Thermina: A minimal model of autonomous agency from the lens of stochastic thermodynamics

Miguel Aguilera[1,2,*] and Xabier E. Barandiaran [3]

[1]BCAM – Basque Center for Applied Mathematics, Bilbao, Spain
[2]IKERBASQUE, Basque Foundation for Science. Bilbao, Spain
[*]sci@maguilera.net
[3]IAS-Research Centre for Life, Mind, and Society, Department of Philosophy, University of the Basque Country (UPV/EHU), Spain

Abstract

We introduce a minimal model of a thermodynamic agent capable of maintaining far-from-equilibrium states by actively harvesting and storing free energy from its environment. Inspired by minimal models of autonomy like *Bittorio* (Varela et al., 1991), our agent —labelled *Thermina*— gives shape to a theoretical framework for studying the interplay between thermodynamics and autonomy. By analytically studying the nonequilibrium steady state of the system, we distinguish between regions of 'autonomous' states —sustaining themselves out-of-equilibrium by harvesting free energy from the environment— and regions of 'non-autonomous' states —close to thermodynamic equilibrium and with very low chances of gathering free energy. Furthermore, we inspect the adaptive mechanisms that allow an agent to regulate its interaction with the environment to robustly maintain its nonequilibrium state. Studying in detail the behaviour of the system, we aim to provide insights into the broader question of how thermodynamic processes contribute to the emergence and maintenance of complex, adaptive behaviour in natural and artificial systems.

Introduction

A central notion to Artificial Life (also to Cognitive Science and, more recently, Artificial Intelligence) is that of *autonomous agency* (Varela and Bourgine, 1992): to which extent a system is a genuine source of its own actions and goals, of its own norms? How independent does it become from the original intent of its designers (or evolution) and increasingly self-determined by its own active history and emergent normativity? It has been argued that autonomous agency is central (perhaps *the* central property) to living and cognitive systems (Varela, 1979; Moreno and Mossio, 2015; Di Paolo et al., 2017). Moreover, autonomous agency is perceived as holding the potential and danger of AI systems detaching from human interest and getting out of control (the alignment problem, Bostrom, 2017; Russell, 2019). Increasingly, scientists and experts are concerned that AI capabilities may advance at a rate which outstrips humanity's collective capacity to adapt to their transformative impacts, including the alleged existential risk it might involve (Roose, 2023). Moreover, talk of 'autonomous agency' in digital scenarios is widespread and considered by many as the new big thing in the 'AI revolution', and there are increasing attempts to expand the recent success of LLM (Large Language Models) to build autonomous agents (Liu et al., 2023; Weng, 2023; Andreas, 2022). However, if the current success of AI is due to systems that are vastly distributed, fundamentally impenetrable to human representational understanding (Preece, 2018), and full of emergent capacities (Wei et al., 2023), how can we assess their autonomy? We believe that many new trends in AI development, future potential and its perceived risks overlap with progress in the understanding of autonomy in bioinspired intelligence, dynamical systems and biological physics, and philosophy. While there is little interaction between these communities to explore the full potential of the concept in a unified and interdisciplinary effort, Artificial Life is perfectly suited to provide bridge tools and concepts. In this contribution, to aim at approaching autonomy from an angle that has often been neglected or assumed as a departure point to be simply left aside once stated: the thermodynamic dimension of autonomy.

Thermodynamic autonomy The ability to maintain organized out-of-equilibrium states is a central requisite for autonomous agents (Wiener, 1965; Kauffman, 2000; Ruiz-Mirazo and Moreno, 2004; Longo and Montévil, 2014). Metabolically speaking, (living) autonomous systems are sophisticated far-from-thermodynamic-equilibrium systems, whose organization is partially explained as a result of self-organizing dissipative structures. Energy is channelled into work that produces constraints (enzymes, membranes, etc.) that is used to produce more work, in a continuous process of keeping chemical processes far-from-equilibrium. But this thermodynamic requirements can also be explored in informational terms, beyond chemistry. For example, the capacities of agents in interaction with their environment (prediction, memory, control) are bounded by out-of-equilibrium quantities in information thermodynamics (Still, 2020; Hartich et al., 2015; Barato et al., 2014).Understanding such out-of-equilibrium properties is of fundamental importance to understanding what are the intrinsic

universal cognitive-informational principles and limits of a given system to behave as an autonomous agent.

Autonomy, in the context of nonequilibrium thermodynamics, requires systems the capacity to recursively maintain dynamic, organized states in far-from-equilibrium conditions (Bickhard, 2000). Unlike processes at thermodynamic equilibrium, which are stable without external input of energy, far-from-equilibrium processes require constant energy input to resist entropy's inexorable march toward equilibrium. A candle flame is an example of far-from-equilibrium stability, which persists only as long as fuel and oxygen are continuously supplied, illustrating the inherently open nature of such systems. In contrast, 'energy well stability' pertains to systems that remain stable in the absence of external energy inputs. For example, the magnetic moments in a ferromagnetic material at low temperatures can align in the same direction and maintain such state, as long as thermal energy from the ambient environment does not disrupt its organization. Unlike the candle flame, autonomous systems present a stronger capacity for self-maintenance, they display *recursive* interactions with the environment carrying information about the conditions of its own nonequilibrium state (Bickhard, 2000). For example, a bacterium that performing chemotaxis reacts to its environment to take the actions that guarantee its metabolism to exist at a nonequilibrium state.

Autonomous agency In the past, minimal models of autonomous agency in bio-cognitive systems have often involved either some explicit reference to metabolic dynamics and their modelling in terms of chemical reaction rates (Ruiz-Mirazo et al., 2004; Barandiaran and Egbert, 2014) or embodied sensorimotor (neuro)dynamics (Ruppin, 2002; Beer, 1990; Aguilera, 2015). These models are meant to be models of agency in virtue of how their variables capture some aspect of real existing agent (or generalizations of such systems).

However, there exists another category of models aimed at capturing autonomous properties in a more abstract and general-purpose fashion, without being bound to a specific domain. One such model is *Bittorio*, initially introduced by Varela (1988) and later incorporated into his seminar work 'The Embodied Mind' 1991. Bittorio is a cellular automaton immersed in a milieu of random binary states (0's and 1's), akin to a cell in a chemical environment. When cells of the automata encounters these states, its own state is replaced by the encountered perturbation. Depending on its specific state and the rules governing Bittorio's behaviour, perturbations (either small or large) may either fully divert the course of its dynamics, or converge to the initial state or attractor. This complex process illustrates Bittorio's structural coupling with its environment, i.e., how it compensates for perturbations and adapts its dynamics accordingly. A similar abstract conceptual model has been analysed in detail by (Beer, 2015, see also previous papers) taking the *glider* in the Game of Life as an 'artificial model organism' to study the dynamics of autopoietic (i.e. autonomous) agents . More abstract mathematical models of autonomy and agency have also been developed, focusing primarily on informational and statistical mechanical aspects of autonomy (Bertschinger et al., 2008; Seth, 2010; Kolchinsky and Wolpert, 2018), conceptualized as measurements of autonomy that can be applied to different specific models. However, there is, to our knowledge, no abstract model that approaches autonomous agency from a thermodynamic point of view.

From a minimalist and most general conceptual level, (deriving from previous formulations like Varela, 1979; Barandiaran et al., 2009) an autonomous system requires at least:

1. A network of processes such that states of component processes recursively depend on each other.

2. Component processes are *precarious*, they would tend to decay and extinguish in the absence of an intervention

3. Interactions between processes compensate for the intrinsic decay of the components.

4. The network as a whole distinguishes from its environment (i.e. with states or processes that do not depend on the activity of the system) from which it also depends on to compensate from its inherent decay.

Autonomous systems are *interactive* when such a network is bidirectionally coupled with its environment and directs energy on this interaction (i.e. it is not simply passively feeding on it). *Autonomous agency*, in turn, requires that the system somewhat regulate its coupling in relation to its viability (Di Paolo, 2005; Barandiaran et al., 2009).

Previous models have tried to satisfy such requirements with references (however vague) to the material specification of the networked processes (neurodynamic, metabolic, immune, etc.). In this paper, we propose a model of maximal generality and minimal complexity based on thermodynamic principles. This is particularly relevant to address the notion of decay, self-maintenance, and the needful relationship with the environment. The precarious nature of autonomous systems is central to explore the nature of norms and the proper sense of individuality that characterized autonomous systems: the system *has* to do certain things in order to survive. However, this precariousness has often been simply assumed, reduced to a single variable (e.g. energy level of a robot), or dynamically encoded without further thermodynamic analysis. The goal of this model is to fill this gap. We start introducing *Thermina*, our model, then we provide different thermodynamic interpretations of its functioning. Next, we carry out a set of experiments to identify autonomous system dynamics and autonomous adaptive

agency on the model. We end up with some general evaluation of the model, its possible extensions and potential future applications.

Thermina: a thermodynamic model of autonomous agency

A minimal model of thermodynamics for autonomous systems In order to capture the energy exchanges of self-organizing dissipative structures, we describe a system driven by an energy function $E(\boldsymbol{s})$, being $\boldsymbol{s}$ its internal state. Physically, $E(\boldsymbol{s})$ can encapsulate the total energy of the system, including contributions from sources such as kinetic energy, potential energy, or other forms. In statistical physics, the energy function determines the probability of states through the Boltzmann distribution, which establishes a connection between energy and probability of states in a system at thermal equilibrium, $p(\boldsymbol{s}) = Z^{-1} \exp\left(-\beta E(\boldsymbol{s})\right)$, where $\beta = (kT)^{-1}$ is an inverse temperature (being k Boltzmann's constant). More generally, in the context of statistical models—e.g. energy based models (LeCun et al., 2006), generalized linear models (Nelder and Wedderburn, 1972), and more generally models in the exponential family (Amari, 2016)— the energy function $E(\boldsymbol{s})$ is often designed such that lower energy values are associated with more plausible or desirable states, while higher energy values are associated with less plausible or desirable states, which are more representative of the data distribution being modelled. The energy function typically serves as the objective function in optimization tasks, where the best fit of the model to the observed data minimizes the average energy.

The model (Fig. 1) is a system capable of microscopic feedback control (i.e., a Maxwell demon), consisting of a rotating dial interacting with a thermal reservoir, a (dimensionless) drag force f, and a 'soup' of bits characterizing an 'energy currency' used to act against the drag. The array of bits is described as $\boldsymbol{a} = \{a_1, \ldots, a_M\}$, with $a_i \in \{0, 1\}$, and serves as a 'battery' for the agent. When $a_i = 1$, the unit contains an amount of energy μ (energy 0 otherwise) which can be used by the agent. Energy units can be placed at a slot in a machine, where converting the state of a_i from $1 \to 0$ activates the state of a variable x from $0 \to 1$. This uses the energy unit μ to rotate the dial clockwise, pulling against the drag force f (the total energy difference is $\mu - f$). Relaxing the state of x from $1 \to 0$ again dissipates an energy f. The opposite happens when x transitions from $1 \to 0$. As we will see later, when situated in a specific environment, the model shows how an agent uses energy to interact with its surroundings. For example, an agent in a viscous medium can overcome friction by expending internal energy, thereby dissipating heat. The agent can sustain continuous movement only if it remains in an out-of-equilibrium state, consistently acquiring and dissipating energy through its actions. Here energy terms μ, f can represent physical quantities — e.g. chemical reaction rates of cellular metabolism. But, more generally, these quantities parameterize transition rates determining the stochastic thermodynamics of probabilistic dynamical systems (Peliti and Pigolotti, 2021).

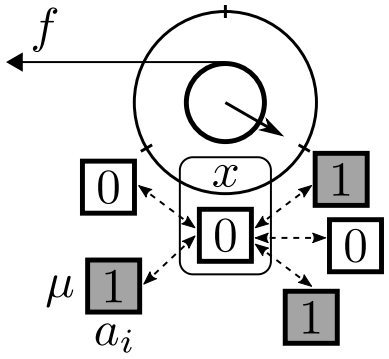

Figure 1: Floating energy units (bits a_i) carrying a free energy μ, can either spontaneously degrade or switch their state with the bit in the machine slot x, rotating the machine (counter-)clockwise when the bit x switches from $0 \to 1$ ($1 \to 0$), pulling against a drag force f.

The total energy of the system is defined by:

$$E(x, A) = fx + \mu A, \tag{1}$$

where $A = \sum_i a_i$ is the energy stored in $\boldsymbol{a}$. We explicitly include drag f in the Energy function—in contrast with non-conservative loads often assumed in molecular motors (Thomas et al., 2001; Schmiedl and Seifert, 2008)—to explicitly observe how the agent spends its surplus energy to compensate dissipation (e.g. due to friction or viscosity).

The dynamics of the model (Fig. 2) can be derived from this energy function, assuming detailed balance condition $k_{\boldsymbol{s}',\boldsymbol{s}} p(\boldsymbol{s}) = k_{\boldsymbol{s},\boldsymbol{s}} p(\boldsymbol{s}')$, being $k_{\boldsymbol{s}',\boldsymbol{s}}$ the rate of jumps from state $\boldsymbol{s}$ to state $\boldsymbol{s}'$ and $p(\boldsymbol{s}')$ the equilibrium distribution.

The value of energy units spontaneously degrade, with bit flipping rates following a ratio between $0 \to 1$ and $1 \to 0$ transitions,

$$\frac{k_a^+}{k_a^-} = \exp\left(-\beta\mu\right). \tag{2}$$

This implies e.g. spontaneous degradation of energy units.

Similarly, variables x, a_i can exchange their values, with rations between $(0, 1) \to (1, 0)$ and $(1, 0) \to (0, 1)$ transitions defined as

$$\frac{k_{xa}^+}{k_{xa}^-} = \exp\left(\beta(\mu - f)\right). \tag{3}$$

Such transitions imply acting against a drag f, e.g. displacing an agent in the opposite direction.

Finally, the value of x can spontaneously flip to $1 - x$, with a ratio of rates of transitions from $0 \to 1$ and $1 \to 0$

$$\frac{k_x^+}{k_x^-} = \exp\left(-\beta f\right). \tag{4}$$

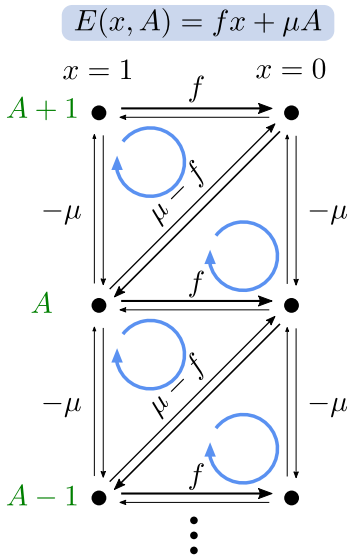

Figure 2: Transition rates in the model defined by detailed balance relations and an energy function in Eq. 1. Labels in arrows define the energy difference in transitions for the direction depicted by the blue cycle arrow.

These transitions account for entropy dissipation, e.g. when heat generated by the drag dissipates.

Taken together, these equations describe a 'bit-eating' engine acting upon a frictional force. As rates in Eqs. 2-4 meet detailed balance, the model will converge to an equilibrium state with reversible dynamics. In equilibrium, the system converges to a Gibbs distribution $p(x, A) = Z^{-1} \exp(-\beta E(x, A))$, with an average value of A is $\langle A \rangle_{\text{eq}} = \frac{1}{2}(1 - \tanh(\beta\mu))$, and average value of x being $\langle x \rangle_{\text{eq}} = \frac{1}{2}(1 - \tanh(\beta J))$. When the value of $A > \langle A \rangle_{\text{eq}}$ the engine will tend to move clockwise (and counterclockwise when $A < \langle A \rangle_{\text{eq}}$).

Interpretation and relation with other models Eq. 4 in our model is related with information engines such as the Mandal-Jarzynski machine (Mandal and Jarzynski, 2012) or the Bennett-Feynmann engine (Feynman, 2018, pp. 146-147). However, in such engines the 'fuel' —an array of bits— is fed into the machine by a frictionless moving tape. In our case, we assume that bits allowing the engine to move are the result of variables in the system with their own stochastic dynamics resulting from the Energy function in Eq. 1, pushing against a drag force (Eq. 3).

In contrast with such information engines, our model is driven by energy currencies units, which can account for the dynamics driving molecular systems, such as molecular motors (Schmiedl and Seifert, 2008; Thomas et al., 2001). For example, Eq. 2 can capture the degradation of adenosine triphosphate (ATP) into adenosine diphosphate (ADP) and inorganic phosphate in a reaction ATP $\rightleftharpoons$ ADP + P,

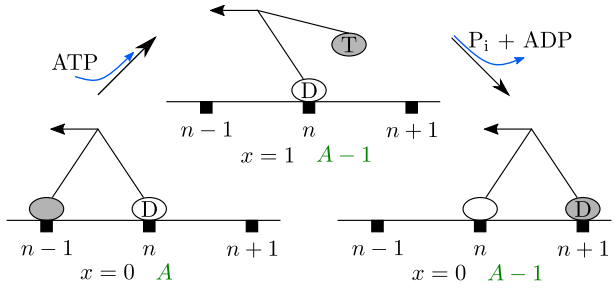

Figure 3: Example of a molecular motor: myosin molecules generate force through a power stroke mechanism, fuelled by the energy released from ATP hydrolysis (Thomas et al., 2001).

being $\mu \equiv \mu_{\text{ATP}} - \mu_{\text{ADP}} - \mu_{\text{P}}$ the difference of chemical potentials associated to each molecule. In addition, Eq. 3 can account for the action of a molecular motor (Fig. 3) harnessing free energy from chemical sources to be converted into mechanical work (Thomas et al., 2001, Eq. 3.7) . Such molecular motors play pivotal roles in essential cellular processes, including intracellular transport, cell division, and muscle contraction.

System-environment coupling: thermodynamic work extraction for sustained motion In order to reach a nonequilibrium steady state, we design an environment where the agent can harvest free energy and use it to sustain an out-of-equilibrium steady-state. We situate the system in a periodic ring of length $N + 1$ (Fig. 4). When the agent is positions $n > 0$, the dynamics are updated according to the rates described in Eqs. 2-4. As in a molecular motor (Fig. 3), the agent can use energy currencies to advance. Thus, each time the agent undergoes a transition of (x, a_i) from $(0, 1) \to (1, 0)$ (or $(1, 0) \to (0, 1)$), associated with rates k^-_{xa} (k^+_{xa}), the agent moves forward to $n + 1$ (or backwards, to $n-1$). Position $n = 0$ acts like a 'charger', injecting energy into the system. When the agent is in position $n = 0$, it stops and transitions to the state $x = 0, a_i = 1, \forall i$, then moving forward to $n = 1$ transforming x from $1 \to 0$, driving the system to a nonequilibrium high-energy state and letting it relax until $n = 0$ again.

Results

In this section, we explore the behaviour of the model for parameters of energy content of units a_i, $\mu = 1$, drag $f = 1$, and an inverse temperature $\beta = 0.5$. We select an environment size $N = 20$, and number of particles a_i, $M = 50$.

Nonequilibrium steady-state For simplicity, we simulate the dynamics of the system as Glauber transition rates, that is, for each transition between states $\boldsymbol{s} = \{x, \boldsymbol{a}, n\}$. We describe transitions to states $\boldsymbol{s}^{a0}, \boldsymbol{s}^{a1}, \boldsymbol{s}^{xa}, \boldsymbol{s}^{x}$ as transitions

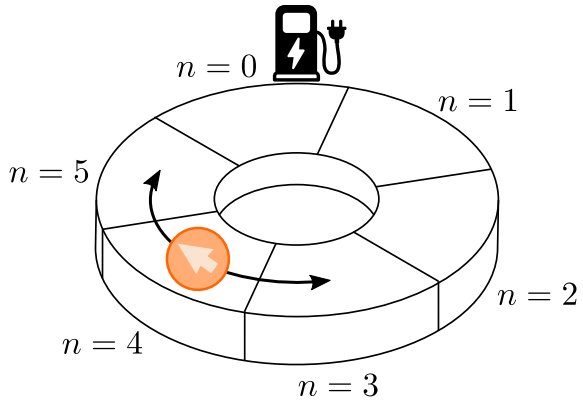

Figure 4: Representation of the agent's environment as a one-dimensional ring of length $N = 6$, where position $n = 0$ recharges energy units to values $a_i = 1$, and the agent is depicted at $n = 4$.

changing the state to $\{x, A-1, n\}, \{x, A+1, n\}, \{1-x, A+2x-1, n-2x+1\}, \{1-x, A, n\}$ respectively, following rations in Eqs. 2-4. Given the ratios between rates, Glauber rates are defined as

$$k_{\boldsymbol{s}^{ab},\boldsymbol{s}} = R_a F_b \left(1 + \exp\left(\beta \Delta E(\boldsymbol{s}^{ab})\right)\right)^{-1} \quad (5)$$

$$k_{\boldsymbol{s}^{xa},\boldsymbol{s}} = R_{xa} F_x \left(1 + \exp\left(\beta \Delta E(\boldsymbol{s}^{xa})\right)\right)^{-1} \quad (6)$$

$$k_{\boldsymbol{s}^{x},\boldsymbol{s}} = R_x \left(1 + \exp\left(\beta \Delta E(\boldsymbol{s}^{x})\right)\right)^{-1} \quad (7)$$

with $\Delta E(\boldsymbol{s}^{\alpha}) = E(\boldsymbol{s}) - E(\boldsymbol{s}^{\alpha})$ the change in energy if variables indicated by α where to flip, $b \in \{0, 1\}$ and $F_b = bA/M + (1-b)(1-A/M)$ selects the faction of $a_i = b$, and R_x, R_a, R_{xa} being parameters associated with the rates of change of each transition. We define $R_x = R_{xa}, R_x/R_a = 4.5$.

The nonequilibrium steady state is calculated integrating the master equation:

$$\begin{aligned}\frac{dp(\boldsymbol{s})}{dt} =& \sum_{\boldsymbol{s}^x} \left(k_{\boldsymbol{s}^x,\boldsymbol{s}} p(\boldsymbol{s}) - k_{\boldsymbol{s},\boldsymbol{s}^x} p(\boldsymbol{s}^x)\right) \\ &+ \sum_{b} \sum_{\boldsymbol{s}^{ab}} \left(k_{\boldsymbol{s}^{ab},\boldsymbol{s}} p(\boldsymbol{s}) - k_{\boldsymbol{s},\boldsymbol{s}^{ab}} p(\boldsymbol{s}^{(a0)})\right) \\ &+ \sum_{\boldsymbol{s}^{xa}} \left(k_{\boldsymbol{s}^{xa},\boldsymbol{s}} p(\boldsymbol{s}) - k_{\boldsymbol{s},\boldsymbol{s}^{xa}} p(\boldsymbol{s}^{xa})\right) \quad (8)\end{aligned}$$

The steady state of the master equation above, $p(x, \boldsymbol{a}, n)$, is equal to all the configurations with similar $A = \sum_i a_i$. Thus, we use for simplicity the marginal distribution $p(x, A, n)$ is represented in (see Fig. 5, where the value of x is averaged out into $p(A, n)$ to facilitate visualization). The result is a bimodal distribution with one cluster of high energy states $\langle A \rangle > \langle A \rangle_{\text{eq}}$ which move in a biased random walk with positive averaged velocity (clockwise rotation) until they are able to recharge their energy levels at $n = 0$. The second cluster is composed of agents with low energy

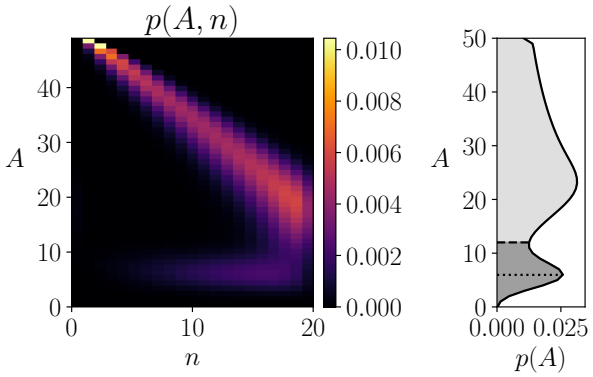

Figure 5: The density of the steady-state state distribution of the agent shows two distinct regions of behaviour. The first region, of 'autonomous states', starts at $n = 0$ with a high level of free energy, and moves sequentially to the right, eventually finding $n = 0$ and replenishing its levels of free energy. The second region, composed of 'non-autonomous-states', displays agents with free energy close to zero, moving as a random walk, with very low probabilities of reaching $n = 0$. The dashed line in the right figure separates both regions. The dotted line represents the value of $\langle A \rangle_{\text{eq}}$.

values ($\langle A \rangle = \langle A \rangle_{\text{eq}}$) which move with near zero average velocity in an unbiased random walk.

The average entropy production rate for each value A for each transitions is measured by the irreversibility of its forward and backward fluxes $J_{\boldsymbol{s}^{\alpha},\boldsymbol{s}} = k_{\boldsymbol{s}^{\alpha},\boldsymbol{s}} p(\boldsymbol{s})$:

$$\sigma_x(A) = \sum_{\boldsymbol{s}^x, x, n} \left(J_{\boldsymbol{s}^x,\boldsymbol{s}} - J_{\boldsymbol{s},\boldsymbol{s}^x}\right) \log \frac{J_{\boldsymbol{s}^x,\boldsymbol{s}}}{J_{\boldsymbol{s},\boldsymbol{s}^x}} \quad (9)$$

$$\sigma_a(A) = \sum_{b, x, n} \sum_{\boldsymbol{s}^{ab}} \left(J_{\boldsymbol{s}^{ab},\boldsymbol{s}} - J_{\boldsymbol{s},\boldsymbol{s}^{ab}}\right) \log \frac{J_{\boldsymbol{s}^{ab},\boldsymbol{s}}}{J_{\boldsymbol{s},\boldsymbol{s}^{ab}}} \quad (10)$$

$$\sigma_{xa}(A) = \sum_{\boldsymbol{s}^{xa}, x, n} \left(J_{\boldsymbol{s}^{xa},\boldsymbol{s}} - J_{\boldsymbol{s},\boldsymbol{s}^{xa}}\right) \log \frac{J_{\boldsymbol{s}^{xa},\boldsymbol{s}}}{J_{\boldsymbol{s},\boldsymbol{s}^{xa}}}. \quad (11)$$

Looking at the entropy dissipation for each value of A (Fig. 6, top), only the first cluster is able to dissipate entropy (i.e. maintain themselves out-of-equilibrium), meaning that states in the second cluster are near equilibrium.

Similarly, we calculate the average velocity (i.e. average rate of change of n) and the absolute average velocity for each value of A

$$\langle v(A) \rangle = \sum_{\boldsymbol{s}^{xa}, x, n} (1 - 2x) \left(J_{\boldsymbol{s}^{xa},\boldsymbol{s}} - J_{\boldsymbol{s},\boldsymbol{s}^{xa}}\right) \quad (12)$$

$$\langle |v(A)| \rangle = \sum_{\boldsymbol{s}^{xa}, x, n} \left(J_{\boldsymbol{s}^{xa},\boldsymbol{s}} - J_{\boldsymbol{s},\boldsymbol{s}^{xa}}\right) \quad (13)$$

In Fig. 6 (bottom) we observe that the agent only consistently moves forward in the case of the first cluster.

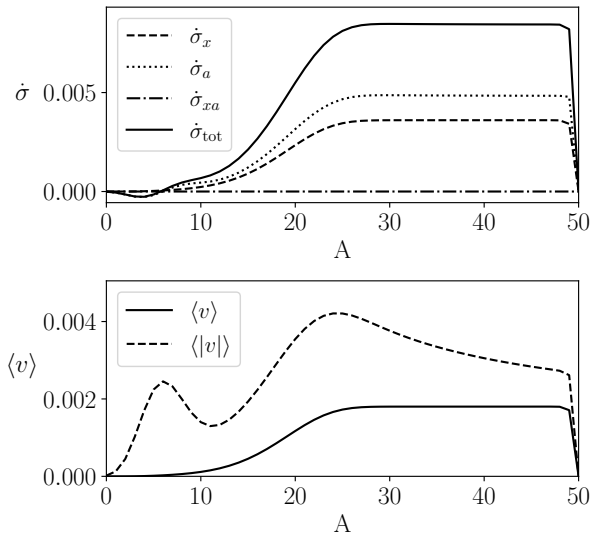

Figure 6: Entropy production rate (top) and average displacement speed (bottom) of agents for different values of A. Agents with high levels of energy A dissipate entropy at independent transitions x, a_i (but not joint x, a_i transitions, where contributions of x and a_i cancel out). Thus, entropy dissipates for both spontaneous degradation of energy units and as a result of the drag force. Agents with low A levels have very low dissipation of entropy (or even negative when $A < \langle A \rangle_{\mathrm{eq}}$. Similarly, both types of systems have positive absolute velocity, but only agents with high A levels are able to move forward, drifting to the right-side of the environment.

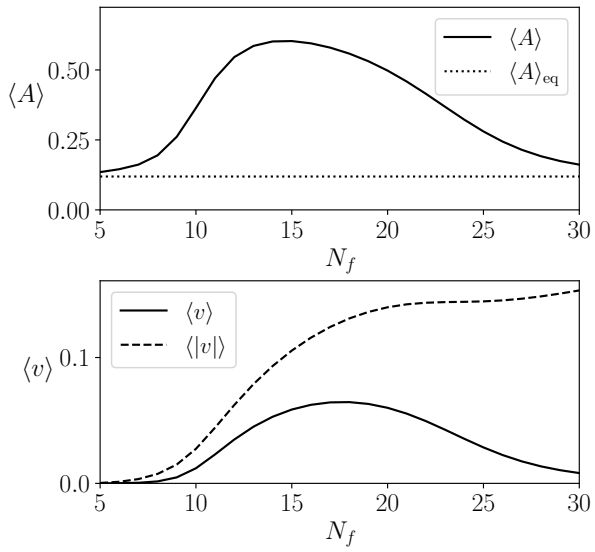

Figure 7: Average entropy dissipation (a) and average displacement speed (b) of agents, depending on the level of stored energy, A. Agents with high levels of energy units dissipate entropy at independent transitions of x and a (but not at joint x, a transitions). Agents with low A levels have very low dissipation of entropy (or even negative when $A < \langle A \rangle_{\mathrm{eq}}$. Similarly, both types of systems have positive absolute velocity, but only agents with high A levels are able to move forward, drifting to the right-side of the environment.

Thus, we observe a bifurcation between systems able to actively maintain themselves out-of-equilibrium, and systems falling into a state of equilibrium. We propose that only the first ones have a level of autonomy, characterized by their entropy dissipation rates.

Adaptive autonomy An autonomous agent is a system that regulates its interaction with the environment according to its own viability conditions. To inspect the adaptive potential of Thermina in the given environment, we allow the model to modify the step size in displacement L, which in turn affects the effective drag $f = FL$ and re-scales the length N of the environment. Smaller steps result in a smaller value of f, but a higher number of steps N required to cover the same distance (for simplicity, we will restrict to integer values of N). We thus describe the resulting environment length for a modified step size and the resulting effective drag f, $N_f = 20/f$, where $f = 1$ recovers the previous environment.

We inspect their viability in a steady-state configuration by measuring $\langle A \rangle$ for each value of N resulting from the effective drag f (Fig. 7, top). Interestingly, we find the maximum value of $\langle A \rangle$ at $N = 15$, which corresponds to $f = 4/3$. This value is larger than the energy provided by energy units, $\mu = 1$. This suggests that the agent is relying on thermal fluctuations to overcome the energy barrier required to advance forward. We plot also the values of $\langle v \rangle, \langle |v| \rangle$ (Fig. 7, bottom), finding that the latter is roughly stable at low and high values for the left and right sides of $\langle A \rangle$'s maximum value. Thus, a simple adaptive rule of the form, if $\langle A \rangle - \langle A \rangle_{\mathrm{eq}} < \theta_A$ and $\langle |v| \rangle < \theta_v$, then decrease f (i.e., increase N_f), and if $\langle A \rangle - \langle A \rangle_{\mathrm{eq}} < \theta_A$ and $\langle |v| \rangle > \theta_v$, then increase f (i.e., decrease N_f), it ensures that the agent is within a viability region where $\langle A \rangle - \langle A \rangle_{\mathrm{eq}} \geq \theta_A$. This rule is robust for different parameters of μ, as represented in Fig. 8.

Discussion

In this paper, we have introduced a minimal model exploring thermodynamic requirements for autonomous agency. Inspired by previous literature (Varela, 1979; Di Paolo, 2009; Barandiaran et al., 2009), we conceive autonomous agents as systems capable of adaptive interaction with their environment to enhance its ability to persist as out-of-equilibrium, precarious processes. The model is inspired in recent models of information-fuelled engines (Mandal and Jarzynski,

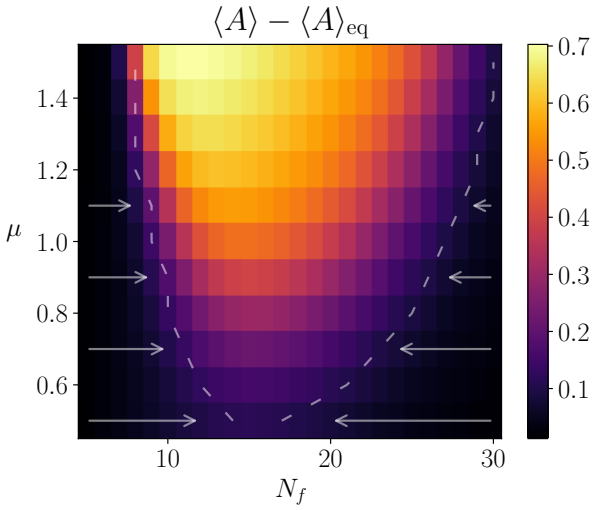

Figure 8: Values of $\langle A\rangle - \langle A\rangle_{\text{eq}}$ for different values of μ and f. A simple rule changing activated for $\langle A\rangle - \langle A\rangle_{\text{eq}} < \theta_A$ (dashed line), decreasing f (increasing N) when $\langle |v| \rangle < \theta_v$ and increasing f (decreasing N_f) when $\langle |v| \rangle < \theta_v$ ensures the viability of the agent, with values $\theta_A = 0.11, \theta_v = 0.06$.

2012) and molecular motors (Thomas et al., 2001), as well as conceptual groundwork on models of autonomy as Bittorio (Varela, 1988).

Out-of-equilibrium self-maintenance is a general property of living organization. Unlike mechanical machines, which just stop functioning when deprived of energy, cellular organization requires constant oxidation to maintain their molecular processes and prevent biochemical disorder (Donnan, 1928). At a larger level, out-of-equilibrium dynamical properties have been suggested as a useful indicator for distinguishing organized states in macroscopic living systems, such as brains during different tasks and states of consciousness (Lynn et al., 2021; de la Fuente et al., 2023). However, it remains unclear how to interpret such nonequilibrium processes when dynamics are partially detached from energetic requirements of slower metabolic processes, as in the case of neural dynamics (Barandiaran and Moreno, 2008).

In his seminal book on cybernetics, Wiener (1965) advocated for viewing cybernetic systems as out-of-equilibrium, non-conservative systems. He emphasized that cybernetic systems are 'effectively coupled to the external world, not merely by their energy flow, their metabolism, but also by a flow of impressions, of incoming messages, and of the actions of outgoing messages'. For Wiener, interaction through sensors, effectors and information transfer led to the emergence 'definite past-future order' associated with irreversibility of nonequilibrium statistical physics. In our case, while the model under consideration is grounded in an energy function, its applicability extends beyond physical systems to encompass general classes of information processing systems according to energy-based statistical models. The ideas explored here could be extended to study the autonomy of neural networks from descriptions of their nonequilibrium thermodynamics (Aguilera et al., 2023), or to analyse the agency of AI models using energy-based descriptions of models like transformers (Vaswani et al., 2017) or generative diffusion models (Du et al., 2023). Such a generalization to energy-based, statistical models enables the exploration of precarious, out-of-equilibrium, emergent behaviours in the interplay between energy-minimization internal dynamics and agent-environment interaction challenging those dynamics (and driving out-of-equilibrium states), offering insights into autonomy across different domains of study.

This model provides a first step on characterizing autonomy from this general thermodynamic perspective. Future developments should include: (i) a more complex and organized internal organization that makes justice to the networked interdependence between component processes in autonomous systems, (ii) a more complex coupling with the environment, and (iii) more complex and rich environments. Scaling up the model would also make it possible to bridge the existing gap between the type of energy analysis that this simplified model affords and the complex thermodynamically-inspired measurements of degrees of autonomy that existing natural and artificial systems require.

Acknowledgements

We thank Artemy Kolchinsky for valuable comments and discussions on this manuscript. MA was funded by a Junior Leader fellowship from 'la Caixa' Foundation (ID 100010434, code LCF/BQ/PI23/11970024), John Templeton Foundation (Grant ID 62828) and the Basque Government under the ELKARTEK 2023 program,(project KK-2023/00085) (project KK-2023/00085). XEB acknowledges IAS-Research group funding IT1668-22 from Basque Government, grant PID2019-104576GB-I00 for project Outonomy funded by MCIN/AEI/10.13039/501100011033. MA and XEB were supported in part by grant MICIU/AEI /10.13039/501100011033 from the Spanish Ministry of Science, Innovation and Universities.

References

Aguilera, M. (2015). *Interaction Dynamics and Autonomy in Cognitive Systems*. Phd thesis, University of Zaragoza.

Aguilera, M., Igarashi, M., and Shimazaki, H. (2023). Nonequilibrium thermodynamics of the asymmetric Sherrington-Kirkpatrick model. *Nature Communications*, 14(1):3685.

Amari, S.-i. (2016). Exponential families and mixture families of probability distributions. In *Information Geometry and Its Applications*, pages 31–49. Springer.

Andreas, J. (2022). Language Models as Agent Models.

Barandiaran, X. and Moreno, A. (2008). Adaptivity: From metabolism to behavior. *Adaptive Behavior*, 16(5):325–344.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior*, 17(5):367–386.

Barandiaran, X. E. and Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life*, 20(1):5–28.

Barato, A. C., Hartich, D., and Seifert, U. (2014). Efficiency of cellular information processing. *New Journal of Physics*, 16(10):103024.

Beer, R. D. (1990). *Intelligence as Adaptive Behaviour: An Experiment in Computational Neuroethology (Perspectives in Artificial Intelligence)*. Academic Press.

Beer, R. D. (2015). Characterizing Autopoiesis in the Game of Life. *Artificial Life*, 21(1):1–19.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, 91(2):331–345.

Bickhard, M. H. (2000). Autonomy, function, and representation. *Communication and Cognition-Artificial Intelligence*, 17(3-4):111–131.

Bostrom, N. (2017). *Superintelligence*. Dunod.

de la Fuente, L. A., Zamberlan, F., Bocaccio, H., Kringelbach, M., Deco, G., Perl, Y. S., Pallavicini, C., and Tagliazucchi, E. (2023). Temporal irreversibility of neural dynamics as a signature of consciousness. *Cerebral Cortex*, 33(5):1856–1865.

Di Paolo, E. (2009). Extended life. *Topoi*, 28(1):9–21.

Di Paolo, E., Buhrmann, T., and Barandiaran, X. E. (2017). *Sensorimotor Life: An enactive proposal*. Oxford University Press.

Di Paolo, E. A. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences*, 4(4):429–452.

Donnan, F. G. (1928). The mystery of life. *Journal of Chemical Education*, 5(12):1558.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. (2023). Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pages 8489–8510. PMLR.

Feynman, R. P. (2018). *Feynman lectures on computation*. CRC Press.

Hartich, D., Barato, A. C., and Seifert, U. (2015). Nonequilibrium sensing and its analogy to kinetic proofreading. *New Journal of Physics*, 17(5):055026.

Kauffman, S. A. (2000). *Investigations*. Oxford University Press US, NY.

Kolchinsky, A. and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface focus*, 8(6):20180041.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Liu, B., Mazumder, S., Robertson, E., and Grigsby, S. (2023). AI Autonomy : Self-Initiated Open-World Continual Learning and Adaptation. arXiv:2203.08994 [cs].

Longo, G. and Montévil, M. (2014). *Perspectives on Organisms*. Lecture Notes in Morphogenesis. Springer Berlin Heidelberg, Berlin, Heidelberg.

Lynn, C. W., Cornblath, E. J., Papadopoulos, L., Bertolero, M. A., and Bassett, D. S. (2021). Broken detailed balance and entropy production in the human brain. *Proceedings of the National Academy of Sciences*, 118(47):e2109889118.

Mandal, D. and Jarzynski, C. (2012). Work and information processing in a solvable model of maxwell's demon. *Proceedings of the National Academy of Sciences*, 109(29):11641–11645.

Moreno, A. and Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384.

Peliti, L. and Pigolotti, S. (2021). *Stochastic thermodynamics: an introduction*. Princeton University Press.

Preece, A. (2018). Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72.

Roose, K. (2023). A.I. Poses 'Risk of Extinction,' Industry Leaders Warn. *The New York Times*.

Ruiz-Mirazo, K. and Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life*, 10(3):235–259.

Ruiz-Mirazo, K., Pereto, J., and Moreno, A. (2004). A universal definition of life: autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere*, 34(3):323–346.

Ruppin, E. (2002). Evolutionary autonomous agents: A neuroscience perspective. *Nature Reviews Neuroscience*, 3(2):132–141.

Russell, S. J. (2019). *Human compatible: artificial intelligence and the problem of control*. Viking.

Schmiedl, T. and Seifert, U. (2008). Efficiency of molecular motors at maximum power. *Europhysics Letters*, 83(3):30005.

Seth, A. K. (2010). Measuring Autonomy and Emergence via Granger Causality. *Artificial Life*, 16(2):179–196. Conference Name: Artificial Life.

Still, S. (2020). Thermodynamic cost and benefit of memory. *Physical Review Letters*, 124(5):050601. Publisher: APS.

Thomas, N., Imafuku, Y., and Tawada, K. (2001). Molecular motors: thermodynamics and the random walk. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1481):2113–2122.

Varela, F. J. (1979). *Principles of biological autonomy*. English. North Holland, New York.

Varela, F. J. (1988). Structural Coupling and the Origin of Meaning in a Simple Cellular Automation. In Sercarz, E. E., Celada, F., Mitchison, N. A., and Tada, T., editors, *The Semiotics of Cellular Communication in the Immune System*, pages 151–161, Berlin, Heidelberg. Springer.

Varela, F. J. and Bourgine, P. (1992). Introduction: Towards a Practice of Autonomous Systems. In Varela, F. J. and Bourgine, P., editors, *Proceedings of the 1st European Conference on Artificial Life*, pages xi–3. MIT Press.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind : cognitive science and human experience*. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Weng, L. (2023). Llm-powered autonomous agents. *Lil'Log*.

Wiener, N. (1965). *Cybernetics: or the Control and Communication in the Animal and the Machine*. The MIT Press, 2 edition.