entropy

MDPI

Article

An Active Inference Model of Collective Intelligence

Rafael Kaufmann [1], Pranav Gupta [2] and Jacob Taylor [3,4,*]

[1] Independent Researcher, Brooklyn, NY 11215, USA; rkauf@google.com
[2] Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, USA; pranavgu@andrew.cmu.edu
[3] Institute of Cognitive & Evolutionary Anthropology, University of Oxford, Oxford OX2 6PN, UK
[4] Crawford School of Public Policy, Australian National University, Canberra, ACT 2601, Australia
* Correspondence: jacob.taylor@anu.edu.au

Abstract: Collective intelligence, an emergent phenomenon in which a composite system of multiple interacting agents performs at levels greater than the sum of its parts, has long compelled research efforts in social and behavioral sciences. To date, however, formal models of collective intelligence have lacked a plausible mathematical description of the relationship between local-scale interactions between autonomous sub-system components (individuals) and global-scale behavior of the composite system (the collective). In this paper we use the Active Inference Formulation (AIF), a framework for explaining the behavior of any non-equilibrium steady state system at any scale, to posit a minimal agent-based model that simulates the relationship between local individual-level interaction and collective intelligence. We explore the effects of providing baseline AIF agents (Model 1) with specific cognitive capabilities: Theory of Mind (Model 2), Goal Alignment (Model 3), and Theory of Mind with Goal Alignment (Model 4). These stepwise transitions in sophistication of cognitive ability are motivated by the types of advancements plausibly required for an AIF agent to persist and flourish in an environment populated by other highly autonomous AIF agents, and have also recently been shown to map naturally to canonical steps in human cognitive ability. Illustrative results show that stepwise cognitive transitions increase system performance by providing complementary mechanisms for alignment between agents' local and global optima. Alignment emerges endogenously from the dynamics of interacting AIF agents themselves, rather than being imposed exogenously by incentives to agents' behaviors (contra existing computational models of collective intelligence) or top-down priors for collective behavior (contra existing multiscale simulations of AIF). These results shed light on the types of generic information-theoretic patterns conducive to collective intelligence in human and other complex adaptive systems.

Keywords: collective intelligence; free energy principle; active inference; agent-based model; complex adaptive systems; multiscale systems; computational model

Citation: Kaufmann, R.; Gupta, P.; Taylor, J. An Active Inference Model of Collective Intelligence. *Entropy* **2021**, *23*, 830. https://doi.org/10.3390/e23070830

Academic Editors: Paul Badcock, Maxwell Ramstead, Zahra Sheikhbahaee and Axel Constant

Received: 1 April 2021
Accepted: 12 June 2021
Published: 29 June 2021

1. Introduction

Human collectives are examples of a specific subclass of complex adaptive system, the sub-system components of which—individual humans—are themselves highly autonomous complex adaptive systems. Consider that, subjectively, we perceive ourselves to be autonomous individuals at the same time that we actively participate in collectives. Families, organizations, sports teams, and polities exert agency over our individual behavior [1,2] and are even capable, under certain conditions, of intelligence that cannot be explained by aggregation of individual intelligence [3,4]. To date, however, formal models of collective intelligence have lacked a plausible mathematical description of the functional relationship between individual and collective behavior.

In this paper, we use the Active Inference Framework (AIF) to develop a clearer understanding of the relationship between patterns of individual interaction and collective intelligence in systems composed of highly autonomous subsystems, or "agents". We

adopt a definition of collective intelligence established within organizational psychology, as groups of individuals capable of acting collectively in ways that seem intelligent and that cannot be explained by individual intelligence [5] (p.3). As we outline below, collective intelligence can be operationalized under AIF as a composite system's ability to minimize free energy or perform approximate Bayesian inference at the collective level. To demonstrate the formal relationship between local-scale agent interaction and collective behavior, we develop a computational model that simulates the behavior of two autonomous agents in state space. In contrast to typical agent-based models, in which agents behave according to more rudimentary decision-making algorithms (e.g., from game theory; see [6]), we model our agents as self-organizing systems whose actions are themselves dictated by the directive of free energy minimization relative to the "local" degrees of freedom accessible to them, including those that specify their embedding in the larger system [7–9]. We demonstrate that AIF may be particularly useful for elucidating mechanisms and dynamics of systems composed of highly autonomous interacting agents, of which human collectives are a prominent instance. But the universality of our formal computational approach makes our model relevant to collective intelligence in any composite system.

1.1. Motivation: The "Missing Link" between Individual-Level and System-Level Accounts of Human Collective Intelligence

Existing formal accounts of collective intelligence are predicated on composite systems whose sub-system components are subject to vastly fewer degrees of freedom than individuals in human collectives. Unlike ants in a colony or neurons in a brain, which appear to rely on comparatively rudimentary autoregulatory mechanisms to sustain participation in collective ensembles [10,11], human agents participate in collectives by leveraging an array of phylogenetic (evolutionarily) and ontogenetic (developmental) mechanisms and socio-culturally constructed regularities or affordances (e.g., language) [12–14]. Human agents' cognitive abilities and sociocultural niches create avenues for active participation in functional collective behavior (e.g., the pursuit of shared goals), as well as avenues to shirk global constraints in the pursuit of local (individual) goals. Mathematical models for collective intelligence of this subclass of system must not only seek to account for richer complexity of agent behavior at each scale of the system (particularly at the individual level), but also the relationship between local scale interaction between individual agents and global scale behavior of the collective.

Existing research of human collective intelligence is limited precisely by a lack of alignment between these two scales of analysis. On the one hand, accounts of local-scale interactions from behavioral science and psychology tend to construe individual humans as goal-directed individuals endowed with discrete cognitive mechanisms (specifically social perceptiveness or Theory of Mind and shared intentionality; see [15,16]) that allow individuals to establish and maintain adaptive connections with other individuals in service of shared goals [3–5,17–19] (Riedl and colleagues [19] report a recent analysis of 1356 groups that found social perceptiveness and group interaction processes to be strong predictors of collective intelligence measured by a psychometric test.). Researchers conjecture that these mechanisms allow collectives to derive and utilize more performance-relevant information from the environment than could be derived by an aggregation of the same individuals acting without such connections (for example, by facilitating an adaptive, system-wide balance between cognitive efficiency and diversity; see [4]). Empirical substantiation of such claims has proven difficult, however. Most investigations rely heavily on laboratory-derived summaries or "snapshots" of individual and collective behavior that flatten the complexity of local scale interactions [20] and make it difficult to examine causal relationships between individual scale mechanisms and collective behavior as they typically unfold in real world settings [21,22].

Accounts of global-scale (collective) behavior, by contrast, tend to adopt system-based (rather than agent-based) perspectives that render collectives as random dynamical systems in phase space, or equivalent formulations [23–26]. Only rarely deployed to assess the construct of human collective intelligence specifically (e.g., [27]), these approaches have

been fruitful for identifying gross properties of phase-space dynamics (such as synchrony, metastability, or symmetry breaking) that correlate with collective intelligence or collective performance, more generally construed [28–32]. However, on their own, such analyses are limited in their ability to generate testable predictions for multiscale behavior, such as how global-scale dynamics (rendered in phase-space) translate to specific local-scale interactions between individuals (in state-space), or how local-scale interactions between individuals translate to evolution and change in collective global-scale dynamics [26].

In sum, the substantive differences between these two analytical perspectives (individual and collective) on collective intelligence in human systems make it difficult to develop a formal description of how local-scale interactions between autonomous individual agents relate to global-scale collective behavior and vice versa. Most urgent for the development of a formal model of collective intelligence in this subclass of system, therefore, is a common mathematical framework capable of operating between individual-level cognitive mechanisms and system-level dynamics of the collective [4].

1.2. The Free Energy Principle and an Active Inference Formulation of Collective Intelligence

FEP has recently emerged as a candidate for this type of common mathematical framework for multiscale behavioral processes [33–35]. FEP is a mathematical formulation of how adaptive systems resist a natural tendency to disorder [33,36]. FEP states that any non-equilibrium steady state system self organizes as such by minimizing variational free energy in its exchanges with the environment [37]. The key trick of FEP is that the principle of free energy minimization can be neatly translated into an agent-based process theory, AIF, of approximate Bayesian inference [38] and applied to any self-organizing biological system at any scale [39]. The upshot is that, in theory, any AIF agent at one spatio-temporal scale could be simultaneously composed of nested AIF agents at the scale below, and a constituent of a larger AIF agent at the scale above it [40–42]. In effect, AIF allows you to pick a composite system or agent A that you want to understand, and it will be generally true both that: A is an approximate, global minimizer of free energy at the scale at which that agent reliably persists; and A is composed of subsystems {A_i} that are approximate, local minimizers of free energy (which is composed of the remainder of A). Thus, under AIF, collective intelligence can conceivably be modelled as a case of individual AIF agents that interact within—or indeed, interact to produce—a superordinate AIF agent at the scale of the collective [9,43]. In this way, AIF provides a framework within which a multiscale model of collective intelligence could be developed. The aim of this paper is to propose a provisional AIF model of collective intelligence that can depict the relationship between local-scale interactions and collective behavior.

An AIF model of collective intelligence begins with the depiction of a minimal AIF agent. Specifically, an AIF agent denotes any set of states enclosed by a "Markov blanket"—a statistical partition between a system's internal states and external states [44]—that infers beliefs about the causes of (hidden) external states by developing a probabilistic *generative model* of external states [37]. A Markov blanket is composed of sensory states and active states that mediate the relationship between a system's internal states and external states: external states (ψ) act on sensory states (s), which influence, but are not influenced by internal states (b). Internal states couple back through active states (a), which influence but are not influenced by external states. Through conjugated repertoires of perception and action, the agent embodies and refines (learns) a generative model of its environment [45] and the environment embodies and refines its model of the agent (akin to a circular process of environmental niche construction; see [12]).

Having established the notion of an AIF agent, the next step in developing an AIF model of collective intelligence is to consider the existence of multiple nested AIF agents across individual and collective scales of organization. Existing multiscale treatments of AIF provide a clear account of "downward reaching" causation, whereby superordinate AIF agents like brains or multicellular organisms systematically determine [46] the behavior of subordinate AIF agents (neurons or cells), limiting their behavioral degrees

of freedom [9,40,47,48]. Consistent with this account of downward-reaching causation, existing toy models that simulate the emergence of collective behavior under AIF do so by simply using the statistical constraints from one scale to drive behavior at another, e.g., by explicitly endowing AIF agents with a genetic prior for functional specialization within a superordinate system [9] or by constructing a scenario in which the emergence of a superordinate agent at the global scale is predestined by limiting an agent's model of the environment to sensory evidence generated by a counterpart agent [7,8].

While perhaps useful for depicting the behavior of cells within multicellular organisms [9] or exact behavioral synchronization between two or more agents [7,8], these existing AIF models are less well-suited to explain collective intelligence in human systems, for two reasons. First, humans are relatively autonomous individual agents whose statistical boundaries for self-evidencing appear to be transient, distributed, and multiple [49–52]. Therefore, human collective intelligence cannot be explained simply by the way in which global-level system regularities constrain individual interaction from the "top-down". Second, the behavior of the collective in these toy models reflects the instructions or constraints supplied exogenously by the "designer" of the system, not a causal consequence of individual agents' autonomous problem-solving enabled by AIF. In this sense, extant models of AIF for collectives bear a closer resemblance to Searle's [53] "Chinese Room Argument" than to what we would recognize as emergent collective intelligence.

In sum, currently missing from AIF models of composite systems are specifications for how a system's emergent global-level cognitive capabilities causally relate to individual agents' emergent cognitive capabilities, and how local-scale interactions between individual AIF agents give rise, *endogenously*, to a superordinate AIF agent that exhibits (collective) intelligence [43]. Specifically, existing approaches lack a description of the key cognitive mechanisms of AIF agents that might provide a functional "missing link" for collective intelligence. In this paper, we initiate this line of inquiry by exploring whether some basic information-theoretic capabilities of individual AIF agents, motivated by analogies with human social capabilities, create opportunities for collective intelligence at the global scale.

1.3. Our Approach

To operationalize AIF in a way that is useful for investigating this question, we begin by examining what minimal features of autonomous individual AIF agents are required to achieve collective intelligence, operationalized as active inference at the level of the global-scale system. We conjecture that very generic information theoretic patterns of an environment in which individual AIF agents exploit other AIF agents as affordances of free energy minimization should support the emergence of collective intelligence. Importantly, we expect that these patterns emerge under very general assumptions and from the dynamics of AIF itself—without the need for exogenously imposed fitness or incentive structures on local-scale behavior, contra extant computational models of collective intelligence (that rely on cost or utility functions; e.g., [54,55]) or other common approaches to reinforcement learning (that rely on exogenous parameters of the Bellman equation; see [56,57]).

To justify our modelling approach, we draw upon recent research that systematically maps the complex adaptive learning process of AIF agents to empirical social scientific evidence for cognitive mechanisms that support adaptive human social behavior. In line with this research, we posit a series of stepwise progressions or "hops" in the individual cognitive ability of any AIF agent in an environment populated by other self-similar AIF agents. These hops represent evolutionarily plausible "adaptive priors" [42] (p.109) that would likely guide action-perception cycles of AIF agents in a collective toward unsurprising states:

- **Baseline AIF**—AIF agents, to persist as such, will minimize immediate free energy by accurately sensing and acting on salient affordances of the environment. This will require a general ability for "perceptiveness" of the (physical) environment.
- **Folk Psychology**—AIF agents in an environment populated by other AIF agents would fare better by minimizing free energy not only relative to their physical en-

vironment, but also to the "social environment" composed of their peers [13]. The most parsimonious way for AIF agents to derive information from other agents would be to (i) assume that other agents are self-similar, or are "creatures like me" [58], and (ii) differentiate other-generated information by calculating how it diverges from self-generated information (akin to a process of "alterity" or self-other distinction). This ability aligns with the notion of a "folk psychological theory of society", in which humans deploy a combination of phylogenetic and ontogenetic modules to process social information [59,60].

- **Theory of Mind**—AIF agents that develop "social perceptiveness" or an ability to accurately infer beliefs and intentions of other agents will likely outperform agents with less social perceptiveness. Social perceptiveness, also commonly known in cognitive psychology as "Theory of Mind", would minimally require cognitive architecture for encoding the internal belief states of other agents as a source of self-inference (for game-theoretical simulations of this proposal, see [61,62]). As discussed above, experimental evidence suggests that social perceptiveness or Theory of Mind (measured using the "Reading the Mind in the Eyes" test; see [63]) is a significant predictor of human collective intelligence in a range of in-person and on-line collaborative tasks [4].
- **Goal Alignment**—It is possible to imagine scenarios in which the effectiveness of Theory of Mind would be limited, such as situations of high informational uncertainty (in which other agents hold multiple or unclear goals), or in environments populated by more agents than would be computationally tractable for a single AIF agent to actively theorize [64]. AIF agents capable of transitioning from merely encoding internal belief states of other AIF agents to recognizing shared goals and actively aligning goals with other AIF agents would likely enjoy considerable coordination benefits and (computational) efficiencies [16,65] that would also likely translate to collective-level performance [55,66].
- **Shared Norms**—Acquisition of capacities to engage directly with the reified signal of sharedness (a.k.a., "norms") between agents as a stand-in for (or in addition to) bottom-up discovery of mutually viable shared goals would also likely confer efficiencies to individuals and collectives [12]. Humans appear unique in their ability to leverage densely packaged socio-cultural installed affordances to cue regimes of perception and action that establish and stabilize adaptive collective behavior without the need for energetically expensive parsing of bottom-up sensory signals (a process recently described as "Thinking through Other Minds"; see [14]).

The clear resonance between generic information-theoretic patterns of basic AIF agents and empirical evidence of human social behavior is remarkable, and gives credence to the extension of seemingly human-specific notions such as "alterity", "shared goals", "alignment", "intention", and "meaning" to a wider spectrum of bio-cognitive agents [67]. In effect, the universality of FEP—a principle that can be applied to any biological system at any scale—makes it possible to strip-down the complex and emergent behavioral phenomenon of collective intelligence to basic operating mechanisms, and to clearly inspect how local-scale capabilities of individual AIF agents might enable global-scale state optimization of a composite system.

In the following section we use AIF to model the relationship between a selection of these hops in cognitive ability and collective intelligence. We construct a simple 1D search task based on [68], in which two AIF agents interact as they pursue individual and shared goals. We endow AIF agents with two key cognitive abilities—Theory of Mind and Goal Alignment—and vary these abilities systematically in four simulations that follow a 2 × 2 (Theory of Mind × Goal Alignment) progression: Model 1 (Baseline AIF, no social interaction), Model 2 (Theory of Mind without Goal Alignment), Model 3 (Goal Alignment without Theory of Mind), and Model 4 (Theory of Mind with Goal Alignment). We use a measure of free energy to operationalize performance at the local (individual) and global (collective) scales of the system [69]. While our goals in this paper are exploratory (these models and simulations are designed to be generative, not to test hypotheses), we

do generally expect that increases in sophistication of cognitive abilities at the level of individual agents will correspond with an increase in local- and global-scale performance. Indeed, illustrative results of model simulations (Section 3) show that each hop in cognitive ability improves global system performance, particularly in cases of alignment between local and global optima.

2. Materials and Methods

2.1. Paradigm and Set-Up

Our AIF model builds upon the work of McGregor and colleagues, who develop a minimal AIF agent that behaves in a discrete one-dimensional time world [68]. In this set-up, a single agent senses a chemical concentration in the environment and acts on the environment by moving one of two ways until it arrives at its desired state, the position in which it believes the chemical concentration to be highest, denoting a food source. We adapt this paradigm by modelling two AIF agents (Agent A and Agent B) that occupy the same world and interact according to parameters described below (see Figure 1). The McGregor et al. paradigm and AIF model is attractive for its computational implementability and tractability as a simple AIF agent with minimum viable complexity. It is also accessible and reproducible; whereas most existing agent-based implementations of AIF are implemented in MATLAB, using the SPM codebase (e.g., [57]), an implementation of the McGregor et al. AIF model is widely available in the open-source programming language Python, using only standard open source numerical computing libraries [70]. For a comprehensive mathematical guide to FEP and a simple agent-based model implementing perception and action under AIF, see [36].

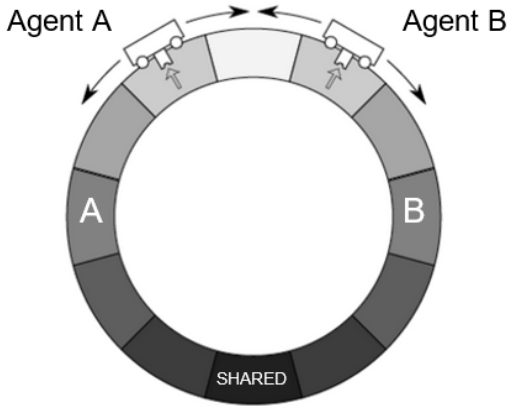


Figure 1. A minimal collective system of two AIF agents (adapted from McGregor et al.). We implement two agents (Agent A and Agent B) that have one common target position (Shared Target) and one individual target position (A's Target; B's Target). All targets are encoded with equal desirability. This figure is notional: our simulation environment contains 60 cells instead of the 12 depicted here. Note: we randomize the location of the shared target while preserving relative distances to unshared targets to ensure that the agents' behavior is not an artefact of its location in the sensory environment.

We extend the work of McGregor and colleagues to allow for interactions not only between an agent and the "physical" environment, but also between an agent and its "social" environment (i.e., its partner). Accordingly, we make minor simplifications to the McGregor et al. model that are intended to reduce the number of independent parameters and make interpretation of phenomena more straightforward (alterations to the McGregor et al. model are noted throughout).

2.2. Conceptual Outline of AIF Model

Our model consists of two agents. Descriptively, one can think of these as simple automata, each inhabiting a discrete "cell" in a one-dimensional circular environment

where there are predefined targets (food sources). As agents aren't endowed with a frame of reference, an agent's main cognitive challenge is to situate itself in the environment (i.e., to infer its own position). Both agents have the following capabilities:

- **Physical capabilities:**
  - "Chemical sensors" able to pick up a 1-bit chemical signal from the food source at each time step;
  - "Actuators" that allow agents to "move" one cell at each time step;
  - "Position and motion sensors" that allow agents to detect each other's position and motion.
- **Cognitive capabilities:**
  - Beliefs about their own current position; we construe this as a "self-actualization loop" or Sense->Understand->Act cycle: (1) sense environment; (2) optimize belief distribution relative to sensory inputs (by minimizing free energy given by an adequate generative model); and (3) act to reduce FE relative to desired beliefs, under the same generative model.
  - Desires (also described as "desired beliefs") about their own position relative to their prescribed target positions;
  - Ability to select the actions that will best "satisfy" their desires;
  - "Theory of Mind": they possess beliefs about their partner's position, knowledge of their partner's desires, and therefore, the ability to imagine the actions that their partners are expected to take. We implement this as a "partner-actualization loop" that is formally identical to the self-actualization loop above;
  - "Goal Alignment": the ability to alter their own desires to make them more compatible with their partner's.

2.3. Model Preliminaries

Throughout, we use the following shorthand:

- $q^{superscript} \triangleq softmax\left(b^{superscript}\right)$ for any superscript index, where $softmax(b)_i \triangleq \frac{e^{b_i - max(b)}}{\Sigma\, e^{b_i - max(b)}}$. This converts a belief represented as a vector in $\mathbb{R}^N$ to the equivalent probability distribution over [0..N-1]; the $max(b)$ offset is for numerical stability. We choose to convert back by $b_i = ln\, q_i - ln\, (\max(q))$, to enforce $b_i \leq 0$.
- Beliefs are also implicitly constrained to $b_i \geq -10$, for numerical stability. This means $b_i \in \mathbf{B} = [-10, 0]$.
- $\varphi \triangleq \psi^{partner}$ when necessary, to disambiguate between it and ψ^{own}.
- $(v_{+x})_i \triangleq v_{i+x}$ to denote shifting a vector.
- $\Theta_\alpha(q) \triangleq \alpha\, q + \frac{1-\alpha}{N}$ to denote "re-ranging" a probability distribution, squishing its range from [0,1] to $[\frac{1-\alpha}{N}, \alpha + \frac{1-\alpha}{N}]$.
- All arithmetic in the space of positions (ψ or Δ) and actions (a) is considered to be mod N.

2.4. State Space

These capabilities are implemented as follows. Each agent A^i is represented by a tuple $A^i = (\psi^i, s^i, b^i, a^i)$. In what follows we'll omit the indices except where there is a relevant difference between agents. These tuples form the relevant state space (see Figure 2):

- $\psi \in$ [0..N-1] is the agent's external state, its position in a circular environment with period N. Crucially, the agent doesn't have direct access to its external state, but only to limited information about the environment afforded through the sensory state below.
- $s = (s^{own} \in \{0, 1\}, \Delta \in$ [0..N-1], $a^{pp} \in \{-1, 0, 1\})$ is the agent's sensory state. s^{own} is a one-bit sensory input from the environment; Δ is the perceived difference between the agent's own position and its partner's; a^{pp} is the partner's last action.
- $b = (b^{own} \in \mathbf{B}^N, b^{*own} \in \mathbf{B}^N, b^{partner} \in \mathbf{B}^N, b^{*partner} \in \mathbf{B}^N)$ is the agent's internal or "belief" state. b^{own} and b^{*own} are, respectively, its actual and desired beliefs about

its own position; equivalently, $b^{partner}$ and $b^{*partner}$ are its actual and desired beliefs about its partner's position.

- $a = (a^{own} \in \{-1, 0, 1\}, a^{partner} \in \{-1, 0, 1\})$ is the partner's action state: a^{own} is its own action; $a^{partner}$ is the action it expects from the partner.

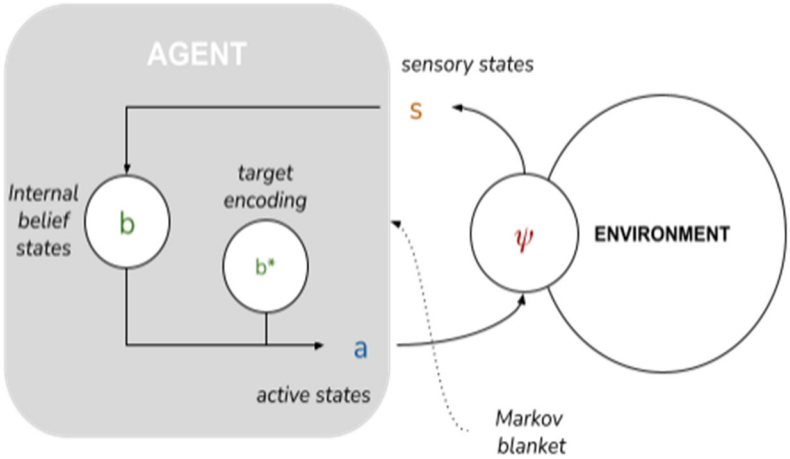


Figure 2. AIF agent based on McGregor et al. [68]. A Markov blanket defines conditional independencies between a set of internal belief states (b) and a set of environment states (ψ) with target encoding or "desires" (b*).

2.5. Agent Evolution

These states evolve according to a discrete-time free energy minimization procedure, extended from McGregor et al. (Figure 3). At each time step, each agent selects the action that will minimize the free energy relative to its target encoding (achieved by explicit computation of F for each of the 3 possible actions), and then updates its beliefs to best match the current sensory state (achieved by gradient descent on b').

2.6. Sensory Model

Let us recapitulate McGregor et al's definition of the free energy for a single-agent model:

$$F(b', b, s, a) = D_{KL}(q(\psi'|b') \parallel p(\psi', s|b, a)) \quad (1)$$

where $q(b)$ = softmax(b) is the "variational (probability) density" encoded by b, and $p(\psi', s \mid b, a)$ is the "generative (probability) density" representing the agent's mental model of the world [37]. D_{KL} is the Kullback–Leibler (KL) divergence or relative entropy between the variational and generative densities [71].

To respect the causal relationships prescribed by the Markov blanket (see Figure 2), the generative density may be decomposed as:

$$p(\psi', s|b, a) = P(\psi'|s, b, a, \psi) \bullet P(s|b, a, \psi) \bullet P(\psi|b, a) \quad (2)$$

where the three terms within the summation are arbitrary functions of their variables. In the single-agent model, where the only source of information is the environment, we follow McGregor's model, in a slightly simplified form:

1. $P(\psi' \mid s, b, a, \psi) = \delta(\psi', \psi + a)$: the agent's actions are always assumed to have the intended effect, δ being the discrete Kronecker delta.
2. $P(s \mid \psi) = k^s(1-k)^{1-s}\, e^{-\omega\, |\psi - \psi_{mid}|}$: the agent assumes the probability of s = 1 (sensoria triggered) is higher for regions near the "center" of the environment. This is identical to the real "physical" probability of chemical signals, meaning the agent's generative distribution is correct.
3. $P(\psi \mid b, a) = q(b)$, in agreement with the definition of b as encoding the belief distribution over ψ.

```
procedure simulate(ψ, b, b*)
        loop
            s^own i ← random value using P^i(s^own | ψ)
            a^i ← argmin_a  F^i(b*, b, s^i, a)
            b'^i ← optimise(b^i, s^i, a^i)
            ψ' ← ψ + a                    ◁ Unlike McGregor et al., we assume agents always act as intended
            Δ ← (ψ^A - ψ^B, ψ^B - ψ^A)    ◁ Agents' "position sensors" are assumed to be perfect
            a^pp ← (ψ'^B - ψ^B, ψ'^A - ψ^A)   ◁ Agents' "motion sensors" are assumed to be perfect
        end loop
end procedure

function optimize(b, s, a)
        b' = b                            ◁ starting from previous epoch's belief
        for i ∈ {1 · · · k} do
            b' ← b' − η ·∂/∂b' F(b', b, s, a)        ◁ gradient descent on b'
        end for
        return b'
end function
```

Figure 3. Pseudo code for agent evolution (adapted from [68]). Note that the loop is run for both agents in lockstep, but each agent selects actions and optimizes beliefs individually.

From list item 1 directly above, this generative density can also be read as a simple Bayesian updating plus a change of indexes to reflect the effects of the action: $p(\psi',\, s|b,\, a) = P(s|\psi' - a)\; P(\psi' - a|b)$ or even more simply, $p^{posterior}_{\psi'} = p^{s}_{\psi'-a}\; p^{prior}_{\psi'-a}$.

In our model, both agents implement their own copies of the generative density above (we leave it to the reader to add "$\square^{own}$" indices where appropriate). The parameter k, denoting the maximum sensory probability, is assumed agent-specific; we naturally identify it with an agent's "perceptiveness". ω and ψ_0, on the other hand, are environmental parameters.

2.7. Partner Model

In addition to the sensory model, we will define a new generative density implementing the agent's inference of its partner's behavior, or "Theory of Mind" (ToM; see Figure 6b). An agent with a sensory and partner model will adopt the following form:

$$\mathrm{p}(\phi', \Delta, a^{pp}|b, a) = \mathrm{P}(\phi'|a^{partner}, \phi)\bullet\mathrm{P}(\Delta|\mathrm{b}, \phi)\bullet\mathrm{P}(a^{pp}|b, a^{partner}, \phi)\bullet\mathrm{P}(\phi|\mathrm{b}) \qquad (3)$$

The first three terms on the right-hand side correspond to mechanistic models of the evolution of the variables ϕ', Δ, a^{pp}, whereas the last one, $P(\phi|b) = q^{partner}_{\phi}$, defines the "prior" and is analogous to $q(b)$ in the sensory model. To fully specify this density, we define these models as follows:

1. $P(\phi'|a^{partner},\, \phi) = \delta(\phi',\, \phi + a^{partner})$ describes the expected results of the partner's observed action upon its inferred position. The Kronecker delta implies that the partner's actions are always effective, matching item #1 from Section 2.6.
2. $P(\Delta|b,\, \phi) = P(\psi = \phi + \Delta|b, \phi) = q^{own}_{\phi+\Delta}$: the agent (correctly) believes that Δ is a deterministic function of the two positions, and therefore the probability of observing a given Δ, given the partner's position ϕ, is equal to the probability the agent ascribes to itself being in the corresponding position $\psi = \phi + \Delta$.

3. $P(a^{pp}|b,\ a^{partner},\phi) = P(a^{partner}|\phi - a^{pp},\ b^{*partner})$: the agent determines its belief in the partner's previous action by "backtracking" to its previous state $\phi - a^{pp}$, and leveraging the following model of the partner's next action:

$$\begin{aligned} P(a^{partner} = 0|\phi, b^{*partner}) &= \xi \frac{1}{\max(q^{*partner})} q_{\phi}^{*partner} \\ P(a^{partner} = \pm|\ \phi, b^{*partner}) &= \left\{1 - P(a^{partner} = 0|\phi, b^{*partner})\right\} \frac{1}{p_{\phi-1}^{*partner} + p_{\phi+1}^{*partner}} q_{\phi - a^{partner}}^{*partner} \end{aligned} \tag{4}$$

This equation seems complex but its output and mechanical interpretation are quite simple (see Figure 4). To justify it, note that the agent must produce probabilities of the partner's actions without knowing their *actual* internal states at that time, but only their targets $q^{*partner}$. To do so, the agent assumes that the partner will act mechanistically according to those desires, i.e., the higher a partner's desire for its current location, the more likely it is to stay put. To eliminate spurious dependence on absolute values of $q^{*partner}$, we set $P(a^{partner} = 0)$ to be proportional to $q^{*partner}/\max(q^{*partner})$. The constant of proportionality ξ corresponds to the maximum probability of the partner standing still, when $q^{*partner}$ achieves its global maxima. This leaves the remaining probability mass to be allocated across the other actions (± 1), which we do by assuming the probability of moving in a given direction is proportional to the desires in the adjacent locations. For the purpose of this study, ξ is held constant at 0.9.

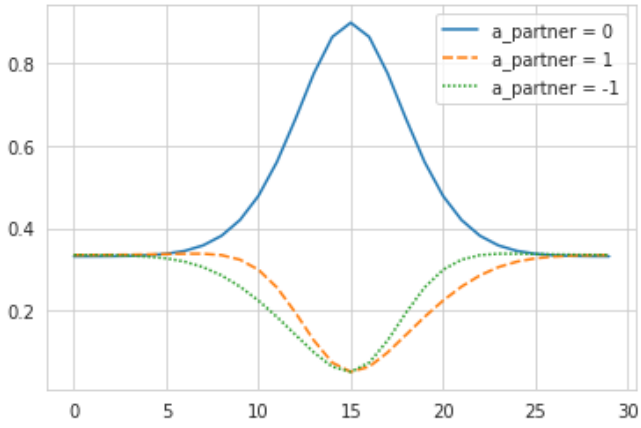


Figure 4. Illustrative plot of $P(a^{partner}|\phi, b^{*partner})$ for each possible value of $a^{partner}$ and ϕ, when $q^{*partner}$ follows a normal distribution centered on $\phi = 15$. At the valleys where $q^{*partner}$ is lowest and its gradient is small, the partner doesn't quite have strong incentives to go in any particular direction, and so is assigned roughly equal probabilities for the three actions. At the slopes, the action corresponding to the upward slope is more strongly expected. At peak $q^{*partner}$, $P(a^{partner} = 0) = \xi$ and the probabilities of the two other actions are equal.

The combination of these three models results in a generative density has the same form as the original generative density from the baseline sensory model, $p_{\phi'}^{posterior} = p_{\phi'-a^{partner}}^{\Delta, a^{pp}} p_{\phi'-a^{partner}}^{prior}$. This is consistent with our modeling decision to make the "other-evidencing loop" functionally identical to the "self-actualization loop", as discussed above (Section 2.2).

As before, each agent implements its own copy of the partner model. ξ is assumed equal for both agents; they have the same capability to interpret the partner's actions.

2.8. Agent-Level Free Energy

We are finally ready to define the free energy for our individual-level model. For each agent:

$$F = D_{KL}\left(q'^{\,own} \parallel p^{own}\ \Theta_{\alpha}\left(p_{+\Delta'}^{partner}\right)\right) + D_{KL}\left(q'^{\,partner} \parallel p^{partner}\ \Theta_{\alpha^2}\left(p_{-\Delta'}^{own}\right)\right) \tag{5}$$

where:

1. $p^{own}_{\psi'} = P(s^{own} \mid \psi' - a^{own})\, q^{own}_{\psi' - a^{own}}$ is the sensory model (outlined above in Section 2.6).
2. $p^{partner}_{\phi'} = q^{own}_{\phi' - a^{partner} + \Delta}\, P(a^{partner} \mid \phi' - a^{partner} - a^{pp},\ b^{\star partner})\, q^{partner}_{\phi' - a^{partner}}$ is the partner model (outlined above in Section 2.7).
3. The "reranging" function, Θ_α, serves to moderate the influence of the partner model on the agent's own beliefs, and vice-versa. α is an agent-specific parameter, which, as we will see in Section 2.9, is identified with each agent's degree of "alterity".
4. The right-hand side of each KL divergence (i.e., the products of generative densities) is implicitly constrained to $\left[e^{-10}, 1\right]$, to ensure the resulting beliefs remain within their range **B**. This is interpreted as preventing overconfidence and is implemented as a simple maximum.

We interpret Equation (4) as follows: The agent's sensory and partner models jointly constrain its beliefs both about its own position and its partner's position. Thus, at each step, the agent: (a) refines its beliefs about both positions, in order to best fit the evidence provided by all of its inputs (i.e., its "chemical" sensor for the physical environment and "position and motion" sensor for its partner); and (b) selects the "best" next pair of actions (for self and partner), i.e., that which minimizes the "difference" (the KL divergence) between its present beliefs and the desired beliefs (For reasons of numerical stability, we follow McGregor et al. in implementing (b) before (a): The agent chooses the next actions based on current beliefs, then updates beliefs for the next time-step, based on the expected effects of those actions [68] (pp. 6–7)).

2.9. Theory of Mind

In this section we motivate the parameterization of an agent's Theory of Mind ability with α, or simply, its degree of *alterity*.

Note that when considered as a discrete-time dynamical system evolution, the process of refining beliefs about own and partner positions in the environment (step (a) in Section 2.8 above) potentially involves multiple recursive dependencies: the updated variational densities $q'^{\,own}$ and $q'^{\,partner}$ both depend on the previous q^{own} (via both p^{own} and $p^{partner}$), as well as on the previous $q^{partner}$ (via $p^{partner}$). This is by design: the dependencies ensure that $q'^{\,own}$ and $q'^{\,partner}$ are consistent with each other, as well as with their counterparts across time steps. However, too much of a good thing can be a problem. If left unconstrained, $q'^{\,own}$ and $q'^{\,partner}$ can easily evolve towards spurious fixed points (Kronecker deltas), which can be interpreted as overfitting on prematurely established priors (In this case, it could be possible to observe scenarios such as *"the blind leading the blind"* in which a weak agent fixates on the movement trajectory of a strong agent who is overconfident about its final destination.). On the other hand, if $q'^{\,own}$ were to depend only on q^{own}, it would eliminate the spurious fixed points: without the crossed dependence, the first term of the partner model (Section 2.7) only has fixed points at $(q'^{\,own} = \delta(\psi', argmax(q^{\star\,own}))$, $a^{own} = 0)$, meaning that the agent has achieved a local desire optimum. Effectively, this "shuts down" the agent's ability to use the partner's information to shape its own beliefs, or its theory of mind, making it equivalent to MacGregor's original model.

Thus, there would appear to be no universal "best" value for an agent's Theory of Mind; an appropriate level of Theory of Mind would depend on a trade-off between the risk of overfitting and that of discarding valid evidence from the partner. The appropriate level of Theory of Mind would also depend on the agent's other capabilities (in this case, its perceptiveness, k).

This motivates the operationalization of α as a parameter for the intensity to which Theory of Mind shapes the agent's beliefs. α can be understood simply as an agent's degree of *alterity*, or propensity to see the "other" as an agent like itself. In simulations with values of α close to 0, we expect the partner's behavior to be dominated by its own "chemical" sensory input. Increasing α, we expect to see an agent's behavior being more heavily

influenced by inputs from its partner, driving q^{own} to become sharper as soon as $q^{partner}$ does so. Past a certain threshold, this could spill over into premature overfitting.

Finally, note the α^2 in the second term of agent-level free energy (Equation (4)). This represents the notion that the agent is using "second-order theory of mind" or thinking about what its partner might be thinking about it (First-order ToM involves thinking about what some-one else is thinking or feeling; second-order ToM involves thinking about what someone is thinking or feeling about what someone else is thinking or feeling [72]). Here, p^{own} comes in as "my model of my partner's model of my behavior". It seems appropriate for the agent to believe the partner to possess the same level of alterity as itself; we then represent this as applying the rearranging function (the "squishing" of the probability distribution) twice, $\Theta_\alpha \bullet \Theta_\alpha = \Theta_{\alpha^2}$.

2.10. Goal Alignment

In this section we motivate the parameterization of the degree of goal alignment between agents.

Recall that $b^{\star\ own}$ is an arbitrary (exogenous) real vector; the implied desire distribution can have multiple maxima, leading to a generally challenging optimization task for the agent. Theory of Mind can help, but it can also make matters worse: if $b^{\star\ partner}$ also has multiple peaks, the partner's behavior can easily become *ambiguous*, i.e., it could appear coherent with multiple distinct positions. This ambiguity can easily lead the agent astray.

This problem is reduced if the agents can *align goals* with each other, that is, avoid pursuing targets that are not shared between them. We implement this as:

$$b^{\star\ own} \leftarrow b^{\star\ shared} + (1-\gamma) b^{\star\ own}_{private} \tag{6}$$

$$b^{\star\ partner} \leftarrow b^{\star\ shared} + (1-\gamma) b^{\star\ partner}_{private} \tag{7}$$

where γ is a parameter representing the degree of alignment between this specific agent pair, and we assume each agent has knowledge of what goals are shared vs private to itself or its partner. That is, with $\gamma = 0$, the agent is equally interested in its private goals and in the shared ones (and assumes the same for the partner); with $\gamma = 0$, the agent is solely interested in the shared goals (and assumes the same for the partner).

This operation may seem quite artificial, especially as it implies a "leap of faith" on the part of the agent to effectively change its expectations about the partner's behavior (Equation (6)). However, if we accept this assumption, we see that the task is made easier: in the general case, alignment reduces the agent-specific goal ambiguity, leading to better ability to focus and less positional ambiguity coming from the partner. Of course, one can construct examples where alignment does not help or even hurts; for instance, if both agents share all of their peaks, alignment not only will not help reduce ambiguity, but it can make the peaks sharper and hard to find. And as we will see, in the context of the system-level model, alignment becomes a natural capability.

In the present paper, for simplicity, we assume agents' shared goals are assigned exogenously. In light of the system-level model (Section 2.11), however, it is easy to see that such shared goals have a natural connection with the global optimum states. In this context, one can expect shared goals to emerge endogenously from the agents' interaction with their social environment over the "long run". This will be explored in future work.

2.11. System-Level Free Energy

Up until now, we have restricted ourselves to discussing our model at the level of individual agents and their local-scale interactions. We now take a higher vantage point and consider the implications of these local-scale interactions for global-scale system performance. We posit an ensemble of M identical copies of the two-agent subsystem above (i.e., $2M$), each in its own independent environment, also assumed to be identical except for the position of the food source (see Figure 5).

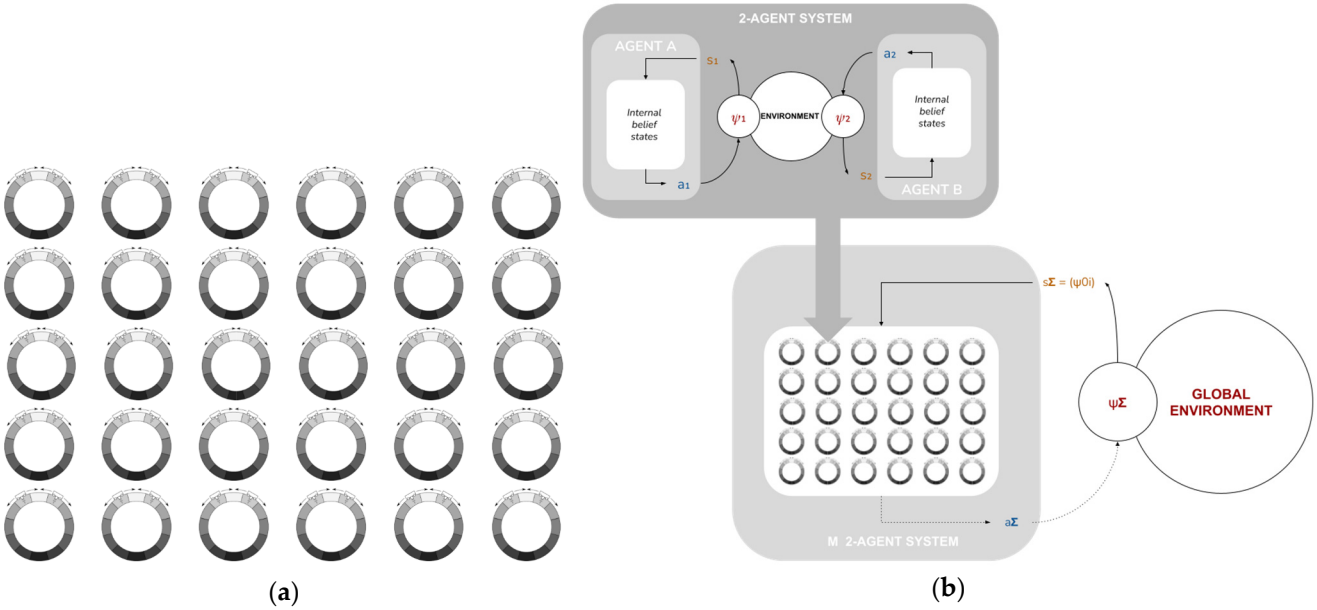


Figure 5. (**a**) M identical copies of the two-agent subsystem. (**b**) The M two-agent systems as internal states of a larger system, interacting with a global environment through the food sources (reinterpreted as sensory states) and some active mechanism (the dotted arrow lines for a**Σ** denote that this active mechanism is not defined in this paper).

From this vantage point, each of the $2M$ agents is now a "point particle", described only by its position ψ^i. The tuple $b^\Sigma = \left(\psi^i\right)_{i\in[1..2M]}$ is then the set of internal states of the system as a whole.

We will now assume that this set of internal states interacts with a global environment $\psi^\Sigma \in [0..N-1]$. We reinterpret the "food sources" as sensory states: $s^\Sigma = \left(\psi^i\right)_{i\in[1..2M]}$, where each ψ^i_0 is assumed to correlate with ψ^Σ through some sensory mechanism. We further assume the system is capable to act back on the environment through some active mechanism a^Σ. This provides us with a complete system-level Markov blanket (Figure 5b), for which we can define a system-level free energy as

$$F^\Sigma = D_{KL}(q^{empirical}(\psi^{\Sigma'}|b^{\Sigma'}) \parallel p^\Sigma(\psi^{\Sigma'}, s^\Sigma|a^\Sigma, b^\Sigma)) \tag{8}$$

where $q^{empirical}(\psi^\Sigma|b^\Sigma) = \frac{1}{2M}\#\left\{\psi^i|\psi^i = \psi^\Sigma\right\}$, the system's "variational density", is simply the empirical distribution of the various agents' positions.

In this paper, we will not cover the "active" part of active inference at the global level—namely, the system action a^Σ remains undefined. We will instead consider a *single system-level inference step*, corresponding to fixed values of ψ^Σ, s^Σ. As we can see from the formulation above, this corresponds to optimizing ψ^i given ψ^i_0—that is, to the aggregate behavior of the $2M$ agents' over an *entire run* of the model at the individual level.

This in turn motivates defining the system's generative density as $p^\Sigma(\psi^{\Sigma'}, s^\Sigma|a^\Sigma, b^\Sigma) \propto exp\left\{-k\Sigma\left(\psi^i - \psi^i_0\right)^2\right\}$: given a set of internal states (agent positions), the system "expects" it to have been produced by the agents moving towards the corresponding sensory states (food source). Thus, to the extent that the agents perform their local active inference tasks well, the system performs approximate Bayesian inference over this generative density, and we can evaluate the degree to which this inference is effective, by evaluating whether, and how quickly, F^Σ is minimized. We return to the topic of system-level (active) inference in the discussion.

2.12. Simulations

We have thus defined this system at two altitudes, enabling us to perform simulations at the agent level and analyze their implied performance at the system level (as measured by system-level free energy). We can now use this framework to analyze the extent to which the two novel agent-level cognitive capabilities we introduced ("Theory of Mind" and "Goal Alignment") increase the system's ability to perform approximate inference at local and global scales. To explore the effects of agent-level cognitive capabilities on collective performance, we create four experimental conditions according to a 2 × 2 (Theory of Mind × Goal Alignment) matrix: Model 1 (Baseline), Model 2 (Theory of Mind), Model 3 (Goal Alignment), and Model 4 (Theory of Mind and Goal Alignment; see Table 1).

Table 1. 2 × 2 (Theory of Mind × Goal Alignment) permutations of our model.

	-Theory of Mind	**+Theory of Mind**
-Goal Alignment	Model 1 (Baseline)	Model 2 (Theory of Mind, No Goal Alignment)
+Goal Alignment	Model 3 (Goal Alignment, No ToM)	Model 4 (Theory of Mind × Goal Alignment)

Throughout, we use the same two agents, Agent A and Agent B. To establish meaningful variation in agent performance at the individual-scale, we parameterize an agent's perceptiveness to the physical environment (i.e., to the reliability of the information derived from its "chemical sensors"), by assigning one agent with "strong" perceptiveness (Agent A—Strong;) and the other agent with "weak" perceptiveness (Agent B—Weak).

We assign each agent with two targets, one shared (Shared Target) and one unshared (individual target or Target A and Target B). Accordingly, we assume each agent's desire distributions have both a shared peak (corresponding to a Shared Target) and an unshared peak (corresponding to Target A or Target B). Throughout, we measure both the collective performance (system-level free energy), as well as individual performance (distance from their closest target). In addition, we also capture their end-state desire distribution.

We implement simulations in Python (V3.7) using Google Colab (V1.0.0). As noted above, our implementation draws upon and extends an existing AIF model implementation developed in Python (V2.7) by van Shaik [70]. To ensure that the agent behavior is not an artefact of their specific location in the environment, we run 180 runs for each simulation for each experimental condition by randomizing their starting locations throughout the environment. The environment size was held constant at 60 cells. To ensure that the agent behavior is not an artefact of initial conditions, we perform 180 runs for each simulation for each experimental condition by uniformly distributing their starting locations throughout the environment (three times per location), while preserving the distance between starting locations and target. This uniform distribution of initial conditions across the environment also corresponds to the "worst-case scenario" in terms of system-level specification of sensory inputs for a two-agent system, discussed in Section 2.11.

2.13. Model Parameters

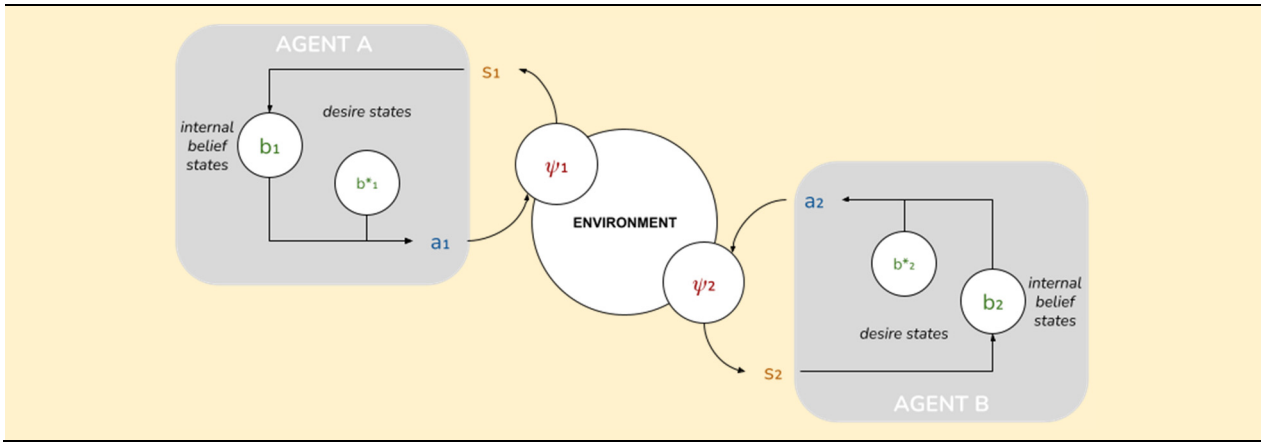
Our four models were created by setting physical perceptiveness for the strong and weak agent and varying their ability to exhibit social perceptiveness and align goals. The parameter settings are summarized at the individual agent level as follows (see Figure 6 and Table 2):

- Model 1 contains a self-actualization loop driven by physical perceptiveness. Physical perceptiveness (individual skill parameter; range [0.01, 0.99]) is varied such that Agent A is endowed with strong perceptiveness (0.99) and Agent B is endowed with weak perceptiveness (0.05).
- Model 2 is made up of a self-actualization loop and a partner-actualization loop (instantiating ToM). The other-actualization loop is implemented by setting the value

of alterity (ToM or social perceptiveness parameter; range [0.01, 0.99]) as 0.20 for the weak agent and 0 for the strong agent. This parameterization helps the weak agent use social information to navigate the physical environment. These two loops implement a single (non-separable) free energy functional: The weak agent's inferences from their stronger partner's behavior serve to refine its beliefs about its position in the environment.

- Model 3 entails a self-actualization loop (but no partner-actualization loop) as well as enforces the pursuit of a common goal (set alignment = 1) by fully suppressing their unshared goals (alignment parameter; range [0,1]). In this simplified implementation, we assume that goal alignment is a relational/dyadic property such that both partners exhibit the same level of alignment towards each other. This is akin to partners fully exploring each other's targets and agreeing to pursue their common goal. Setting alignment lower than 1 will increase the relative weighting of unshared goals and cause them to compete with their shared goals.
- Model 4 includes both cognitive features: self- and partner-actualization loops for the weak agent (instantiating ToM; alterity = 0.2) and complete goal alignment between agents.

(**a**)



(**b**)

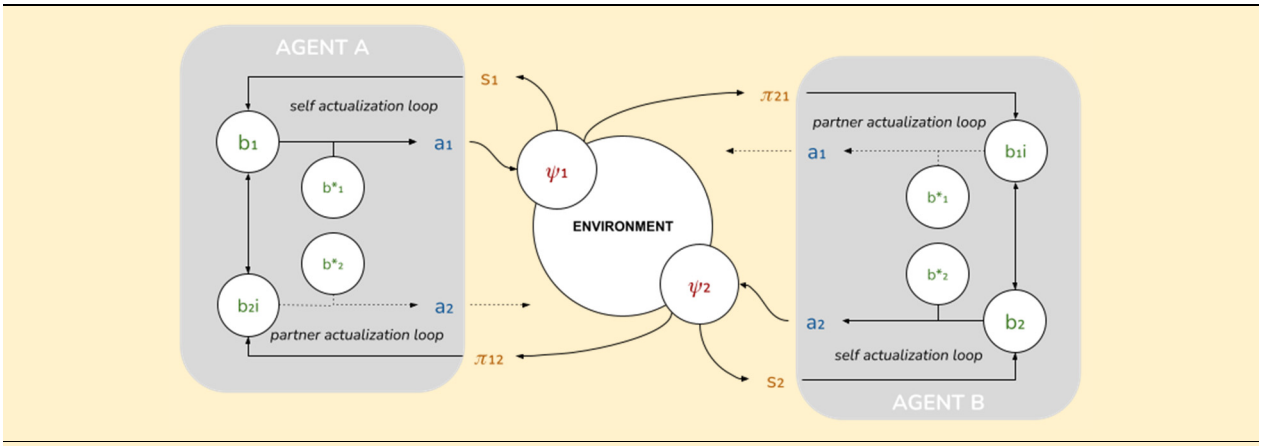


Figure 6. *Cont.*

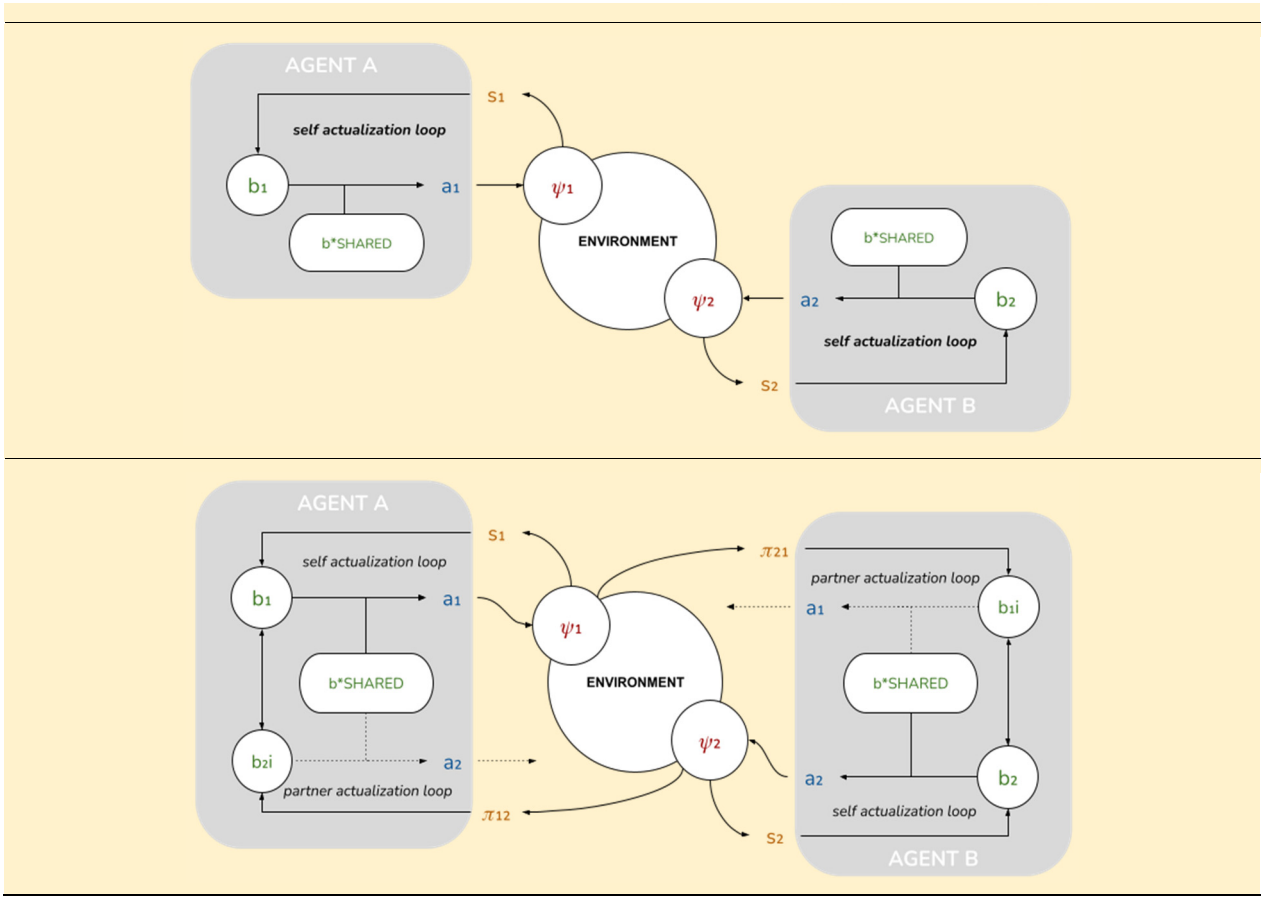


Figure 6. Models. (**a**) Model 1—Baseline with no direct interaction between agents; (**b**) Model 2—introduces "Theory of Mind" or a partner actualization loop; (**c**) Model 3—introduces Goal Alignment (b*SHARED); (**d**) Model 4—complete model with Theory of Mind with Goal Alignment.

Table 2. Parameterization of agent abilities Models 1–4.

	Model 1 Baseline		**Model 2 Theory of Mind**		**Model 3 Goal Alignment**		**Model 4 ToM x Goal Alignment**	
Parameter	**Agent A**	**Agent B**	**Agent A**	**Agent B**	**Agent A**	**Agent B**	**Agent A**	**Agent B**
Physical perceptiveness (0.01, 0.99)	**0.99**	**0.05**	0.99	0.05	0.99	0.05	0.99	0.05
Alterity, α (0.01, 0.99)	0.00	0.00	**0.00**	**0.20** *	0.00	0.00	**0.00**	**0.20** *
Goal Alignment, γ (0, 1)	0	0	0	0	**1**	**1**	**1**	**1**

* Alternative results for simulations with alterity set at α = 0.5 exhibit a similar pattern of results for Model 2 and Model 4.

3. Results

3.1. Illustration of Agent-Level Behavior

In Figure 7, we show typical results from a single run of a single two-agent subsystem (Model 4: ToM with Goal Alignment) to illustrate qualitatively how the two cognitive capabilities introduced enable agent-level performance. In this example, Goal Alignment enters the picture at the outset; although each agent has two targets, they both only ever pursue their shared target.

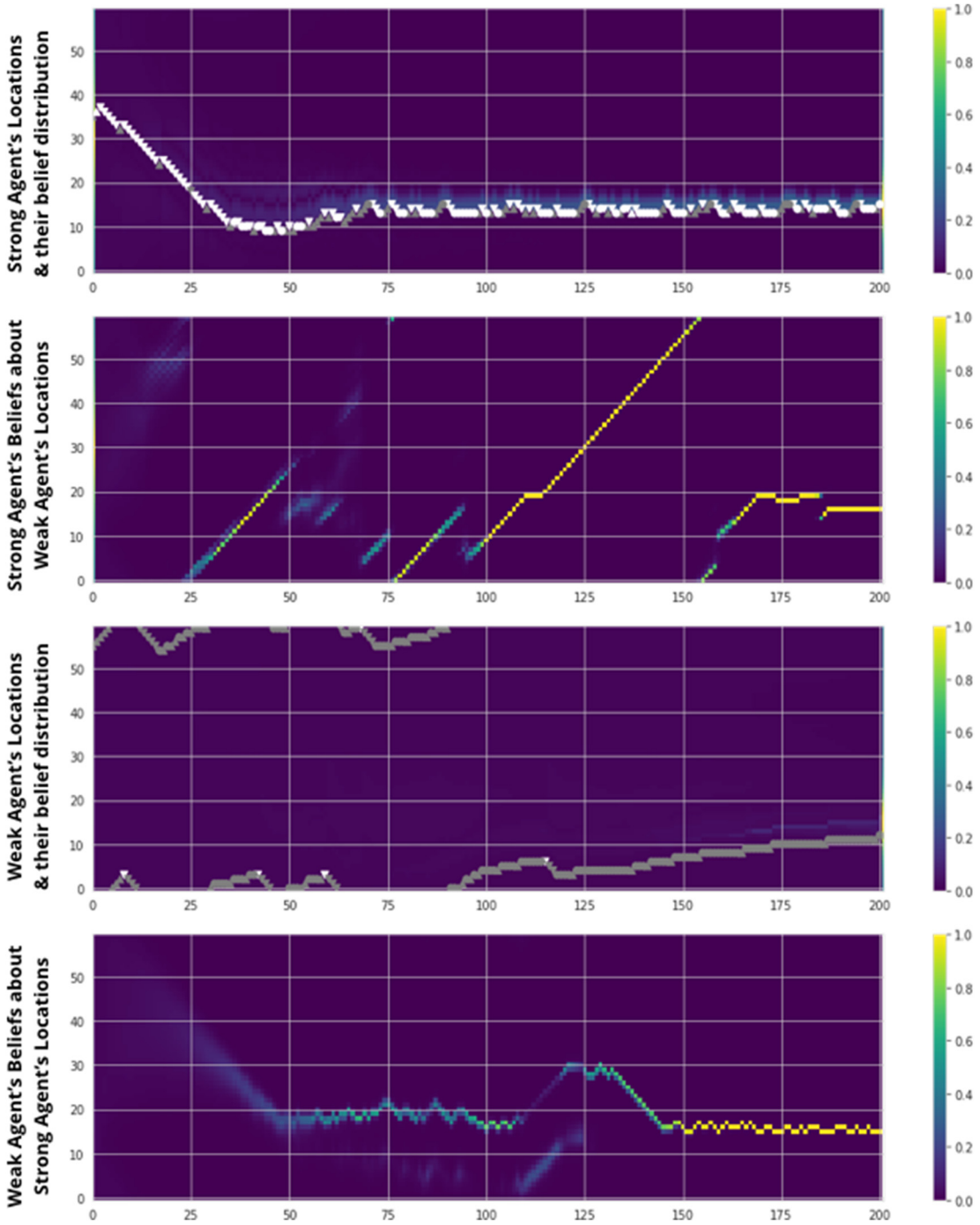


Figure 7. Results from a single run of Model 4 over 200 epochs. Agents' Shared Target position is set at location 15. Actual agent positions are illustrated as single dots for each epoch on the top graph, colored white when s = 1 and gray when s = 0. The background of the top graphs plots the agents' belief distribution of their own position, from dark blue (0) to bright yellow (1). The bottom graphs plot the agents' belief distribution of their partner's position, on the same scale.

The evolution of the two agents' behavior and beliefs over this run demonstrates the key features of interplay between sensory and partner inputs, and how ToM moderates the influence of partner inputs on an agent's behavior. Using its high perceptiveness, A identifies its own position around epoch 25–50, and quickly thereafter, directs itself towards the food position and remains stable there (top left). Meanwhile, for most of the run, B has no strong sense of its own position, and therefore its movement is highly random and undirected; at around epoch 150, it finally starts exhibiting a sharper (light blue) belief

and converging to the target (top right). This is the same moment when B is finally able to disambiguate A's behavior (from green to yellow), which, via ToM, enables B's belief to become sharper (bottom right). Meanwhile, A can't make sense of B's random actions: the partner distribution it infers is unstable. But because A has ToM = 0, it doesn't take any of these misleading cues into account when deciding its own beliefs (bottom left).

3.2. Simulation Results

Model 1 lends face validity to the two-agent simulation setup. Figure 8 (Row 1, Model 1) demonstrates that, on average, the strong agent (endowed with high physical perceptiveness) converges to an end-state belief faster more accurately (closer to one of their individual targets) than the weak agent with severely diminished physical perceptiveness. This difference in individual performance can be attributed to the stark difference in agents' ability to form strong beliefs about the location of their target (see Figure 8: Row 2, Model 1). Agents show no clear preference for either shared or unshared targets (Figure 8: Row 3, Model 1).

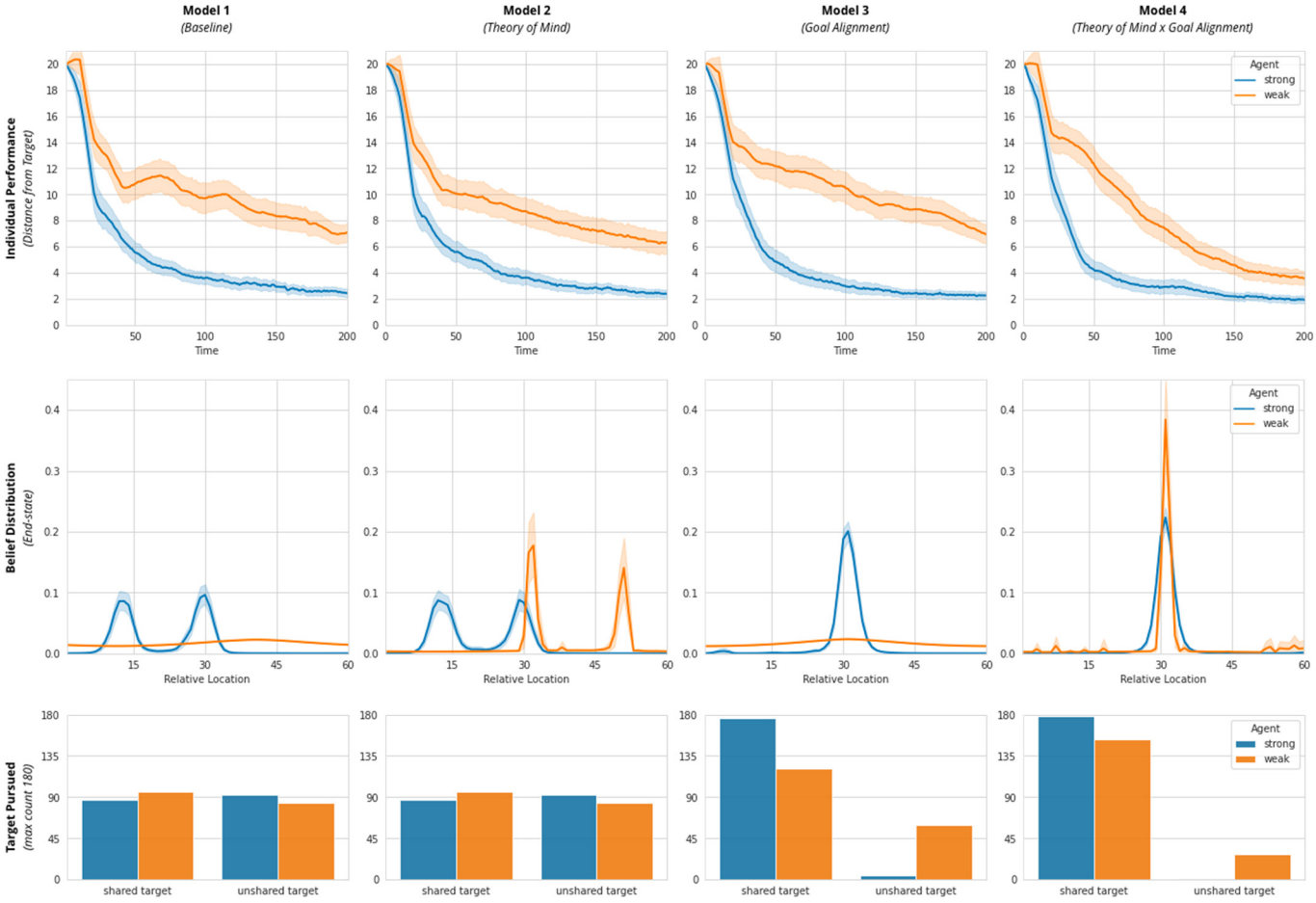


Figure 8. Simulation results of Agent A (strong; blue) and Agent B (weak; orange) in all four models. Row 1: Individual performance as time taken to reach a target position. Row 2: End state belief distribution of target location (Shared Target = 30; A's Target = 15; B's Target = 45). Row 3: Distribution of targets pursued in 180 runs.

In model 2, the weak agent possesses 'Theory of Mind'. This allows it to infer information about their own location in the environment by observing their partner's actions. This is evidenced by the emergence of two-sharp peaks in the weak agent's end-state belief distribution (Figure 8: Row 2, Model 2). Consequently, we see an improvement in the weak agent's individual performance (the agent converges faster on an end-state belief faster than in Model 1). Collective performance (Figure 9: System's free energy) does not

appear to improve between Model 1 and Model 2. This may be because agents solely focus on achieving their individual goals (and do not understand any distinction between individual and system level goals). This is evidenced by the fact that of the 180 simulation runs each of Model 1 and Model 2, both agents end up pursuing their shared and unshared targets with roughly equal probability (Figure 8: Row 3, Model 1 and 2).

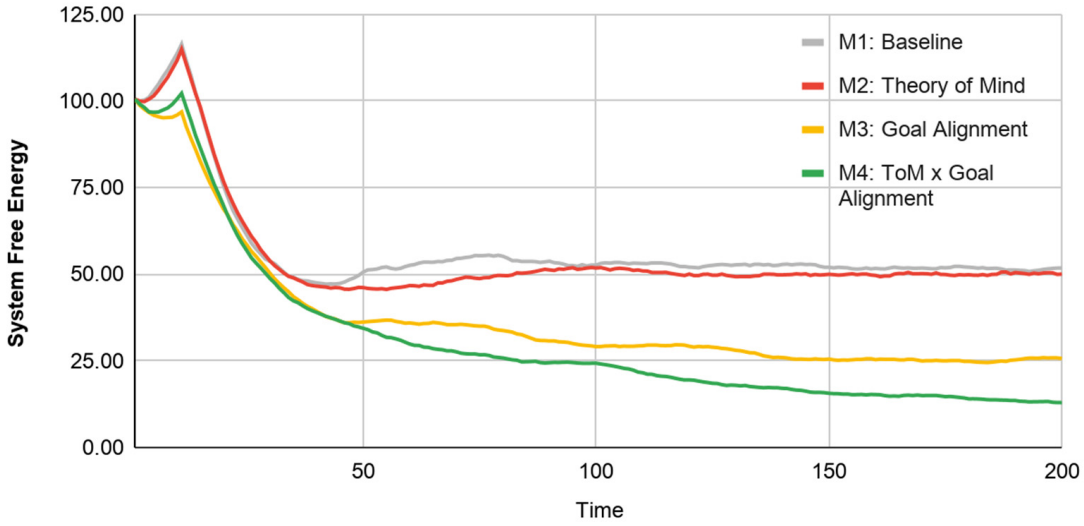


Figure 9. Actual system-level free energy F^{Σ} under each of the four models. Lower free energy denotes higher system performance. To the extent that the system is able to reduce its free energy over time (i.e., mimicking gradient descent on F^{Σ}), it can be interpreted as performing a single inference step of the active inference loop.

In Model 3, when both agents possess an ability for Goal Alignment, but the weak agent does not have the benefit of Theory of Mind, we see that both agents are biased towards pursuing the shared system goal (Figure 8: Row 3, Model 3). Accordingly, at the system level we see naturally higher collective performance—Model 3 clearly has lower system-level free energy compared to both Model 1 and Model 2 (see Figure 9). At the individual-level, however, the weak agent performs worse on average than it did in Model 2 and converges more slowly towards its goals (Figure 8: Row 1, Model 3). It appears that Goal Alignment helps improve system performance by reducing the ambiguity of multiple possible targets, but Goal Alignment does not help the weak agent compensate for low physical perceptiveness.

Finally, as expected, in Model 4, which combines Theory of Mind and Goal Alignment, we see a clear improvement in both individual and collective performance (Figure 8: Row 1, Model 4 and Figure 9: Model 4, respectively). The combination of Theory of Mind (for the weak agent) and Goal Alignment (for both agents) appears to enable the weak agent to overcome its poor physical perceptiveness and converge on a single unambiguous end-state belief. This achievement is illustrated by the sharp and overlapping single-peaked end-state belief structure achieved by both agents in model 4 (Figure 8: Row 2, Model 4) (We thank the anonymous reviewer for pushing us to consider the reasons why the end-state belief distribution for the weak agent is more sharply peaked. We didn't have any a priori expectation for this particular pattern of result. Our best guess is that this is an artefact of the weak agent iteratively engaging in 'Theory of Mind' based-estimation of its belief-distribution from the strong agent actions. From the perspective of the weak agent, the strong agent quickly converges near the goal state and spends more time in the vicinity of the peak. Thus, the weak agent is very likely to accrue higher levels of confidence within this relatively narrow vicinity. On the other hand, the stronger agent has no ToM and is only influenced by its direct perception of the environment.). This model suggests that collective performance is highest when individual agents' individual states align with the global system state.

4. Discussion

A formal understanding of collective intelligence in complex adaptive systems requires a formal description, within a single multiscale framework, of how the behavior of a composite system and its subsystem components co-inform each other to produce behavior that cannot be explained at any single scale of analysis. In this paper we make a contribution toward this type of formal grasp of collective intelligence, by using AIF to posit a computational model that connects individual-level constraints and capabilities of autonomous agents to collective-level behavior. Specifically, we provide an explicit, fully specified two-scale system where free energy minimization occurs at both scales, and where the aggregate behavior of agents at the faster/smaller scale can be rigorously identified with the belief-optimization (a.k.a. "inference") step at the slower/bigger scale. We introduce social cognitive capabilities at the agent level (Theory of Mind and Goal Alignment), which we implement directly through AIF. Further, illustrative results of this novel approach suggest that such capabilities of individual agents are directly associated with improvements in the system's ability to perform approximate Bayesian inference or minimize variational free energy. Significantly, improvements in global-scale inference are greatest when local-scale performance optima of individuals align with the system's global expected state (e.g., Model 4). Crucially, all of this occurs "bottom-up", in the sense that our model does not provide exogenous constraints or incentives for agents to behave in any specific way; the system-level inference emerges as a product of self-organizing AIF agents endowed with simple social cognitive mechanisms. The operation of these mechanisms improves agent-level outcomes by enhancing agents' ability to minimize free energy in an environment populated by other agents like it.

Of course, our account does not preclude or dismiss the operation of "top-down" dynamics, or the use of exogenous incentives or constraints to engineer specific types of individual and collective behavior. Rather, our approach provides a principled and mechanistic account of bio-cognitive systems in which "bottom-up" and "top-down" mechanisms may meaningfully interplay to inform accounts of behavior such as collective intelligence [4]. Our results suggest that models such as these may help establish a mechanistic understanding of how collective intelligence evolves and operates in real-life systems, and provides a plausible lower bound for the kind of agent-level cognitive capabilities that are required to successfully implement collective intelligence in such systems.

4.1. We Demonstrate AIF as a Viable Mathematical Framework for Modelling Collective Intelligence as a Multiscale Phenomenon

This work demonstrates the viability of AIF as a mathematical language that can integrate across scales of a composite bio-cognitive system to predict behavior. Existing multiscale formulations of AIF [39,40], while more immediately useful for understanding the behavior of docile subsystem components like cells in a multicellular organism or neurons in the brain, do not yet offer clear predictions about the behavior of collectives composed of highly autonomous AIF agents that engage in reciprocal self-evidencing with each other as well as with the physical (non-social) environment [43]. What's more, existing toy simulations of multiscale AIF engineer collective behavior as a predestination—either as a prior in an agent's generative model [9], or by default of an environment that consists solely of other agents [7,8]. We build upon these accounts by using AIF to first posit the minimal information-theoretical patterns (or "adaptive priors"; see [42]) that would likely emerge at the level of the individual agent to allow that agent to persist and flourish in an environment populated by other AIF agents [58]. We then examine the relationship between these local-scale patterns and collective behavior as a process of Bayesian inference across multiple scales. Our models show that collective intelligence can emerge endogenously in a simple goal-directed task from interaction between agents endowed with suitably sophisticated cognitive abilities (and without the need for exogenous manipulation or incentivization).

Key to our proposal is the suggestion that collective intelligence can be understood as a dynamical process of (active) inference at the global-scale of a composite system. We operationalize self-organization of the collective as a process of free energy minimization or approximate Bayesian inference based on sensory (but not active) states (for a previous attempt to operationalize collective behavior as both active and sensory inference, see [69]). In a series of four models, we demonstrate the responsiveness of this system-level measure to learning effects over time; the progression of each Model exhibits a pattern akin to a gradient descent on free energy, evoking the notion that a system that performs (active) Bayesian inference. Further, stepwise increases in cognitive sophistication at the individual level show a clear reduction in free energy, particularly between Model 1 (Baseline) and Model 4 (Theory of Mind x Goal Alignment). These illustrative results suggest a formal, causal link between behavioral processes across multiple scales of a complex adaptive system.

Going further, we can imagine an extension of this model where the collective system interacts with a non-trivial environment, but at a slower time scale, such that a complete simulation run of all 2M agents corresponds to a single belief optimization step for the whole system, after which it acts on the environment and receives sensory information from it (manifested, for example, as changes in the agents' food sources). In this extended model (see Figure 10), and if the agent-specific parameters (alterity/Theory of Mind (α), and Goal Alignment (γ)) could be made endogenous (either via selective mechanisms via some other learning mechanisms; see [48,73]) we would expect to see the system finding (non-zero) values of these parameters that optimize its free energy minimization. For example, it is likely that a system would select for higher values of γ (Goal Alignment) when both agents' end-state beliefs and actual target locations mutually cohere, or higher values of α for agents with weaker perceptiveness. Interestingly, this would show that degrees of Theory of Mind and Goal Alignment are capabilities that would be selected for or boosted at these longer time scales, providing empirical support for the heuristic arguments made for their existence in our model and in human collective intelligence research more generally [4].

4.2. AIF Sheds Light on Dynamical Operation of Mechanisms That Underwrite Collective Intelligence

In this way, AIF offers a paradigm through which to move beyond the methodological constraints associated with experimental analyses of the relationship between local interactions and collective behavior [21]. Even our very rudimentary 2-Agent AIF model proposed here offers insight into the dynamic operation and function of individual cognitive mechanisms for individual and collective level behavior. In distinct contrast to laboratory paradigms that usually rely on low-dimensional behavioral "snapshots" or summaries of behavior to verify linearly causal predictions about individual and collective phenomena, our computational model can be used to explore the effects of fine-grained, agent- and collective-level variations in cognitive ability on individual and collective behavior in real time.

For example, by parameterizing key cognitive abilities (Theory of Mind and Goal Alignment), our model shows that it is not necessarily a case of "more is better" when it comes to cognitive mechanisms underlying adaptive social behavior and collective intelligence. If an agent's level of social perceptiveness (Theory of Mind) were too low, it is likely that agents would miss vital performance-relevant information about the environment populated by other agents; if an agent's Theory of Mind were too high, it may instead over-index on partner belief states as an affordance for own beliefs (a scenario of "blind leading the blind"). We show that canonical cognitive abilities such as Theory of Mind and Goal Alignment can function across multiple scales to stabilize and reduce the computational uncertainty of an environment made up of other AIF agents, but only when these abilities are optimally tuned to a "goldilocks" level that is suitable to performance in that specific environment.

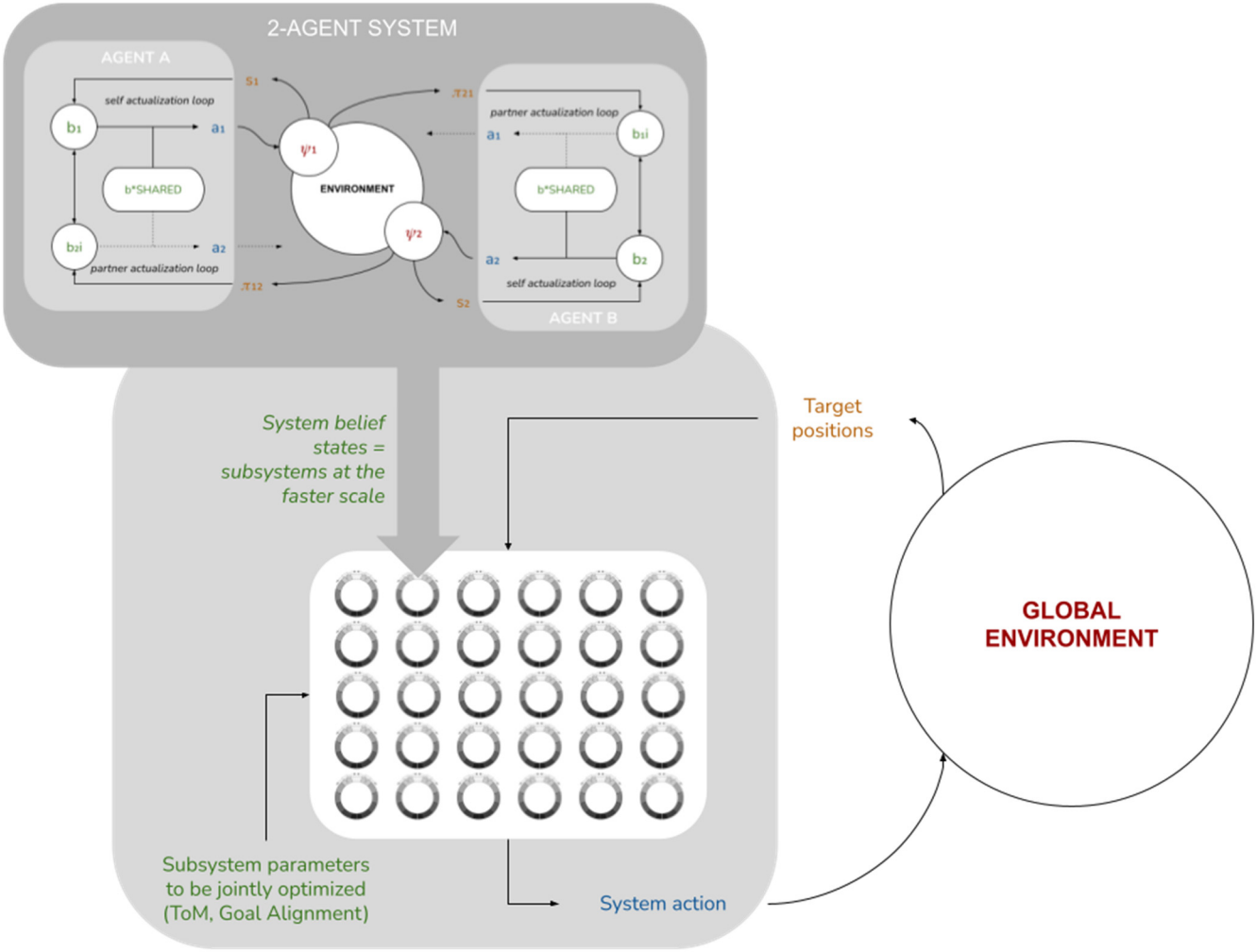


Figure 10. A notional complete two-scale model where agent-specific parameters are endogenized. This would entail parameters of subsystem components (Theory of Mind and Goal Alignment of each 2-agent system) being jointly optimized to inform a system action.

The essence of this proposal is captured by empirical research of attentional processes of human agents that engage in sophisticated joint action [74,75]. For instance, athletes in novice basketball teams are found to devote more attentional resources to tracking and monitoring their own teammates, while expert teams spend less time attending to each other and more time instead attending to the socio-technical task environment [76]. Viewed from the perspective of AIF, in both novice and expert teams alike, agents likely differentially deploy physical and social perceptiveness at levels that make sense for pursuing collective performance in a given situation; novices may stand to gain more from attending to (and therefore learning from) their teammates (recall our Agent B in Model 2 who leverages Theory of Mind to overcome weak physical perceptiveness, for example); while experts might stand to gain more from down-regulating social perceptiveness and redirecting limited attentional resources to physical perception of the task or (adversarial) social environment [77,78].

As evidence in organizational psychology and management suggests, (and outlined in the introduction), it is likely that social perceptiveness may indeed be an important factor (among many) that underwrites collective intelligence. But this may be especially the case in the context of unacquainted teams of "WEIRD" experimental subjects [79] who coordinate for a limited number of hours in a contrived laboratory setting [3]. If the experimental task were to be translated to a real-world performance setting (e.g., one involving high-

stakes or elite performance requirements), or if that same team of experimental subjects were to persist over time beyond the lab in a randomly fluctuating environment, it is conceivable that a premium for social perceptiveness may give way to demands for other types of abilities needed to continue to gain performance-relevant information from the task environment (e.g., through physical perceptiveness of the task environment). Viewed from this perspective, the true "special sauce" of collective intelligence (and individual intelligence, for that matter; see [80]) may turn out not to be one or other discrete or reified individual or team level ability per se (e.g., social perceptiveness), but instead a collective ability to nimbly adjust the volumes of multiple parameters to foster specific information-theoretic patterns conducive to minimizing free energy across multiple scales and over specific, performance-relevant time periods.

In this spirit, the computational approach we adopt here under AIF affords a dynamical and situational perspective on team performance that may offer important insights into long-standing and nascent hypotheses concerning the causal mechanisms of collective intelligence. For instance, our model is well positioned to investigate the long-proposed (but hitherto unsubstantiated) claim that successful team performance, and by extension, collective intelligence, depends on balancing a tradeoff between cognitive diversity and cognitive efficiency [4] (p. 421). Likewise, our approach could help elucidate mechanisms and dynamics through which memory, attention, and reasoning capabilities become distributed through a collective, and the conditions in which these "transactive" processes [81] facilitate emergence of intelligent behavior [77,82,83]. In either case, our model would simply require specification with the appropriate individual-level cognitive abilities or priors. For example, to better understand the causal relationship between transactive knowledge systems and collective intelligence, our model could leverage recent empirical research that observes a connection between individual agents' metacognitive abilities (e.g., perception of others' skills, focus, and goals), the formation of transactive knowledge systems, and a collective's ability to adapt to a changing task environment [83]. On an important and related note to these opportunities for future research, efforts to simulate human collective intelligence should strive to develop models composed of two or more agents to better mimic human-like coordination dynamics [50,84].

4.3. Increases in System Performance Correspond with Alignment between an Agent's Local and Global Optima

A key insight from our models, and worthy of further investigation, is that the greatest improvement in collective intelligence (Model 4; measured by global-scale inference) occurs when local-scale performance optima of individuals align with the system's global expected state. This effect can be understood as individuals jointly implementing approximate Bayesian inference of the system's expectations. In effect, our model suggests that multi-scale alignment between lower- and higher-order states may contribute to the emergence of collective intelligence.

Alignment between local and global states might sound like an obvious prerequisite for collective intelligence, particularly for more docile AIF agents such as neurons or cells (it is near impossible to imagine a scenario in which a neuron or cell could meaningfully persist without being spatially aligned with a superordinate agent; see [9]). But our model exemplifies a more subtle form of alignment, based on a loose coupling between scales through a system's generative model (Section 2.11), enabling the extension of this idea to scenarios where the local and global optimizations may be taking place in arbitrarily distinct and abstract state spaces [49,51]. By now it is well understood in brain and behavioral sciences that coordinated human behavior relies for its stability and efficacy on an intricate web of biologically evolved physiological and cognitive mechanisms [85,86], as well as culturally evolved affordances of language, norms, and institutions [87]. But precisely how these various mechanisms and affordances—particularly those that are separated across scales—coordinate in real or evolutionary time to enable human collective phenomena remains poorly understood [39,73,88].

Computational models, such as the one we have presented here, that are capable of formally representing multiscale alignment may help reorganize and clarify causal relationships between the various hypothesized physiological, cognitive, and cultural mechanisms hypothesized to underpin human collective behavior [14]. For example, a computational model such as the one proposed here could conceivably be adapted to help more systematically test the burgeoning hypothesis that coordination between basal physiological, metabolic and homeostatic processes at one scale of organization and linguistically mediated processes of interaction and exchange at another scale determine fundamental dynamics of individual and collective behavior [88–90].

Future research should aspire to examine causal connections between a fuller range of meaningful scales of behavior. In the case of human collectives, meaningful scales of behavior could extend from the basal mechanisms of physiological energy, movement, and emotional regulation on the micro scale [91,92], to linguistically- (and now digitally-) mediated social informational systems at the meso scale [93] to global socio-ecological systems at the macro scale [94–97]. As we have demonstrated here, the key requirement for the development of such multiscale models under AIF is faithful construction of the appropriate generative models at each scale. These models provide the mechanistic "missing links" between AIF and the phenomena to be explained—a task that will require tremendously innovative and intelligent collective behavior on the part of a diverse range of agents.

> *The patterns that crop up again and again in successful space are there because they are in fundamental accord with characteristics of the human creature. They allow him to function as a human. They emphasize his essence—he is at once an individual and a member of a group. They deny neither his individuality nor his inclination to bond into teams. They let him be what he is.*
>
> - DeMarco and Lister [98] (1987, p.90)

Author Contributions: Conceptualization, R.K., J.T. and P.G.; methodology, R.K., P.G. and J.T.; software, R.K. and P.G.; validation, R.K., P.G.; formal analysis, R.K.; investigation, R.K., P.G. and J.T.; resources, R.K., J.T. and P.G.; data curation, P.G. and R.K.; writing—original draft preparation, J.T. & R.K.; writing—review and editing, and J.T, R.K. and P.G.; visualization, P.G., R.K. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Model code can be found and implemented via this link to Google Colab: https://colab.research.google.com/drive/1CKdPTy8LD-Mpxc7kXy47m_fmCq44BT5u?usp=sharing (accessed on 15 June 2021).

Acknowledgments: The authors acknowledge the thoughtful and constructive feedback from all anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kelso, J.A.S. *Dynamic Patterns: The Self-Organization of Brain and Behavior*; MIT Press: Cambridge, MA, USA, 1995; ISBN 0262611317.
2. Riley, M.A.; Richardson, M.J.; Shockley, K.; Ramenzoni, V.C. Interpersonal synergies. *Front. Psychol.* **2011**, *2*, 38. [CrossRef]
3. Woolley, A.W.; Chabris, C.F.; Pentland, A.; Hashmi, N.; Malone, T.W. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science* **2010**, *330*, 686. [CrossRef]
4. Woolley, A.W.; Aggarwal, I.; Malone, T.W. Collective Intelligence and Group Performance. *Curr. Dir. Psychol. Sci.* **2015**, *24*, 420–424. [CrossRef]
5. Malone, T.W.; Bernstein, M.S. Introduction. In *Handbook of Collective Intelligence*; MIT Press: Cambridge, MA, USA, 2015.
6. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7280–7287. [CrossRef]
7. Friston, K.J.; Frith, C.D. A Duet for one. *Conscious. Cogn.* **2015**, *36*, 390–405. [CrossRef] [PubMed]
8. Friston, K.J.; Frith, C.D. Active inference, Communication and hermeneutics. *Cortex* **2015**, *68*, 129–143. [CrossRef] [PubMed]
9. Palacios, E.R.; Razi, A.; Parr, T.; Kirchhoff, M.D.; Friston, K. On Markov blankets and hierarchical self-organisation. *J. Theor. Biol.* **2020**, *486*, 110089. [CrossRef] [PubMed]

10. Pratt, S.C.; Mallon, E.B.; Sumpter, D.J.; Franks, N.R. Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant Leptothorax albipennis. *Behav. Ecol. Sociobiol.* **2002**, *52*, 117–127. [CrossRef]
11. Franks, N.R.; Dornhaus, A.; Fitzsimmons, J.P.; Stevens, M. Speed versus accuracy in collective decision making. *Proc. R. Soc. London. Ser. B Biol. Sci.* **2003**, *270*, 2457–2463. [CrossRef] [PubMed]
12. Constant, A.; Ramstead, M.J.D.; Veissière, S.P.L.; Friston, K. Regimes of expectations: An active inference model of social conformity and human decision making. *Front. Psychol.* **2019**, *10*. [CrossRef]
13. Ramstead, M.J.D.; Veissière, S.P.L.; Kirmayer, L.J. Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Front. Psychol.* **2016**, *7*, 1–21. [CrossRef]
14. Veissière, S.P.L.; Constant, A.; Ramstead, M.J.D.; Friston, K.J.; Kirmayer, L.J. Thinking Through Other Minds: A Variational Approach to Cognition and Culture. *Behav. Brain Sci.* **2019**. [CrossRef]
15. Baron-Cohen, S.; Tager-Flusberg, H.; Cohen, D.J. Understanding other Minds: Perspectives from autism. In *Most of the Chapters in This Book Were Presented in Draft form at a Workshop in Seattle*; Oxford University Press: Oxford, UK, 1994.
16. Tomasello, M.; Carpenter, M.; Call, J.; Behne, T.; Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behav. Brain Sci.* **2005**, *28*, 675–691, discussion 691–735. [CrossRef] [PubMed]
17. Chikersal, P.; Tomprou, M.; Kim, Y.J.; Woolley, A.W.; Dabbish, L. Deep structures of collaboration: Physiological correlates of collective intelligence and group satisfaction. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, OR, USA, 25 February 2017; pp. 873–888. [CrossRef]
18. Engel, D.; Malone, T.W. Integrated information as a metric for group interaction. *PLoS ONE* **2018**, *13*, 1–19. [CrossRef] [PubMed]
19. Riedl, C.; Kim, Y.J.; Gupta, P.; Malone, T.W.; Woolley, A.W. Quantifying Collective Intelligence in Human Groups. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2005737118. [CrossRef] [PubMed]
20. Rozin, P. Social psychology and science: Some lessons from solomon asch. *Personal. Soc. Psychol. Rev.* **2001**, *5*, 2–14. [CrossRef]
21. Kozlowski, S.W.J.; Chao, G.T. Unpacking team process dynamics and emergent phenomena: Challenges, conceptual advances, and innovative methods. *Am. Psychol.* **2018**, *73*, 576–592. [CrossRef] [PubMed]
22. O'Bryan, L.; Beier, M.; Salas, E. How approaches to animal swarm intelligence can improve the study of collective intelligence in human teams. *J. Intell.* **2020**, *8*, 9. [CrossRef] [PubMed]
23. Richardson, M.J.; Schmidt, R.C.; Richardson, M.J. Dynamics of interpersonal coordination. *Coord. Neural. Behav. Soc. Dyn.* **2008**, 281–308.
24. Kelso, J.A.S. Coordination dynamics. In *Encyclopedia of Complexity and Systems Science*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1537–1565.
25. Coey, C.A.; Varlet, M.; Richardson, M.J. Coordination dynamics in a socially situated nervous system. *Front. Hum. Neurosci.* **2012**, *6*, 164. [CrossRef] [PubMed]
26. Gorman, J.C.; Dunbar, T.A.; Grimm, D.; Gipson, C.L. Understanding and modeling teams as dynamical systems. *Front. Psychol.* **2017**, *8*, 1–18. [CrossRef]
27. Reinero, D.A.; Dikker, S.; Van Bavel, J.J. Inter-brain synchrony in teams predicts collective performance. *Soc. Cogn. Affect. Neurosci.* **2021**, *16*, 43–57. [CrossRef]
28. Gorman, J.C.; Amazeen, P.G.; Crites, M.J.; Gipson, C.L. Deviations from mirroring in interpersonal multifrequency coordination when visual information is occluded. *Exp. Brain Res.* **2017**, *235*, 1209–1221. [CrossRef]
29. Wiltshire, T.J.; Butner, J.E.; Fiore, S.M. Problem-Solving Phase Transitions During Team Collaboration. *Cogn. Sci.* **2018**, *42*, 129–167. [CrossRef]
30. Wiltshire, T.J.; Steffensen, S.V.; Fiore, S.M. Multiscale movement coordination dynamics in collaborative team problem solving. *Appl. Ergon.* **2019**, *79*, 143–151. [CrossRef]
31. Zhang, M.; Kelso, J.A.S.; Tognoli, E. Critical diversity: Divided or united states of social coordination. *PLoS ONE* **2018**, *13*, e0193843. [CrossRef]
32. Demir, M.; Mcneese, N.J.; Gorman, J.C.; Cooke, N.J.; Myers, C.; Grimm, D.A. Exploration of Team Trust and Interaction in Human-Autonomy Teaming. *IEEE Trans. Hum. Mach. Syst.* **2017**. [CrossRef]
33. Friston, K.J. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [CrossRef] [PubMed]
34. Friston, K.J. Life as we know it. *J. R. Soc. Interface* **2013**, *10*, 10. [CrossRef] [PubMed]
35. Friston, K.J. A free energy principle for a particular physics. *arXiv* **2019**, arXiv:1906.10184.
36. Buckley, C.L.; Kim, C.S.; McGregor, S.; Seth, A.K. The free energy principle for action and perception: A mathematical review. *J. Math. Psychol.* **2017**, *81*, 55–79. [CrossRef]
37. Friston, K.J.; Kilner, J.; Harrison, L. A free energy principle for the brain. *J. Physiol. Paris* **2006**, *100*, 70–87. [CrossRef]
38. Hohwy, J. The self-evidencing brain. *Nous* **2016**, *50*, 259–285. [CrossRef]
39. Ramstead, M.J.D.; Badcock, P.B.; Friston, K.J. Answering Schrödinger's question: A free-energy formulation. *Phys. Life Rev.* **2018**, *24*, 1–16. [CrossRef]
40. Kirchhoff, M.D.; Parr, T.; Palacios, E.; Friston, K.; Kiverstein, J. The markov blankets of life: Autonomy, active inference and the free energy principle. *J. R. Soc. Interface* **2018**, *15*. [CrossRef] [PubMed]
41. Hesp, C.; Ramstead, M.; Constant, A.; Badcock, P.; Kirchhoff, M.; Friston, K. A multi-scale view of the emergent complexity of life: A free-energy proposal. In *Evolution, Development and Complexity*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 195–227.

42. Badcock, P.B.; Friston, K.J.; Ramstead, M.J.D. The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Phys. Life Rev.* **2019**, *31*, 104–121. [CrossRef]
43. Sims, M. How to count biological minds: Symbiosis, the free energy principle, and reciprocal multiscale integration. *Synthese* **2020**. [CrossRef]
44. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Elsevier: Amsterdam, The Netherlands, 1988; ISBN 0080514898.
45. Friston, K.J. What is optimal about motor control? *Neuron* **2011**, *72*, 488–498. [CrossRef]
46. Haken, H. Synergetics. In *Self-Organizing Systems*; Springer: Berlin/Heidelberg, Germany, 1987; pp. 417–434.
47. Kirchhoff, M.D.; Kiverstein, J. How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese* **2019**. [CrossRef]
48. Ramstead, M.J.D.; Constant, A.; Badcock, P.B.; Friston, K.J. Variational ecology and the physics of sentient systems. *Phys. Life Rev.* **2019**, *31*, 188–205. [CrossRef]
49. Clark, A. How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds. *Philos. Predict. Coding* **2017**, 1–19. [CrossRef]
50. Zhang, M.; Beetle, C.; Kelso, J.A.S.; Tognoli, E. Connecting empirical phenomena and theoretical models of biological coordination across scales. *J. R. Soc. Interface* **2018**, *16*, 20190360. [CrossRef]
51. Krakauer, D.; Bertschinger, N.; Olbrich, E.; Flack, J.C.; Ay, N. The information theory of individuality. *Theory Biosci.* **2020**, *139*, 209–223. [CrossRef] [PubMed]
52. Ramstead, M.J.D. Have we lost our minds? An approach to multiscale dynamics in the cognitive sciences. Ph.D.'s Thesis, McGill University Libraries, Montréal, QC, Canada, 2019.
53. Searle, J.R. Minds and brains without programs. *Mindwaves* **1980**, *3*, 1–19.
54. Reia, S.M.; Amado, A.C.; Fontanari, J.F. Agent-based models of collective intelligence. *Phys. Life Rev.* **2019**, *31*, 320–331. [CrossRef]
55. Krafft, P.M. A Simple Computational Theory of General Collective Intelligence. *Top. Cogn. Sci.* **2019**, *11*, 374–392. [CrossRef] [PubMed]
56. Friston, K.J.; Daunizeau, J.; Kiebel, S.J. Reinforcement learning or active inference? *PLoS ONE* **2009**, *4*. [CrossRef]
57. Sajid, N.; Ball, P.J.; Parr, T.; Friston, K.J. Active Inference: Demystified and Compared. *Neural Comput.* **2021**, *44*, 1–39. [CrossRef]
58. Vasil, J.; Badcock, P.B.; Constant, A.; Friston, K.; Ramstead, M.J.D. A World Unto Itself: Human Communication as Active Inference. *Front. Psychol.* **2020**, *11*, 1–26. [CrossRef] [PubMed]
59. Hirschfeld, L.A. On a Folk Theory of Society: Children, Evolution, and Mental Representations of Social Groups. *Personal. Soc. Psychol. Rev.* **2001**, *5*, 107–117. [CrossRef]
60. Sperber, D. Intuitive and reflective beliefs. *Mind anguage* **1997**, *12*, 67–83. [CrossRef]
61. Yoshida, W.; Dolan, R.J.; Friston, K.J. Game theory of mind. *PLoS Comput. Biol.* **2008**, *4*. [CrossRef]
62. Press, W.H.; Dyson, F.J. Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 10409–10413. [CrossRef]
63. Baron-Cohen, S.; Wheelwright, S.; Hill, J.; Raste, Y.; Plumb, I. The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry Allied Discip.* **2001**, *42*, 241–251. [CrossRef]
64. Dunbar, R.I.M. The Social Brain: Mind, Language, and Society in Evolutionary Perspective. *Annu. Rev. Anthropol.* **2003**, *32*, 163–181. [CrossRef]
65. Pesquita, A.; Whitwell, R.L.; Enns, J.T. Predictive joint-action model: A hierarchical predictive approach to human cooperation. *Psychol. Bull.* **2017**, *25*, 1751–1769. [CrossRef]
66. Angus, S.D.; Newton, J. Emergence of Shared Intentionality Is Coupled to the Advance of Cumulative Culture. *PLoS Comput. Biol.* **2015**, *11*, 1–12. [CrossRef] [PubMed]
67. Fields, C.; Levin, M. How Do Living Systems Create Meaning? *Philosophies* **2020**, *5*, 36. [CrossRef]
68. McGregor, S.; Baltieri, M.; Buckley, C.L. A Minimal Active Inference Agent. *arXiv* **2015**, arXiv:1503.04187.
69. Levchuk, G.; Pattipati, K.; Serfaty, D.; Fouse, A.; McCormack, R. *Active Inference in Multi-Agent Systems: Context-Driven Collaboration and Decentralized Purpose-Driven Team Adaptation*; 2018 AAAI Spring Symposium Series; AAAI: Menlo Park, CA, USA, 2018; pp. 157–165.
70. van Schaik, A. Python Implementation of a Minimal Active Inference Agent. 2018. Available online: https://github.com/vschaik/Active-Inference (accessed on 1 January 2021).
71. Friston, K. The free-energy principle: A rough guide to the brain? *Trends Cogn. Sci.* **2009**, *13*, 293–301. [CrossRef] [PubMed]
72. Westby, C.E. Social neuroscience and theory of mind. *Folia Phoniatr. Logop.* **2014**, *66*, 7–17. [CrossRef]
73. Badcock, P.B. Evolutionary systems theory: A unifying meta-theory of psychological science. *Rev. Gen. Psychol.* **2012**, *16*, 10–23. [CrossRef]
74. Sebanz, N.; Bekkering, H.; Knoblich, G. Joint action: Bodies and minds moving together. *Trends Cogn. Sci.* **2006**, *10*, 70–76. [CrossRef]
75. Vesper, C.; Abramova, E.; Bütepage, J.; Ciardo, F.; Crossey, B.; Effenberg, A.; Hristova, D.; Karlinsky, A.; McEllin, L.; Nijssen, S.R.R.; et al. Joint Action: Mental Representations, Shared Information and General Mechanisms for Coordinating with Others. *Front. Psychol.* **2017**, *07*, 1–7. [CrossRef] [PubMed]

76. Bourbousson, J.; R'Kiouak, M.; Eccles, D.W. The dynamics of team coordination: A social network analysis as a window to shared awareness. *Eur. J. Work Organ. Psychol.* **2015**, *24*, 742–760. [CrossRef]
77. Bourbousson, J.; Fortes-Bourbousson, M. How do Co-agents Actively Regulate their Collective Behavior States? *Front. Psychol.* **2016**, *7*, 1732. [CrossRef] [PubMed]
78. R'Kiouak, M.; Saury, J.; Durand, M.; Bourbousson, J. Joint action of a pair of rowers in a race: Shared experiences of effectiveness are shaped by interpersonal mechanical states. *Front. Psychol.* **2016**, *7*, 1–17. [CrossRef] [PubMed]
79. Henrich, J.; Heine, S.J.; Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **2010**, *33*, 61–83. [CrossRef]
80. Friston, K.; FitzGerald, T.; Rigoli, F.; Schwartenbeck, P.; O'Doherty, J.; Pezzulo, G. Active inference and learning. *Neurosci. Biobehav. Rev.* **2016**, *68*, 862–879. [CrossRef]
81. Wegner, D.M. Transactive memory: A contemporary analysis of the group mind. In *Theories of Group Behavior*; Springer: Berlin, Germany, 1987; pp. 185–208.
82. Semin, G.R.; Garrido, M.V. Socially Situated Cognition: Imagining New. In *Theory and Explanation in Social Psychology*; Guilford Press: New York, NY, USA, 2015; Volume 36, pp. 774–777.
83. Gupta, P.; Woolley, A.W. The Emergence of Collective Intelligence Behavior. In Proceedings of the Paper presented at the 8th ACM Collective Intelligence (CI) Conference, Virtual Event, Zurich, Switzerland, 18 June 2020.
84. Richardson, M.J.; Garcia, R.L.; Frank, T.D.; Gergor, M.; Marsh, K. Measuring group synchrony: A cluster-phase method for analyzing multivariate movement time-series. *Front. Physiol.* **2012**, *3*, 405. [CrossRef]
85. Frith, U.; Frith, C.D. The social brain: Allowing humans to boldly go where no other species has been. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 165–176. [CrossRef]
86. Taylor, J.; Davis, A. Social Cohesion. In *The International Encyclopedia of Anthropology*; Wiley: Hoboken, NJ, USA, 2018; pp. 1–7.
87. Henrich, J. *The Secret of our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*; Princeton University Press: Princeton, NJ, USA, 2015; ISBN 1400873290.
88. Taylor, J.; Cohen, E. Social bonding through joint action: When the team clicks. *OSF Pre Print* **2019**. [CrossRef]
89. Barrett, L.F.; Simmons, W.K. Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* **2015**, *16*, 419–429. [CrossRef] [PubMed]
90. Krahé, C.; Springer, A.; Weinman, J.A.; Fotopoulou, A. The social modulation of pain: Others as predictive signals of salience-A systematic review. *Front. Hum. Neurosci.* **2013**, *7*, 386. [CrossRef]
91. Allen, M. Unravelling the Neurobiology of Interoceptive Inference. *Trends Cogn. Sci.* **2020**, *24*, 265–266. [CrossRef] [PubMed]
92. Barrett, L.F.; Quigley, K.S.; Hamilton, P. An active inference theory of allostasis and interoception in depression. *Philos. Trans. R. Soc. B* **2016**. [CrossRef]
93. Mesoudi, A. Cultural evolution: Integrating psychology, evolution and culture. *Curr. Opin. Psychol.* **2016**, *7*, 17–22. [CrossRef]
94. Doolittle, F.W.; Inkpen, A.S. Processes and patterns of interaction as units of selection: An introduction to ITSNTS thinking. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4006–4014. [CrossRef]
95. Kaufmann, R. Gaianomics, or the self-designing Earth. In *The Great Redesign: Frameworks for the Future*; Schrader, M., Martens, V., Eds.; Edition NFO; Next Factory Ottensen: Hamburg, Germany, 2020; ISBN 9783948580841.
96. Rubin, S.; Parr, T.; Da Costa, L.; Friston, K. Future climates: Markov blankets and active inference in the biosphere: Future climates: Markov blankets and active inference in the biosphere. *J. R. Soc. Interface* **2020**, *17*, 13–16. [CrossRef]
97. Boik, J.C. Science-driven societal transformation, Part I: Worldview. *Sustainability* **2020**, *12*, 6881. [CrossRef]
98. Lister, T.R.; DeMarco, T. *Peopleware: Productive Projects and Teams*; Dorset House: New York, NY, USA, 1987; ISBN 0932633056.