PLOS ONE

RESEARCH ARTICLE

Human-AI ecosystem with abrupt changes as a function of the composition

Pierluigi Contucci[1☯], **János Kertész**[2☯], **Godwin Osabutey**[1☯*]

1 Department of Mathematics, University of Bologna, Bologna, Italy, **2** Department of Network and Data Science, Central European University, Vienna, Austria

☯ All these authors are contributed equally to this work.
* godwin.osabutey2@unibo.it

Citation: Contucci P, Kertész J, Osabutey G (2022) Human-AI ecosystem with abrupt changes as a function of the composition. PLoS ONE 17(5): e0267310. https://doi.org/10.1371/journal.pone.0267310

Editor: Ning Cai, Beijing University of Posts and Telecommunications, CHINA

Received: January 7, 2022

Accepted: April 5, 2022

Published: May 27, 2022

Data Availability Statement: We do not use empirical data for the study. The obtained results are simulations performed with the model discussed in the manuscript and are reproducible by standard techniques. The parameter values used for the simulation and production of all the results can be found on top of each figure.

Funding: This project was supported by the EU H2020 ICT48 project Humane AI Net, through a grant awarded to PC and JK (952026). The European Union – Horizon 2020 Program also supported this study, under the scheme INFRAIA-

Abstract

The progressive advent of artificial intelligence machines may represent both an opportunity or a threat. In order to have an idea of what is coming we propose a model that simulate a Human-AI ecosystem. In particular we consider systems where agents present biases, peer-to-peer interactions and also three body interactions that are crucial and describe two humans interacting with an artificial agent and two artificial intelligence agents interacting with a human. We focus our analysis by exploring how the relative fraction of artificial intelligence agents affect that ecosystem. We find evidence that for suitable values of the interaction parameters, arbitrarily small changes in such percentage may trigger dramatic changes for the system that can be either in one of the two polarised states or in an undecided state.

Introduction

Artificial Intelligence (AI) is the property of human-made systems with specific or general goals and the ability to perceive and process information from their environment, to take action towards achieving their goals. Fields of applications include understanding human speech, pattern recognition, recommendation systems, navigation, health care, military and many more [1]. Such systems have become ubiquitous during the recent decades and are influencing our lives in many ways. Consciously or not, human-AI interactions have become natural in several domains like e-commerce, medical diagnostics, self-driven vehicles or online social networks.

The economic, sociological and ethical aspects of this emergent, complex human-AI ecosystem are of eminent interest, yet, we are only at the beginning of their investigation and even finding the right questions and approaches are challenging tasks. Unavoidably, complexity science should have a significant role in the endeavor of exploring this "terra incognita". Using its tools ranging from non-linear dynamics to network science and the theory of collective phenomena we have learned much about the human society [2]. Extending the results to the human-AI system and discovering the related new phenomena will be the goals of the period ahead.

Our main focus here stems from a paramount consideration of historical perspective. It is expected that the advent of machines that cover part of the intellectual activities of humans (AI

01-2018-2019 – Integrating Activities for Advanced Communities, through a grant given to JK (871042). This study was also funded by the CHIST-ERA grant CHIST-ERA-19-XAI-010, FWF, project SAI: Social Explainable Artificial Intelligence, awarded to JK (I 5205).

Competing interests: The authors have declared that no competing interests exist.

agents) is as unavoidable as the advent of those machines that came to help our physical work during the industrial revolution. At that time, in fact, the production of work from energy started to be shifted from the human and animal body to that of combustion engines. The progressive lowering of costs of both energy sources and work production from those machines pushed toward an unavoidable increase in their number. The lowering costs of information processing and the parallel increase of their abilities up to the modern deep learning marvels is analogously pushing the increase of their number and frequency of use. A specific problem we are interested in is understanding how the Human-AI ecosystem reacts therefore to an increase of the relative fraction of machines α. In particular, will the ecosystem react smoothly or abruptly? In the first scenario humans can optimize the outcomes by feedback while in the second there could be irreversible consequences.

Our plan is to describe the Human-AI ecosystem using the complex systems, statistical physics approach which has a well developed framework to spot abrupt changes and phase transitions.

Complex systems consist of interacting agents spanning a network. In social systems the nodes are individuals and the links between pairs of them stand for the different kinds of relationships humans may have. Recently the importance of so-called higher order interactions has been emphasized [3–5], where interactions beyond pairs are considered and processes on such networks have been studied, including opinion dynamics. From the mathematical point of view such generalization implies going from a simple graph-theoretical setting of vertices and edges to a more complex hyper-graph environment where also three-body terms *faces* or higher are considered. For example, in social systems the binding of a triple is significantly different from the sum of three links between its components. Analogously the presence of different types of agents, i.e. the nature of the agent inhabiting a node, represents a further element of complexity to be considered (see for instance [6]).

In the present paper we focus on ecosystems with both humans and AI agents with the presence of binary and higher order interactions. There are several reasons why such a system cannot be discussed in its full complexity. There is no generally accepted model of the problem, and we are not aware of the details of the interactions. In such a situation there are two main and in some sense opposite routes of theoretical approaches. First, one can set up a complex agent-based model where the numerous parameters can be target of a complex fitting procedure [7] to mimic behaviour which is thought to be reasonable. Second, one can chose an oversimplified model, where some basic elements of the original problem are present and where the analytical approach can be pushed to its limits. We can expect from such a model insight into the qualitative behaviour of the system, information about possible phases and the transitions between them. We follow in this paper the second route.

The first approach of this type can be traced back to Daniel McFadden's Discrete Choice theory [6]. It is possible to show indeed that such theory is equivalent to a multi-populated model without interaction or, in other words, a model of independent agents belonging to a finite number of groups [8]. In that model each agent aims to optimize its utility function up to some fluctuation of logistic type. In [9–11] the model, for a single group case, was generalized to include a mean field interaction between agents. A full interacting generalization of the McFadden model to the multi-group case came in [12–15] where specific case studies were investigated. It is important to emphasize that, for all the mentioned examples, the Gibbs distribution is only a working hypothesis to be used as a possible guide and, at best, to be tested against data: a distribution is assumed and the free parameters it depends on are statistically inferred. For an information theoretical perspective see [16–18].

As the starting point we have chosen the Ising model [19], the fundamental model of statistical physics, which was originally designed to describe the paramagnetic-ferromagnetic phase

transition. It consists of agents with binary state variables represented by nodes of a network, interacting with their neighbors, where interactions prefer alike or opposite states. A transition probability of the state variables depending on the neighbors' states and the noise level (temperature) completes the definition of the model. Later on we will define the model in mathematical terms. The Ising model has been solved analytically in the limit of large number of agents on the complete graph (mean-field solution) and on two-dimensional lattices. Here solution means the description of the system in its equilibrium, stationary state. It is worth noticing that Ising models with cubic interaction were studied in [20] in the context of occupation number variables (i.e. taking 0 and 1 values). In that study, the phase diagram and the segregation kinetics of a system of particles on a two-dimensional square lattice were investigated. The Ising model has been extensively used for modelling social phenomena [21].

In this work we extend the original mean-field Ising model in two ways. First, we have introduced two kinds of agents for humans and AI units, both having binary state variables, but the interaction between the different types of agents can be different. It should be noted that such a generalization to a two-component model has been studied earlier in a different context [22, 23]. Furthermore, we introduce higher order interactions to investigate the effects mentioned above. Our task is to calculate the phases and the transitions between them as a function of the coupling parameters and, especially, of the relative size of the two components that we call $\alpha \in [0, 1]$. When $\alpha = 1/2$ the two groups, for instance Human and AI would be of the same size, $\alpha = 0$ would mean only Human agents are present and vice versa $\alpha = 1$ only AI. Our aim is to gain insight into the possible behaviour of the human-AI ecosystem.

The paper is organized as follows. In the next section, we define the mathematical model with three-body interaction. The third section is devoted to the exploration of the results. We describe the solution with one component, and solve the two-component model. The analysis of the phases, and the determination of the order of the transitions are also discussed in the same section. General remarks and perspectives are discussed in the final section Conclusion and Outlook.

Materials and methods

The model and mathematical results

The human-AI ecosystem is an interacting system with two components, i.e., two kinds of entities. There are many ways to model mathematically such a system, depending on the characterization of the entities, the types of interactions, and the topology of the interactions. Our choice has been governed by the following viewpoints. We wanted to have an analytically tractable model, as such a solution enables general conclusions and qualitative predictions. Furthermore, we wanted to work with a model, which is closely related to standard theoretical models to be able to tie in with traditional techniques and results. Our choice has been a specific class of solvable Ising models that we can analytically control and make qualitative predictions with. Applications of the same class of models have been used in different fields such as health [12], immigration [8, 14, 24], education [15], energy conservation [25], and protein structure [26], among others. The paradigm we're considering therefore doesn't have to be understood as strictly typical to AI and Human agents; it is rather a flexible analytical guide to a variety of phenomena characterised by ecosystems of multi-components interacting agents.

The interactions that will be taken into account are the binary, quadratic ones (H-H, H-AI and AI-AI), the triplet or cubic (AI-AI-AI, AI-AI-H, AI-H-H, H-H-H) while the higher order ones (quartic, etc.) will be ignored. Furthermore, we consider the state (or opinion) variables of the agents as binary having in mind a task with possible binary outcome, like should a patient be operated or not. The nature of the interaction is that two or more agents in contact

may like (or dislike) each others "opinion" and they change their state variables such that the system finds a stationary state. We assume that all agents are in contact with each other, i.e., they sit on the nodes of a complete graph—an assumption needed for the analytical solvability of the system.

The model described above can be formulated mathematically as follows. Let us consider a system of N interacting agents, where each agent i has an internal degree of freedom described by a spin variable $\sigma_i \in \{-1, 1\}$ that represents the agent's opinion. A configuration of the system is then determined by a vector $\sigma \in \Sigma_N = \{-1, 1\}^N$. The class of models we are interested in is described by a *Hamiltonian function* or *cost function* of the following form

$$\mathcal{H}_N(\sigma) = -\sum_{i,j,k=1}^{N} K_{i,j,k}\,\sigma_i\sigma_j\sigma_k - \sum_{i,j=1}^{N} J_{i,j}\sigma_i\sigma_j - \sum_{i=1}^{N} h_i\sigma_i\,, \tag{1}$$

where $K_{i,j,k}$, $J_{i,j}$ are families of real parameters that tune the interactions among agents and the h_i tunes the bias of each agent. Eq (1) is the Hamiltonian of the Ising model with two-body and three-body interactions. Eq (1) is very general and it naturally includes the possibility of describing a two-component system, like the Human-AI ecosystem.

The parameter $J_{i,j}$ tunes the interaction among the couple of agents, while the $K_{i,j,k}$ those among triples. The positivity of those parameters imply, for the GKS inequalities [27–29], that the configurations with aligned opinions are the favored ones. The collective properties of the equilibrium state of the system are codified by a Boltzmann-Gibbs type probability distribution, related to the cost function (1)

$$\mu_N(\sigma) = \frac{e^{-\mathcal{H}_N(\sigma)}}{Z_N} \tag{2}$$

where the normalization factor, Z_N also called the partition function, is used to compute the generating function of the moments for the previous distribution (2)

$$p_N = \frac{1}{N}\log Z_N\,, \tag{3}$$

which coincides, up to a multiplicative factor, with the free energy of the model.

Clearly, the large N behaviour or thermodynamic limit (TDL) of p_N depends on the choice of the parameters of the Hamiltonian (1). In the sequel we will show how to represents the large number limit of p_N as a finite dimensional variational problem in two specific mean-field cases.

Results and discussion

The one component cubic mean-field model

In the sequel we consider only mean-field interaction among agents, i.e. we make the assumption based on the full permutation symmetry among agents, $K_{i,j,k} = K/3N^2$, $J_{i,j} = J/2N$ and $h_i = h$. The resulting model turns out to be a cubic mean-field model:

$$\mathcal{H}_N(\sigma) = -N\left(\frac{K}{3}m_N^3(\sigma) + \frac{J}{2}m_N^2(\sigma) + hm_N(\sigma)\right), \tag{4}$$

where $m_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i$ is average opinion. In Eq (4) K and J are the cubic and binary couplings, respectively, and h is the uniform bias. A remarkable fact about the considered model is that it may be solved exactly [30] by means of the *large deviations* technique, a method developed in [31]. One can prove that the large limit number of the generating functional (3) related

to the Hamiltonian (4) admits the following variational representation:

$$p(K,J,h) \coloneqq \lim_{N\to\infty} p_N = \sup_{m\in[-1,1]} \phi(m), \tag{5}$$

where $\phi(m) = U(m) - I(m)$ with

$$U(m) = \frac{K}{3}\, m^3 + \frac{J}{2} m^2 + hm \tag{6}$$

is the energy contribution and

$$I(m) = \frac{1-m}{2} \log\left(\frac{1-m}{2}\right) + \frac{1+m}{2} \log\left(\frac{1+m}{2}\right) \tag{7}$$

is the entropy contribution, expressing the logarithm of the number ways the value m can be produced with different σ configurations.

The structure of the probability measure identified by the variational principle (5) select stable solutions, i.e. a small stochastic disturbance of the system will produce small changes on the opinions unless the system is close to a second order phase transition.

The solutions of the variational principle (5) must satisfy the stationarity condition

$$\bar{m} = \tanh(K\bar{m}^2 + J\bar{m} + h), \tag{8}$$

the equation can be solved [30] by means of the local fixed-point method. Among those solutions we are only interested in the ones that realize the supremum of ϕ in (5) since they represent the overall opinion of the system with respect to the measure (2) in the equilibrium state. The quantity $\bar{m}$ is called the order parameter of the model. The overall phase picture that emerges presents novel features.

Unlike the quadratic mean field model that, for $h = 0$, has a second order continuous phase transition in J, the cubic case analyzed here displays a remarkable *discontinuous first order phase transition* in K when $J = h = 0$ shown in Fig 1. Starting from small absolute valued K-s and increasing or decreasing it, the value $\bar{m}$ characterizing the stable stationary solution remains at zero until $K = K_c \approx \pm 2.016295$ where suddenly we observe a jump in the order parameter.

The behaviour of the order parameter in the planes $(K, J, 0)$ and $(K, 0, h)$ is shown in Fig 2, while the case $(0, J, h)$ correspond to the classical two-body mean-field model. From panel (a) one can observe the presence, for $J < 1$, of three distinct phases: the one with negative average opinion (in blue), the one with zero average opinion (in gray) and the one with positive average opinion (red). In that region therefore a progressive increase in K from negative to positive values encounter two consecutive jumps.

The two component cubic mean-field model

Our main interest is to understand the behaviour of a system consisting of two kinds of agents, corresponding to AI and H. To this end we investigate the two component version of the model in Eq (1), defined as follows. Let partition the system of N agents into two subsystems AI and H of sizes N_1 and N_2 respectively, such that AI $\cap$ H $= \emptyset$ and $N_1 + N_2 = N$. Let $m_S(\sigma) = \frac{1}{|S|}\sum_{i\in S}\sigma_i$ be the average opinion of agents in a subsystem S and denote m_1 and m_2 as the average opinion for the subsystems AI and H respectively. Further, we define as the relative sizes of AI and H agents $\alpha_1 = \frac{N_1}{N}$ and $\alpha_2 = \frac{N_2}{N}$ respectively. The two component cubic mean-field model

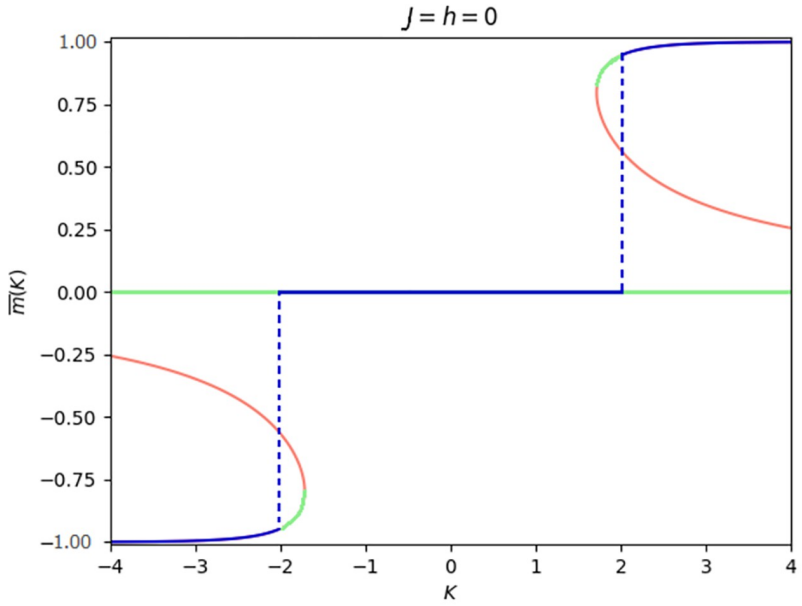

Fig 1. Average opinion, $\bar{m}$, of the system for $J = h = 0$ as a function of K. There is a transition in $\bar{m}$ from zero to positive average opinion when crossing $K = 2.016295$ from below and a transition from zero to negative average opinion when crossing $K = -2.016295$ from above. The curves represent all the solutions of the stationary condition (8), the blue ones corresponds to the global stable ones, i.e., the solution that realize the supremum of ϕ, the green ones are locally stable solutions and the red ones are the unstable solutions.

https://doi.org/10.1371/journal.pone.0267310.g001

has the following energy contribution:

$$U(m_1, m_2) = \frac{1}{3}\left[K_{111}\alpha_1^3 m_1^3 + 3K_{112}\alpha_1^2\alpha_2 m_1^2 m_2 + 3K_{122}\alpha_1\alpha_2^2 m_1 m_2^2 + K_{222}\alpha_2^3 m_2^3\right] + \frac{1}{2}\left[J_{11}\alpha_1^2 m_1^2 + 2J_{12}\alpha_1\alpha_2 m_1 m_2 + J_{22}\alpha_2^2 m_2^2\right] + \left[h_1\alpha_1 m_1 + h_2\alpha_2 m_2\right]. \tag{9}$$

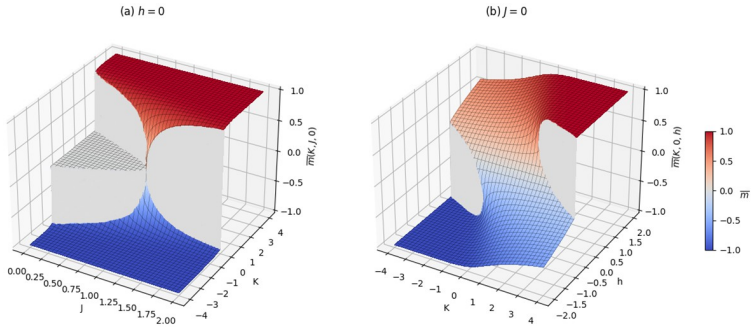

Fig 2. Average opinion surfaces of the cubic mean-field model. In panel **(a)**, $h = 0$ while in **(b)** $J = 0$. When $J = 0$ in **(a)**, we observe the global stable solution found in Fig 1 which indicates jumps at K_c and when $K = 0$, we obtain the solution of the simple Ising model without cubic interaction. For fixed $J < 1$ and moving along K the system presents two jumps separated by a plateau at zero. Those two jumps coalesce into a single one when J cross the unit value. In **(b)** we observe a discontinuity in the average opinion for two separated jumps when h and K falls within certain thresholds.

https://doi.org/10.1371/journal.pone.0267310.g002

The variational form of the large number limit of the generating functional (3) associated to the two component cubic mean-field model (9) is [30]:

$$\sup_{m\in[-1,1]^2} \Phi(m) = \sup_{m\in[-1,1]^2} [U(m_1, m_2) - (\alpha_1 I(m_1) + \alpha_2 I(m_2))] \tag{10}$$

where $I(m_1)$ and $I(m_2)$ are the entropy associated to the average opinions of the subsystems and they sum up to the total number of configurations as a product of the individual ones. The stationary solutions $\bar{m}$ of Φ are as follows

$$\bar{m}_l = \tanh\left(h_l + \sum_{p,q=1}^{2} \alpha_p (J_{lp} + \alpha_q K_{lpq} \bar{m}_q)\bar{m}_p\right) \quad \text{for } l = 1, 2, \tag{11}$$

from which the global stable ones are to be selected. In the rest of this work, we assign $\alpha = \alpha_1$ and $(1 - \alpha) = \alpha_2$ then $\alpha \in [0, 1]$ and the total average opinion $\bar{m} = \alpha\bar{m}_1 + (1 - \alpha)\bar{m}_2$ will be used as combined order parameter. It is worth recalling that, when $\alpha = 0$ then there are only Human agents in the population and when $\alpha = 1$ there are only AI agents in the population. For the rest of the work, we adopt the re-parameterisation of h_1 and h_2 found in [22]. In this sense, the parameters h_1 and h_2 are thought of as dependent on the internal average opinion (given by m_1^* and m_2^*) and interaction within each subsystem without interaction with the other agents. Hence following the fixed point method, we define h_1 and h_2 as follows;

$$\begin{aligned} h_1 &= \tanh^{-1}(m_1^*) - K_{111}m_1^*2 - J_{11}m_1^* \\ h_2 &= \tanh^{-1}(m_2^*) - K_{222}m_2^*2 - J_{22}m_2^* \end{aligned} \tag{12}$$

Surfaces of the solution of (11) that gives rise to the global maxima of Φ in Eq (10), with respect to the free parameters α and $K_{112} = K_{122} = K$ for fixed values of the other parameters are shown as Fig 3.

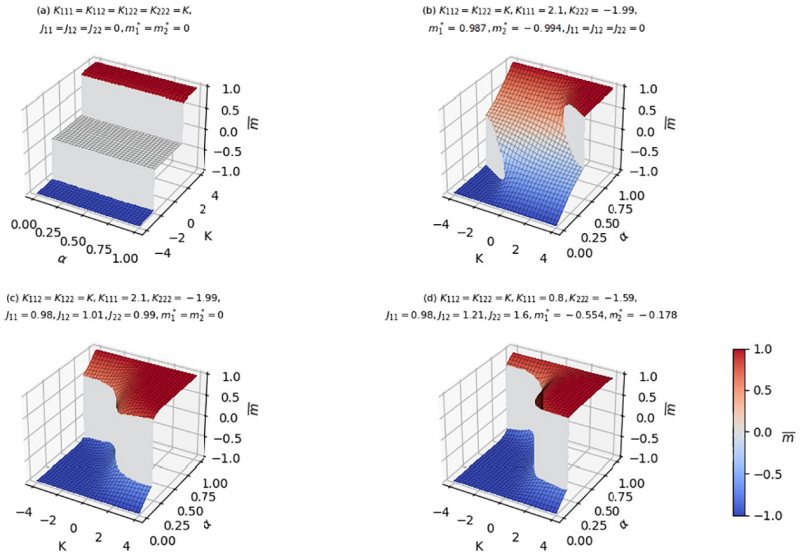

Fig 3. Total average opinion surfaces of the two component cubic mean-field model. In panel **(a)** we observe first order phase transitions at K_c. Here, α is constant in K. When the cubic interactions are fixed (i.e.$K_{111} = K_{112} = K_{122} = K_{222} = K$) the proportion of AI and Human agents present in the system has no effect on their average opinion as observed in panel **(a)**. Two distinct jumps in $\bar{m}$ are observed in panel **(b)** for certain values of K and α. For panel **(c)** and **(d)** α varies smoothly for the total average opinion and then observe sudden jump to another phase.

https://doi.org/10.1371/journal.pone.0267310.g003

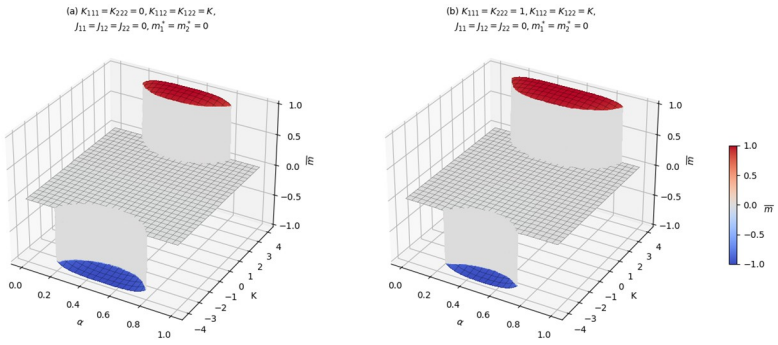

Fig 4. Average opinion for $K_{111} = K_{222} = 0$ and $K_{111} = K_{222} = 1$ with $K_{112} = K_{122} = K$ varying. In the left panel, panel **(a)**, the cubic in-group interaction for the AI agents and that of humans are set to zero (i.e. $K_{111} = K_{222} = 0$) and in **(b)** to one (i.e. $K_{111} = K_{222} = 1$) with varying inter-group interaction.

https://doi.org/10.1371/journal.pone.0267310.g004

When cross cubic interactions (i.e. $K_{112} = K_{122} = K$) are fixed, as observed in panels **(b), (c)** and **(d)** of Fig 3, there are jumps in the average opinion of the agents depending on their relative fractions. Smaller values of α (i.e. more Human agents), may lead to an inclination of the minus opinion while positive opinion inclination may result from larger values of α (i.e. more AI agents). Hence, a larger proportion of the AI agent population may lead to abrupt changes in behaviour of the ecosystem.

AI machines are made to assist and work with humans, therefore they are assumed to interact with humans. Fig 4 gives the scenario of an interacting system where only Humans or only AI agents are not interacting among themselves (see panel **(a)**) and when we assume that the cubic interaction among humans and AI agents are equal (see panel **(b)**). In both cases we observe transitions for large enough fraction of the AI agents in the order parameter when interaction bias and mutual interaction are absent.

Exploration of the effect of the composition

In this section we present phase diagrams of the model relating the parameter α and one among the interacting ones. The black continuous line is used to separate the opinion phases and in particular it emphasises the first order phase transitions, i.e. sudden jumps of the opinion resulting in abrupt changes of color in the picture. This is illustrated in Fig 5. The α-value found in correspondence of the black line indicates the proportion of AI agents required for Human opinion to lose its prevalence over the entire population.

The simulation of the results obtained for suitable values of the model parameter in Fig 5**(a)** suggest that even in the case where there is very small fraction of AI agents we can still observe abrupt behaviours in opinion formation within the Human AI ecosystem. This observation is likewise similar to that of panel **(d)** of Fig 5, which illustrate that for a system where there is less interaction among Human agents (K_{222}), smaller fraction of the AI agents may lead to phase transition and hence prevalent opinion formation.

Conclusion and outlook

The rapid spread of AI raises sharply the question of the relationship between humans and machine intelligence. How far can we control this development? Will we become vulnerable or will we succeed in humanising this new form of collaboration [32]? In fact, the problems are already all around us. What happens when AI and human participants in a medical consultation take different positions? Can trust or reticence develop between representatives of

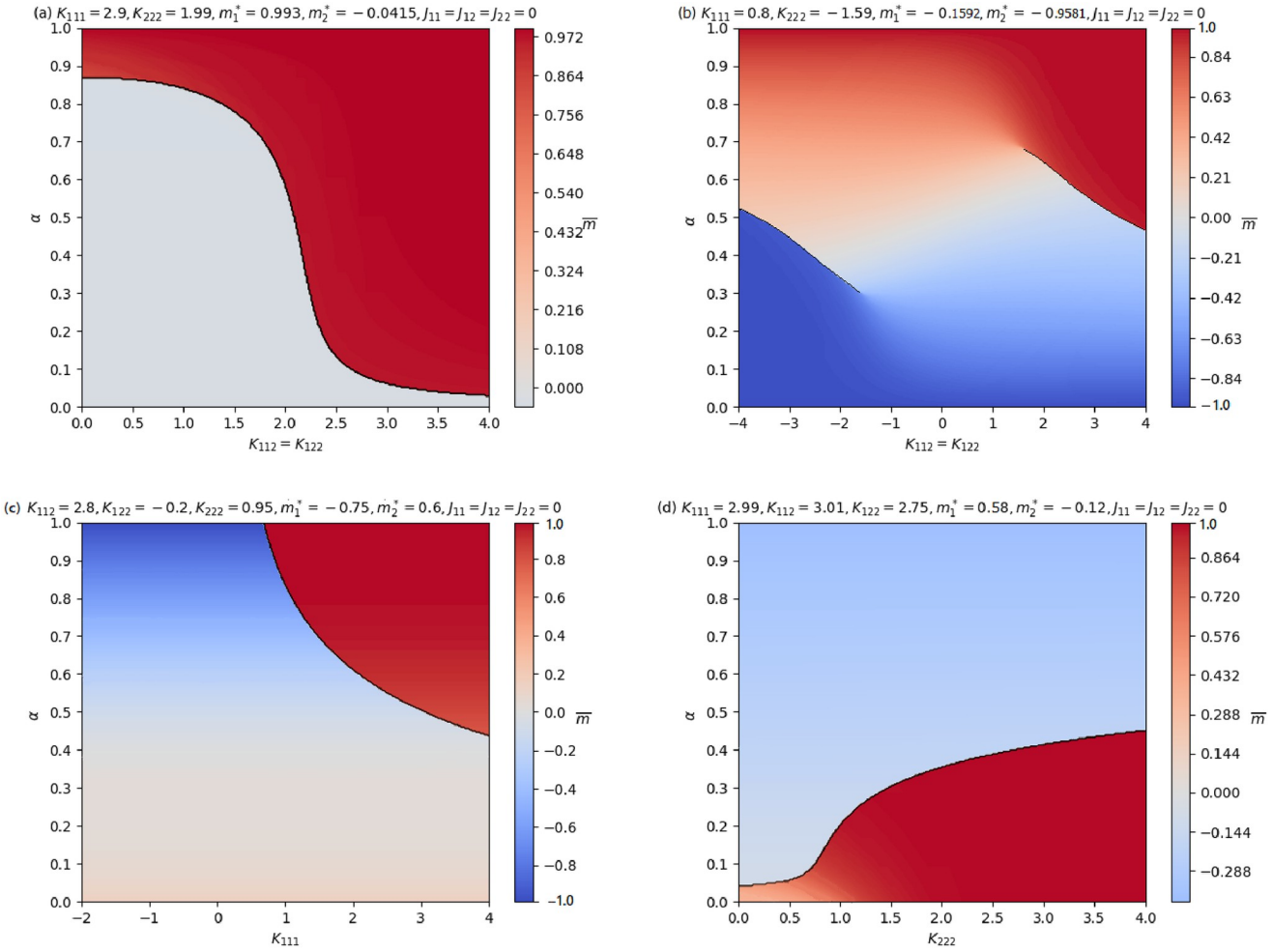

Fig 5. Phase diagram for fixed parameters of the cubic mean-field model. In panel **(a)**, when $K_{112} = K_{122}$ is small the system require a larger fraction (i.e. α) of the AI agents to observe a phase transition and as $K_{112} = K_{122}$ increases, the proportion of AI agents required for a phase transition decreases. In this scenario, the relative fraction of the AI agents corresponding to the black line is a decreasing function of $K_{112} = K_{122}$. While panel **(b)** illustrate two separate jumps in the average opinion depending on the values of $K_{112} = K_{122}$ and α. We observe from panel **(c)** that when interaction among AI agents (K_{111}) increases their proportion needed to observe a phase transition decreases. While in panel **(d)**, when interaction among Human agents (K_{222}) increases, the fraction of AI agents needed to effect a phase transition increases and vice versa.

https://doi.org/10.1371/journal.pone.0267310.g005

different origins? Can group interactions play a role? Our simple model aims to clarify such questions.

With the advancement of AI systems, there are growing concerns among scientists and psychologists about the creation of a machine world with AI and its eventual replacement of people in the labour force [33, 34]. In light of this possible occurrence, the following works [35–40] advise users and developers, as well as humans in general, to take an ethical approach to the design and use of these AI technologies, ensuring that humans are at the center of the algorithms with AI machines.

Notwithstanding these important factors, AI machines are needed to meet our daily needs as humans. To illustrate the significance of this study, let us first consider the undoubted benefit of AI machines to our healthcare system. Starting from chronic diseases and cancer to radiography and risk assessment, there are virtually endless opportunities to leverage technology to deliver more precise, efficient, and forceful interventions at precisely the appropriate

moment in a patient's care. Given the massive amount of data accessible, artificial intelligence is clearly positioned to be the primary engine that will propel our healthcare system forward as need dictates.

In this study we propose a simple, analytically solvable, mathematical model with higher order interaction that seeks to model the complex interactions between humans and AI agents. Using this model we predict and emphasize some crucial features of the system that are triggered on its composition i.e. the proportion of AI participating in a joint decision with us. This aspect is especially important in the light of the new tendency toward distributed AI, when the interaction of a large number of AI units is be more the standard than the exception [41]. The model utilized here has the potential to forecast the likely outcome in opinion formation (such as whether a patient should be operated on or not) for a certain proportion of AI or Human agents as they interact. Abrupt changes, known as discontinuous phase transitions in statistical physics, may predict thresholds beyond which AI machines may gain control prevalence. Experts in human-computer interaction, industries, institutions, psychologists, and others who work with AI machines may rely on the use of these models to set boundaries in order to achieve the desired and ethical benefit to humans while reducing its negative impact. The simulations conducted with our model (see Fig 5) show that a seemingly modest proportion of AI agents may lead to their prevalence in opinion formation over humans.

A noteworthy feature of the cubic mean-field model stems from the fact that we can observe three distinct phases depending on the parameter values in the absence of interaction bias of the agent(s) in the ecosystem (see for instance Figs 2(a), 3(a) and 4). Unlike the quadratic mean-field model, where we observe a jump from negative to positive state, instead one can observe a jump from the negative average opinion to a zero average opinion and a jump to a positive average opinion when three body interaction is considered. The zero average opinion, which is a stable paramagnetic state, is an indication of symmetry in opinion such that the agents have no preference for one over the other. As K increases or decreases, the symmetry in opinion is broken, and the total average opinion of the agents in the ecosystem shifts to either a positive or negative state.

The results illustrated in this work are some of the possible simulation for a wide class of values of the parameters. The model used and the whole statistical mechanics approach presented might also be used to infer the values of those parameters starting from real data as it was done in [12, 13–15].

Clearly, our approach has limitations. First, the dynamics leading to the stationary statistical distribution described by Eq (3) is a special one, while in reality opinion dynamics may be quite different as reflected in the numerous models introduced to study it [42]. However, we believe that our simplified model is sufficient for calling the attention to a possible source of criticality, namely that the dependence of the outcome in an AI/H ecosystem depends very non-linearly on the composition of the participants and this may have severe consequences.

A further aspect is that in realistic settings time should play an important role, which has been completely ignored here. With reference to the example above about the decision making process in a critical situation in health care, there is probably not enough time to achieve a complete equilibrium state of the participating opinion carriers. Another source of non-stationarity could be that the system is driven by a continuous flow of data. Therefore, more realistic models have to be dynamic in nature.

The above critical points give a guide to us in which direction one should continue the research on the Human-AI ecosystem. An important step should be to collect and use data of related systems as a starting point for the developments of adequate models.

Acknowledgments

G.O. thanks Emanuele Mingione for useful discussion during the beginning of this work.

Author Contributions

Conceptualization: Pierluigi Contucci, János Kertész, Godwin Osabutey.

Formal analysis: Pierluigi Contucci, János Kertész, Godwin Osabutey.

Funding acquisition: Pierluigi Contucci, János Kertész.

Methodology: Pierluigi Contucci, János Kertész, Godwin Osabutey.

Writing – original draft: Pierluigi Contucci, János Kertész, Godwin Osabutey.

Writing – review & editing: Pierluigi Contucci, János Kertész, Godwin Osabutey.

References

1. Russell S, Norvig P. Artificial intelligence: A modern approach. 4th ed. Upper Saddle River, NJ: Pearson; 2020.
2. Ball P. Why society is a complex matter: Meeting twenty-first century challenges with a new kind of science. 2012th ed. Berlin, Germany: Springer; 2012.
3. Alberici D, Contucci P, Mingione E, Molari M. Aggregation models on hypergraphs. Ann Phys (N Y). 2017; 376:412–424. https://doi.org/10.1016/j.aop.2016.12.001
4. Battiston F, Cencetti G, Iacopini I, Latora V, Lucas M, Patania A, et al. Networks beyond pairwise interactions: Structure and dynamics. Phys Rep. 2020; 874:1–92. https://doi.org/10.1016/j.physrep.2020.05.004
5. Bianconi G. Higher-Order Networks. Cambridge: Cambridge University Press; 2021. (Elements in Structure and Dynamics of Complex Networks).
6. McFadden D. Economic choices. Am Econ Rev. 2001; 91:351–378. https://doi.org/10.1257/aer.91.3.351
7. Murase Y, Jo H-H, Török J, Kertész J, Kaski K. Deep learning based parameter search for an agent based social network model. arXiv [physics.soc-ph]. 2021. Available from: http://arxiv.org/abs/2107.06507.
8. Gallo I, Barra A, Contucci P. A minimal model for the imitative behaviour in social decision making: theory and comparison with real data. Math Models Methods Appl Sci. 2009; 19.
9. Durlauf SN. of NBER, technical working paper series. Statistical Mechanics Approach to Socioeconomic Behavior. 1996;203.
10. Durlauf SN. How can statistical mechanics contribute to social science? Proc. Proc Natl Acad Sci USA. 1996; 96:10582–10584. https://doi.org/10.1073/pnas.96.19.10582
11. Brock WA, Durlauf SN. Discrete choice with social interactions Rev. Rev Econ Stud. 2001; 68:235–260. https://doi.org/10.1111/1467-937X.00168
12. Burioni R, Contucci P, Fedele M, Vernia C, Vezzani A. Enhancing participation to health screening campaigns by group interactions. Sci Rep [Internet]. 2015; 5:9904. https://doi.org/10.1038/srep09904 PMID: 25905450
13. Contucci P, Ghirlanda S. Modeling society with statistical mechanics: an application to cultural contact and immigration. Qual Quant. 2007; 41:569–578. https://doi.org/10.1007/s11135-007-9071-9
14. Barra A, Contucci P, Sandell R, Vernia C. An analysis of a large dataset on immigrant integration in Spain. The Statistical Mechanics perspective on Social Action. Sci Rep. 2015; 4. https://doi.org/10.1038/srep04174
15. Opoku AA, Osabutey G, Kwofie C. Parameter evaluation for a statistical mechanical model for binary choice with social interaction. J Probab Stat. 2019; 2019:1–10. https://doi.org/10.1155/2019/3435626
16. Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M, et al. Statistical mechanics for natural flocks of birds. Proc Natl Acad Sci U S A. 2012; 109:4786–4791. https://doi.org/10.1073/pnas.1118633109 PMID: 22427355
17. Jaynes ET. Information theory and statistical mechanics. Phys Rev. 1957; 106:620–630. https://doi.org/10.1103/PhysRev.106.620

18. Mckay D. Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press; 2003.

19. Brush SG. History of the Lenz-Ising model. Reviews of modern physics. 1967; 39. https://doi.org/10.1103/RevModPhys.39.883

20. Subramanian B, Lebowitz J. The study of a three-body interaction Hamiltonian on a lattice. J Phys A Math Gen. 1999; 32:6239–6246. https://doi.org/10.1088/0305-4470/32/35/302

21. Galam S. Sociophysics: A review of Galam models. Int J Mod Phys C. 2008; 19:409–440. https://doi.org/10.1142/S0129183108012297

22. Contucci P, Gallo I, Menconi G. Phase transitions in social sciences: two-population mean field theory, Int. Int Jou Mod Phys B. 2008; 22:1–14.

23. Gallo I, Contucci P. Bipartite Mean Field Spin Systems. Existence and Solution. Math Phys E J. 2008; 14.

24. Contucci P, Vernia C. On a statistical mechanics approach to some problems of the social sciences. Front Phys. 2020; 8. https://doi.org/10.3389/fphy.2020.585383

25. Osabutey G, Opoku AA, Gyamfi S. A statistical mechanics approach to the study of energy use behaviour. J Appl Math. 2020; 2020:1–14. https://doi.org/10.1155/2020/7384053

26. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 2011; 108:E1293–301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

27. Griffiths RB. A Proof that the Free Energy of a Spin System is Extensive. J Math Phys. 1964; 5:1215–1222. https://doi.org/10.1063/1.1704228

28. Griffiths RB. Correlations in Ising Ferromagnets. I. J Math Phys. 1967; 8:478–483. https://doi.org/10.1063/1.1705219

29. Kelly DG, Sherman S. General Griffiths' Inequalities on Correlations in Ising Ferromagnets. J Math Phys. 1968; 9:466–484. https://doi.org/10.1063/1.1664600

30. Osabutey G. *PhD Thesis in preparation*. Alma Mater Studiorum—University of Bologna, Italy.

31. Ellis RS. Entropy, large deviations, and statistical mechanics. 2006th ed. Berlin, Germany: Springer; 2005.

32. Xu W, Dainoff MJ, Ge L, Gao Z. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. arXiv [cs.HC] [Preprint] 2021. Available from: http://arxiv.org/abs/2105.05424.

33. Hawking S, Musk E, Wozniak S. Autonomous weapons: an open letter from AI and robotics researchers. Future of Life Institute. 2015.

34. Russell S, Dewey D, Tegmark M. Research priorities for robust and beneficial artificial intelligence. AI Mag. 2015; 36:105–114. https://doi.org/10.1609/aimag.v36i4.2577

35. Lau N, Fridman L, Borghetti BJ, Lee JD. Machine learning and human factors: Status, applications, and future directions. Proc Hum Factors Ergon Soc Annu Meet. 2018; 62:135–138. https://doi.org/10.1177/1541931218621031

36. Hancock PA. Some pitfalls in the promises of automated and autonomous vehicles. Ergonomics. 2019; 62:479–495. https://doi.org/10.1080/00140139.2018.1498136 PMID: 30024303

37. Stephanidis C, Salvendy G, Antona M, Chen JYC, Dong J, Duffy VG, et al. Seven HCI grand challenges. Int J Hum Comput Interact. 2019; 35:1229–1269. https://doi.org/10.1080/10447318.2019.1619259

38. Salmon PM. The horse has bolted! Why human factors and ergonomics has to catch up with autonomous vehicles (and other advanced forms of automation): Commentary on Hancock (2019) Some pitfalls in the promises of automated and autonomous vehicles. Ergonomics. 2019; 62:502–504. https://doi.org/10.1080/00140139.2018.1563333 PMID: 30957703

39. Shneiderman B. Human-centered artificial intelligence: Reliable, safe and trustworthy. Int J Hum Comput Interact. 2020; 36:495–504. https://doi.org/10.1080/10447318.2020.1741118

40. Shneiderman B. Design lessons from AI's two grand goals: Human emulation and useful applications. IEEE Trans Technol Soc. 2020; 1:73–82. https://doi.org/10.1109/TTS.2020.2992669

41. Eisenstadt V, Espinoza-Stapelfeld C, Mikyas A, Althoff K-D. Explainable distributed case-based support systems: Patterns for enhancement and validation of design recommendations. Case-Based Reasoning Research and Development. Cham: Springer International Publishing; 2018. p. 78–94.

42. Xia H, Xuan Z. Opinion Dynamics: A Multidisciplinary Review and Perspective on Future Research. International Journal of Knowledge and Systems Science. 2011; 2:72–91. https://doi.org/10.4018/jkss.2011100106