufficient refinements on infinite trees. As such this is not computationally feasible, but gives a theoretical doorway towards designing algorithms.

refinement of an equivalence relation E, if $E' \subseteq E$. We say that an equivalence relation E' is a *minimal sufficient refinement* of E, if E' is sufficient, $E' \subseteq E$ and for all sufficient refinements $E'' \subseteq E$ we have $E'' \subseteq E'$. ⊣

The following is Theorem 4.19 of [30].

Theorem 15 *Suppose $\mathcal{S} = (S, A, \tau)$ is a DTS. Let E be an equivalence relation on S. Then the minimal $\mathcal{S}$-sufficient refinement of E exists and is unique. We denote it by* $\mathrm{MSR}(E) = \mathrm{MSR}_{\mathcal{S}}(E)$. □

Definition 16 Let $\mathcal{S}_i = (S_i, A, \tau_i)$ be a DTS for $i \in \{0, 1\}$. A function $h\colon S_0 \to S_1$ is called a *homomorphism from $\mathcal{S}_0$ to $\mathcal{S}_1$*, if for all $(s, a) \in S_0 \times A$ we have

$$\tau_1(h(s), a) = h(\tau_0(s, a)).$$

It is called an *epimorphism*, if it is onto, and *isomorphism* if it is a bijection. ⊣

Definition 17 Let E be an equivalence relation on S. If $s \in S$, denote by $[s]$, or by $[s]_E$, the equivalence class of s which is the set $\{s' \in S \mid (s, s') \in E\}$. The set of all equivalence classes is denoted by S/E. Suppose that $h\colon S \to Y$ is a function and E is an equivalence relation on Y. Define

$$h^{-1}(E) = \{(s, s') \in S \times S \mid (h(s), h(s')) \in E\}.$$

It is easy to verify that $h^{-1}(E)$ is an equivalence relation on S. Denote by E_h the equivalence relation $h^{-1}(\mathrm{id}_Y)$. Equivalently, $(s, s') \in E_h \iff h(s) = h(s')$. A function $h\colon S \to Y$ is *E-invariant*, if E is a refinement of E_h. We say that an equivalence relation E is *h-closed*, if E_h is a refinement of E. ⊣

We leave the following two observations for the reader to prove:

Lemma 18 *Let $\mathcal{S} = (S, A, \tau)$ be DTS and $h\colon S \to Y$. Then E_h is h-closed, and h is E_h-invariant.* □

Lemma 19 *Suppose $\mathcal{S} = (S, A, \tau)$ is a DTS, E is an $\mathcal{S}$-sufficient equivalence relation, and $h\colon S \to Y$. If E is h-closed, then $(h/E)\colon \mathcal{S}/E \to Y$ defined by $(h/E)([s]) = h(s)$ is well-defined.* □

Lemma 20 *(Lemma 4.5 of [30]) Suppose $\mathcal{S} = (S, A, \tau)$ is a DTS, and E is an $\mathcal{S}$-sufficient equivalence relation. Let $S' = S/E$ and define $\tau'\colon S' \times A \to S'$ by $\tau'([s], a) = [\tau(s, a)]$ (also denoted $\tau' = \tau/E$). Then τ' is well-defined and $\mathcal{S}' = (S', A, \tau')$ is a DTS (also denoted by $\mathcal{S}' = \mathcal{S}/E$).* □

Proposition 21 (Pullback) *Suppose $\mathcal{S}_i = (S_i, A, \tau_i)$ is a DTS for $i \in \{0, 1\}$. Suppose that E_1 is an $\mathcal{S}_1$-sufficient equivalence relation on S_1. Suppose that $h\colon S_0 \to S_1$ is an epimorphism. Then $E_0 = h^{-1}(E_1)$ is an $\mathcal{S}_0$-sufficient equivalence relation on S_0 and $\mathcal{S}_0/E_0 \cong \mathcal{S}_1/E_1$.*

Proof. See Appendix. □

Corollary 22 (Universality) *Let $\mathcal{F}(A)$ be as in Definition 3. The free DTS of is* universal *in the sense that if $\mathcal{S} = (S, A, \tau)$ is any strongly connected DTS, then there is an equivalence relation E on $A^{<\mathbb{N}}$ such that $\mathcal{F}(A)/E \cong \mathcal{S}$.*

Proof. Fix $s_0 \in S$. Define the map $h\colon A^{<\mathbb{N}} \to S$ by induction on the length of $\bar{a}$ as follows. Let $h(\varnothing) = s_0$, and if $h(\bar{a})$ is defined, let $h(\bar{a}^\frown a) = \tau(h(\bar{a}), a)$. Let $E_1 = \mathrm{id}_S$. Then it is clearly $\mathcal{S}$-sufficient. Let $E_0 = h^{-1}(E_1)$. By strong connectedness, h is an epimorphism. Now the result follows from Proposition 21. □

Corollary 22 means that finding a DTS is equivalent to finding an equivalence relation on $\mathcal{F}(A)$. It follows that finding an appropriate internal system for a robot's brain can be reformulated as finding such an equivalence relation. The condition of sufficiency is essentially the same as surprise minimization: by definition of sufficiency the equivalence class of a state predicts the equivalence class of the following state. This condition can be evaluated within the internal system, but we will show analogously to Theorem 13 that if the robot can achieve sufficiency, then it achieves equivalence. In fact, minimal sufficiency will imply isomorphism between the internal and external systems under some conditions.

Definition 23 Suppose $H \subseteq S \times S$ is any set. Then $\langle H \rangle$ is the *equivalence relation generated by H*. It is defined to be the set of all pairs $(s_0, s_1) \in S \times S$ such that there exists a finite sequence $(x_0, \ldots, x_k)$ of elements of S such that $s_0 = x_0$, $s_1 = x_k$, and for all $i \in \{0, \ldots, k-1\}$ we have $(x_i, x_{i+1}) \in H$ or $(x_{i+1}, x_i) \in H$. ⊣

Lemma 24 *Suppose $\mathcal{S}_i = (S_i, A, \tau_i)$ is DTS for $i \in \{0, 1\}$, $E, E^0, E^1, \ldots, E^{m-1}$ are equivalence relations on S_0, E_1 an equivalence relation on S_1, and $h\colon S_0 \to S_1$ a function. Then the following hold:*

1. *Each E^i is a refinement of $\langle E^0 \cup \cdots \cup E^{m-1} \rangle$,*
2. *If E^i are $\mathcal{S}_0$-sufficient for all $i < m$, then so is $\langle E^0 \cup \cdots \cup E^{m-1} \rangle$,*
3. *If E^i is a refinement of E for all $i < m$, then $\langle E^0 \cup \cdots \cup E^{m-1} \rangle$ is also a refinement of E.*
4. *If E is a refinement of E^i for all $i < m$, then E is a refinement of $\langle E^0 \cup \cdots \cup E^{m-1} \rangle$.*
5. *If E^i is h-closed for all $i < m$, then so is $\langle E^0 \cup \cdots \cup E^{m-1} \rangle$.*
6. *E_h is a refinement of $h^{-1}(E_1)$.*
7. *If h is onto, and E is h-closed, then*

$$h(E) = \{(h(s), h(s')) \in S_1 \times S_1 \mid (s, s') \in E\}$$

   is an equivalence relation on S_1.
8. *If h is onto, E is h-closed, and E is a refinement of $h^{-1}(E_1)$, then $h(E)$ is a refinement of E_1.*
9. *If h is an epimorphism, E is $\mathcal{S}_0$-sufficient and h-closed, then $h(E)$ is $\mathcal{S}_1$-sufficient.*

10. If h is a homomorphism and E_1 is $\mathcal{S}_1$-sufficient, then $h^{-1}(E_1)$ is $\mathcal{S}_0$-sufficient,

11. If h is a homomorphism, then E_h is $\mathcal{S}_0$-sufficient.

Proof. See Appendix. □

The following theorem is the key to understanding why internal processing is "enough". It shows that the operator MSR commutes with any epimorphism. One can compute the minimal sufficient refinement of a relation and then take the inverse image of the result. Or, one can first take the inverse image, and then compute the minimal sufficient refinement of that. According to Theorem 25 both give the same result. This is valuable for us, because the inverse image of the relation (before applying MSR) corresponds to what the robot can sense. Then, the robot can internally apply MSR, and can be sure to get the same as if MSR was applied in the external system first. We explore the applications of Theorem 25 to the robot system in Section 4.

Theorem 25 *Let $\mathcal{S}_i = (S_i, A, \tau_i)$ be DTS for $i \in \{0,1\}$, and suppose that $h\colon S_0 \to S_1$ is an epimorphism. Suppose E_1 is an equivalence relation on S_1. Then*

$$\mathrm{MSR}_{\mathcal{S}_0}(h^{-1}(E_1)) = h^{-1}(\mathrm{MSR}_{\mathcal{S}_1}(E_1)).$$

Proof. Denote $E_0 = h^{-1}(\mathrm{MSR}_{\mathcal{S}_1}(E_1))$. By Proposition 21 the relation E_0 is sufficient. It is also a refinement of $h^{-1}(E_1)$ because $\mathrm{MSR}_{\mathcal{S}_1}(E_1) \subseteq E_1$ implies $h^{-1}(\mathrm{MSR}_{\mathcal{S}_1}(E_1)) \subseteq h^{-1}(E_1)$. By Theorem 15 it is now enough to show that E_0 is a minimal sufficient refinement of $h^{-1}(E_1)$. Suppose for a contradiction that there exists an $\mathcal{S}_0$-sufficient refinement E_0' of $h^{-1}(E_1)$ such that $E_0' \not\subseteq E_0$. Let $E_1' = h(\langle E_0 \cup E_0' \cup E_h \rangle)$. By Lemma 24(2) and 24(11) the relation $\langle E_0 \cup E_0' \cup E_h \rangle$ is $\mathcal{S}_0$-sufficient, and by Lemma 24(1) and 24(4) it is h-invariant. So by 24(9) E_1' is $\mathcal{S}_1$-sufficient. Since $E_0' \not\subseteq E_0$, we have by 24(1) that

$$\langle E_0 \cup E_0' \cup E_h \rangle \not\subseteq E_0$$

and so

$$E_1' = h(\langle E_0 \cup E_0' \cup E_h \rangle) \not\subseteq h(E_0) = \mathrm{MSR}_{\mathcal{S}_1}(E_1). \tag{2}$$

On the other hand E_0 and E_0' are refinements of $h^{-1}(E_1)$, so by 24(3) and 24(6) $\langle E_0 \cup E_0' \cup E_h \rangle$ is a refinement of $h^{-1}(E_1)$. By 24(8) we now have that E_1' is a refinement of E_1. Combining this with (2) above, and the sufficiency of E_1', we have a contradiction with the minimality of $\mathrm{MSR}_{\mathcal{S}_1}(E_1)$. □

Lemma 26 *If E is a sufficient equivalence relation on $\mathcal{S} = (S, A, \tau)$, then for all $\bar{a} \in A^{<\mathbb{N}}$ and all $s, s' \in S$ we have*

$$(s, s') \in E \quad \Longrightarrow \quad (s_0 * \bar{a}, s' * \bar{a}) \in E.$$

Proof. By induction on the length of $\bar{a}$. □

4 Application to Internal and External Systems

Recall that we conventionally fix U and Y to be the sets of motor commands and sensor data, respectively.

In this section we will work with the universal exploratory system (Definition 2) $\mathcal{I} = \mathcal{F}(U)$ in which the initial state is always the empty sequence. Given an external system $\mathcal{X} = (X, U, f, h)$, and initial state $x_0 \in X$, let $\hat{f}_{x_0} : U^{<\mathbb{N}} \to X$ denote the function $\hat{f}_{x_0}(\bar{u}) = x_0 * \bar{u}$. Note that $\hat{f}_{x_0}$ is a homomorphism from $\mathcal{F}(U)$ to $\mathcal{X}$.

Theorem 27 *Let $\mathcal{X} = (X, U, f, h)$ be a strongly connected external system with the initial state $x_0 \in X$. Suppose that E is any equivalence relation on X. Let $E_{\mathcal{I}} = \mathrm{MSR}_{\mathcal{F}(U)}(\hat{f}_{x_0}^{-1}(E))$, and $E_{\mathcal{X}} = \mathrm{MSR}(E)$. Then*

$$\mathcal{I}/E_{\mathcal{I}} \cong \mathcal{X}/E_{\mathcal{X}}.$$

Proof. Since $\mathcal{X}$ is connected, $\hat{f}_{x_0}$ is an epimorphism. By Theorem 25 we have

$$E_{\mathcal{I})} = \mathrm{MSR}_{\mathcal{F}(U)}(\hat{f}_{x_0}^{-1}(E)) = \hat{f}_{x_0}^{-1}(\mathrm{MSR}_{\mathcal{X}}(E)) = \hat{f}_{x_0}^{-1}(E_{\mathcal{X}}).$$

The result now follows from Proposition 21. □

It is useful to note that $\mathrm{MSR}(E_h)$ can be a measure of symmetry of the environment $(X, U \times Y, f, h)$. A bisimulation is *trivial*, if it is the identity relation. An *autobisimulation* is a bisimulation of a DTS with itself, a relation on $X \times X$. The following is Theorem 4.4 of [30].

Theorem 28 *An environment $\mathcal{X} = (X, U, f, h)$, has a non-trivial autobisimulation if and only if* $\mathrm{MSR}(E_h) \neq \mathrm{id}_X$. □

The notion of autobisimulation is a weak version of automorphism. There is also a one-sided version of this theorem for automorphisms:

Theorem 29 *If there is a non-trivial automorphism of $\mathcal{X}$, then* $\mathrm{MSR}(E_h) \neq \mathrm{id}_X$.

Proof. An automorphism is a special case of an autobisimulation, so the result follows from Theorem 28. □

So we can call the environment *chiral*, if $\mathrm{MSR}(E_h) = \mathrm{id}_X$. The equivalence relation $\hat{f}_{x_0}^{-1}(E_h)$ on $\mathcal{F}(U)$ equates those paths which lead to identical sensor data. So we may call it the relation of *sensory indistinguishability*. The following corollary shows that if the environment is chiral, then taking the quotient of the history information space by the minimal sufficient refinement of the sensory indistinguishability yields a structure isomorphic to the environment.

Corollary 30 *Suppose (X, U, f, h) is a strongly connected environment, and $x_0 \in X$. Suppose that* $\mathrm{MSR}_{\mathcal{X}}(E_h) = \mathrm{id}_X$. *Then*

$$\mathcal{F}(U)/\mathrm{MSR}(\hat{f}_{x_0}^{-1}(E_h)) \cong \mathcal{X}.$$

Proof. Since $\mathcal{X}/\mathrm{id}_{\mathcal{X}} \cong \mathcal{X}$, the result follows from Theorem 27. □

Taking the minimal sufficient refinement can be interpreted as minimizing surprise while also minimizing computational resources. The minimization of surprise is due to the definition of sufficiency which say that the current equivalence class predicts the next one. Minimization of resources is captured by the minimality of the refinement, as it yields the smallest possible quotient space.

An interesting class of chiral environments consists of those environments in which E_h has one equivalence class which is a singleton. Call such an equivalence relation *pointed*. We will show that environments with pointed E_h are chiral under some very minimal assumptions on $\mathcal{X}$. An automaton (S, A, τ) is *minimally distinguishing*, if for all $s_0, s_1, s_2 \in S$ and all $a \in A$ we have that if $\tau(s_1, a) = \tau(s_2, a) = s_0$, then one of the following holds: $s_0 = s_1$, $s_1 = s_2$ or $s_0 = s_2$.

Theorem 31 *Let $\mathcal{S} = (S, A, \tau)$ be a strongly connected and minimally distinguishing. If E is a pointed equivalence relation on X, then $\mathrm{MSR}(S) = \mathrm{id}_X$.*

Proof. See Appendix. □

We say that a sensor mapping $h\colon X \to Y$ is *pointed*, if E_h is a pointed equivalence relation which is equivalent to saying that there is some $y \in Y$ such that $h^{-1}(y)$ is a singleton.

Corollary 32 *Let (X, U, f, h) be a strongly connected minimally distinguishing external system. Let $x_0 \in X$, and suppose that $h\colon X \to Y$ is a pointed sensor mapping. Let $E = \mathrm{MSR}_{\mathcal{F}(U)}(\hat{f}_{x_0}^{-1}(E_h))$. Then*

$$\mathcal{F}(U)/E \cong \mathcal{X}.$$

Proof. By Theorem 31, $\mathrm{MSR}_{\mathcal{X}}(S) = \mathrm{id}_{\mathcal{X}}$. Then apply Corollary 30. □

Remark 2. Suppose the state space consists of positions of the robot's body in the environment. Then it follows that the robot does not need any "external" sensors to successfully mirror the environment internally. It is enough to have proprioceptive feedback to single one specific position of its own body. Striving for sufficiency (surprise minimization) takes care of the rest by Corollary 32. ⊣

Finally, we give an application for a mathematically ideal situation where the motor commands generate a group acting on the external state space:

Corollary 33 *Suppose that $(G, +)$ is a group generated by $U \subseteq G$, and suppose that $\tau\colon X \times G \to X$ is a transitive action of G on X. Consider the environment $\mathcal{X} = (X, U, f, h)$ where $f = \tau \restriction (X \times U)$ and h is pointed. Let $E = \mathrm{MSR}_{\mathcal{F}(A)}(\hat{f}_{x_0}^{-1}(E_h))$. Then*

$$\mathcal{F}(A)/E \cong \mathcal{X}.$$

Proof. By Corollary 32 it is enough to show that $\mathcal{X}$ is connected and minimally distinguishing. Since the action is transitive, $\mathcal{X}$ is connected. Suppose $x_1 * u = x_2 * u = x_0$ for some $x_0, x_1, x_2 \in X$ and $u \in U$. Multiplying by the inverse of u on the right this implies that $x_1 = x_2$. □

5 Examples

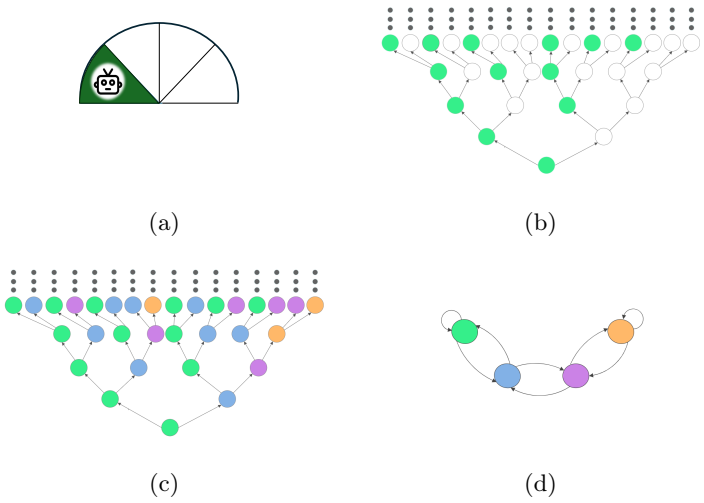

Fig. 1: The environment (a) has four states. Agent can move clockwise and counter-clockwise, but in states 1 and 4 nothing happens, if it tries to go left or right respectively. In (b) the binary tree of action observation sequences is depicted. In most nodes the sensor data is "white" but when the agent is in the left-most state, the data is "green". In (c) we show the minimal sufficient refinement of (b), and in (d) we have taken the quotient of (c) with respect to the refinement. It turns out to be isomorphic to (a), as predicted by Corollary 32.

Example 1. Simple examples of finite external systems which can be "learned" by taking minimal sufficient refinements over the exploratory internal state space are depicted in Figures 1 and 2. It is easy to see that in both cases the external state space is minimally distinguishing and the sensor mapping is pointed, so we can apply Corollary 32. ⊣

In the final example below we show that very limited proprioceptive feedback is enough for the emergent of internal structure isomorphic to the environment.

Example 2. (Robot arm) In Figure 3 there is a robot arm with three joints. For the sake of applicability of our theory, assume that the state space is discrete. Each joint can turn by, say, $\pm 1°$. If the arm is pushing against an obstacle while applying a rotation of any of the joints, then the external state stays the same as before. The configuration space is a subset of the (discrete) 3-torus. Assume that there is one (and possibly only one!) position of the arm where it receives a proprioceptive feedback. It could be the original position of the arm where

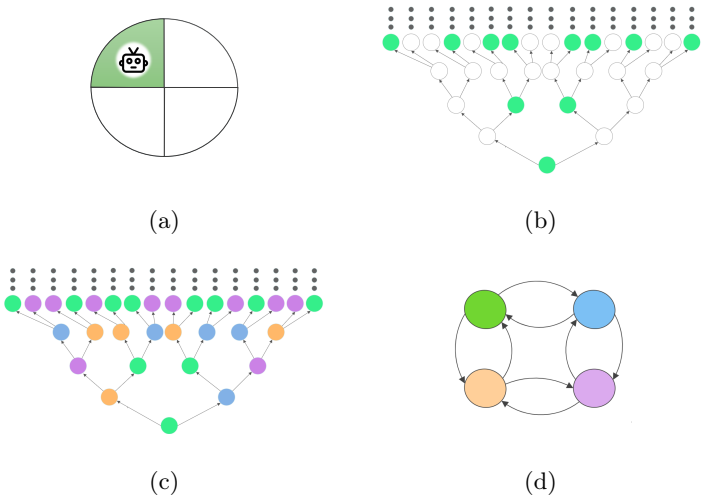

Fig. 2: Same idea, as in Figure 1, but with a circular environment. In this case Corollary 33 could be applied, with the group $\mathbb{Z}/4\mathbb{Z}$ acting on $\mathcal{X}$.

it receives a "click" and no sensor feedback in any other position. Then this sensor mapping is pointed. It is also not hard to see that the robot arm satisfies minimal distinguishability: Suppose the arm is in some positions x_1 or x_2 and u $\pm 1°$ rotation of one of the joints. If this rotation does not result in hitting an obstacle for either x_1 or x_2, then $x_1 \neq x_2$ implies $x_1 * u \neq x_2 * u$ because the rotation u acts "homeomorphically" on the torus. If, however, say, x_1 faces an obstacle when rotating by u, then $x_1 = x_1 * u$, so we are done. Applying Corollary 32 we see that if such robot arm succeeds in minimizing the surprise of "when is the proprioceptive 'click' feedback received", then it will end up building an isomorphic copy of the external state space. ⊣

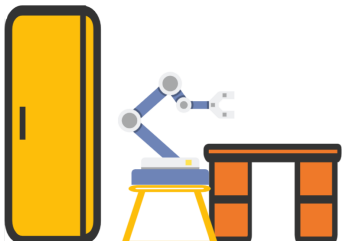

Fig. 3: Robot arm with three joints and obstacles.

6 Discussion

Our aim was to introduce a mathematical principle that the robot can internally evaluate, and if satisfied, ensures that the internal system becomes isomorphic or bisimulation equivalent to the environment. In doing so, we proposed a framework without pre-assigned meanings or assumed correlations between the internal and the external systems. Yet, a correlation emerges between them due to structural coupling. The proposed mathematical principle is sufficiency, which can be interpreted as the lack of surprise. In this way, our framework can be seen as an extension of the FEP framework to the combinatorial realm of finite or countable automata. We propose to explore the ideas of FEP further within this framework. One future direction is to explore the notion of Markov blankets in the context of discrete non-probabilistic systems. After all, the notion of independence is more ubiquitous than that dictated by probability theory [3, Ch.2], [21,29].

One drawback of our results is that both our assumptions and conclusions are very strong. The assumption of sufficiency is unrealistic, and the resulting similarity between the internal and external models is too strong for most applications. Thus, another direction for future research is to explore significantly weaker notions. For example, instead of minimizing surprise globally, the agent may focus on minimizing surprise relative to its goals. We envision replacing minimal sufficient refinements by equivalences that retain only task-relevant information. This approach will guide the theory towards game-theoretic aspects, incorporating von Neumann's concepts of turn-taking and discrete strategies, as well as Nash's payoff functions and Aumann's epistemology works.

We are also working on extending the framework to incorporate multiple sensing modalities within one agent's brain, as well as multiple agents, thereby uncovering new pathways in communication theory.

Acknowledgments. Authors 1, 3, and 5 were supported by a European Research Council Advanced Grant (ERC AdG, ILLUSIVE: Foundations of Perception Engineering, 101020977).

References

1. Antsaklis, P.: Autonomy and metrics of autonomy. Annual Reviews in Control **49**, 15–26 (2020)
2. Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine **34**(6), 26–38 (2017)
3. Baldwin, J.T.: Fundamentals of Stability Theory. Cambridge University Press (Mar 2017)
4. Berger, A.A.: Semiotics and society. Society **51**(1), 22–26 (Feb 2014 2014/02//)
5. Conant, R.C., Ashby, W.R.: Every good regulator of a system must be a model of that system. International Journal of Systems Science **1**(2), 89–97 (1970)
6. Francis, B., Wonham, W.: The internal model principle of control theory. Automatica **12**(5), 457–465 (1976)

7. Friston, K.: A theory of cortical responses. Philosophical Transactions - Royal Society. Biological Sciences **360**(1456), 815–836 (2005)
8. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**, 127–138 (2010)
9. Friston, K.: A free energy principle for biological systems. Entropy **14**(11), 2100–2121 (2012)
10. Gallagher, S.: Enactivist Interventions: Rethinking the Mind. Oxford University Press (2017)
11. Heins, C., Millidge, B., Costa, L.D., Mann, R.P., Friston, K.J., Couzin, I.D.: Collective behavior from surprise minimization. Proceedings of the National Academy of Sciences of the United States of America **121**(17) (2024)
12. Hutto, D.D., Myin, E.: Radicalizing enactivism: Basic minds without content. MIT Press (2012)
13. Hutto, D.D., Myin, E.: Evolving Enactivism. MIT Press (2017)
14. Kantaros, Y., Zavlanos, M.M.: Sampling-based optimal control synthesis for multirobot systems under global temporal tasks. IEEE Transactions on Automatic Control **64**(5), 1916–1931 (2019)
15. Lanillos, P., Meo, C., Pezzato, C., Meera, A.A., Baioumy, M., Ohata, W., Tschantz, A., Millidge, B., Wisse, M., Buckley, C.L., et al.: Active inference in robotics and artificial agents: Survey and challenges. arXiv preprint arXiv:2112.01871 (2021)
16. LaValle, S.M.: Planning Algorithms. Cambridge University Press, Cambridge, U.K. (2006), also available at http://lavalle.pl/planning/
17. LaValle, S.M., Center, E.G., Ojala, T., Pouke, M., Prencipe, N., Sakcak, B., Suomalainen, M., Timperi, K.G., Weinstein, V.: From virtual reality to the emerging discipline of perception engineering. Annual Review of Control, Robotics, and Autonomous Systems **7**(1) (Nov 2023)
18. LeCun, Y., Courant: A path towards autonomous machine intelligence version 0.9.2, 2022-06-27 (2022)
19. Littman, M., Sutton, R.S.: Predictive representations of state. Advances in neural information processing systems **14** (2001)
20. Ma, B., Tang, J., Chen, B., Pan, Y., Zeng, Y.: Tensor optimization with group lasso for multi-agent predictive state representation. Knowledge-based Systems **221**, 106893 (2021)
21. Paolini, G.: Independence logic and abstract independence relations. Mathematical Logic Quarterly **61**(3), 202–216 (May 2015)
22. Parr, T., Pezzulo, G., Friston, K.J.: Active inference: The Free Energy Principle in Mind, Brain and Behaviour (2022)
23. Ramstead, M.J.D., Friston, K.J., Hipólito, I.: Is the free-energy principle a formal theory of semantics? from variational density dynamics to neural and phenotypic representations. Entropy **22**(8), 889 (2020)
24. Rivest, R.L., Schapire, R.E.: Diversity-based inference of finite automata. Journal of the Association for Computing Machinery **41**(3), 555–589 (1994)
25. Sakcak, B., Weinstein, V., LaValle, S.M.: The limits of learning and planning: Minimal sufficient information transition systems. In: 2022 International Workshop on the Algorithmic Foundations of Robotics (WAFR) (2022)
26. de Saussure, F., Baskin, W., Meisel, P., Saussy, H.: Course in General Linguistics. Columbia University Press (2011)
27. Tumova, J., Dimarogonas, D.V.: Multi-agent planning under local ltl specifications and event-based synchronization. Automatica **70**, 239–248 (2016)
28. Varela, F., Rosch, E., Thompson, E.: The Embodied Mind: Cognitive Science and Human Experience. Cambridge, Massachusetts: MIT Press (1992)

29. Väänänen, J.: Dependence Logic: A New Approach to Independence Friendly Logic. London Mathematical Society Student Texts, Cambridge University Press (2007)
30. Weinstein, V., Sakcak, B., LaValle, S.M.: An enactivist-inspired mathematical model of cognition. Frontiers in Neurorobotics **16** (2022)