ARAYA RESEARCH

Mathematical approaches to the study of agents

Manuel Baltieri [1,2,†], Keisuke Suzuki [3]

[1] Araya Inc., Tokyo, Japan
[2] University of Sussex, Brighton, UK
[3] Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN), Hokkaido University, Sapporo, Japan

The definition of life remains one of science's most profound challenges, with contemporary approaches usually focusing on two research programmes: Darwinian evolution and self-maintenance in chemical systems. While evolution has been successfully abstracted and mathematically modelled, the concept of a self-sustaining system has so far resisted a comparable level of formalisation. This paper tackles this challenge by reframing the concept of self-sustaining system within a more abstract framework to study *agents*: goal-directed systems acting in an environment. We build on an existing conceptual framework comprising three requirements for agents: individuality, normativity (or goal-directedness), and interactional asymmetry. We then provide a systematic analysis, under a unified notation, of several mathematical approaches aiming to formalise these requirements, including the free energy principle, integrated information theory and dynamical systems. Unlike this conceptual framework, which commits to an intrinsic perspective on agency, we focus on a less ontologically committed *as-if* stance. Using this, we discuss links between identity and normativity, and a way to understand actions as if they were produced by causal interventions. Taken together, our systematic analysis clarifies the limitations of current proposals and reveals how they can work synergistically within a unified, mathematical account of agency across natural and artificial domains.

Keywords: agency, individuality, normativity, asymmetry

1. Introduction

Artificial Life (ALife) formulates an approach to the study of living systems based on the study of "life as it could be" [1]. ALife can be seen as a form of "comparative biology", that extends research in biology and origins of life to artificial universes, allowing us to study the properties of life, individuality and evolution from a more general, substrate-independent perspective [2]. This line of research is tightly connected to inquiries about Terran life, i.e. life as it is here on Earth, and its origins. To this point, the working definition of life provided by NASA, "a self-sustaining chemical system capable of undergoing Darwinian evolution" [3], elegantly captures the two central pillars that drive modern inquiries into the origins of living systems. The second part of this definition, "[C]apable of undergoing Darwinian evolution", has been the subject of extensive and successful formal abstraction, particularly in evolutionary biology [4], computer science [5] and artificial life [6]. Computational systems such as Avida [7, 8] and Tierra [9] have demonstrated that core principles of evolution including selection, replication, and mutation can be instantiated in silico, yielding complex emergent phenomena that mirror biological evolution without recapitulating the intricacies

†Correspondence e-mail: `manuel_baltieri@araya.org`

of terrestrial biochemistry. The implementation and study of modern, complex artificial worlds like Lenia [10, 11] further show the power of this approach, revealing a universe of digital creatures with their own evolving dynamics (see also [12] for a historical review of this and other related approaches in studies of cellular automata).

On the other hand, the first part, "a self-sustaining chemical system [. . .]", has received comparatively less attention from a formal, mathematical and computational perspective. While we have powerful models describing the dynamics and structure of evolution over time given an entity of interest, i.e. some kind of self-sustaining (chemical) system, we mostly lack a corresponding formal perspective to characterise the spectrum of possible entities of interest. The main goal of this paper is to provide a unifying presentation of different proposals for the study of such entities by focusing on the concept of *agents*, which we informally describe, for now, as *goal-directed systems acting in an environment*.

This is in contrast to other ways to describe, or explain away the existence of agents. In much of physics for instance, the world is explained by dynamical laws that contain no explicit variables that can be attributed to agents or actions. On the other hand, thermodynamic and complex-systems views link agency to irreversibility and information use [13], but risk conflating generic non-equilibrium systems (like a steam engine [14]) with living ones. By contrast, fields like computer science pragmatically assumes the existence of agents: software agents act in digital environments [15], agent-based models explore macro-patterns from micro-rules [16], while AI extends the notion to learning systems, including contemporary "agentic" models and reinforcement learning's agent-environment loop [17, 18]. Control theory likewise presupposes goals and actions, formalising controller-plant-environment loops that steer systems toward targets and thereby bridge physics and AI using pragmatic notions of goal-directedness [19]. Beyond these, philosophy [20, 21, 22, 23], psychology [24, 25, 26, 27], biology [28, 29, 30], sociology [31], economics [32], and law [33] mostly proceed as if agents exist, even when their ontological status is debated, studying sense of agency, organismic organisation, and social or institutional agency. These literatures thus split between traditions that assume agents (natural/social sciences) and those that tend to deny, bracket them or explain them away (much of physics/chemistry).

In this work we adopt a pragmatic stance that seeks to identify the *minimal* requirements for possible definition of agents, similar in spirit to the quest to identify the minimal requirements of cognition, or rather minimally cognitive behaviour [34]. We start by informally addressing agents as goal-directed systems acting in an environment. *Goal-directedness* will provide a vast generalisation and abstraction to the idea of self-sustainment, *system* will address relevant parts of the notion of individuality and include entities beyond chemical systems, *action* will be used to try and differentiate an agent's output from mere physical happenings, and *environment* as a notion complementary to the agent that is useful to frame or carve out an identity for agents as distinct entities.

To operationalise the core components of a mathematical theory of agency, we will then take inspiration from an existing conceptual framework proposed in [35]. We will use, in particular, their three requirements of individuality, normativity (or goal-directedness) and interactional asymmetry to organise and discuss the role of various mathematical proposals found in the literature that are relevant for a definition of agents. Similarly to the work done in artificial life for evolution, we will then focus on formal methods that abstract the concept of agent, for instance, going beyond the particular chemical or physical instantiations of systems subject to Darwinian evolution such as those part of Terran Life. In our analysis we will thus consider different formal proposals through the lenses of a unified conceptual framework, with a unifying mathematical notation, to see how they can be seen as complementing each other. This will provide a more cohesive picture of the differences and complementary roles of these proposals, and in particular will highlight two important points that we will elaborate at the end of the work based on an *as-if* approach to agents previously introduced in [36, 37] and adopted here. Firstly, we will look at a possible *individuality-normativity correspondence* motivated by recent work showing how individuality is deeply linked to goal-directedness. Secondly, we will consider the exact formal meaning currently attributed to actions using interventionist notions [38] of causality (i.e. with the agent seen as a causal individual acting on the environment).

To capture the main parts of a definition of agents discussed here, in Section 2 we review an established conceptual proposal for a theory of agents, based on the ideas of individuality, normativity, and interactional asymmetry. In Sections 3 to 5 we then proceed to classify, under a unified mathematical language, some of the most prominent existing proposals for mathematical theories of agents that capture at least one of the main aspects highlighted in Section 2. Finally, in Section 6 we discuss some of the assumptions in our classification and some ideas raised by our unification attempt, highlighting especially the strong overlap of individuality and normativity, based on a mathematical argument (for now, working on a relatively simple class of systems) stating how identity and goal-directedness are deeply intertwined.

2. A conceptual framework for agents

Our starting point for a conceptual framework of agents is the set of desiderata for a definition of agents in [35]. We introduce them here to show how they have effectively driven, or at least captured, most of the research on mathematical notions of agency discussed in the next few sections. This will allow us to reframe existing proposals in a unifying language and systematically compare different approaches, showing how they need not be mutually exclusive. While most of them will turn out to have implications and consequences beyond our somewhat rigid classification, we believe it is nonetheless important and useful to first and foremost characterise their main goal, background and repercussions.

The three requirements advocated by [35] are:

- *Individuality*, which gives the basic unit of inquiry, the agent, as a system with some spatio-temporal persistent properties, i.e. a notion of identity. The starting point is that an agent must be a distinct entity with its own boundaries (physical, functional, etc.) that distinguishes it from its environment. In other words, given a full system, there is a way to factorise it into an agent and an environment. There could be multiple parallel or nested agents and environments, but the minimum requirement is to have at least one agent and one environment, which can then be further factorised if needed.
- *Normativity*, or goal-directedness, which states that an agent's interactions with the environment ought to be regulated according to specific norms or goals. Goals can correspond to ways in which an agent comes to be, or keeps on persisting, but crucially are not limited to those cases. Importantly, agents can seemingly fail to achieve goals, and still be agents. Arguably, failure is the main characterising feature of agents which puts them in contrast to non-agents, for instance machines of some kind, which on the other hand cannot fail to behave according to some goal.
- *Interactional asymmetry*, stating that an agent must be the active source of its activity, its actions, adapting its coupling with the environment rather than being merely a passive receiver of external forces. Agency involves the system's capacity to adapt the parameters and constraints that govern its interaction with the environment.

Note that in [35] there is a strong emphasis on these requirements being *intrinsic* to the agent. We will not consider this here, and discuss our alternative, as-if stance in Section 6 based on [36, 37]. In the next three sections, we will look at how these three requirements have been explicitly, and at times implicitly, the main driving force for formal frameworks proposing definitions for what constitutes an agent. Some proposals focus on a single requirement, while others attempt to encompass two, or even three at times. We do however believe that all the frameworks overviewed here have a clear, main goal in mind that well aligns with one of these three requirements (however see the individuality-normativity correspondence in Section 6), which will be used to map their contributions and connections to other proposals.

In this work uppercase letters (e.g. X_t, Y_t) are used to denote indexed [1] random variables and lowercase (e.g. x_t, y_t) their realisations. Deterministic variables are included in this convention as degenerate random variables, i.e. X_t such that $\Pr(X_t = x_t) = 1$ for some constant x_t. Calligraphic letters (e.g. $\mathcal{X}, \mathcal{Y}$) denote the sets over which they take values, and the symbol Δ (as in $\Delta(\mathcal{X}), \Delta(\mathcal{Y})$) is used to denote the collection of all distributions over those sets. We use the shorthand notation $p(x_t \mid y_t) = \Pr(X_t = x_t \mid Y_t = y_t)$ to express probabilities when there is no risk of ambiguity, and assume that equalities of the form $p(x_t \mid y_t, z) = p(x_t \mid y_t)$ hold for all realisations that occur with non-zero probability. We use the following abbreviations: $X_{a:b} = (X_a, \ldots, X_b)$, $X_{:b} = X_{0:b}$, $X_{a:} = X_{a:\infty}$, and when considered as a whole $X_{0:\infty}$, we usually drop the index altogether, $X = X_{0:\infty}$.

3. Relational methods to formalise individuality

This first class of frameworks attempts to formally define an agent starting from the relation it ought to have with its environment, a mode of coupling between the agent and its environment. We associate this class of proposals with the first requirement, individuality: for an individual to exist, it must be distinct from its environment but at the same time interact with it, since complete independence would make it a quite uninteresting entity and undermine the very concept of action.

The *dynamical systems approach* pioneered in [39, 40] is an attempt to formalise this idea, moving the explanatory focus from the agent's internal mechanisms to the continual mutual interaction between the agent and, importantly, its environment. The core theoretical commitment is to model the agent with states S taking values in $\mathcal{S}$, and the environment with states E taking values in $\mathcal{E}$, as two open dynamical systems whose states are governed by equations of the form:

$$\begin{aligned} \mathrm{upd}_{\mathcal{S}} &: \mathcal{S} \times \mathcal{O} \to \mathcal{S}, \quad \mathrm{out}_{\mathcal{S}}\colon \mathcal{S} \to \mathcal{A} \\ \mathrm{upd}_{\mathcal{E}} &: \mathcal{E} \times \mathcal{A} \to \mathcal{E}, \quad \mathrm{out}_{\mathcal{E}} : \mathcal{E} \to \mathcal{O} \end{aligned} \tag{1}$$

These correspond to classical Moore machines, and can be generalised to systems without the factorisation between updates and output mappings (Mealy machines) as in, e.g. [41, 42]. For simplicity, in the remainder of this work we will refer to these open dynamical systems as *systems*, and only explicitly bring out different definitions if necessary. We will also overload the notation and just refer to them using their state space when obvious from context, i.e. we will talk about an agent $\mathcal{S}$ and an environment $\mathcal{E}$, assuming that each system is equipped with update and output maps of the above form unless otherwise stated. For a more general, structural treatment of systems we refer to [43, 44]. The systems above are assumed to be deterministic, and in the original work [39, 40] also in continuous time. Here, for simplicity, we introduce them in their discrete time counterpart to simplify the exposition, since

[1] Usually by time $t \in \mathcal{T}$, unless otherwise stated.

generalisations to other kinds of systems (including those in continuous time), don't affect the overall arguments [43]. The central insight is that these two systems can be coupled, i.e. connected to form a single autonomous dynamical system, $\mathcal{X}$, whose state space is an appropriate composition of the agent and environment state spaces. The relevant behaviour of an agent is not something generated by the agent alone, but is a property of the trajectories of this entire coupled system. Agency is not located in a controller issuing commands, but distributed across the entire feedback loop. The notion of *adaptive fit* provides the criterion for successful agency. An agent is considered adaptively fit as long as trajectories of the coupled system $\mathcal{X}$ remain within a specific constraint, or viability volume, C in the total state space $\mathcal{X}$, i.e. as long as agent and environment interact in a way that preserves certain properties represented by C. For a biological organism, this constraint could mean survival, i.e. maintaining the integrity of its autopoietic (self-producing) processes [28]. For an agent more in general, the constraint is defined by the goal it needs to accomplish (e.g. for a robotic vacuum, the floor remaining clean). Agency is thus the capacity to interact with the environment to ensure the coupled system's trajectory satisfies the constraint C. This makes agency a relational property, defined not by what an agent is, but by its ability to maintain a specific kind of relationship with its environment over time. This perspective is explored in detail using gliders in the Game of Life as a toy model [45, 46, 47, 48, 49, 50]. By treating the Game of Life update rule as an artificial physics, [47] formalises a glider's self-maintaining organisation as a closed process dependency network. This provides a bottom-up, relational definition of the glider's identity. Related work [46] then characterises the agent-environment relationship by exhaustively mapping the interaction graph between a glider and its surroundings. This graph defines a macrodynamic function, $\text{upd}_{\mathcal{G}} : \mathcal{G} \times \mathcal{D} \to \mathcal{G}$, where $\mathcal{G}$ is the set of glider states and $\mathcal{D}$ is a set of non-lethal perturbations, see also [51]. The concept of structural coupling is then formalised by finding the mutually consistent trajectories of the agent and the environment, yielding the set of all possible glider lives.

The *free energy principle* [52, 53, 54] defines an agent as a system that persists over time by maintaining a statistical boundary that separates it from its environment. This requires the full system to be described as a stochastic differential equation

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t \tag{2}$$

which possesses a non-equilibrium steady-state density, $p(x)$. Here, W_t is a standard Brownian motion in $\mathbb{R}^z$, and the flow $f \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}^{n \times z}$ are smooth vector and matrix fields, respectively. [2] On this view, the defining feature of an agent is the statistical boundary known as a Markov blanket, claimed to induce a factorisation of the states of the full system, X, into four subsets. These subsets include internal states, S, that constitute the agent, external states, E, that constitute the environment, and blanket states, $B = (O, A)$, that mediate between the internal and external states and are composed of sensory states, O, that are influenced by external states but not internal states, and active states, A, that are influenced by internal states but not external states. It is however unclear at this stage what conditions are required for this factorisation to hold over time, i.e. once we consider S_t, E_t, O_t, A_t instead of S, E, O, A [55, 56, 57, 58, 59]. According to its proponents, this structure of influences, often addressed as sparse coupling in the system's flow, $f(X_t)$, ought to guarantee (but see criticisms above) a crucial conditional independence at non-equilibrium steady-state: internal and external states are statistically independent when conditioned on the blanket states. This is formally expressed

[2] Note that proponents of the free energy principle often emphasise "generalised coordinates" as a label for state augmentation in continuous time for, with Brownian motion in higher dimension to reduce non-Markovian processes to Markovian ones [53].

as a conditional independence, $\Pr(S, E \mid O, A) = \Pr(S \mid O, A)\Pr(E \mid O, A)$, or more compactly $(S \perp E) \mid O, A$, which holds within a time slice (i.e. for a time-synchronous blanket) but generally fails once we consider the full history of observations and actions (i.e. for a time-unrolled blanket) [59, 58]. This statistical insulation is the formal definition of what constitutes an agent, it is the boundary that carves the agent out from the rest of the system. It is often claimed that a consequence of this boundary is that internal states must infer the state of the external world, based on classical results from cybernetics [60] and control theory [61]. However these claims remain hard to verify given the role of different possible Markov blankets (within a time slice, or over trajectories) and of another fundamental assumption of the free energy principle, the existence of a synchronisation map, $\zeta : \mathcal{S} \to \mathcal{E}$. This map captures an explicit relation between an agent and its environment, such that for two given functions $g_{\mathcal{S}} : \mathcal{O} \times \mathcal{A} \to \mathcal{S}$ and $g_{\mathcal{E}} : \mathcal{O} \times \mathcal{A} \to \mathcal{E}$, within a time slice:

$$\begin{aligned} g_{\mathcal{S}}(O_t, A_t) &\coloneqq \mathbb{E}_{\Pr(S_t \mid O_t, A_t)}[S_t] \\ g_{\mathcal{E}}(O_t, A_t) &\coloneqq \mathbb{E}_{\Pr(E_t \mid O_t, A_t)}[E_t] \end{aligned} \tag{3}$$

we have

$$\zeta(S_t) \coloneqq g_{\mathcal{E}}\left(g_{\mathcal{S}}^{-1}(S_t)\right). \tag{4}$$

The synchronisation map is a function that links an agent's (expected) internal states to the (expected) external states they supposedly infer. The existence of this map is not guaranteed, as it depends on the condition that any two blanket states mapping to the same expected internal state must also map to the same expected external state, see [62] for a discussion of the linear case. This map is what allows an agent to ostensibly infer its environment, since an agent is not directly influenced by states E_t. It can only register its effects through sensory states O_t. The presence of this map implies that the internal states S_t parametrise a variational density, $Q_{\mathcal{S}}(E_t)$, which serves as a probabilistic model of the external states, given the blanket states, for a definition using the synchronisation map above see [63, Eq. 85]. The persistence of this entire factorised system at non-equilibrium steady-state then is thought to imply that its dynamics must follow a specific path, one that minimises a functional called variational free energy, $F(O_t, A_t, S_t)$ that is defined as:

$$F(O_t, A_t, S_t) = \mathbb{E}_{Q_{\mathcal{S}}(E_t)}[\ln Q_{\mathcal{S}}(E_t) - \ln \Pr(E_t, O_t, A_t, S_t)]. \tag{5}$$

This can be expressed as an upper bound on the surprisal (negative log-probability) of the agent's own states, actions and observations $F(O_t, A_t, S_t) = -\ln \Pr(O_t, A_t, S_t) + D_{\mathrm{KL}}[Q_{\mathcal{S}}(E_t) || \Pr(E_t \mid O_t, A_t, S_t)]$ and according to this proposal, agents are systems that minimise free energy, since that is what ensures that their surprisals remain within acceptable bounds over time.

The *Bayesian interpretation map* framework [63, 41] aims to overcome some of the possible limitations of the free energy principle, highlighted for instance in [57, 55, 56, 58, 64, 65, 66]. Following Dennett's intentional stance [67], this framework provides a formalisation of what it means to interpret a physical system as having beliefs and for its dynamics to be interpreted as a process of Bayesian belief updating [63]. Furthermore, it also provides criteria to interpret a physical system as taking actions according to its beliefs so to achieve a goal [41]. The core of the framework is the interpretation map, a function $\psi : \mathcal{S} \to \Pr(\mathcal{H})$ proposed by an observer to link the physical states S_t of an agent [3], taking values in $\mathcal{S}$, to abstract belief states, i.e. probability distributions, over hidden variables of a

[3] More precisely, a system that can be interpreted to be an agent.

model H_t, taking values in $\mathcal{H}$. The agent's physical state transitions are governed by a machine kernel $\mathrm{upd}_{\mathcal{S}} : \mathcal{O} \times \mathcal{S} \to \Pr(\mathcal{S})$, which gives the probability of transitioning to a new state S_{t+1} given the current state S_t and a sensory input O_t. An interpretation includes both an interpretation map and a model kernel $\kappa : \mathcal{H} \times \mathcal{A} \to \Pr(\mathcal{H} \times \mathcal{O})$ which represents the agent's (potentially incorrect) model of how its actions A_t influence the hidden state and produce sensory inputs. Note that this kernel combines update and output maps into one, assuming that they aren't necessarily factorisable, unlike the dynamical systems approach discussed earlier. Interestingly, the probabilistic system $\mathcal{H}$ can, but need not be the environment, $\mathcal{E}$. In many interesting cases it however will be at least a model, to some extent veridical, of the environment [68] or of the whole agent/environment combined system [69]. An interpretation is said to be consistent if the physical dynamics of the agent and the belief dynamics align perfectly. This is captured by a consistency equation, which, unpacked from its category-theoretic form, states that for any possible next physical state S_{t+1}, the belief it represents, $\psi(H_{t+1} \mid S_{t+1})$, must be equal to the Bayesian posterior belief, calculated from the prior belief $\psi(H_t \mid S_t)$ and the evidence O_t. Formally, this means that

$$\psi(H_{t+1} \mid S_{t+1}) = \frac{\sum_{H_t \in \mathcal{H}} \kappa\big(H_{t+1}, O_t \mid H_t, \pi(S_t)\big)\psi(H_t \mid S_t)}{\sum_{H_t, H_{t+1} \in \mathcal{H}} \kappa\big(H_{t+1}, O_t \mid H_t, \pi(S_t)\big)\psi(H_t \mid S_t)}, \tag{6}$$

where $\pi(S_t)$ is the action taken in state S_t. This framework has been used to show deep connections to control theory and concepts of agency in that field, proving that any system satisfying the assumptions of the classic internal model principle [69] and of the good regulator [68] can be interpreted as performing Bayesian updates on a model built by an observer, which can sometimes correspond to the environment.

Cartesian frames [70] provide on the other hand a mathematical framework for modelling agency that is designed to overcome the limitations of traditional models that treat the agent-environment boundary as a fixed, primitive reality. Cartesian frames constitute a way to carve an agent-centric perspective of choice onto an objective, physical set of all possible outcomes of a particular full system, allowing for an analysis of agency across different levels of description. The fundamental building block of this framework is the Cartesian frame, a Chu space $\mathcal{J} = (\mathcal{S}, \mathcal{E}, \cdot)$ over a set $\mathcal{X}$, a structure that generalises a decision matrix. Here, $\mathcal{S}$ is the set of all possible ways the agent can choose to be, e.g. its states, $\mathcal{E}$ is the environment's set of possible states, and the outcome function $\cdot : \mathcal{S} \times \mathcal{E} \to \mathcal{X}$ maps agent-environment pairs to a set of possible worlds of the full system $\mathcal{X}$. An agent is defined as a system $\mathcal{S}$ that can be obtained from the factorisation of a larger system $\mathcal{X}$ whose state can be composed with the environment $\mathcal{E}$ to produce $\mathcal{X}$ itself. An agent can be carved out to be a separate system from the environment, but its identity and interface with the environment are subject to constant change. Note that here there is no explicit notion of dynamics, although one could understand the outcome function as mapping between trajectories, therefore folding time into the sets $\mathcal{S}$ and $\mathcal{E}$, nor of an agent's inputs (observations) $\mathcal{O}$ or its outputs (actions) $\mathcal{A}$, which are instead derived as follows. Observations are defined as:

$$\mathcal{O} = \left\{\mathcal{X}' \subseteq \mathcal{X} \mid \forall s_0, s_1 \in \mathcal{S}, \exists s \in \mathcal{S}, s \in \mathrm{if}\left(\mathcal{X}', s_0, s_1\right)\right\} \tag{7}$$

where $\mathrm{if}\left(\mathcal{X}', s_0, s_1\right)$ denotes the set of all $s \in \mathcal{S}$ such that for all $e \in \mathcal{E}$, $(s \cdot e \in \mathcal{X}') \to (s \cdot e = s_0 \cdot e)$ and $(s \cdot e \notin \mathcal{X}') \to (s \cdot e = s_1 \cdot e)$. Agents in this setting observe events, which are either true or false, not variables in full generality. We will say that given a frame $\mathcal{J}$, its observables, $\mathcal{O}$, are the set of all $\mathcal{X}'$ such that $\mathcal{J}$'s agent, S, can observe whether $x' \in \mathcal{X}'$ is true. Actions are defined as:

$$\mathcal{A} = \left\{\mathcal{X}' \subseteq \mathcal{X} \mid \exists s \in \mathcal{S}, \forall e \in \mathcal{E}, s \cdot e \in \mathcal{X}'\right\} \cap \left\{\mathcal{X}'' \subseteq \mathcal{X} \mid \exists s \in \mathcal{S}, \forall e \in \mathcal{E}, s \cdot e \notin \mathcal{X}''\right\} \tag{8}$$

where $\mathcal{X}'$ gives a set of world states an agent can ensure, and $\mathcal{X}''$ a set of world states an agent can prevent. The framework's power comes from its use of morphisms or maps between frames. A morphism from frame $\mathcal{J} = (\mathcal{S}, \mathcal{E}, \cdot)$ to $\mathcal{J}' = (\mathcal{S}', \mathcal{E}', *)$ is a pair of functions $(g : \mathcal{S} \to \mathcal{S}, h : \mathcal{E}' \to \mathcal{E})$ that satisfies the adjointness condition: $s \cdot h(e') = g(s) * e$ for all acts $s \in \mathcal{S}$ and environment states $e' \in \mathcal{E}'$. This mathematical structure allows for formal operations like moving between different levels of description (coarsening or refining the set of possible worlds $\mathcal{X}$) and, most notably, composing or decomposing agents by modelling *subagency* of different kinds, *additive* to describe agents after some commitment to particular subsets of choices or states, and *multiplicative* to define agents within a larger agent.

4. Prediction-based methods to formalise normativity

This second class of frameworks posits that agents are systems that possess some kind of predictive power, about themselves, their environment or both. On this view, agents are portrayed as systems that can be characterised in terms of their ability to predict features that are crucial for their goals. At this stage we remain agnostic about whether this ability to predict is intrinsic for the agent or in the eye of the beholder [71] and return to this in Section 6. We also note that some of the proposals in the previous section, chiefly the free energy principle, make strong claims about agents and their ability to predict features of their environment. This is related to another point we will discuss at the end, a correspondence between individuality and normativity.

The *information theory of individuality* [72] provides a quantitative way to identify individuals at any scale by defining them as aggregates that maintain temporal integrity by propagating information from their past into their future. Building on earlier work on information-theoretic autonomy [73], this approach analyses the total predictive information that the present state of an agent, S_t, and the state of the environment, E_t, jointly provide about the agent's next state, S_{t+1}, captured by the mutual information $I(S_{t+1}; S_t, E_t)$. The framework's key insight is that this total information can be decomposed in two distinct ways using the chain rule for mutual information:

$$I(S_{t+1}; S_t, E_t) = \underbrace{I(S_{t+1}; S_t)}_{\text{organismal individuality}} + \underbrace{I(S_{t+1}; E_t \mid S_t)}_{\text{environmentally determined individuality}} = I(S_{t+1}; E_t) + \underbrace{I(S_{t+1}; S_t \mid E_t)}_{\text{colonial individuality}} \tag{9}$$

These decompositions define a taxonomy of individuality. Organismal individuality, proposed in [73] as a measure of autonomy for systems that control their environment, measures the information that an agent's own past provides about its future, ignoring the environment. It captures highly autonomous individuals that are largely in control of their own dynamics. Colonial individuality, see again [73], instead measures autonomy for agents that are driven by their environment. More specifically, the additional information the system's past provides about its future, once the influence of the environment's past is already taken into account. [4] This ought to capture more loosely-coupled individuals, like microbial colonies, whose persistence depends heavily on ongoing interaction. To see this, [72] employs tools from partial information decomposition to more carefully unpack the influence of environment and agent. We do not however venture into this level of detail here. Environmental determined individuality is on the other hand defined by the information flow from the environment, which identifies entities whose structure is largely imposed by external forces.

[4] This also corresponds to Granger autonomy, or G-autonomy as proposed in [74] for Gaussian variables, see [75].

The *agent description* of [76] provides a different formalisation of Dennett's intentional stance [67], framing the question of agency as a Bayesian model comparison problem. In this setup, an observer is presented with a system's behaviour, i.e. a trajectory of outputs, or actions for a candidate agent, $A_{:t}$, given inputs, or observations for a candidate agent, $O_{:t}$, and must decide which of two explanatory models is more probable. The model M_d describes the system as a device, a reactive machine whose behaviour is explained by a direct, mechanistic input-output mapping, which is associated to Dennett's physical stance. The likelihood of the trajectory under this model is a weighted average over all possible device descriptions:

$$M_d(A_{:t} \mid O_{:t}) = \sum_{d \in \mathcal{M}_d} d(A_{:t} \mid O_{:t})\, \omega_d, \tag{10}$$

where $d : (\mathcal{O} \times \mathcal{A})^* \to \Delta(\mathcal{A})$ produces a probability distribution of outputs given an interaction history of inputs $O_{:t}$ and outputs $A_{:t}$, and ω_d is a prior over devices. The model M_s identifies instead the system as an agent whose behaviour appears as the result of optimising a goal or utility function $u \in \mathcal{U}$, proposed to correspond to Dennett's intentional stance. The likelihood is derived using inverse reinforcement learning:

$$M_s(A_{:t} \mid O_{:t}) = \sum_{u \in \mathcal{U}} \pi_u(A_{:t} \mid O_{:t})\, \omega_u, \tag{11}$$

where $\pi_u : (\mathcal{O} \times \mathcal{A})^* \to \Delta(\mathcal{A})$ is the optimal (or near-optimal) policy for that utility function, and ω_s is a prior over devices. The framework then uses Bayes' rule to calculate the posterior probability that the system is an agent:

$$p(\text{agent} \mid A_{:t}, O_{:t}) = \frac{M_s(A_{:t} \mid O_{:t})}{M_d(A_{:t} \mid O_{:t}) + M_s(A_{:t} \mid O_{:t})} \tag{12}$$

assuming equal priors. The prior weights (ω_d, ω_u), which can be based on the Kolmogorov complexity of describing the device or the utility function, provide an intuition for this definition by thinking about the compression aspect that arises from the principles of algorithmic information theory. In particular, a goal-based description is the better description if the complexity of describing the goal is lower than the complexity of describing the intricate mechanistic policy that achieves it. This framework thus quantifies the intuition that agency is a useful concept when it provides a more compressed, predictive explanation for a system's behaviour, and while reliant on Bayesian model comparison, extensions can make use of different objectives, see for instance [77].

Complete local integration [29, 78, 79] seeks to describe an important property of agents within a dynamical system by quantifying the internal coherence of the parts that constitute them. The core idea is that an important aspect of being an agent is that a system can be characterised as an (ι-)entity, defined in [78, 79] as a spatiotemporal pattern where every part (which is itself another spatiotemporal pattern) makes every other part more probable. This is formalised through the measure of specific local integration. For a given spatiotemporal pattern, i.e. a set of fixed values of a candidate agent's states $S_{\mathcal{W}} = s_{\mathcal{W}}$ indexed over time and space by $\mathcal{W} \subseteq \mathcal{T} \times \mathcal{R}$, and a partition ξ of the set of variables $\mathcal{W}$ into blocks b, the specific local integration is defined as:

$$\mathrm{mi}_{\xi}(S_{\mathcal{W}}) = \log \frac{\Pr(S_{\mathcal{W}})}{\prod_{b \in \xi} \Pr(S_b)}. \tag{13}$$

A positive specific local integration indicates that the whole pattern $S_{\mathcal{W}}$ is more probable than would be expected if its parts (defined by the partition ξ) were independent. To identify truly integrated entities that can serve as candidate agents, the framework introduces then complete local integration, which is the minimum specific local integration value over all possible non-unit partitions of the pattern:

$$\iota(S_{\mathcal{W}}) = \min_{\xi \in \mathfrak{L}(\mathcal{W}) \setminus 1_{\mathcal{W}}} \mathrm{mi}_{\xi}(S_{\mathcal{W}}), \tag{14}$$

where $\mathfrak{L}(\mathcal{W})$ is the partition lattice of $\mathcal{W}$, i.e. the set of partitions of $\mathcal{W}$ partially ordered by refinement, and $1_{\mathcal{W}}$ is the unit partition. A spatiotemporal pattern is then defined as an ι-entity if its complete local integration is greater than zero, $\iota(S_{\mathcal{W}}) > 0$. The claim here is thus that agents, as individuals, are systems with positive complete local integration. This means that a pattern is integrated with respect to every possible way of decomposing it. The importance of positive complete local integration is formally established through a *disintegration theorem*, demonstrating that ι-entities are the fundamental, maximally disintegrated components of a system's trajectory. Positive complete local integration also implies that, from an information-theoretic perspective, the surprisal (information content) of the whole entity is lower than the sum of the surprisal of the parts (cf. the minimisation of surprisal as a central point of the free energy principle [52, 53, 54]). This tells us that the whole is spatially more predictable (lower surprisal) than the parts, and temporally more predictable because the presence of an entity at a certain point in time makes it more probable that a corresponding entity will exist at a subsequent moment. From a coding-theory perspective, the fact that a ι-entity has positive integration across all possible partitions means that the most efficient, compressed description of the pattern is always the one that treats it as a single, integrated whole. Any attempt to describe it as a composition of independent parts results in a less efficient, longer description, meaning that a ι-entity captures in some sense a pocket of predictability that arises from a system's dynamics.

Rosen's theory of *anticipatory systems* [80, 81, 82] usually refers to the study of biological systems, but similarly to what we do here, focuses on aspects of living systems not directly related to their ability to undergo (Darwinian) evolution. Because of this, we will treat this as a framework attempting to capture aspects of agency, in particular claiming that agents are anticipatory systems. Anticipatory systems are closed systems characterised by a modelling relation expressed as a commuting diagram between a system, say an agent $\mathcal{S}$, often addressed as *natural system* in this literature, and a system, $\mathcal{H}$, usually called *formal system* which serves as its epistemic, predictive model or abstraction, and is built by an observer. The maps representing state transitions are called causal entailment for the physical laws in the agent, e.g. its dynamics, $c : \mathcal{S} \to \mathcal{S}$, and inferential entailment for the logical rules of the formal system, e.g. the dynamics, $i : \mathcal{H} \to \mathcal{H}$. The mappings between the systems are the encoding, $\varepsilon : \mathcal{S} \to \mathcal{H}$, which represents measurement of the natural system, and the decoding, $\delta : \mathcal{H} \to \mathcal{S}$, which represents an interpretation of properties of a formal model in the natural world. A modelling relation between $\mathcal{S}$ and $\mathcal{H}$ exists if the following diagram commutes [83, 4.14] in the appropriate setup [5]:

$$\begin{tikzcd} \mathcal{S} \arrow[r, "\varepsilon"] \arrow[d, "c"'] & \mathcal{H} \arrow[d, "i"] \\ \mathcal{S} & \mathcal{H} \arrow[l, "\delta"'] \end{tikzcd} \tag{15}$$

meaning that the evolution in the natural system $\mathcal{S}$ must be equivalent to the path of encoding, inferring, and decoding, i.e. $c = \delta \circ i \circ \varepsilon$, and

$$i = \varepsilon(c), \tag{16}$$

[5] We can imagine these objects to be sets in the category of sets and relations for instance, **Rel**, with arrows given by relations or multi-valued functions, since this is the starting point of the categorical setup of anticipatory systems [82, 84].

a condition stating that the updates of the formal model must be related to the updates of the natural system, or agent. [6] The system $\mathcal{H}$ can be seen as a predictive model because from the perspective of $\mathcal{S}$, one can first encode its state via ε, then process it as part of the model using i and finally decode the result with δ, checking that it is equivalent to its own evolution [82]. We note the resemblance with some of the arguments formalised in [69], in particular when the system $\mathcal{H}$ is a model of the agent and its environment, however the exact connection remains unclear and won't be explored here.

5. Causality-based methods to formalise interactional asymmetry

The third class of frameworks seeks to operationalise agent causation by arguing that an agent's internal states are the genuine causes of its actions and their effects in the world. These frameworks aim to identify and quantify this causal power, overcoming some of the philosophical challenges of defining causality. We note that for these proposals too, it is unclear whether causality ought to be portrayed as an intrinsic feature of agents (it often is portrayed as such) and return to this in Section 6.

Using *integrated information theory (IIT)* [86, 87], traditionally portrayed as a theory of consciousness, an agent has been defined as a physical system that constitutes itself as a single, irreducible causal entity [88, 89, 90, 91]. The existence of an agent is assumed to be determined by the system's internal cause-effect power [90, 91], defined as the degree of causal irreducibility and quantified by integrated information, Φ. We start with a full system with states $X_{\mathcal{K}}$ indexed over time and state $\mathcal{K} \subseteq \mathcal{T} \times \mathcal{V}$, corresponding to a multivariate stochastic process. For each different full system variable X_t' (i.e. indexing over $\mathcal{T}$ only), integrated information is defined as the following (note that this varies across versions [87]):

$$\phi(X_t') = \min_{t\pm 1}\left(\max_{Q}\left(\min_{\xi}\left(\mathrm{Dist}\left[r(Q_{t\pm 1} \mid X_t'); \xi\left(r(Q_{t\pm 1} \mid X_t')\right)\right]\right)\right)\right) \tag{17}$$

such that the subset of variables X_t' for which integrated information is positive, $\phi(X_t' = x_t') > 0$, are called mechanisms. We unpack this equation in more detail here. To define mechanisms, this framework requires that variables X' specify information about both their causes and effects (it is not enough to do one or the other), this is captured by the first $\min_{t\pm 1}$ in the equation (if no information is specified about either causes or effects, this will return 0, making this system not an agent). Variables Q are another subset of the full system variables, called purview, over which the cause and effect repertoires of a mechanism in a state are calculated, and are chosen to maximise $\phi(X_t')$ via $\max_Q$. ξ is a partition of variables in the cause/effect repertoires, see also the description given for complete local integration [29, 78, 79] which is inspired by this work, and selecting the one that makes the least difference via $\min_\xi$ gives the so-called minimum information partition. To calculate this, we need a distance $\mathrm{Dist}[-;-]$ which captures the degree to which a system is integrated. A system is said to be integrated if its causal dynamics cannot be decomposed into independent parts (for the specific choice of a measure in different versions, see [87]). Formally, this means that larger distances give larger ϕ, capturing the idea that a system's joint causal-effect structure cannot be factorised into separate substructures without loss of information. Using the do-operator from do-calculus [38], introduced to simulate a notion of physical intervention that deletes certain functions from a model, replacing them with a constant $X = x$, the *effect repertoire* is given by $r(Q_{t+1} \mid X_t') = \prod_i \Pr(Q_{i,t+1} \mid \mathrm{do}(X_t' = x_t'))$,

[6] Different attempts to formalise this using a categorical setup, as a functor between categories of natural systems and of formal models, can be found in, for instance, [83, 85, 82], but the details won't be covered here.

and marginalises purviews over non-X' variables leaving only possible dependencies on X'. On the other hand, the *cause repertoire* is given by $r(Q_{t-1} \mid X'_t) = \frac{1}{L} \prod_v \Pr(Q_{t-1} \mid \mathrm{do}(X'_{v,t} = x'_{v,t}))$ with L a normalising constant, which marginalises instead over variables X' to remove common inputs from non-Q variables. Variables X' with positive integrated information thus specify information about purviews $Q_{t\pm1}$ to the extent that conditioning on them, $r(Q_{t\pm1} \mid X'_t)$, constrains the state of the purviews when compared to their unconstrained probabilities $r(Q_{t\pm1})$. Rather than simply assuming (top-down) a system that can be factorised in mechanisms and non-mechanisms, this framework can also be used to define (bottom-up) a system of interest, for instance an agent $\mathcal{S}$, as a collection of mechanisms X', the minimal collection that can't be factorised without loss of integrated information. This can be done by defining a system-level integrated information [91]:

$$\Phi(X'_t) = \min_{\xi} \left(\mathrm{Dist}\left[\mathcal{C}(X'_t); \mathcal{C}\left(\xi(X'_t)\right) \right] \right) \tag{18}$$

where the cause-effect structure $\mathcal{C}(X'_t)$ is a set of mechanisms. This definition is quite relevant for [35]'s notion of interactional asymmetry. In particular, following [88, 92, 89], an agent's causal boundary can be defined as a local maximum of Φ: as an emergent macro-level entity with greater causal power than its underlying micro-constituents. For large systems, this highly integrated state is achieved by being poised near a critical point (a phase transition), where integration diverges. Based on this, [88, 92, 89] introduce a notion of asymmetry to define autonomous agents as systems that must be the (sub)system which causally drives their interaction with the environment, rather than being determined by it. This is shown for example in [88], where autonomy is measured in a model by the agent's ability to dynamically modulate its causal coupling with the environment. Here, the agent, which is only formed of sensors $\mathcal{O}$ and motor $\mathcal{A}$ units, can in some cases be the most integrated system, above and beyond the full system $\mathcal{X}$ including the environment $\mathcal{E}$, i.e. $\Phi_{\mathcal{O}\mathcal{A}} > \Phi_{\mathcal{O}\mathcal{A}\mathcal{E}}$, and thus can be seen as setting its boundary to demarcate itself from the environment. In other situations, the full agent-environment system becomes the most integrated unit ($\Phi_{OAE} > \Phi_{OA}$), with the agent recruiting (parts of) the environment to change its identity and causal presence on the remaining parts of the environment.

Crucially for this section, IIT has been especially used to investigate when an agent can be said to be the actual cause of its own actions [93, 91], and therefore asymmetrically driving its interactions with the environment. Using a similar formalism, derived from the same principles underlying integrated information [90], one can define the causal strength R, a measure of irreducible information effects. To this point, action variables A_t can be used to specify information about their cause, a subset of an agent's states S'_{t-1} taking values in $\mathcal{S}' \subseteq \mathcal{S}$, quantifying (in bits) how much information is lost when applying a partition ξ that splits causes and effects into subcomponents to test their local causal links:

$$R(A_t) = \max_{S'} \left(\min_{\xi} \left(\log_2 \left(\frac{r(S'_{t-1} \mid A_t)}{\xi\left(r(S'_{t-1} \mid A_t)\right)} \right) \right) \right)$$

where $r(S'_{t-1} \mid A_t)$ is the causal repertoire defined above. [7]

Mechanised Causal Graphs [94] provide another proposal of causality-based methods. The core idea is that "agents are systems that would adapt their policy if they were aware that their decisions influenced the world in a different way". The starting point is a mechanised structural causal game,

[7] In [93] we find simple probabilities in place of cause repertoires, but we believe this to be a typo since in the original reference (on which this work is based) seems to define causal strength using cause repertoires [90].

a tuple $\mathcal{M} = \langle N, P_N, G, \Pr(P_N) \rangle$. As in the standard definition of structural causal models, N is a set of endogenous variables indexed over state $\mathcal{V}$, $\{N_v\}_{v \in \mathcal{V}}$, P_N is a set of exogenous variables indexed by endogenous variables, $\{N_v = g_{N_v}(N, P_{N_v})\}$ is a set of structural equations indexed by endogenous variables, and $\Pr(P_N)$ is a distribution over the exogenous variables. The mechanised version however extends the standard definition in two ways. The first extension comes from the definition of mechanised structural causal models and is given by a factorisation of endogenous variables into object-level and mechanism variables, $N = X \cup \tilde{X}$. For each object-level variable X, a corresponding mechanism $\tilde{X}$ parametrises the relative structural equation. The second extension is the assumption that endogenous variables can be factorised into chance (environment's state), decision (actions) and utility variables, which combined with the first extension give $X = E \cup A \cup U$ and $\tilde{X} = \tilde{E} \cup \tilde{A} \cup \tilde{U}$. The crucial idea behind this framework is that when an object-level variable is a decision, A, its associated mechanism, $\tilde{A}$, is formally treated as an agent's decision rule or policy. The mechanism is the reasoning that produces the decision, which is then assumed to be the action. This framework thus provides a way to build mechanised structural causal games and identify decision nodes and their associated mechanisms. To do it, it distinguishes between two fundamental types of interventions [38]: standard object-level interventions, $\mathrm{do}(X = x)$, which set a variable's value, and mechanism interventions, $\mathrm{do}(\tilde{X} = \tilde{x})$, which alter the function g^X itself, and must satisfy the following condition: given an object-level variable X_v and a related mechanism $\tilde{X}_v$

$$\Pr\left(X_v \mid \mathrm{Pa}^{X_v}, \mathrm{do}(\tilde{X}_v = \tilde{x}_v)\right) = \Pr\left(X_v \mid \mathrm{do}(\tilde{X}_v = \tilde{x}_v)\right). \tag{19}$$

For decisions, this means that one can distinguish between interventions on the action itself, $\mathrm{do}(A = a)$, and interventions on the agent's policy, $\mathrm{do}(\tilde{A} = \tilde{a})$, which force the agent to adapt. The latter corresponds to a change in the world's dynamics that a prospective agent can be made aware of and adapt to. Agency thus is discovered by performing interventions on mechanism variables, in particular when these are deemed to be policies. To determine this, [94] defines a specific causal structure, the terminal mechanism edge. This edge links a utility's mechanism to a decision's mechanism, $\tilde{U} \to \tilde{A}$ and is defined by two conditions tested via mechanism interventions, interventions that sever an object-level node from its consequences. The first condition determines if a node is a utility node. The rule is that an agent's policy $\tilde{A}$ must remain sensitive to the utility's mechanism $\tilde{U}$ even when the utility's own consequences are nullified. For any structural intervention on the children of U, $\mathrm{do}(\mathrm{Ch}^U)$), there must exist a mechanism intervention $\mathrm{do}(\tilde{U} = \tilde{u})$ such that:

$$\Pr\left(\tilde{A} \mid \mathrm{do}(\tilde{u}, \mathrm{Ch}^U)\right) \neq \Pr\left(\tilde{A} \mid \mathrm{do}(\mathrm{Ch}^U)\right) \tag{20}$$

This identifies U as a utility node because the agent values it for its own sake, not for its downstream effects. The second condition tests if a node is a decision node. To be one, the agent's policy, $\tilde{A}$ must become insensitive to the utility's mechanism when the decision's own consequences are nullified. For any structural intervention on the children of A, denoted $\mathrm{do}(\mathrm{Ch}^A)$:

$$\Pr\left(\tilde{A} \mid \mathrm{do}(\tilde{u}, \mathrm{Ch}^A)\right) = \Pr\left(\tilde{A} \mid \mathrm{do}(\mathrm{Ch}^A)\right) \tag{21}$$

This identifies A as a decision node because the policy only adapts for the sake of achieving the decision's consequences. Notice how this approach remains agnostic with respect to time and to a full factorisation of a system that includes an agent's states and its observations, which could perhaps be represented by some mechanism variables, but whose status is otherwise unclear.

Using different measures of causality and a different notion of intervention from that of Pearl [38], *semantic information* is defined as the information that an agent has about its environment that is causally necessary for the agent to maintain its own existence over time [95]. This information ought to be causally necessary for an agent because often an agent needs to take actions in the environment to ensure its existence, and a mechanism to disentangle relevant from irrelevant information is thought to be a good candidate explanation for how this can be achieved efficiently [96]. The theory begins by defining an agent as a far-from-equilibrium system $\mathcal{S}$ that actively works to maintain its own existence by keeping itself in a low-entropy state. This degree of existence is quantified by a viability function, V. The primary function proposed is the negative Shannon entropy, $\mathfrak{H}$, of the system's marginal state distribution, $\Pr(S_t)$, at a time t, $V\big(\Pr(S_t)\big) = -\mathfrak{H}(S_t)$. The core claim is that an agent maintains its viability by using semantic information, which is formally defined as the portion of syntactic (Shannon) information the system has about its environment, $\mathcal{E}$, that is causally necessary for its self-maintenance. The trajectory of the system under its actual dynamics is compared to *intervened* trajectories where correlations between the system and environment have been scrambled. Here, "intervening" does not mean a surgical Pearl-style $\mathrm{do}(\cdot)$ that sets variables to fixed values [38]. Instead, the framework perturbs only the information channel from environment to system by coarse-graining what the system can distinguish about $\mathcal{E}$ ("pre-garbling" of the channel), while leaving the environment's own dynamics otherwise unchanged. Concretely, one replaces the response kernel from environment to system with a version that depends only on a coarse-grained signal $\mu(E_t)$:

$$\hat{\Pr}^{\mu}(S_{t+1} \mid S_t, E_t) := \Pr(S_{t+1} \mid S_t, \mu(E_t)), \tag{22}$$

so the system "sees" $\mu(E_t)$ rather than E_t, a degradation of input distinctions. By varying μ, one obtains a family of *partial interventions* that "scramble" correlations between $\mathcal{S}$ and $\mathcal{E}$ to controlled degrees. If μ is the identity, the intervened and actual processes coincide, if μ is constant ("fully scrambled"), all transfer entropy from $\mathcal{E}$ to $\mathcal{S}$ vanishes. The value of information, ΔV, is the loss in viability that results from such an intervention at time t:

$$\Delta V_t = V\big(\Pr(S_t)\big) - V\big(\hat{\Pr}^{\mu}(S_t)\big), \tag{23}$$

where $\Pr(S_t)$ is the actual distribution and $\hat{\Pr}^{\mu}(S_t)$ is an intervened one, for a coarse-graining map of environment states $\mu : \mathcal{E} \to \mathcal{E}$.[8] To isolate the essential, meaningful information, the theory defines an optimal intervention, $\hat{\Pr}^{\text{opt}}$. This is the intervention that preserves the minimum amount of initial mutual information, $I\,(S_0; E_0)$, required to achieve the exact same viability as the unintervened system at a time τ:

$$\hat{\Pr}^{\text{opt}} \in \underset{\hat{\Pr}^{\mu}}{\arg\min}\, I_{\hat{\Pr}^{\mu}}(S_0; E_0) \quad \text{s.t. } \Delta V_\tau = 0. \tag{24}$$

All information that survives this scrambling process is assumed to causally necessary and therefore semantic. The amount of semantic information is the quantity of syntactic information, Σ, preserved under this optimal intervention. For information stored in the initial state, this is:

$$\Sigma_{\text{stored}} = I_{\hat{\Pr}^{\text{opt}}}(S_0; E_0). \tag{25}$$

This framework considers also a notion of observed semantic information that can be acquired. This is thought to be more characteristic of active agents and is derived by intervening on the dynamic

[8] In the original presentation the coarse-graining is a deterministic function f. More general stochastic coarse-grainings (Markov kernels) are a natural extension for input *degradation* of the environment-to-system channel.

flow of information, as measured by the amount of transfer entropy that remains under this optimal intervention [9]:

$$\Sigma_{\text{observed}} = \sum_{t=1}^{\tau} \mathrm{TE}_{\hat{\mathrm{Pr}}^{\text{opt}}}(E_{:t-1} \to S_t) = I_{\hat{\mathrm{Pr}}^{\text{opt}}}(S_t; E_{:t-1} \mid S_{:t-1}). \tag{26}$$

This is the minimal rate of information flow that is causally necessary for the agent to regulate its coupling with the environment, and to take appropriate actions to maintain its existence. It quantifies how an agent's perceptions are meaningful and essential for guiding its behaviour over time. These measures lead to a rigorous definition of an agent: a physical system is an autonomous agent to the extent that it possesses a high value of semantic information, Σ, both in terms of stored and observed semantic information. This notion is derived from the system's own dynamics and it's relative to a goal, in this case the imperative to maintain a low-entropy existence, but could in principle be defined for different goals.

6. Discussion

Our work provides a characterisation of different formal approaches to the study of agents through the lenses of a unifying conceptual framework [35]. This framework relies on three requirements: individuality, normativity or goal-directedness and interactional asymmetry, which we believe to capture desirable aspects of a theory of agents. This framing allows us to consider the strengths of each proposal, and what they are missing given these three requirements. Under this light, it appears that none of the existing theories currently analysed here meets all the highlighted requirements. Some theories can be seen as implementing two of them, especially in the sense we will see below for individuality and normativity, but none appear to embrace them all. This can either mean that none of these theories analysed here is currently a complete theory of agency (as directly stated in several cases) or that some of the requirements adopted here are in fact not necessary. For the latter, the major candidate is perhaps interactional asymmetry. This is because it is hard to imagine a way to talk about agents that doesn't include a way to formalise individuality, defining the unit of investigation, and goal-directedness, giving a way to describe what distinguishes agents from *mere* physical systems. It is also the case because while this requirement seems to make sense on the surface, is in fact not formulated very precisely, see [37]. It is also perhaps more suggestive, if at all, of a definition of higher-order agents, agents that may require some notion of causality [96], as opposed to the form of minimal agency advocated here as part of a definition of, among others, living systems.

One of the core features of the formalisation of agency proposed in [35] is that agents are systems whose individuality, normativity and asymmetry are *intrinsic* defining properties of what it means to be an agent. Identity should not be imposed by an external observer, as with an artifact, but ought to be an intrinsic and ongoing achievement of the system itself. An agent's norms should also not be arbitrary or merely assigned by a designer, but rather generated from within the system itself, based on its need to maintain its own organisation. Similarly, interactional asymmetry ought to be induced by the internal causal structure of an agent, and not be dependent on external factors. Among the approaches we analysed and reframed in this work, several however explicitly assume that agency is, at least in part, an observer-dependent property [78, 76, 52, 63]. Other proposals could also be interpreted as making a similar assumption, albeit more implicitly. For instance by choosing a

[9] Note that [95] assumes Markovianity as a simplifying assumption, but as explicitly stated it is not strictly necessary, so we give here the version of transfer entropy that considers all the histories.

good state space over which information theoretic measures can be applied [72], by building a model space over which structural causal models are well defined and meaningful for questions related to agency [94], or by requiring knowledge of other systems over which to calculate the maximum degree of causal power [97]. To fill this explanatory gap between intrinsic and extrinsic, observer-dependent proposals, following [36, 37] we adopt a more pragmatic "as-if" approach to the study of agents. This means that an adequate account of an agent should, in our opinion, include the requirement that for an observer, a candidate agent ought to have beliefs like an agent, act on the environment like an agent, and follow goals like an agent. Intrinsic aspects of agency could still be deemed constitutive of a full definition, but would just further be refining this intuition. No intrinsic aspect of agency should appear in systems that cannot be interpreted to be agents (whether they are not) by an observer. In this sense, we consider as-if agency a necessary, but perhaps not sufficient if intrinsic aspects ought to be included, condition for agency.

This as-if stance has two consequences in the present work. Firstly, the fact that at least in an as-if sense, individuality and normativity are deeply connected. To see this, we first note that at a high level the relational methods used to describe individuality and the predictive methods seem to overlap considerably. For instance, the theory of individuality [72] and complete local integration [29, 78, 79], which we described as predictive and relevant for normativity, were originally introduced to define individuality. On the other hand, relational methods such as the free energy principle [52, 53, 54] and the Bayesian interpretation map [63, 41] seem to suggest the necessity of a model that is in some sense predictive for an individual. This is a consequence of the a possible *identity-normativity correspondence*. Informally, this captures the idea that identity, as a relation between agent and environment, necessarily implies the presence of goals as relevant properties of the coupled agent-environment system that the agent can predict and achieve. Conversely, this also means that specific goals give rise to ways to carve particular identities out of a (dynamical) system. Formally, this has been shown for a simple class of systems in [68], similar to the one of [39, 40], except for the fact that the environment need not factorise into separate update and output maps, see Eq. (1), and that we have now included a well-defined notion of beliefs for a candidate agent. Under the assumptions of [68], one can say that any system achieving a goal, described as the ability for itself and its coupled system to stay within a viability volume C (the "good set"), can be interpreted as an agent performing Bayesian inference or filtering (depending on the context) on another component, often just (a model of) the environment. Or with a slight abuse of terminology, as a system *predicting* its own environment. This Bayesian interpretation relies on the existence of an interpretation map from states of the first system to beliefs on states of the second system that obeys a consistency equation equivalent to Bayesian updates under different assumptions. In other words, it describes a *relation* between agent and environment that satisfies some formal requirements and builds of the presence of an individual, the agent. Thus, given a full system that achieve a goal in the above sense, the full system can in principle be factorised into an agent and environment coupled systems. This factorisation is not unique beyond perhaps some trivial case, so an interpretation depends (among other things) on where the boundary between agent and environment is drawn. However, a valid interpretation must obey some consistency constraints, hence factorisations are not arbitrary.

As explained in [68, 69] this correspondence follows from the old: "Every good regulator of a system must be a model of that system" [60], roughly claiming to provide a mathematical statement about how systems that achieve some goal must do so by modelling their environment. The mathematical grounding of this statement is however unclear, and later only partially formalised with the internal model principle in control theory [61], under rather strong assumptions as explained in [69]. As suggested by [98], this statement also forms the basis of frameworks such as the free energy principle [52, 53, 54], stating roughly the same idea but for a larger class of systems than the ones covered by [60]. We believe, however, that [68] makes this claim more formal and mathematically precise, and while limited for now to a class of systems arguably smaller than the one used by the free energy principle, within this class the claim applies to systems under quite general conditions.

The second consequence is a more coherent treatment of actions. In general, determining whether an agent's actions can be formally regarded as the consequence of causal interventions can be conceptually and formally challenging. This is because there is a difference between an action that *is* produced by an intervention, and an action that can be treated *as-if* it was produced by an intervention. In Pearl's structural causal models for instance, the first perspective implicitly assumes that an agent's output, its actions, are "local surgeries" performed by some interventions, i.e. one has to assume that some other entity (another agent?) implements these interventions as manipulations of a model via the do-operator [99]. In other words, under this interpretation some entity's actions (for instance a scientist's) will correspond to interventions on a model, $\mathcal{A} = \{\text{interventions on model } \mathcal{M}\}$. The role of this entity is however unclear: it has to somehow replace a variable's structural equation and severe its causal dependencies so that an agent can perform an action. The second perspective just re-describes the process of some agent choosing its own actions as an intervention, with its policy, i.e. the map from states/beliefs to actions, seen as soft or policy intervention [100], one where not all causal dependencies are severed (to allow for dependencies on the agent's previous states/beliefs and actions) [101]. This perspective doesn't need to introduce extra entities, but in some sense it could be said that it simply re-describes existing knowledge using ideas from causal reasoning. It is however non-trivial: this less ontologically committed perspective allows us to consider also cases where actions cannot in fact be seen as some kind of intervention. For instance, in off-line/off-policy reinforcement learning, where observations can be originated by something other than the agent itself (a teacher, or nature itself) and thus causal interventions by an agent don't really give a natural language to describe policies it generates [102, 100].

An aspect we mostly overlooked in this work is that mathematical theories of agents ought to capture agency that can appear at different levels of abstraction. This ties directly with theories of *emergence* describing intuitions and formal characterisations of how certain properties and functions can appear at some scale while not being well defined at some other. The idea of emergence appears in several fields [103] and while in some cases it could serve for a definition of individuality that is relevant for agents [73, 74, 104, 75, 105], it doesn't necessarily speak to individuality as induced by the relation between an agent and its environment, in the way we phrased individuality in Section 3 for coupled systems. On the other hand, one could interpret emergence to be more closely aligned to the prediction-based methods used to describe normativity in Section 4, assuming that different kinds of coarse-grainings necessary to define some form of emergence can be derived from objectives that describe the predictive power of some meso or macro scale of a system (by assuming, or not, some level of supervenience) [104, 75, 105]. This exploration is however left to future work.

7. Conclusion

In this work, we started from the standard definition of life provided by NASA, "a self-sustaining chemical system capable of undergoing Darwinian evolution" [3] and sought to explore how the first half of this definition, "a self-sustaining chemical system", can be abstracted and mathematically formalised in the study of life as it could be, ALife, and other related fields. We suggested the study of agents as a step in this direction, and initially operationalised agents as goal-directed systems acting in an environment. Building on a conceptual framework that identifies individuality, normativity or goal-directedness, and interactional asymmetry as key requirements of a theory of agents, we then analysed and reframed different mathematical approaches under a shared formal notation, enabling a more direct comparison of different proposals, their strengths and possible missing parts. We then saw how adopting an as-if stance on agency resolves a few conceptual tensions and reveals a deep link between individuality and normativity: an agent's identity strongly depends on its goal-directed behaviour and vice-versa, a relation shown so far formally for a specific class of systems but likely to generalise. Within this as-if perspective, actions can be interpreted as produced by interventions that capture the causal asymmetry between agent and environment. This perspective allows us to evade questions about who or what, if not the agent, ought to formally implement interventions in a more ontologically committed interpretation of actions in the language causal reasoning. Overall, this synthesis exposes complementary strengths and gaps among existing frameworks, discussing mathematical and philosophical perspectives of agency and outlining a path toward an integrated, mathematically explicit theory of agents based on a common conceptual framing.

Acknowledgments

This manuscript is based on M.B.'s lectures for the 2023 CHAIN Winter School on "Minimal Cognition and Agency", organised by K.S, and on a lecture at the "Life As We Don't Yet Know It," Breakthrough Discuss 2025, in the session on "Forms of Non-Terrestrial Life," by M.B. The current material reflects a revised, improved and edited version, with K.S., that makes use of feedback received in these and other events. The authors would also like to thank in particular Martin Biehl for feedback on a previous version of the manuscript. M.B. was supported by JST, Moonshot R&D, Grant Number JPMJMS2012. K.S. was supported by JST, CREST Grant Number JPMJCR21P4 and JSPS KAKENHI Grant Numbers 24H01534, Japan.

References

[1] Christopher Langton. *Artificial Life: Proceedings Of An Interdisciplinary Workshop On The Synthesis And Simulation Of Living Systems*. Routledge, New York, 1989.

[2] John Maynard Smith. Byte-sized evolution. *Nature*, 355(6363), 1992.

[3] Steven A. Benner. Defining Life. *Astrobiology*, 10(10):1021–1030, December 2010.

[4] Martin A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard university press, 2006.

[5] David E. Goldberg. *The Design of Innovation*, volume 7 of *Genetic Algorithms and Evolutionary Computation*. Springer US, Boston, MA, 2002.

[6] Christoph Adami. *Introduction to Artificial Life*. Springer Science & Business Media, 1998.

[7] Richard E. Lenski, Charles Ofria, Robert T. Pennock, and Christoph Adami. The evolutionary origin of complex features. *Nature*, 423(6936):139–144, May 2003.

[8] Charles Ofria and Claus O. Wilke. Avida: A Software Platform for Research in Computational Evolutionary Biology. *Artificial Life*, 10(2):191–229, 2004.

[9] Thomas Ray. An Approach to the Synthesis of Life. In *Proceedings of Artificial Life II (p. 371).* Reading, MA: Addison-Wesley, 1991.

[10] Bert Wang-Chak Chan. Lenia: Biology of Artificial Life. *Complex Systems*, 28(3), 2019.

[11] Erwan Plantec, Gautier Hamon, Mayalen Etcheverry, Bert Wang-Chak Chan, Pierre-Yves Oudeyer, and Clément Moulin-Frier. Flow-Lenia: Emergent Evolutionary Dynamics in Mass Conservative Continuous Cellular Automata. *Artificial Life*, 31(2):228–248, May 2025.

[12] Hiroki Sayama and Chrystopher L. Nehaniv. Self-Reproduction and Evolution in Cellular Automata: 25 Years After Evoloops. *Artificial Life*, 31(1):81–95, February 2024.

[13] Carlo Rovelli. Agency in Physics, July 2020.

[14] Manuel Baltieri and Ryota Kanai. Steam-engine naturalism, October 2025.

[15] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, 3rd edition, 2009.

[16] Sameera Abar, Georgios K. Theodoropoulos, Pierre Lemarinier, and Gregory M.P. O'Hare. Agent Based Modelling and Simulation tools: A review of the state-of-art software. *Computer Science Review*, 24:13–33, May 2017.

[17] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B. Divya. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access*, 13:18912–18936, 2025.

[18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018.

[19] Karl J. Åström and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton Univ. Press, Princeton, NJ, 2008.

[20] Gertrude Elizabeth Margaret Anscombe. *Intention*. Harvard University Press, 1963.

[21] Donald Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.

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

[23] Luca Ferrero. An introduction to the philosophy of agency. In *The Routledge Handbook of Philosophy of Agency*, pages 1–18. Routledge, 2022.

[24] Patrick Haggard. Sense of agency in the human brain. *Nature Reviews Neuroscience*, 18(4):196–207, April 2017.

[25] Keisuke Suzuki, Peter Lush, Anil K. Seth, and Warrick Roseboom. Intentional Binding Without Intentional Action. *Psychological Science*, 30(6):842–853, June 2019.

[26] Wen Wen and Hiroshi Imamizu. The sense of agency in perception, behaviour and human–machine interactions. *Nature Reviews Psychology*, 1(4):211–222, April 2022.

[27] Michael Tomasello. *The Evolution of Agency: Behavioral Organization from Lizards to Humans*. MIT Press, 2022.

[28] Humberto R Maturana and Francisco J Varela. *Autopoiesis and Cognition: The Realization of the Living*. Springer Science & Business Media, 1980.

[29] Martin Biehl, Takashi Ikegami, and Daniel Polani. Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. In *ALIFE 2016, the Fifteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 722–729. MIT Press, July 2016.

[30] Phillip Ball. Organisms as agents of evolution. Technical report, John Templeton Foundation: West Conshohocken, PA, USA, 2023.

[31] Mustafa Emirbayer and Ann Mische. What Is Agency? *American Journal of Sociology*, 103(4):962–1023, January 1998.

[32] Stephen A. Ross. The Economic Theory of Agency: The Principal's Problem. *The American Economic Review*, 63(2):134–139, 1973.

[33] Roderick Munday. *Agency: Law and Principles*. OUP Oxford, 2010.

[34] Randall D Beer. Some historical context for minimal cognition. *Adaptive Behavior*, 29(1):89–92, February 2021.

[35] Xabier E. Barandiaran, Ezequiel Alejandro Di Paolo, and Marieke Rohde. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior*, 17(5):367–386, October 2009.

[36] Simon McGregor. A More Basic Version of Agency? As If! In *Lecture Notes in Computer Science*, pages 183–194, 2016.

[37] Simon McGregor. The Bayesian stance: Equations for 'as-if' sensorimotor agency. *Adaptive Behavior*, 25(2):72–82, 2017.

[38] Judea Pearl. *Causality - Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[39] Randall D. Beer. A dynamical systems perspective on agent-environment interaction. *Artificial intelligence*, 72(1-2):173–215, 1995.

[40] Randall D. Beer. The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, 20(2-4):257–289, June 1997.

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

[42] Fernando Rosas, Alexander Boyd, and Manuel Baltieri. AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability. In *Proceedings of the Second Reinforcement Learning Conference*, April 2025.

[43] David Jaz Myers. *Categorical Systems Theory*. 2021.

[44] Matteo Capucci. Notes on categorical systems theory. Technical report, 2024.

[45] Randall D. Beer. Autopoiesis and Cognition in the Game of Life. *Artificial Life*, 10(3):309–326, June 2004.

[46] Randall D. Beer. The cognitive domain of a glider in the game of life. *Artificial Life*, 20(2):183–206, 2014.

[47] Randall D. Beer. Characterizing Autopoiesis in the Game of Life. *Artificial Life*, 21(1):1–19, February 2015.

[48] Randall D Beer. Bittorio revisited: Structural coupling in the Game of Life. *Adaptive Behavior*, 28(4):197–212, August 2020.

[49] Randall D. Beer. An integrated perspective on the constitutive and interactive dimensions of autonomy. In *Artificial Life Conference Proceedings 32*, pages 202–209. MIT Press, 2020.

[50] Randall D. Beer. An investigation into the origin of autopoiesis. *Artificial Life*, 26(1):5–22, 2020.

[51] Martin Biehl and Nathaniel Virgo. The game of life in a glider's frame of reference. 2023.

[52] Karl J. Friston. A free energy principle for a particular physics, June 2019.

[53] Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, March 2022.

[54] Karl J. Friston, Lancelot Da Costa, Noor Sajid, Conor Heins, Kai Ueltzhöffer, Grigorios A. Pavliotis, and Thomas Parr. The free energy principle made simpler but not too simple. *Physics Reports*, 1024:1–29, June 2023.

[55] Jelle Bruineberg, Krzysztof Dołęga, Joe Dewhurst, and Manuel Baltieri. The Emperor's New Markov Blankets. *Behavioral and Brain Sciences*, 45:e183, 2022.

[56] Jelle Bruineberg, Krzysztof Dołęga, Joe Dewhurst, and Manuel Baltieri. The Emperor Is Naked: Replies to commentaries on the target article. *Behavioral and Brain Sciences*, 45, 2022.

[57] Martin Biehl, Felix A. Pollock, and Ryota Kanai. A Technical Critique of Some Parts of the Free Energy Principle. *Entropy*, 23(3):293, February 2021.

[58] Nathaniel Virgo, Fernando E. Rosas, and Martin Biehl. Embracing sensorimotor history: Time-synchronous and time-unrolled Markov blankets in the free-energy principle. *Behavioral and Brain Sciences*, 45:e215, 2022.

[59] Fernando E. Rosas, Pedro A.M. Mediano, Martin Biehl, Shamil Chandaria, and Daniel Polani. Causal blankets: Theory and algorithmic framework. In *International Workshop on Active Inference*, pages 187–198. Springer, 2020.

[60] Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970.

[61] W. M. Wonham. Towards an Abstract Internal Model Principle. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):735–740, November 1976.

[62] Lancelot Da Costa, Karl J. Friston, Conor Heins, and Grigorios A. Pavliotis. Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2256):20210518, December 2021.

[63] Nathaniel Virgo, Martin Biehl, and Simon McGregor. Interpreting Dynamical Systems as Bayesian Reasoners. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, December 2021.

[64] Miguel Aguilera, Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. How particular is the physics of the free energy principle? *Physics of Life Reviews*, 40:24–50, March 2022.

[65] Ezequiel Alejandro Di Paolo, Evan Thompson, and Randall D Beer. Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, 2022.

[66] Vicente Raja, Dinesh Valluri, Edward Baggs, Anthony Chemero, and Michael L. Anderson. The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, 39:49–72, December 2021.

[67] Daniel C. Dennett. *The Intentional Stance*. A Bradford Book. MIT Press, 1989.

[68] Nathaniel Virgo, Martin Biehl, Manuel Baltieri, and Matteo Capucci. A "good regulator theorem" for embodied agents, August 2025.

[69] Manuel Baltieri, Martin Biehl, Matteo Capucci, and Nathaniel Virgo. A Bayesian Interpretation of the Internal Model Principle, March 2025.

[70] Scott Garrabrant, Daniel A. Herrmann, and Josiah Lopez-Wild. Cartesian Frames, September 2021.

[71] Manuel Baltieri, Christopher L. Buckley, and Jelle Bruineberg. Predictions in the eye of the beholder: An active inference account of Watt governors. In *The 2020 Conference on Artificial Life*, pages 121–129, Online, 2020. MIT Press.

[72] David Krakauer, Nils Bertschinger, Eckehard Olbrich, Jessica C. Flack, and Nihat Ay. The information theory of individuality. *Theory in Biosciences*, 139(2):209–223, June 2020.

[73] Nils Bertschinger, Eckehard Olbrich, Nihat Ay, and Jürgen Jost. Autonomy: An information theoretic perspective. *Biosystems*, 91(2):331–345, February 2008.

[74] Anil K. Seth. Measuring Autonomy and Emergence via Granger Causality. *Artificial Life*, 16(2):179–196, April 2010.

[75] Lionel Barnett and Anil K. Seth. Dynamical independence: Discovering emergent macroscopic processes in complex dynamical systems. *Physical Review E*, 108(1):014304, July 2023.

[76] Laurent Orseau, Simon McGregor McGill, and Shane Legg. Agents and devices: A relative definition of agency, 2018.

[77] Matt MacDermott, James Fox, Francesco Belardinelli, and Tom Everitt. Measuring Goal-Directedness. In *Advances in Neural Information Processing Systems*, volume 27, 2024.

[78] Martin Biehl. *Formal Approaches to a Definition of Agents*. PhD thesis, University of Hertfordshire, 2018.

[79] Martin Biehl, Takashi Ikegami, and Daniel Polani. Specific and Complete Local Integration of Patterns in Bayesian Networks. *Entropy*, 19(5):230, May 2017.

[80] R Rosen. On anticipatory systems: I. When can a system contain a predictive model of another? *Journal of Social and Biological Systems*, 1(2):155–162, April 1978.

[81] R Rosen. On anticipatory systems: II. The nature of the modelling relation between systems. *Journal of Social and Biological Systems*, 1(2):163–180, April 1978.

[82] Aloisius H. Louie. Mathematical Foundations of Anticipatory Systems. In Roberto Poli, editor, *Handbook of Anticipation*, pages 1–29. Springer International Publishing, Cham, 2017.

[83] Aloisius H. Louie. *More Than Life Itself: A Synthetic Continuation in Relational Biology*. DE GRUYTER, December 2009.

[84] Fosco Loregian. Rosen's no-go theorem for regular categories, May 2021.

[85] Elias Zafiris. Rosen's modelling relations via categorical adjunctions. *International Journal of General Systems*, 41(5):439–474, July 2012.

[86] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, 2014.

[87] Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjørn E. Juel, Shuntaro Sasai, Keiko Fujii, Isaac David, Jeremiah Hendren, Jonathan P. Lang, and Giulio Tononi. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*, 19(10):e1011465, October 2023.

[88] Miguel Aguilera, Carlos Alquézar, and Manuel G. Bedia. Agency and Integrated Information in a Minimal Sensorimotor Model. In *The 2018 Conference on Artificial Life*, pages 396–403, Tokyo, Japan, 2018. MIT Press.

[89] Miguel Aguilera and Ezequiel Alejandro Di Paolo. Integrated information in the thermodynamic limit. *Neural Networks*, 114:136–146, June 2019.

[90] Larissa Albantakis, William Marshall, Erik Hoel, and Giulio Tononi. What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, 21(5):459, May 2019.

[91] Larissa Albantakis, Francesco Massari, Maggie Beheler-Amass, and Giulio Tononi. A Macro Agent and Its Actions. In *Top-down Causation and Emergence*, pages 135–155. Springer, 2021.

[92] Miguel Aguilera and Ezequiel Alejandro Di Paolo. Integrated Information and Autonomy in the Thermodynamic Limit. In *The 2018 Conference on Artificial Life*, pages 113–120, Tokyo, Japan, 2018. MIT Press.

[93] Bjørn Erik Juel, Renzo Comolatti, Giulio Tononi, and Larissa Albantakis. When is an action caused from within? Quantifying the causal chain leading to actions in simulated agents. In *Artificial Life Conference Proceedings 31*, 2019.

[94] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *Artificial Intelligence*, 322:103963, September 2023.

[95] Artemy Kolchinsky and David H. Wolpert. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6):20180041, December 2018.

[96] Filippo Torresan and Manuel Baltieri. Disentangled representations for causal cognition. *Physics of Life Reviews*, 51:343–381, December 2024.

[97] H.H. Morch. Is Consciousness Intrinsic? A Problem for the Integrated Information Theory. *Journal of Consciousness Studies*, 26(1-2):133–162, January 2019.

[98] Anil K Seth. The Cybernetic Bayesian Brain. In Wanja Wiese and Thomas K Metzinger, editors, *Open MIND*, pages 9–24. Frankfurt am Main, Germany: MIND Group, 2015.

[99] Judea Pearl. Causation, Action, and Counterfactuals. In Maria Luisa Dalla Chiara, Kees Doets, Daniele Mundici, and Johan Van Benthem, editors, *Logic and Scientific Methods*, pages 355–375. Springer Netherlands, Dordrecht, 1997.

[100] Elias Bareinboim, Junzhe Zhang, Sanghack Lee, and Snu Ac Kr. An Introduction to Causal Reinforcement Learning, 2021.

[101] Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal Reinforcement Learning: A Survey, November 2023.

[102] Oliver Schulte and Pascal Poupart. When Should Reinforcement Learning Use Causal Reasoning? *Transactions on Machine Learning Research*, 2025.

[103] Ross H. McKenzie. Emergence: From physics to biology, sociology, and computer science, August 2025.

[104] Erik P. Hoel. Agent Above, Atom Below: How Agents Causally Emerge from Their Underlying Microphysics. In Anthony Aguirre, Brendan Foster, and Zeeya Merali, editors, *Wandering Towards a Goal*, pages 63–76. Springer International Publishing, Cham, 2018.

[105] Fernando E. Rosas, Bernhard C. Geiger, Andrea I. Luppi, Anil K. Seth, Daniel Polani, Michael Gastpar, and Pedro A. M. Mediano. Software in the natural world: A computational approach to hierarchical emergence, June 2024.