Designing ecosystems of intelligence from first principles

Karl J Friston1,2, Maxwell JD Ramstead1,2, Alex B Kiefer1,3, Alexander Tschantz1, Christopher L Buckley1,4, Mahault Albarracin1,5, Riddhi J Pitiya1,6, Conor Heins1,7,8,9, Brennan Klein1,10, Beren Millidge1,11, Dalton AR Sakthivadivel1,12,13,14, Toby St Clere Smithe1,6,15, Magnus Koudahl1,16, Safae Essafi Tremblay1,17, Capm Petersen1, Kaiser Fung1, Jason G Fox1, Steven Swanson1, Dan Mapes1 and Gabriel René1

1VERSES Research Lab, Los Angeles, CA, USA (KJF, MJDR, ABK, AT, CLB, MA, RJP, CH, BK, BM, DARS, TSCS, MK, SET, CP, KF, JGF, SS, DM, GR)
2Wellcome Centre for Human Neuroimaging, University College London, London, UK (KJF, MJDR)
3Department of Philosophy, Monash University, Melbourne, VIC, Australia (ABK)
4Department of Informatics, University of Sussex, Brighton, UK (CLB)
5Department of Computer Science, Université du Québec à Montréal, Montréal, QC, Canada (MA)
6Department of Experimental Psychology, University of Oxford, Oxford, UK (RJP, TSCS)
7Department of Collective Behaviour, Max Planck Institute of Animal Behavior, Konstanz, Germany (CH)
8Department of Biology, University of Konstanz, Konstanz, Germany (CH)
9Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany (CH)
10Network Science Institute, Northeastern University, Boston, MA, USA (BK)
11Brain Network Dynamics Unit, University of Oxford, Oxford, UK (BM)
12Department of Mathematics, Stony Brook University, New York, NY, USA (DARS)
13Department of Physics and Astronomy, Stony Brook University, New York, NY, USA (DARS)
14Department of Biomedical Engineering, Stony Brook University, New York, NY, USA (DARS)
15ToPos Institute, Berkeley, CA, USA (TSCS)
16Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands (MK)
17Department of Philosophy, Université du Québec à Montréal, Montréal, QC, Canada (SET)

Abstract
This white paper lays out a vision of research and development in the field of artificial intelligence for the next decade (and beyond). Its denouement is a cyber-physical ecosystem of natural and synthetic sense-making, in which humans are integral participants—what we call “shared intelligence.” This vision is premised on active inference, a formulation of adaptive behavior that can be read as a physics of intelligence, and which inherits from the physics of self-organization. In this

Corresponding author:
Dalton AR Sakthivadivel, VERSES Research Lab, 5877 Obama Blvd, Los Angeles, CA 90016, USA.
Email: dalton.sakthivadivel@verses.ai

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
context, we understand intelligence as the capacity to accumulate evidence for a generative model of one’s sensed world—also known as self-evidencing. Formally, this corresponds to maximizing (Bayesian) model evidence, via belief updating over several scales, that is, inference, learning, and model selection. Operationally, this self-evidencing can be realized via (variational) message passing or belief propagation on a factor graph. Crucially, active inference foregrounds an existential imperative of intelligent systems; namely, curiosity or the resolution of uncertainty. This same imperative underwrites belief sharing in ensembles of agents, in which certain aspects (i.e., factors) of each agent’s generative world model provide a common ground or frame of reference. Active inference plays a foundational role in this ecology of belief sharing—leading to a formal account of collective intelligence that rests on shared narratives and goals. We also consider the kinds of communication protocols that must be developed to enable such an ecosystem of intelligences and motivate the development of a shared hyper-spatial modeling language and transaction protocol, as a first—and key—step towards such an ecology.

Keywords
Active inference, free energy principle, artificial intelligence, belief updating, belief propagation

Introduction
This white paper presents active inference as an approach to research and development in the field of artificial intelligence (AI), with the aim of developing ecosystems of natural and artificial intelligences. The path forward in AI is often presented as progressing from systems that are able to solve problems within one narrow domain—so-called “artificial narrow intelligence” (ANI)—to systems that are able to solve problems in a domain general-manner, at or beyond human levels: what are known as “artificial general intelligence” (AGI) and “artificial super-intelligence” (ASI), respectively (Bostrom, 2014). We believe that approaching ASI (or, for reasons outlined below, even AGI) likely requires an understanding of networked or collective intelligence. Given the growing ubiquity of things like autonomous vehicles, robots, and arrays of edge computing devices and sensors (collectively, the internet of things), the zenith of the AI age may end up being a distributed network of intelligent systems, which interact frictionlessly in real time, and compose into emergent forms of intelligence at superordinate scales. The nodes of such a distributed, interconnected ecosystem may then be human users as well as human-designed artifacts that embody or implement forms of intelligence.

In order to enable such ecosystems, we must learn from nature. While acknowledging neuroscience as a key inspiration for AI research, we argue that we must move beyond brains, and embrace the active and nested characteristics of natural intelligence, as it occurs in living organisms and as it might be implemented in physical systems more generally. In our view, this entails asking not only “How does intelligence present to us, as researchers?” but also, crucially, the complementary question “What is it that intelligence must be, given that intelligent systems exist in a universe like ours?” To address this challenge, we aim to deduce fundamental properties of intelligence from foundational considerations about the nature of persisting physical systems (i.e., “first principles”).

In so doing, we foreground active inference, which combines the virtues of such first-principles, physics-based approach to AI with Bayesian formulations, thus reframing and, in some key respects, extending the methods found in Bayesian approaches to machine learning, which provide the foundations of state-of-the-art AI systems. Active inference is an account of the inevitable existence of agency in physical worlds such as ours, which motivates a definition of intelligence as the capacity of systems to generate evidence for their own existence. This encompasses cognition (i.e., problem-solving via action and perception) and curiosity, as well as the capacity for creativity, which underwrites the current interest in generative AI (Sequoia Capital, 2022). Active inference offers a formal definition of intelligence for AI research, and entails an explicit mechanics of the beliefs of agents and groups of agents—known as Bayesian mechanics (Ramstead et al., 2023; Da Costa et al., 2021)—which is uniquely suited to the engineering of ecosystems of intelligence, as it allows us to write down the dynamics of sparsely coupled systems that self-organize over several scales or “levels” (Friston et al., 2015; Friston, 2013; Kuchling et al., 2020; Ramstead et al., 2021). We argue that the design of intelligent systems must begin from the physicality of information and its processing at every scale or level of self-organization. The result is AI that “scales up” the way nature does: by aggregating individual intelligences and their locally contextualized knowledge bases, within and across ecosystems, into “nested intelligences”—rather than by merely adding more data, parameters, or layers to a machine learning architecture.

We consider the question of how to engineer ecosystems of AI using active inference, with a focus on the problem of communication between intelligent agents, such that shared forms of intelligence emerge, in a nested fashion, from these interactions. We highlight the importance of shared narratives
and goals in the emergence of collective behavior, and how active inference helps account for this in terms of sharing (aspects of) the same generative model. We close our discussion with a sketch of stages of development for AI using the principles of active inference. Our hypothesis is that taking the multi-scale and multi-level aspects of intelligence seriously has the potential to be transformative with respect to the assumptions and goals of research, development, and design in the field of AI, with potentially broad implications for industry and society: that technologies based on the principles described in this paper may be apt to foster the design of an emergent ecosystem of intelligences spanning spatial and cognitive domains (a hyper-spatial web).

A first-principles approach to multi-scale artificial intelligence

The field of artificial intelligence has from the outset used natural systems, whose stunning designs have been refined over evolutionary timescales, as templates for its models. Neuroscience has been the most significant source of inspiration, from the McCulloch-Pitts neuron (Maass, 1997) to the parallel distributed architectures of connectionism and deep learning (LeCun et al., 2015; Lillicrap et al., 2019), to the contemporary call for “Neuro-IA” as a paradigm for research in AI, in particular machine learning (Zador, 2022). Indeed, the definitive aspect of deep learning inherits from the hierarchical depth of cortical architectures in the brain (Zeki and Stewart, 1988). More recently, machine learning has come, in turn, to influence neuroscience (Richards, 2019; Yamins et al., 2014; Yamins and DiCarlo, 2016).

Academic research as well as popular media often depict both AGI and ASI as singular and monolithic AI systems, akin to super-intelligent, human individuals. However, intelligence is ubiquitous in natural systems—and generally looks very different from this. Physically complex, expressive systems, such as human beings, are uniquely capable of feats like explicit symbolic communication or mathematical reasoning. But these paradigmatic manifestations of intelligence exist along with, and emerge from, many simpler forms of intelligence found throughout the animal kingdom, as well as less overt forms of intelligence that pervade nature.

Examples of “basal cognition” abound—and often involve distributed, collective forms of intelligence. Colonies of slime molds, for example, can—as a group—navigate two-dimensional spatial landscapes, and even solve mathematical problems that are analytically intractable (Murugan et al., 2021). Certain forms of cognition and learning are (at least arguably) observable in plants (Calvo and Friston, 2017), and we know that plants grow in a modular fashion, as a structured community of tissues that self-organize into a specific configuration (Leyser, 2011).

Communication between organisms is often mediated by network structures, which themselves consist of other organisms; for instance, it is known that mycorrhizal networks are able to facilitate communication, learning, and memory in trees (Simard, 2018). Mobile groups of schooling fish can, as a collective, sense light gradients over a wide spacetime window, even as the individuals that comprise the group can only detect local light intensity (Berdahl et al., 2013). Perhaps most germanely, in morphogenesis (i.e., pattern formation in multicellular organisms), the communication of collectives of cells implements a search for viable solutions in a vast problem space of body configurations (Fields and Levin, 2022; Davies and Levin, 2022; Fields et al., 2020). This is not merely a metaphorical extension or use of the word “intelligence,” as it is no different, at its core, from our experience of navigating three-dimensional space (Kuchling et al., 2020).

Thus, at each physical spatiotemporal scale of interest, one can identify systems that are competent in their domain of specialization, lending intelligence in physical systems a fundamentally multi-scale character (Levin, 2019; Fields et al., 2021). Observation of nature suggests, moreover, that simpler and more complex forms of intelligence are almost always related compositionally: appreciably intelligent things tend to be composed of systems that are also intelligent to some degree. Most obviously, the intelligence of individual human beings, to the extent that it depends on the brain, implements the collective intelligence of neurons—harnessed by many intervening levels of organization or modularity, and observed by organelles at the cellular level. This communal or collective aspect of intelligence is reflected in the etymology of “intelligence”—from inter- (which means between) and legere (which means to choose or to read)—literally inter-legibility, or the ability to understand one another.

Since intelligence at each scale supervenes on, or emerges from, simpler (though still intelligent) parts, the multi-scale view of natural intelligence implies not a mysterious infinite regress, but a recursive, nested structure in which the same functional motif (the action-perception loop) recurs in increasingly ramified forms in more complex agents (Lycan, 2008). The emergence of a higher level intelligence—from the interaction of intelligent components—depends on network structure (e.g., the organization of the nervous system, or communication among members in a group or population) and sparse coupling (i.e., the fact that things are defined by what they are not connected to (Sakhivadivel, 2022a)), which together often lead to functional specialization among the constituents (Parr et al., 2020a).

But how do we engineer systems like these? We argue that instead of focusing merely on empirical descriptive adequacy and “reasoning by analogy” (e.g., the Turing test or imitation game (Harnad, 2003), one can leverage the fundamental organizing principles that underwrite the operation of intelligence in nature, separating them from the contingent details of particular biological systems. Such an approach has its origins
in the cybernetics movement of the 1940s, which set out to describe the general properties of regulatory and purposive systems—that is, properties not tied to any given specific architecture—and from which we draw now commonplace principles of system design, such as feedback and homeostasis (Ashby, 1956; Rosenblueth et al., 1943). Perhaps the most well-known example of this is the good regulator theorem (Conant and Ashby, 1970), later developed as the internal model principle in control theory (Francis and Wonham, 1976), according to which systems that exist physically must contain structures that are homomorphic to whatever environmental factors they are capable of controlling. A precursor to the good regular theorem is Ashby’s “law of requisite variety” (Ashby, 1958), according to which a system that controls an environment (as represented by a set of observations) must possess at least as many degrees of freedom (in probabilistic terms, as much entropy) as the phenomenon controlled.

Contemporary developments in the statistical mechanics of far-from-equilibrium systems (and in particular, multi-scale, living systems) allow us to formalize these insights—that of early cybernetics—as a physics of self-organization, which enables the study of intelligence itself as a basic, ubiquitous, physical phenomenon. This has been called a physics of sentient systems; where “sentient” means “responsive to sensory impressions” (Friston, 2019; Ramstead et al., 2019). More specifically, we argue that one can articulate the principles and corollaries of the core observation that intelligent systems (i.e., agents) exist in terms of a “Bayesian mechanics” that can be used to describe or simulate them (Ramstead et al., 2023; Da Costa et al., 2021).

We note that physical implementation is the ultimate constraint on all forms of engineered intelligence. While this claim might sound trivial, it has been a core locus of recent progress in our understanding of the physics of information itself. According to Landauer’s principle (Landauer, 1961; Bennett, 2003; Jarzynski, 1997; Evans, 2003), there is an energy cost to irreversibly read-write any information in a physical medium. Thus, the physicality of information and its processing at every scale of self-organization should be accounted for in the design of intelligent systems. Apart from being principled, forcing models to respect constraints or conservation laws—of the kind furnished by physical implementation—often improves their performance or even enables unique capabilities. Our core thesis is that all of this is naturally accommodated by an approach to AI grounded in the physics of intelligence.

**Active inference**

*"Model evidence is all you need"*

We approach the challenges just outlined from the perspective of active inference, a first-principles or physics-based approach to intelligence that aims to describe, study, and design intelligent agents from their own perspective (Eliasmith, 2005). Active inference shares the same foundations as quantum, classical, and statistical mechanics, and derives a scale-free theory of intelligence by adding an account of the individuation of particular things within their environments (Friston, 2019).

We begin with the observation that individual physical objects can be defined by the typical absence of influence of some parts of the universe on others (e.g., air temperature directly impacts my skin, but not my internal organs). In sparse causal networks, some nodes act as informational bottlenecks that serve both as mediating channels and as (probabilistic) boundaries (Holland, 2014), on which the variability of states on either side is conditioned. The persistence of such stable boundaries in a changing world (i.e., away from thermodynamic equilibrium) is possible only to the extent that the boundary conditions can be predicted and controlled, leveraging an implicit statistical model—a generative model of how they are caused by external changes.

To exist as an individuated thing is thus to gather observational evidence for such a model (“self-evidencing” (Hohwy, 2016)). This “model evidence” can be scored by a scalar value that conveys the degree to which some observations conform to (i.e., are predictable from) the model. To account for perception, one can update variables in order to maximize model evidence (e.g., update beliefs to match the data). To account for learning, one can update parameters in order to maximize model evidence (e.g., update models to match the data). To account for action, one can select actions in order to maximize (expected) model evidence (assuming that the model encodes preferences in terms of prior beliefs) (Ramstead et al., 2019; Friston, 2011). From this perspective, model evidence is the only thing that needs to be optimized.

Importantly, model evidence can be approximated in a form that has interesting decompositions, into quantities that map onto distinct facets of intelligence. For instance, a lower bound on the model evidence (equivalently, an upper bound on surprise)—called variational free energy (Winn and Bishop, 2005)—can always be written as complexity minus accuracy. When a system minimizes free energy, in so doing, it automatically maximizes the predictive accuracy of its model while minimizing its complexity (implementing a version of Occam’s razor). This means that self-evidencing mandates an accurate account of sensory exchanges with the world that is minimally complex, which serves to limit overfitting and poor generalization (Sengupta and Friston, 2018).

Active inference builds on these insights. If inference entails maximizing accuracy while minimizing complexity, it follows that self-evidencing should minimize the inaccuracy and complexity that is expected following upon a
course of action. It transpires that expected complexity is exactly the same quantity minimized in optimal control theory (Kappen et al., 2012; Todorov, 2008); namely, risk, while expected inaccuracy is just the ambiguity inherent in the way we sample data (e.g., resolved by switching the lights on in a dark room). Perhaps more interestingly, the ensuing expected free energy can be rearranged into expected information gain and expected value, where value is just the (log) preference for an outcome. This result captures exactly the dual aspects of Bayes optimality; namely, optimal Bayesian experimental design (Balietti et al., 2021; Lindley, 1956; MacKay, 1992) and decision theory (Berger, 1985). In essence, it favors choices that ensure the greatest resolution of uncertainty, under the constraint that preferred outcomes are realized. In other words, it mandates information and preference-seeking behavior, where one contextualizes the other. The ensuing curiosity or novelty-seeking thus emerges as an existential imperative (Barto et al., 2013; Lindley, 1956; MacKay, 1992; Schmidhuber, 2006; Schwartenbeck et al., 2019; Still and Preucel, 2012)—to the extent that one could say that to be intelligent is (in part) to be curious, and to balance curiosity against preferences or reward in an optimal fashion.

Crucially, the approach to existence as modeling just outlined can be applied recursively, in a nested fashion, to systems as well as their components, providing the foundations for mathematical theories of collective intelligence at any scale, from rocks to rockstars. Indeed, if existing in a systems as well as their components, providing the foundation can be applied recursively, in a nested fashion, to systems as well as their components, providing the foundations for mathematical theories of collective intelligence at any scale, from rocks to rockstars.4!

There are several types of uncertainty at play when learning from data. First, there may be irreducible noise in the measurement process itself. Examples of such noise include pixel blur in images. Second, the values of the hidden variables being estimated from data may be ambiguous (e.g., “Is the image I’m viewing of a duck or a rabbit?” or “It looks like rain: should I bring an umbrella?”). Third, there may be noise in the model of the function being learned (e.g., “What do rabbits look like? How do hidden variables map to data?”). Overcoming and accounting for these different types of uncertainty is essential for learning.

Non-probabilistic approaches to AI encounter these forms of uncertainty but do not represent them explicitly in the structure or parameters of their functions. These methods thus hope to learn successfully without quantifying uncertainty, which is variably feasible depending on the specific data and output being learned. AI systems that are not purpose-built to select actions in order to reduce uncertainty in an optimal manner will struggle to assign confidence to their predictions. Further, as users of these kinds of AI systems, we have no way of knowing how confident they are in their assignments of probability—they are “black boxes.”

Taken together, the probabilistic approach provides a normative theory for learning—starting from the first principles of how AI should deal with data and uncertainty. The downside to probabilistic modeling is that it induces severe computational challenges. Specifically, such models must marginalize all the variables in the model in order to arrive at exact “beliefs” about a given variable. Thus, the main computational task in probabilistic inference is marginalization, whereas in traditional AI it is the optimization
of parameters. As such, a focus on optimization per se in contemporary AI research and development may be misplaced to some extent. Current state-of-the-art AI systems are essentially general-purpose optimization machines, built to handle a specific task domain. But optimization in and of itself is not the same as intelligence. Rather, in an intelligent artifact, optimization should be a method in the service of optimizing our beliefs about what is causing our data. Fortunately, there are mathematical tricks, such as variational inference, which convert the (intractable) problem of marginalization into a (tractable) optimization problem, allowing probabilistic approaches to utilize the wealth of techniques available for optimization while retaining the benefits of uncertainty quantification.

Comparison to current state-of-the-art approaches

Active inference is a very general formulation of intelligence, understood as a self-organizing process of inference. Yet the generality of the formulation is integrative, rather than adversarial or exclusive: it formally relates or connects state-of-the-art approaches (e.g., it has been shown that all canonical neural networks minimize their free energy; (Isomura et al., 2022)), showcasing their strengths, enabling cross-pollination, and motivating refinements.

Managing complexity. In the context of machine learning, the complexity term derivable from model evidence (a.k.a., information gain) is especially interesting (Spiegelhalter et al., 2002), since it means that active inference puts predictive accuracy and complexity on an equal footing. In brief, self-evidencing bakes complexity into the optimization of beliefs about the world in a way that automatically finesse
ts many problems with machine learning schemes that focus solely on accuracy (Hochreiter and Schmidhuber, 1997). To take a salient example from recent discussions, many of the considerations that seem to motivate non-generative

Figure 1. Belief updating on a statistical manifold.
approaches—to learning world models (LeCun, 2022)—stem from considering only the likelihood in generative models, rather than model evidence or marginal likelihood—whereas, the inclusion of complexity encourages a model to find parsimonious explanations of observations, abstracting from useless detail. In other words, accuracy comes at a (complexity) cost, which must be discounted.

Complexity minimization also speaks to the importance of dimensionality reduction and coarse-graining as clear ways to learn the structure of generative models (Gershman and Niv, 2010). This is motivated by the intuition that, while sensor data itself is extremely high-dimensional, noisy, unpredictable, ambiguous, and redundant, there is a description of the data in terms of its generating causes (e.g., the world of objects with defined properties) that is lower-dimensional, more predictable, and less ambiguous. Such a description furnishes a compressed, and therefore more efficient, account of the data at hand. Thus, while scaling up data spaces, one may have to scale down the number of latent states generating those data, to the extent that doing so does not sacrifice accuracy.

Relatedly, active inference provides a promising framework for learning representations in which distinct generative factors are disentangled (Higgins et al., 2018), via the sensorimotor contingencies associated with controllable latent factors (Hinton et al., 2011; Tschantz et al., 2020; Van de Maelle et al., 2021). Low-dimensional disentangled representations, in addition to being useful for an AI system itself in achieving its own ends, are more explainable and human-interpretable than generic latent representations.

Finally, by encouraging only as much complexity or uncertainty in the model as is needed to account for the source entropy (i.e., the entropy of the generative distribution over states) (Miller et al., 2022), the variational free energy objective satisfies Ashby’s law of requisite variety, while also ensuring that no more complexity exists in a system than is needed to accurately predict observations. Moreover, the need for efficient factorizations of the variational density favor hierarchical, multi-scale systems of the kind we have been describing.

In such systems, patterns or ensembles at the superordinate level will possess a certain degree of complexity (as measured by available degrees of freedom or generative factors) that is requisite to model observations at that scale. This entails variability in the internal states of a system, sufficient to model relevant sources of external variability (this can also be motivated as a version of the principle of (constrained) maximum entropy (Jaynes, 1957; Sakhivadivel, 2022b): the optimal ‘beliefs’—encoded by internal states—are maximally entropic, given constraints supplied by the generative model). But the internal states at one scale just are what individuals at the lower scale are trying to model—so we should expect diversity among the agents participating in any such collective. Simultaneously, sociality, as implemented via belief sharing (which is necessary to the degree that we are to be predictable to one another) limits variety or entropy, and amounts to an accuracy constraint. In such a system, the sharing of beliefs broadens the evidence base available to each agent (I learn as much about the world by listening to others as by direct observation), but with built-in constraints on both conformity and eccentricity of belief (radical or unusual beliefs may emerge, but they cannot by definition be the norm in the population)—as agents both “infer together” (as part of a larger model) and “infer each other” (as each constitutes part of the external environment for the others) (Parr, 2020).

Minimizing complexity (i.e., compression) thus points to a direction of travel to the Holy Grail of a generic and robust AI; a move from “big data” to “smart data” or frugal data sampling, based on the principles of multiscale biological intelligence. This has important implications for hardware and energy efficiency. Because the complexity cost has an accompanying thermodynamic cost—via Landauer’s principle and the Jarzynski equality (Jarzynski, 1997)—there is a lower bound on the thermodynamic cost of any belief updating that can, even in principle, be realized with the right updating scheme. Using active inference, belief updating can be implemented with biomimetic efficiency, without the need for traditional, GPU-based high-performance computing and accompanying costs.

**Reinforcement learning and active inference.** State-of-the-art AI designed for action selection typically implements reinforcement learning, a set of methods for maximizing the expected sum of rewards under a sequence of actions. From a Bayesian perspective, however, curiosity and exploration are as fundamental to intelligence as maximizing reward. Specifically, the epistemic, exploratory, curiosity-driven aspect of intelligence motivates actions expected to reduce uncertainty in the variables and parameters that define your model, which, in the active inference formulation, corresponds to inference and learning, respectively (Barto et al., 2013; Lindley, 1956; MacKay, 1992; Schmidhuber, 2006; Schwartenbeck et al., 2019; Still and Precup, 2012).

In line with the above discussion of self-evidencing, rather than select actions that maximize some arbitrary state-dependent reward (or equivalently, minimize an arbitrary cost function), an intelligent system ought to generate observations or sensory data consistent with its characteristic (i.e., preferred) exchanges with the sensed world, and thus with its own continued existence. That is, an intelligent agent ought to maximize the evidence for its generative model. Active inference thereby generalizes the notion of reward, and labels every encountered outcome (and implicitly every latent state) in terms of how likely it is that “this would happen to me.” This is scored in terms of prior
preferences over outcomes, which are part of the generative model. Preferences over some kinds of outcomes are precise (e.g., not being eaten or embarrassed), others less so (“I prefer coffee in the morning, but tea is nice”). To summarize, preferences provide constraints that define the “kind of thing I am,” with more precise preferences playing a similar role, for example, to the “intrinsic costs” in the actor-critic system proposed in (LeCun, 2022).

In this view, Bayesian reinforcement learning is a special case of active inference, in which the preference for all outcomes is very imprecise—apart from one privileged outcome called reward, for which there is a very precise preference. The perspective from active inference moves our notion of intelligence away from a monothematic reward optimization problem, towards a multiple-constraint-satisfaction problem, where the implicit “satisficing” (Gigerenzer, 2010) just is self-evidencing.

Multi-scale considerations. Another key difference concerns the multi-scale architecture of active inference. First, active inference commits to a separation of temporal scales, which allows it to finesse key issues in AI research. On the present account, learning is just slow inference, and model selection is just slow learning. All three processes operate in the same basic way, over nested timescales, to maximize model evidence.

Second, active inference predicts, and provides a formalism for describing, the multi-scale character of intelligence in nature; see also (Krakauer et al., 2020). Although this has generally not been a focus of research in machine learning, work in the field consonant with this perspective includes the complex internal structure of LSTM cells (Hochreiter and Schmidhuber, 1997), the repetition of the split-transform-merge strategy across scales in the ResNeXt architecture (Xie et al., 2017), capsule networks (Sara et al., 2017), in which individually complex nodes engage in a form of self-organization, the Thousand Brains theory of the cooperation of cortical columns to produce global representations (Hawkins et al., 2017), or the perspective on restricted Boltzmann machines as products of experts (Hinton, 2002).

Relatedly, beyond fixing certain hyperparameters of system design (such as the general character of the objective function to be optimized), active inference is itself silent on the way in which model evidence maximization is implemented in particular systems. For this reason, it is crucial that work within this framework be informed by, and participate in, ongoing threads of research in machine learning and empirical neuroscience. Predictive coding, for example (Rao, 1999), is a way of realizing active inference in brains, and perhaps in similar systems with many internal degrees of freedom and shorter-timescale plasticity. Many other aspects of complex intelligence, including quite essential ones with roots deep in evolutionary history, may depend on details of that history difficult to predict from first principles alone—for example, mechanisms within the hippocampal/entorhinal system known to enable spatial navigation and localization may constitute much more general-purpose high-level design patterns for neural systems (Safron et al., 2022).

Shared narratives

We have noted that intelligence as self-evidencing is inherently perspectival, as it involves actively making sense of and engaging with the world from a specific point of view (i.e., given a set of beliefs). Importantly, if the origins of intelligence indeed lie in the partitioning of the universe into subsystems by probabilistic boundaries, then intelligence never arises singly but always exists on either side of such a boundary (Constant et al., 2018; Veissière et al., 2020). The world that one models is almost invariably composed of other intelligent agents that model one in turn.

This brings us back to the insight that intelligence must, at some level, be distributed over every agent and over every scale at which agents exist. Active inference is naturally a theory of collective intelligence. There are many foundational issues that arise from this take on intelligence; ranging from communication to cultural niche construction: from theory of mind to selfhood (Constant et al., 2018, 2019; Veissière et al., 2020; Laland et al., 1999; Vasil et al., 2020). On the active inference account, shared goals emerge from shared narratives, which are provided by shared generative models (Bouizegarene and Ramstead, 2020).

Furthermore—on the current analysis—certain things should then be curious about each other.

The importance of perspective-taking and implicit shared narratives (i.e., generative models or frames of reference) is highlighted by the recent excitement about generative AI (Sequoia Capital, 2022), in which generative neural networks demonstrate the ability to reproduce the kinds of pictures, prose, or music that we expose them to. Key to the usage of these systems is a dyadic interaction between artificial and natural intelligence, from the training of deep neural networks to the exchange of prompts and generated images with the resulting AI systems, and the subsequent selection and sharing of the most apt “reproductions” among generated outputs. In our view, a truly intelligent generative AI would then become curious about us—and want to know what we are likely to select. In short, when AI takes the initiative to ask us questions, we will have moved closer to genuine intelligence, as seen through the lens of self-evidencing.

From Babel to binary

Human intelligence and language have co-evolved, such that they both scaffold, and are scaffolded by, one another
(Heyes, 2018; Tomasello, 2016). The core functional role of language is to enable communication and shared understanding: language has been optimized for sharing with other intelligent creatures (as a language that can be easily passed on reaches further generations). Language has thus facilitated the emergence of more complex interactions and shared customs between agents, which has in turn allowed for the emergence of intensive human collaboration at multiple communal scales (Henrich, 2016). Relatedly, language provides a reference for how to “carve nature at its joints” (e.g., into objects, properties, and events), facilitating learning about the world and the way it works. Finally, it has allowed humans to build an external store of knowledge far beyond the epistemic capacity of any human individual. Human beings both benefit from—and contribute to—this store of knowledge, which, like language itself, has co-evolved with our intelligence.

Across cultures, the earliest recorded narratives of our species have emphasized the astounding integrative power of shared communication systems along with their flipside: the discord and disarray wrought by miscommunication and a lack of mutual understanding. This is illustrated potently in the biblical story of the Tower of Babel, which tells of a mighty civilization that attempted to build a glorious city with a tower that rose to the heavens. These lofty aspirations fell to ruin after a divine disruption that eliminated their common language, shattering it into a thousand unintelligible dialects. In their confusion and miscomprehension, they were unable to complete the Tower and were thus scattered across the Earth, forced to survive in the clustered tribes that shared their regional vernacular.

Today, humans cope with a “post-Babel” world via a combination of increasing multilingualism, rallying (for better or worse) behind hegemonic languages like English, and, recently, increasingly effective machine translation (Wu et al., 2016). Digital computers do share a common or universal machine language (i.e., binary representation). If situations can be represented adequately in an appropriate machine syntax, they can be subjected to the operations of mathematical logic, formalized and thereby processed in an unambiguous way. At a higher level, it may be said that “vectorese” is the universal language of AI, in that vectors (i.e., ordered lists of numbers representing a point in an abstract space) constitute the input, output, and medium of data storage and retrieval for most AI algorithms.

Vectors are analogous to the medium of action potentials in the brain—they are capable of representing anything we can think of, but nearly all the interesting (and representationally load-bearing) structure lies in the (learned) transformations and accompanying transition dynamics of the underlying dynamical system. Often, an output vector space can be considered as an embedding or transformation of the input space, and mappings among vector spaces are much like translations among languages. However, vectors themselves may only provide a base structure or medium (analogous to sound or light) for higher-level languages.

It has been clear from the early days of neural language modeling that vector space representations can in principle be learned that capture both the semantic and syntactic regularities implicit in the co-occurrence statistics of natural language corpora (Collobert et al., 2011; Mikolov et al., 2013). Despite this, we lack anything like a common high-level language that AIs can use to communicate with one another and with humans—other than, arguably, human natural languages themselves, which can be used to interface with AIs via modern language models. The fact that reinforcement learning agents trained to produce prompts for such models often produce unintelligible nonsense strings, however Webson and Pavlick (2022), shows that even where large language models use English, they do not use or understand it in the way humans do; this raises the question whether natural languages can really play the role of a shared human–machine language without modification.

Moreover, while the necessity of serializing thought into discrete token sequences for the purposes of communication helps enforce the kind of sparsity structure that we have argued is essential to intelligence and complexity itself, a more direct form of information transfer is also conceivable in which the richness of a latent vector representation (or “thought”) is directly externalized as a data structure. While current state-of-the-art AI can learn the language of vector space embeddings, the science of inter-AI communication and shared latent spaces is in its infancy. For the most part, each AI must learn to carve up the world from scratch, and is unable to share its knowledge fluidly or update it in collaboration with other AIs.6

We argue that the future evolution of AI would benefit greatly from a focus on optimization for shareability (i.e., gathering evidence for a model of an intrinsically social creature.) This might take the form of a shared external store of knowledge about how to communicate with relevant others, or a structured communication protocol that can act as the lingua franca of AI. A general framework that ties together different embedding spaces and inter-AI messaging over a shared network architecture would, among other things, enable AI agents to learn to offload certain tasks or predictions to other, more specialized AI agents.

**Active inference and communication**

An underlying theme thus far is that intelligence at any scale requires a shared generative model and implicit common ground. There are many ways to articulate this theme; from ensemble learning to mixtures of experts (Hinton, 2002), from distributed cognition to Bayesian model averaging (FitzGerald et al., 2014).

Imagine that someone has locked you in a large dark room. As a self-evidencing and curious creature, you would
be compelled to feel your way around to resolve uncertainty about your situation. Successive palpations lead you to infer that there is a large animal in the room—by virtue of feeling what seem to be a tail, a succession of legs, and eventually a trunk. Your actions generate accumulated evidence for the hypothesis “I am in a room with an elephant.” Now, imagine an alternative scenario in which you and five friends are deployed around the same room, and can report what you feel to each other. In this scenario, you quickly reach the consensus “We are in a room with an elephant.” The mechanics of belief updating are similar in both scenarios. In the first, you accumulate evidence and successively update your posterior belief about latent states. In the second, the collective assimilation of evidence is parallelized across multiple individuals.

Is the latter equivalent to having one brain with twelve hands? Not quite. The second kind of belief updating rests upon a shared generative model or hypothesis space that enables you to assimilate the beliefs of another. For example, you share a common notion of a “trunk,” a “leg,” and a “tail”—and crucially, you have access to a shared language for communicating such concepts. Sharing a generative model allows each agent to infer the causes of its sensations and disentangle the causes that are unique to the way the world is sampled—for example, “where I am looking”—and causes that constitute the shared environment (e.g., “what I am looking at”) (Friston et al., 2015; Veissière et al., 2020; Ramstead et al., 2016). Just as importantly, any dyad or ensemble of self-evidencing agents will come to share a generative model (or at least some factors of a generative model) via their interactions (Sakhivadivel, 2022a) (see (Friston and Frith, 2015; Isomura et al., 2019) for numerical experiments in active inference that illustrate this phenomenon, and Table A.1 for related applications.)

What results is a shared intelligence (i.e., a kind of collective super-intelligence) that emerges from an ensemble of agents. Heuristically, maximizing model evidence means making the world as predictable as possible (Albarracin et al., 2022; Kastel and Hesp, 2021). This is assured if we are both singing from the same hymn sheet, so to speak—so that I can predict you and you can predict me. Mathematically, this is evinced as a generalized synchrony between the dynamics on our respective statistical manifolds (Friston and Frith, 2015; Palacios et al., 2019). This generalized synchrony (or synchronicity) is special because it unfolds in a (shared) belief space, meaning it can be read as mutual understanding: that is, coming to align our beliefs, via a shared language and a shared generative model. This sharedness is arguably the basis of culture and underpins the existence of our civilization. Our challenge, which we take to be a necessary step toward ASI or even AGI, is to expand the sphere of culture to include artificial agents.

**Belief propagation, graphs, and networks**

Operationally, ecosystems of shared intelligence can be described in terms of message passing on a factor graph (Dauwels, 2007; Friston et al., 2017; Winn and Bishop, 2005; Yedidia et al., 2005), a special kind of graph or network in which nodes correspond to the factors of a Bayesian belief or probability distribution. Factors are just probabilistic beliefs that one multiplies together to get a joint distribution (i.e., a generative model). For example, one could factorize beliefs about the latent states of an object into “what” and “where.” These beliefs jointly specify a unique object in extrapersonal space; noting that knowing what something is and knowing where it is are largely independent of each other (Ungerleider and Haxby, 1994). The edges of a factor graph correspond to the messages passed among factors that underwrite belief updating. In the implementations of active inference that we have been describing, they comprise the requisite sufficient statistics that summarize the beliefs of other nodes.

Technically, this is useful because for any generative model there is a dual or complementary factor graph that prescribes precisely the requisite message passing and implicit computational architecture. In our setting, this architecture has an interesting aspect: we can imagine the nodes of a vast graph partitioned into lots of little subgraphs. Each of these would correspond to an agent updating its beliefs via the propagation of internal messages. Conversely, external messages would correspond to communication and belief sharing that rests upon certain factors being distributed or duplicated over two or more subgraphs (i.e., agents or computers). This kind of architecture means that, in principle, any subgraph or agent can see, vicariously, every observable in the world—as seen through the eyes of another agent. But what is the functional and structural form of the generative model that underwrites such an architecture?

Taking our lead from human communication, the most efficient (minimum description length or minimum-complexity) generative model of worldly states should be somewhat simplified (i.e., coarse-grained), leveraging discrete representations with only as much granularity as is required to maintain an accurate account of observations (Heins et al., 2022; Klein and Hoel, 2020). There are many motivations for this kind of generative model. First, it is continuous with the approach to thing-ness or individuation described above, according to which individuals are defined by the sparsity of their interactions. Concepts should evince a sparse structure, both because they are themselves “things” (and so should have sparse connections to other similar “things”), and because they are accurate representations of a world characterized by sparsity. Second, belief updating can, in this case, be implemented with simple linear operators, of the sort found in quantum computation (Fields et al., 2021, 2022; Parrondo et al., 2015).
Furthermore, this kind of discretization via coarse-graining moves us into the world of the theory of signs or semiotics (Roy, 2005; Sewell, 1992), Boolean logic and operators, and the sort of inference associated with abductive reasoning (Peirce, 1931). Finally, it finesse the form of message passing, since the sufficient statistics of discrete distributions can be reduced to a list of the relative probabilities of being in the states or levels of any given factor (Ghahramani and Jordan, 1997), enabling AI systems to flexibly switch contexts and acquire knowledge from others quickly and adaptively, based on a repository of shared representations.

**Intelligence at scale**

A subtle aspect of active inference, in this setting, is the selection of which messages to listen or attend to. In principle, this is a solved problem—in the simple case, each agent (i.e., subgraph) actively selects the messages or viewpoints that afford the greatest expected information gain (Friston et al., 2020). The neurobiological homologue of this would be attention: selecting the newsworthy information that resolves uncertainty about things you do not already know, given a certain context. There are many interesting aspects of this active (action-oriented) aspect of message passing; especially when thinking about nested, hierarchical structures in a global (factor) graph. In these structures—and in simulations of hierarchical processing in the brain—certain factors at higher hierarchical levels can control the selection of messages by lower levels (Parr and Friston, 2017; Smith et al., 2019). This motivates exploration of the multi-scale aspects of shared intelligence.

The emerging picture is of a message passing protocol that instantiates variational message passing on graphs of discrete belief spaces. But what must these messages contain? Clearly, on the present proposal, they must contain vectors of sufficient statistics; but they also have to identify themselves in relation to the (shared) factors to which they pertain. Furthermore, they must also declare their origin, in much the same way as neuronal populations in the brain receive spatially addressed inputs from other parts of the brain.

In a synthetic setting, this calls for spatial addressing, leading to the notion of a spatial message passing protocol and modeling language—of the sort being developed as open standards in the Institute of Electrical and Electronics Engineers (IEEE) P2874 Spatial Web Working Group (Standard for Saptial Web Protocol, Architecture and Governance). In short, the first step—toward realizing the kind of distributed, emergent, shared intelligence we have in mind—is to construct the next generation of modeling and message passing protocols, which include an irreducible spatial addressing system amenable to vectorization, and allowing for the vector-based shared representation of much of human knowledge.

**Ethical and moral considerations**

We conclude our discussion of large-scale collective intelligence with a brief discussion of the relevant areas of ethical discussion—and contention. First, it is important to note that the kind of collective intelligence evinced by eusocial insects (e.g., ant colonies), in which most individuals are merely replaceable copies of one another, is not the only paradigm for shared intelligence—nor is it a suitable one for systems in which individual nodes embody complex generative models. We believe that developing a cyber-physical network of emergent intelligence in the manner described above not only ought to, but for architectural reasons must, be pursued in a way that positively values and safeguards the *individuality* of people (as well as potentially non-human persons).

This idea is not new. Already in the late 1990s, before the widespread adoption of the internet as a communication technology, a future state of society had been hypothesized in which the intrinsic value of individuals is acknowledged in part because knowledge is valuable and *knowledge and life are inseparable* (Levy and Bonanno, 1997)—that is, each person has a distinct and unique life experience and, as such, knows something that no one else does. This resonates deeply with our idea that every intelligence implements a generative model of its own existence. The form of collective intelligence that we envision can emerge only from a network of essentially unique, epistemically and experientially diverse agents. This useful diversity of perspectives is a special case of functional specialization across the components of a complex system.

Much discussion in the field of AI ethics focuses on the problem of AI alignment; that is, aligning our value systems with those of hypothetical conscious AI agents, which may possibly evince forms of super-intelligence (Russell, 2019; Allen et al., 2005; Russell et al., 2015); for critical discussion, see (Marcus and Davis, 2019). This can be discussed under the broader rubric of the capacity for empathy or sympathy—what one might call sympathetic intelligence—which concerns the ability of agents to share aspects of their generative models, to take the perspective of other agents, and to understand the world in ways similar enough to enable coordinated action. This likely requires avoiding undesirable equilibria, such as those evincing pathologies of alignment (e.g., the elimination of a healthy diversity of perspectives), as well as those resembling the predator-prey systems found in nature*. Whether the emergence of shared intelligence in such a network structure entails the emergence of a new, collective mind is an open question.

Current state-of-the-art AI systems are largely “black boxes.” Such an approach to the design of AI ultimately puts severe limits on its transparency, explainability, and auditability. In addition, their capacity to engage in genuine
collaboration with humans and other AI is limited, because they lack the ability to take the perspective of another. Moving to multi-scale active inference offers a number of technical advantages that may help address these problems. One is that leveraging explicit generative models, which carve the world into discrete latent states, may help us to identify and quantify bias in our models. Such architectures feature increased auditability, in that they are explicitly queryable and their inferences can be examined forensically—allowing us to address these biases directly. Shared generative models also effectively equip AI with a theory of mind, facilitating perspective-taking and allowing for genuinely dyadic interactions.

Much like a brain, with its many layers and connections, the multi-scale architecture for collective intelligence that we propose could be equipped with nodes and layers to enable a kind of collective self-monitoring and self-organization of salience. However, this raises the question of authority and power: this kind of approach to the design of AI must account for the plurality and vulnerability of individual perspectives, and the need to understand and counterbalance potential abuses of power. More broadly and perhaps more fundamentally, we note that the approach to AI that we have presented here does not obviate the dangers associated with bias in AI technologies, especially when deployed at industrial scale in commercial settings, for example, (Birhane, 2021). The general idea is that the deployment of AI technologies in societies that have preexisting hierarchies of power and authority can have problematic consequences. For example, discriminatory bias encoded in data will result in unfairly biased AI systems (e.g., Birhane, 2022) regardless of the specific technologies used to build that AI. It is highly probable that the use of AI technologies premised on such data will sustain social biases and practices that are harmful, or may represent future harm, the consequences of which are not yet fully known—or may be unknowable—regardless of the intentions of the creators. These concerns are well founded and cannot be resolved through narrowly technical means. As such, some combination of novel social policies, government regulations, and ethical norms are likely to be required to ensure that these new technologies harness and reflect our most essential and persistent values.

We are not pessimistic. Nature provides us with endless demonstrations of the success of emergent, shared intelligence across systems at every scale. Looking back to the elegant design of the human body, we find bountiful examples of diverse systems of nested intelligences working together to seek out a dynamic harmony and balance. As an integrated system, the body is capable of achieving multi-scale homeostasis and allostatics, notably via the incredible coordination and communicative power of the nervous system, allowing it to adapt to novel environmental conditions and to regulate its needs in real time. We reiterate our conviction that the design of AI should be informed by, and aligned with, these time-tested methods and design principles. Furthermore, we believe that the class of sympathetic and shared intelligences that we have described in this paper offers a responsible and desirable path to achieving the highest technical and ethical goals for AI, based on a design of ecosystems of intelligence from first principles.

Conclusion: Our proposal for stages of development for active inference as an artificial intelligence technology

The aim of this white paper was to present a vision of research and development in the field of artificial intelligence for the next decade (and beyond). We suggested that AGI and ASI will emerge from the interaction of intelligences networked into a hyper-spatial web or ecosystem of natural and artificial intelligence. We have proposed active inference as a technology uniquely suited to the collaborative design of an ecosystem of natural and synthetic sense-making, in which humans are integral participants—what we call shared intelligence. The Bayesian mechanics of intelligent systems that follows from active inference led us to define intelligence operationally, as the accumulation of evidence for an agent’s generative model of their sensed world—also known as self-evidencing. This self-evidencing can be implemented using message passing or belief propagation on (factor) graphs or networks. Active inference is uniquely suited to this task because it leads to a formal account of collective intelligence. We considered the kinds of communication protocols that must be developed to enable such an ecosystem of intelligences, and argued that such considerations motivate the development of a generalized, hyper-spatial modeling language and transaction protocol. We suggest that establishing such common languages and protocols is a key enabling step towards an ecosystem of naturally occurring and artificial intelligences.

In closing, we provide a roadmap for developing intelligent artifacts and message passing schemes as methods or tools for the common good. This roadmap (summarized in Table 1) is inspired by the technology readiness levels (TRLs) that have been adopted as a framework for understanding progress in technical research and development by institutions such as the European Commission, the International Organization for Standardization (ISO), and the National Aeronautics and Space Administration agency (NASA).

Stages of development for active inference

S0: Systemic Intelligence. This is contemporary state-of-the-art AI, namely, universal function approximation—mapping from input or sensory states to outputs or action states—that optimizes some well-defined value function or cost of (systemic) states. Examples include deep learning and Bayesian reinforcement learning.
S1: Sentient intelligence. Sentient behavior or active inference based on belief updating and propagation (i.e., optimizing beliefs about states as opposed to states per se); where “sentient” means “responsive to sensory impressions.”

This entails planning as inference, namely, inferring courses of action that maximize expected information gain and expected value, where value is part of a generative (i.e., world) model, namely, prior preferences. This kind of intelligence is both information-seeking and preference-seeking. It is quintessentially curious.

S2: Sophisticated intelligence. Sentient behavior—as defined under S1—in which plans are predicated on the consequences of action for beliefs about states of the world, as opposed to states per se. That is, a move from “what will happen if I do this?” to “what will I believe or know if I do this?” (Friston et al., 2021; Hesp et al., 2020). This kind of inference generally uses generative models with discrete states that “carve nature at its joints”; namely, inference over coarse-grained representations and ensuing world models.

This kind of intelligence is amenable to formulation in terms of modal logic, quantum computation, and category theory. This stage corresponds to “artificial general intelligence” in the popular narrative about the progress of AI.

S3: Sympathetic (or Sapient) intelligence. The deployment of sophisticated AI to recognize the nature and dispositions of users and other AI and—in consequence—recognize (and instantiate) attentional and dispositional states of self, namely, a kind of minimal selfhood (which entails generative models equipped with the capacity for Theory of Mind). This kind of intelligence is able to take the perspective of its users and interaction partners—it is perspectival, in the robust sense of being able to engage in dyadic and shared perspective-taking.

S4: Shared (or Super) intelligence. The kind of collective that emerges from the coordination of Sympathetic Intelligence (as defined in S3) and their interaction partners or users—which may include naturally occurring intelligence such as ourselves, but also other sapient artifacts. This stage corresponds, roughly speaking, to “artificial super-intelligence” in the popular narrative about the progress of AI—with the important distinction that we believe that such intelligence will emerge from dense interactions between agents networked into a hyper-spatial web. We believe that the approach that we have outlined here is the most likely route toward this kind of hypothetical, planetary-scale, distributed super-intelligence (Frank et al., 2022).

Implementation

Theoretical. The basis of belief updating (i.e., inference and learning) is underwritten by a formal calculus (e.g., Bayesian mechanics), with clear links to the physics of self-organization of open systems far from equilibrium.

Proof of principle. Software instances of the formal (mathematical) scheme, usually on a classical (i.e., von Neumann) architecture.

Deployment at scale. Scaled and efficient application of the theoretical principles (i.e., methods) in a real-world setting

Table 1. Stages of AI premised on active inference.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Theoretical</th>
<th>Proof of principle</th>
<th>Deployment at scale</th>
<th>Biomimetic</th>
<th>Timeframe</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0: Systemic</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>N/A</td>
<td>Now</td>
</tr>
<tr>
<td>S1: Sentient</td>
<td>Established</td>
<td>Established</td>
<td>Established</td>
<td>Provisional</td>
<td>6 months</td>
</tr>
<tr>
<td>S2: Sophisticated</td>
<td>Established</td>
<td>Established</td>
<td>Provisional</td>
<td>Provisional</td>
<td>1 year</td>
</tr>
<tr>
<td>S3: Sympathetic</td>
<td>Established</td>
<td>Established</td>
<td>Provisional</td>
<td>Provisional</td>
<td>2 years</td>
</tr>
<tr>
<td>S4: Shared</td>
<td>Established</td>
<td>Established</td>
<td>Provisional</td>
<td>Aspirational</td>
<td>8 years</td>
</tr>
</tbody>
</table>

(e.g., edge-computing, robotics, and variational message passing on the web).

**Biomimetic hardware.** Implementations that elude the von Neumann bottleneck, on biomimetic or neuromorphic architectures. For example, photonics, soft robotics, and belief propagation, that is, message passing of the sufficient statistics of (Bayesian) beliefs.

**Acknowledgments**

The authors thank Rosalyn Moran and George Percivall for useful discussions. Table A.1 in Appendix A has been reproduced from (Da Costa et al., 2020) under a CC BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/).

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: All work on this paper was funded by VERSES. KF is supported by funding for the Wellcome Centre for Human Neuroimaging (Ref: 2050103/Z/16/Z) and a Canada-UK Artificial Intelligence Initiative (Ref: ES/T01279X/1). CH is supported by the U.S. Office of Naval Research (Ref: N00014-19-1-2556). BK & CH acknowledge the support of a grant from the John Templeton Foundation (Ref: 61780). The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. BM was funded by Rafal Bogacz with a BBSRC grant (Ref: BB/s006338/1) and a MRC grant (Ref: MC UU 00003/1). SET is supported in part by funding from the Social Sciences and Humanities Research Council of Canada (Ref: 767-2020-2276).

**ORCID iDs**

Alex B Kiefer [https://orcid.org/0000-0003-1957-797X](https://orcid.org/0000-0003-1957-797X)  
Brennan Klein [https://orcid.org/0000-0001-8326-5044](https://orcid.org/0000-0001-8326-5044)  
Dalton AR Sakthivadivel [https://orcid.org/0000-0002-7907-7611](https://orcid.org/0000-0002-7907-7611)

**Supplemental Material**

Supplemental material for this article is available online.

**Notes**

1. Researchers in AI have often borrowed tools from physics, such as Hamiltonian mechanics, to finesse the inference problems that they face, leading to tools like the Hamiltonian Monte Carlo algorithm, which massively speeds up certain kinds of inferential problem-solving (Neal, 2011). Conversely, AI has been used in physics, chemistry, and biochemistry to great effect, allowing us to simulate the containment of plasma in Tomahawk nuclear fusion reactors (Delgráve et al., 2022), or predict the ways in which proteins will fold, as the famous AlphaFold system enables (Jumper et al., 2021). What we have in mind, however, is not to borrow techniques or formalisms from physics to solve the problem of intelligent systems design, or to use AI to help finesse problems from physics; but rather, in a complementary fashion, to treat the study of intelligence itself as a chapter of physics.

2. Simulated neural networks, for example, often overfit and fail to generalize if they are not forced to learn compressed representations of their inputs (MacKay, 1995; Olshausen and Field, 1996; Schmidhuber, 2010; Wallace and Dowe, 1999). Relatedly, ubiquitous forms of regularization can be motivated from physical considerations about the finite bandwidth of neurons (Sengupta et al., 2010), and schemes such as predictive coding and sparse coding by considerations about efficient signal transmission (Olshausen and Field, 1996; Elias, 1955; Rao, 1999; Lance, 1987; Barlow and Rosenblith, 1961; Simoncelli and Olshausen, 2001).

3. In the context of scientific modeling, a statistical model is a mathematical object that encodes the way that things change, relative to the way that other things change. Formally, the structure that encodes such contingencies is called a joint probability distribution. This is the generative model.

4. Even rocks, while not agents per se, track the state of their environment: for instance the interior of a rock “knows” that the environment must be well below the melting point of rock (albeit not under that English description). As systems become more elaborate, they can represent more about the things to which they couple (Dennett, 1983).

5. The importance of fluid exchange between artificial and human intelligence in this paradigm is evinced by the rapidly growing interest in prompt engineering, i.e., an increasingly self-aware and theory-driven approach to the role that prompts play in co-creating the outputs of these types of systems (Liu et al., 2022), which has recently been extended to the optimization of text prompts by distinct AI agents (Deng et al., 2022).

6. An important exception is the proliferation of fine-tuned copies of large monolithic pre-trained models such as BERT. This is not obviously relevant to our interest in (possibly real-time) communication and mutual updating among persistent, physically situated AI systems, though it may constitute a form of evolution of populations of AI systems with partially divergent learning histories.

7. See (Albarracin et al., 2022) for more complex cases where agents have preferences for certain kinds of interaction partners, resulting in the formation of “echo chambers.”

8. We thank George Percivall for raising these points.

9. To preempt any worries, we emphasize that we do not mean that sentient intelligent systems are necessarily conscious, in the sense of having qualitative states of awareness; e.g., as the word was used in the recent controversy surrounding Google’s AI system LaMDA (Thopilán, 2022). It is standard to use the word “sentient” to mean “responsive to sensory impressions” in the
literature on the free energy principle; e.g., in (Ramstead et al., 2019).

References


