Bionoetics vol. 14, no. 1, 2018

UNIVERSAL PRINCIPLES OF INTELLIGENT SYSTEM DESIGN[1]

Kirill Popov, Aamod Shanker and Debanjan Bhowmik

ABSTRACT: We propose a universal definition of intelligence of a physical system and discuss its implications on design of intelligent systems. The definition proposed is universally valid and does not invoke teleological or anthropic concepts. We discuss the relationship of intelligence to energetic properties by invoking recent results in inequilibrium thermodynamics and computational mechanics. Intelligent system design is reformulated in three natural problems: selective forgetting, memory maintenance and self-recording. We conclude with highlighting the relationship of energetic and informational optimality principles involved in designing intelligent physical systems or their spontaneous emergence.

KEYWORDS: Intelligence; Information; Prediction; Causality; Memory; Design; Dynamical systems; Thermodynamics; Learning; epistemology

I. INTRODUCTION

We live in the era of yet another technological revolution. If the previous one was associated with the development of energy conversion tools to produce work, harnessing increasingly vast resources over the course of history, the world nowadays is being rapidly transformed by the information processing tools. Starting with arithmetic devices that support conditional clauses and store intermediate results, we have arrived at machines emulating major functions that have before typically been attributed only to living beings or specifically to humans — learning, knowledge management, pattern recognition and control. The source of hopes and fears these days is being called «artificial intelligence»,

[1] Editor's note: *Foundations of Mind*, the independent research group that has provided the papers for this

special edition, has never taken either corporate or state money and is financed entirely by donations. Authors keep copyright without paying. The typical fee for this charged by open-access journals such as those published by PLOS, is around $2k. If you value this project, and wish to see further such proceedings from this group, we ask you to consider donating to *Foundations of Mind* – as little as $5 per download, through their website: http://www.foundationsofmind.org/donate. This will ensure there will be further published proceedings on the foundations of mind like this one for you and others to enjoy free.

potentially dramatically enhancing the life conditions for humans and at the same time outcompeting us in the labor market. There is also quite an amount of speculation and research concerning theoretically possible super-intelligences posing existential risk for our species, and whether it is possible to align their interests with ours. Despite the massive effort in the field, the big question remains standing in the background: What is intelligence?

There have been numerous attempts at defining intelligence. In those proposals, the concept of intelligence often involves at least two distinct notions — one concerning the performance of an entity in a given fixed environment, and another one concerning its ability to learn new environments. The emphasis depends on the preferences of individual authors. It is also notable that these definitional inherently teleological, describing the intelligent entity as having a goal or receiving rewards. The most elegant and general definition [1] is as well set in the framework of reinforcement learning and involves uncomputable universal probability distribution which precludes its practical use. The notion of intelligence in a system is strongly associated with its information processing architecture [2,3]. Luckily, information is a well-defined property of a signal. Moreover, since the advent of information era, a number of connections were established, relating information theory to statistical physics and thermodynamics. By now the energy-information duality is fundamental to our understanding of organization in complex systems and emergence of functionality. Intelligence is a functional property of systems, therefore it is supposed to have complementary energetic and informational descriptions. Using the connection between information theory and physics provided by computational mechanics, the question of intelligence becomes physically tractable.

In this paper we propose a universal definition of intelligence for a physical system and discuss its implications on design of intelligent systems.

II. GUIDING PRINCIPLES

We want our definition of intelligence to have certain properties:

1. *Physical:* intelligence is well-defined in terms of physical observables of the system. 2. *Universal*: the definition should be applicable to any realizable physical system without referring to anthropic metaphors (e.g. goals or rewards) or particularities of the realization. 3. *Observable:* intelligence of an entity should be in principle observable by the third parties. 4. *Practical:* intelligence of an entity should be determinable in finite time using only

the observation of its behavior. Note that the word behavior here is used in a more general sense than its traditional use in psychology, and refers to the path of system in its configuration space as it evolves in spacetime, or some coarse- grained version of it that can be obtained from measurement. 5. *Partial order:* we want to be able to compare intelligences of different systems, grading them as more or less intelligent. 6. *Causal:* intelligence of an entity should depend only on the regions of spacetime causally accessible from it. 7. *Object-agnostic:* intelligence should be defined for any non-pathological 4-dimensional open manifold in spacetime. 8. *Instantaneous:* intelligence should be a pointwise property, manifesting at a given event in spacetime. 9. *Boundary-dependent:* intelligence at a given point should be a property of the boundaries defining an entity and not only the point of evaluation. It corresponds to the intuition that intelligence of a single bee is different then the intelligence of the bee hive, even if the evaluation of the latter occurs in a measurement of a behavior of a particular bee. 10. *Useful:* the usage of the term «intelligence» should correspond to it current usage in language, referring to intelligent living beings and artificial intelligence. 11. *Objective*: intelligence should not depend on any subjective criteria.

We adopt the macro scale view of the phenomena, using relativistic spacetime as the background. This also allows us to use thermodynamics, and connect it to statistical mechanical description of the dynamical systems present in the object. Within the statistical paradigm, the connection to information theory arises, which allows us to discuss information processing properties of the system and its computational ability in relation to intelligence [4-10]. The definition we construct in the next two sections satisfies all of the properties proposed here.

III. DEFINING THE ENTITY AND CAUSAL PATHWAYS

Consider an open connected 4-dimensional manifold M the closure of which is compact. Suppose M is embedded in Minkowski spacetime and represents the region that encloses (and defines) the entity the intelligence of which is being evaluated. We also require that M admits a 1-dimesional worldline L that is time-like at every point and has continuous intersection with M, i.e. $\tau\tau(L \cap M) = (\tau\tau_{min}, \tau\tau_{max})$ in any proper time parametrization $\tau\tau$: $L \to \mathbb{R}$ diffeomorphic to the real line and oriented from past to the future. We set $\tau\tau_{min} = 0$ to simplify further notation. To avoid pathological cases, assume that intersection of L with the boundary of M is transversal. We refer to the two points of $L \cap \partial\partial M$ as *birth* and *death,* since a point in four-dimensional

spacetime corresponds to an event. A worldline L satisfying the above conditions with respect to M will be called a *lifeline in M.*

We identify the worldline with a path taken by an *observer*, which is a point-like entity able to receive, store manipulate and record information. Our observer is rather special in that it only receives signals that originate in M. The recording of information and its processing occurs in internal state of the observer. Since observer is defined in purely geometric and computational terms, we interpret the internal state of the observer to be the state of a spacetime element, describing its position in the available configuration space (e.g. intensity spectrum of EM field, contents of various distinct species, temperature, etc). We denote the state of a spacetime element at event x as $S(x)$.

At each point $x \in L$, let us denote the time-like region of past as $TP_L(x)$, time-like region of future as $TF_L(x)$ and the 4-dimensional ball of radius t as $B(x, t)$. The boundary of the ball in Minkowski metric corresponds to a two-sheet hyperboloid, with its components lying in $TP_L(x)$ and $TF_L(x)$ respectively [see Figure 1].

Figure 1. Two-dimensional cut of the space-time diagram of an entity *M* with observer worldline *L*, representing instantaneous causal structure at *x*. For illustrative purposes we assumed that L is a subset of the cut.

We define the *causal M-past* at *x* along *L* as a union of all intersections of time-like regions of past over points of $L \cap M$ up to *x*:

$$P_M(L, x) := \bigcup_{t \in (0,\, \tau\tau(x))} [TP_L(y) \cap B(y, t) \cap M] = M \cap \left(\bigcup_{t \in (0,\, \tau\tau(x))} [TP_L(y) \cap B(y, t)]\right), \text{ where } y \text{ is a point in } L \text{ with } \tau\tau(y) = t.$$

The causal *M*-past is empty at and before birth: $P_M(L, x_{birth}) = \varnothing$. In English, the birth event of an observer of the entity *M* is independent of any part of *M*. The locus of the worldline between birth and any $x \in L \cap M$ is always contained in the causal *M*-past at *x*, reflecting the intuition that an observer moving along *L* is always able to be influenced by its own past, independently of the choice of entity *M*. Note that $P_M(L, x) \subset P_M(L, y)$ for any points $x, y \in L$ with $\tau\tau(x) \leq \tau\tau(y)$. This reflects that an observer at later proper times is influenced

by at least the events that she was influenced with before.

Analogously, we define the *causal M-future* at x along L:

$$F_M(L, x) := \cup_{t \in (\tau\tau(x),\, \tau\tau_d)} [TF_L(y) \cap B(y, \tau\tau_d\text{-}t) \cap M] = M \cap \big(\cup_{t \in (0,\, \tau\tau(x))} [TF_L(y) \cap B(y, \tau\tau_d\text{-}t)]\big),$$

where y is a point in L with $\tau\tau(y) = t$,

and $\tau\tau_d = \tau\tau(x^{\text{death}})$.

The set of events $F_M(L, x)$ represents the events in M that the observer at x can influence over the course of its lifetime after x. The locus of the worldline between any $x \in L \cap M$ and death is always contained in the causal M-future at x, reflecting the intuition that an observer moving along L is always able to influence its own future till its death, independently of the choice of entity M. Note that causal M-future is empty at and after death: $F_M(L, x_{death}) = \varnothing$. Note that $F_M(L, x) \subset F_M(L, y)$ for any points $x, y \in L$ with $\tau\tau(y) \leq \tau\tau(x)$. In other words, the causal future of points further on the worldline is contained in the future of earlier points.

To summarize, we consider the observer to travel along L, with internal clock counting proper time $\tau\tau$. It receives signals from some events in $P_M(L, x)$, processes them and sends signals to some events in $F_M(L, x)$. Therefore an observer at x is able to operate at most with the information that is contained in $P_M(L, x)$, and is able to influence at most the information that is (or will be, if a dynamic view is adopted) contained in $F_M(L, x)$.

IV. DEFINING INTELLIGENCE

We want the intelligence of M to be defined it terms of prediction capability of the most

predictive observer worldline L accommodated by M. We would say that the quality of prediction is quantified by two features: how well does the observer know the future, and how far into the future can it look ahead. Also we suggest that «irrational knowledge» should not be considered intelligent. In more rigorous terms, we only consider the predictive information that observer has regarding the signals that she is potentially able to interact with later and compare with its predictions.

It suffices to require the intelligent system to be able to *predict the events that happen on its boundary.* The reason for this choice is motivated by the current usage of the world when referring to living beings or artificial computational machines. As the system M is 4-dimensional, its boundary $\partial\partial M$ is 3-dimensional. Hence the intersection of $\partial\partial M$ with the hyperplane $H_L(x)$ of events simultaneous with $x \in L$ in L-observer frame is 2-dimensional almost everywhere. For natural intelligence of living beings, this geometric object corresponds to instantaneous (constant L-time slice) positions of the sensory input surfaces, e.g. retina, eardrum, skin, at smaller scales - membranes of individual sensory neurons. For silicon-based artificial intelligence, the surfaces at question are the cross- sections of the wires providing the input streams into CPU-RAM complex.

We advocate for the extreme epistemic philosophical standpoint, suggesting that the intelligent models of reality do not even in principle need to represent the ontological state of reality. On the contrary, intelligent beings only need to be able to maintain a model that allows them to predict the future of their interactions with their immediate environments. Such model is not explicitly required to be representative of potentially existing objective reality.

The particular property that we attribute intelligence to is *shortcut processing* — intuitively, we will say that a system is intelligent if and only if it constructs a good predictive model of what will happen to it before it actually happens. In more rigorous terms, the measure of intelligence of a system is how much information does it contain about what will happen to its boundary over the remaining course of its lifetime.

Consider an observer at point $x \in L$. In its future at point $y \in L : \tau\tau(x) \leq \tau\tau(y)$, it will receive signals from some subset of $P_M(L, y) \setminus P_M(L, x)$. Replacing M with the boundary of M in the definition of causal M-past, we obtain the *causal past interface increase* between x and y:

$$\mathrm{CPI}_{M,L}(x,y) := \bigcup_{t \in (\tau\tau(x),\, \tau\tau(y))} [TP_L(z) \cap B(z, t) \cap \partial\partial M], \text{ where } \tau\tau(z) = t.$$

We want to establish how much does the observer know about its future while at x. The rigorous way is to calculate the mutual information $I(S(x); S(\mathrm{CPI}_{M,L}(y, z)))$ between the state of the observer at x and the state of the (3-dimensional) boundary that the observer will be receiving signals from as it traverses L from y to z. Then we can

differentiate the above quantity with respect to z and obtain *boundary information density* $\mathrm{BID}(x; y)$ along L. This quantity tells us how predictive is the observer state at x of the signals that will happen to it upon infinitesimal proper time period starting at point y. To obtain the *intelligent information* the observer has regarding all of the future, this density should be integrated:

$$II\,(M, L, x) := \mathrm{I}\,(\,\mathrm{S(x)};\,\mathrm{S}(\mathrm{CPI}_{\mathrm{M,L}}(x,\,x^{death}))\,) = \int_{y \in (\mathrm{x},\,\mathrm{x}_{\mathrm{death}})} \mathrm{BID}(x;\,y)\,dy$$

The last step in formally defining intelligence of entity *M* at event *x* is finding the worldline that satisfies the constraints of section III and maximizes intelligent information defined above:

$$\text{Intelligence}\,(M, x) := \sup_{\{L \text{ - lifelines in M} \mid x \in L\}} II\,(M, L, x)$$

This definition satisfies all of the constraints proposed in section II by construction. However, the calculations to perform are not straightforward. In the next section we try to make the procedure more explicit.

V. PRACTICAL INTELLIGENCE MEASUREMENT

The procedure outline in section IV involves calculating the mutual information between the state of an observer at a fixed point in spacetime and the compound state of the causal past interface increase between two points in the future. How can this be done?

In general, the performance of the measurements with necessary precision can be impossible without destroying the dynamical order in the system. Therefore the suggested use of the definition is for dynamical systems that can be converted into symbolic systems by choosing a proper state space partition. This would allow for an ε- machine reconstruction, which can be used to produce distributions identical to those generated by continuous dynamics in actual system, yet by performing symbolic manipulations. Hence in its current state the calculation of intelligence of a system requires to be able to reproduce its dynamics as optimally as possible.[4,19,24]

VI. ENERGETIC CONSIDERATIONS

We propose a view that learning capability is not a prerequisite of intelligence, but the other way around. Intelligence, as defined by prediction capability, implies that observer is able to select which information out of the input stream reaching it should be discarded and when it should be discarded in order for predictive information to arise — given energy constraints. Here we discuss the problem of intelligence in the universe

from the design perspective: i.e., what conditions shall be true for a given system to ensure its large intelligence?

Recently the connection between thermodynamics and prediction capacity has been established firmly [20-25]. The memory of the environment's past by the system (mutual information between system present and environment past) should not be much larger compared to the predictive power of the system (mutual information between system present and environment future) in order to minimize work dissipation. The «useless nostalgia» is proportional to average work being dissipated instantaneously, and is a lower bound on total work dissipated over lifetime. Even more fundamentally, the Landauer's principle is modified by «nostalgia»: it appears as a new term in the lower bound for heat released. Inferential model with low efficiency (i.e. high «nostalgia») can not be made energy efficient. Maximally energy-efficient functional system (with memory) has to be predictive and not nostalgic.

Therefore we suggest three natural problems arising in design of intelligent systems. 1) Selective forgetting: out of all data received by the system, how do the system dynamics figure out which parts shall be forgotten due to uselessness, and which shall be retained?

2) Memory maintenance: out of all data maintained by the system, how do the system dynamics figure out which parts to forget and when?

3) Self-recording: out of all data sent by the system, how do the system dynamics figure out which parts shall it record itself for future reference?

Note that forgetting costs dissipation bounded from below by sum of how much nostalgia does the system currently contains, so for low dissipation it is best to forget when non-predictive information is already low. In other words, if a system has learnt the skill of selective forgetting well in the past, it will make it a better learner in the future.

In Landauer bound, forgetting refers to self-information of the system, not mutual information with environment. Hence increase in variability of one's internal state also counts as forgetting. But it can be beneficial for an agent, and potentially increase its predictive capability. Hence an agent wants to have that capacity — preferably in an efficient manner. Energetic cost of the process is bounded from below by increase in self- information plus current nostalgia. Therefore to be good at increasing your entropy (Landauer's forgetting), you need to be minimally nostalgic.

Note that for a piece of data stored, a highly intelligent system would store it in the form that is the most convenient, given the estimated lifetime till erasure. Therefore intelligent system also contains information on its own future behavior, thus explicitly exhibiting reflexivity.

CONCLUSION

We have constructed a definition of intelligence for a physically instantiated system and provided a roadmap to practical calculations of intelligence in dynamical systems. Connecting information theory to thermodynamics, we discuss the relationship of optimality principles constraining the design of intelligent systems.

Kirill Popov, University of California, Berkeley
Aamod Shanker, University of California, Berkeley
Debanjan Bhowmik, Indian Institute of Technology
Correspondence: popkir@gmail.com (K.Popov)

BIBLIOGRAPHY

Shane Legg and Marcus Hutter. Universal Intelligence: A Definition of Machine

Intelligence. arXiv:0712.3329v1 [cs.AI] 20 Dec 2007 Chiara Santolin and Jenny R. Saffran. Constraints on Statistical Learning Across Species. Trends in Cognitive Sciences, January 2018, Vol. 22, No. 1, DOI: 10.1016/j.tics.2017.10.003 Max Tegmark. Consciousness as a state of matter. Chaos, Solitons & Fractals, Volume

76, July 2015, Pages 238-270. DOI: 10.1016/j.chaos.2015.03.014 David Feldman. A Brief Introduction to: Information Theory, Excess Entropy and Computational Mechanics. April 1998 (Revised October 2002). http://hornacek.coa.edu/dave/ Ya. G. Sinai. Metric Entropy of Dynamical System. March 20, 2007 Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield. Above and Beyond the Landauer Bound: Thermodynamics of Modularity . Santa Fe Institute Working Paper L. Brillouin «The Negentropy Principle of Information» Journal of Applied Physics 24,

1152 (1953) Norman Margolus, Lev B.Levitin «The maximum speed of dynamical evolution» Physica D: Nonlinear Phenomena, Volume 120, Issues 1–2, 1 September 1998, Pages 188-195 Jean-Bernard Brissaud «The meanings of entropy» Entropy 2005, 7[1] , 68-96 M. N. Bera et al. «Thermodynamics as a Consequence of Information Conservation».

arXiv:1707.01750 Paul M. Riechers, James P. Crutchfield. «Fluctuations When Driving Between

Nonequilibrium Steady States». C. Aghamohammadi and J.P. Crutchfield, «Thermodynamics of Random Number

Generation», Physical Review E 95:6 (2017) 062139 A.B. Boyd, D. Mandal, and J.P. Crutchfield, «Above and Beyond the Landauer Bound:

Thermodynamics of Modularity». J. M. R. Parrondo et al. «Thermodynamics of information» Nature Physics, 2015 Chiara Marletto «Constructor Theory of Thermodynamics» 2017 A.B. Boyd, D. Mandal, and J.P. Crutchfield, «Identifying Functional Thermodynamics in

Autonomous Maxwellian Ratchets», New Journal of Physics 18 (2016) A. B. Boyd, D. Mandal, and J.P. Crutchfield. «Correlation-powered information engines

and the thermodynamics of self-correction» Phys. Rev. E 95, 012152 (2017) James P Crutchfield. «Hierarchical Thermodynamics» Workshop on Information Engines at the Frontier of Nanoscale Thermodynamics. Telluride Science Research Center, 3-11 August 2017. Cosma Rohilla Shalizi and James P. Crutchfield «Computational Mechanics: Pattern and Prediction, Structure and Simplicity» Journal of Statistical Physics, Vol. 104, Nos. 3/4, 2001 Hong et al. «Experimental test of Landauer's principle in single-bit operations on

nanomagnetic memory bits» Sci. Adv. 2016;2:e1501492 11 March 2016 Antoine Berut, Artyom Petrosyan and Sergio Ciliberto «Experimental test of Landauer's

principle in single-bit operations on nanomagnetic memory bits» arxiv 2015 Charles H. Bennett «Notes on Landauer's principle, reversible computation, and Maxwell's Demon» Studies in History and Philosophy of Modern Physics 34 (2003) 501–510 Jeffrey Bub «Maxwell's Demon and the Thermodynamics of Computation» arxiv 2002 James P. Crutchfield «The Origins of Computational Mechanics: A Brief Intellectual History and Several Clarifications» Santa Fe Institute Working Paper 17-10- XXX arxiv.org:1710.06832 [cond-mat.stat-mech] Susanne Still, David A. Sivak, Anthony J. Bell, and Gavin E. Crooks. Thermodynamics of Prediction. Physical Review Letters 109, 120604 (2012), DOI: 10.1103/PhysRevLett.109.120604