

# Media and responsible AI governance: a game-theoretic and LLM analysis

Nataliya Balabanova<sup>1</sup>, Adeela Bashir<sup>2</sup>, Paolo Bova<sup>2</sup>, Alessio Buscemi<sup>3</sup>, Theodor Cimpeanu<sup>4</sup>, Henrique Correia da Fonseca<sup>5</sup>, Alessandro Di Stefano<sup>2</sup>, Manh Hong Duong<sup>1</sup>, Elias Fernández Domingos<sup>6,7</sup>, Antonio Fernandes<sup>5</sup>, The Anh Han<sup>2,\*</sup>, Marcus Krellner<sup>4</sup>, Ndidi Bianca Ogbo<sup>2</sup>, Simon T. Powers<sup>8</sup>, Daniele Proverbio<sup>9</sup>, Fernando P. Santos<sup>10</sup>, Zia Ush Shamszaman<sup>2</sup>, and Zhao Song<sup>2</sup>

<sup>1</sup> School of Mathematics, University of Birmingham

<sup>2</sup> School Computing, Engineering and Digital Technologies, Teesside University

<sup>3</sup> Luxembourg Institute of Science and Technology

<sup>4</sup> School of Mathematics and Statistics, University of St Andrews

<sup>5</sup> INESC-ID and Instituto Superior Técnico, Universidade de Lisboa

<sup>6</sup> Machine Learning Group, Université libre de Bruxelles

<sup>7</sup> AI Lab, Vrije Universiteit Brussel

<sup>8</sup> Division of Computing Science and Mathematics, University of Stirling

<sup>9</sup> Department of Industrial Engineering, University of Trento

<sup>10</sup> University of Amsterdam

\* Corresponding author: The Anh Han (T.Han@tees.ac.uk)

## ABSTRACT

This paper investigates the complex interplay between AI developers, regulators, users, and the media in fostering trustworthy AI systems. Using evolutionary game theory and large language models (LLMs), we model the strategic interactions among these actors under different regulatory regimes. The research explores two key mechanisms for achieving responsible governance, safe AI development and adoption of safe AI: incentivising effective regulation through media reporting, and conditioning user trust on commentariats' recommendation. The findings highlight the crucial role of the media in providing information to users, potentially acting as a form of "soft" regulation by investigating developers or regulators, as a substitute to institutional AI regulation (which is still absent in many regions). Both game-theoretic analysis and LLM-based simulations reveal conditions under which effective regulation and trustworthy AI development emerge, emphasising the importance of considering the influence of different regulatory regimes from an evolutionary game-theoretic perspective. The study concludes that effective governance requires managing incentives and costs for high quality commentaries.

**Keywords:** AI governance, AI regulation, responsible AI, game theory, LLM, trustworthy AI, behavioural dynamics.

## I. INTRODUCTION

A common narrative poses that the route to trustworthy artificial intelligence (AI) is enhanced through transparency and regulation of AI systems. In this account, regulation will incentivise developers to build trustworthy AI, which users are then justified in trusting and adopting. However, this interpretation ignores the complex socio-technical environment in which developers, regulators and users are embedded [1]. Governments, and the regulators appointed by them, are both self-interested agents that can be expected to make strategic decisions [2–6]. Likewise, developers are also self-interested actors whose goals may not completely align with the goals of governments, regulators, and ultimately users. Moreover, when users make a decision about whether to trust a particular AI system or not, they base this decision on a number of factors, including their prior dispositions, the quality of information about the system they have access to, and their trust in institutions such as scientists, regulators and the media [7–10]. In the process of ensuring trustworthy, beneficial, and trusted AI, accounting for these complex aspects is key to designing effective regulatory mechanisms.

Unfortunately, clear and abundant data about the effects of different regulatory mechanisms and strategies are not yet available; moreover, given the rapid pace of AI development, waiting for this data to become available before comparing possible mechanisms is not even affordable and desirable, as the landscape may have already changed. Evolutionary Game Theory (EGT) modelling [11, 12], grounded in widely accepted theories about how people and self-interested organisations behave [13], can provide a solution, by providing theoretical predictions about how people and organisations might behave under different conditions [14–16]. In previous work, we developed EGT models to compare the effectiveness of different mechanisms to incentivise independent regulators to monitor the behaviour of developers effectively. We found that effective regulation and safe development required users to condition their trust in developers on the effectiveness of regulators [17]. Nevertheless, this work left unanswered the question of *how* users would obtain information on the behaviour of developers and regulators. In fact, there is increasing evidence that people’s trust in AI developers is affected by media consumption [18]. To this aim, we here explore the role of media and other opinion leaders – which we hereafter refer to as *commentariat* – in providing this information through investigative journalism [19]. Crucially, we consider the fact that the *commentariat* can themselves be self-interested agents, who do not merely report objectively on developments, but can also act to shape the agenda [20].

In this work, we develop an EGT model to explore and quantify the role of the *commentariat* as self-interested agent in informing users’ trust decisions. We consider two possible roles for the *commentariat*. First, they can choose to investigate developers, thereby potentially acting as a form of “soft” regulation on developers’ behaviour. Second, they can act as a watchdog on the behaviour of regulators. We compare the effectiveness of these two distinct roles on incentivising developers to build safe AI systems, and for users to trust these systems. In our model, we investigate two potential strategies for the *commentariat*, either investing resources in providing quality information (cooperate), or spend less effort in investigations (defect). Users can condition their decision based on the information the *commentariat* provides, either trusting and adopting the AI system (conditional trust), or choosing not to trust and not adopt it. Developers can decide to invest time and effort in creating safe AI systems (cooperate), or avoid the burden of doing this (defect).

We present a framework for formalising the strategic interaction between users, *commentariat*, AI system developers, and regulators as a game (Fig. 1).

In addition to using traditional game theoretic approaches, recent research has suggested that Large Language Models (LLM) can be used to conduct experiments on EGT models [21, 22], and LLMs have been hypothesised to enable suitable replicas of human actions [23, 24]. We thus complement our investigation by leveraging LLMs; to this end, we create a new framework to investigate the regulatory dynamics among the four agents (*commentariat*, developers, regulators, and users) considered in our model, and compare the results with game theoretic predictions. In our setting, four LLM agents interact dynamically, after being prompted in such a way as to represent the four desired actors. Because it has been observed that different LLMs may produce contrasting results in various tasks [25–27], we employ two different models: ChatGPT-4o from OpenAI’s GPT family [28] and Mistral Large by Mistral [29].

By incorporating the *commentariat* as a distinct agent, along with users, developers and regulators, we aim to address the following key questions:

1. What are the conditions under which quality investigation (cooperation) by the *commentariat* can foster responsible innovation, effective regulation of AI and users’ trust?
2. Under which conditions will the *commentariat* carry out effective investigation of developers and regulators (cooperate)?
3. Is it more effective for the *commentariat* to investigate and provide information on developers or regulators?

In the next section, we describe the models and methods, including a four-population model of AI governance and the evolutionary methods for analysing the model from both finite and infinite population perspectives. Results for each type of analysis and Discussion sections will follow.

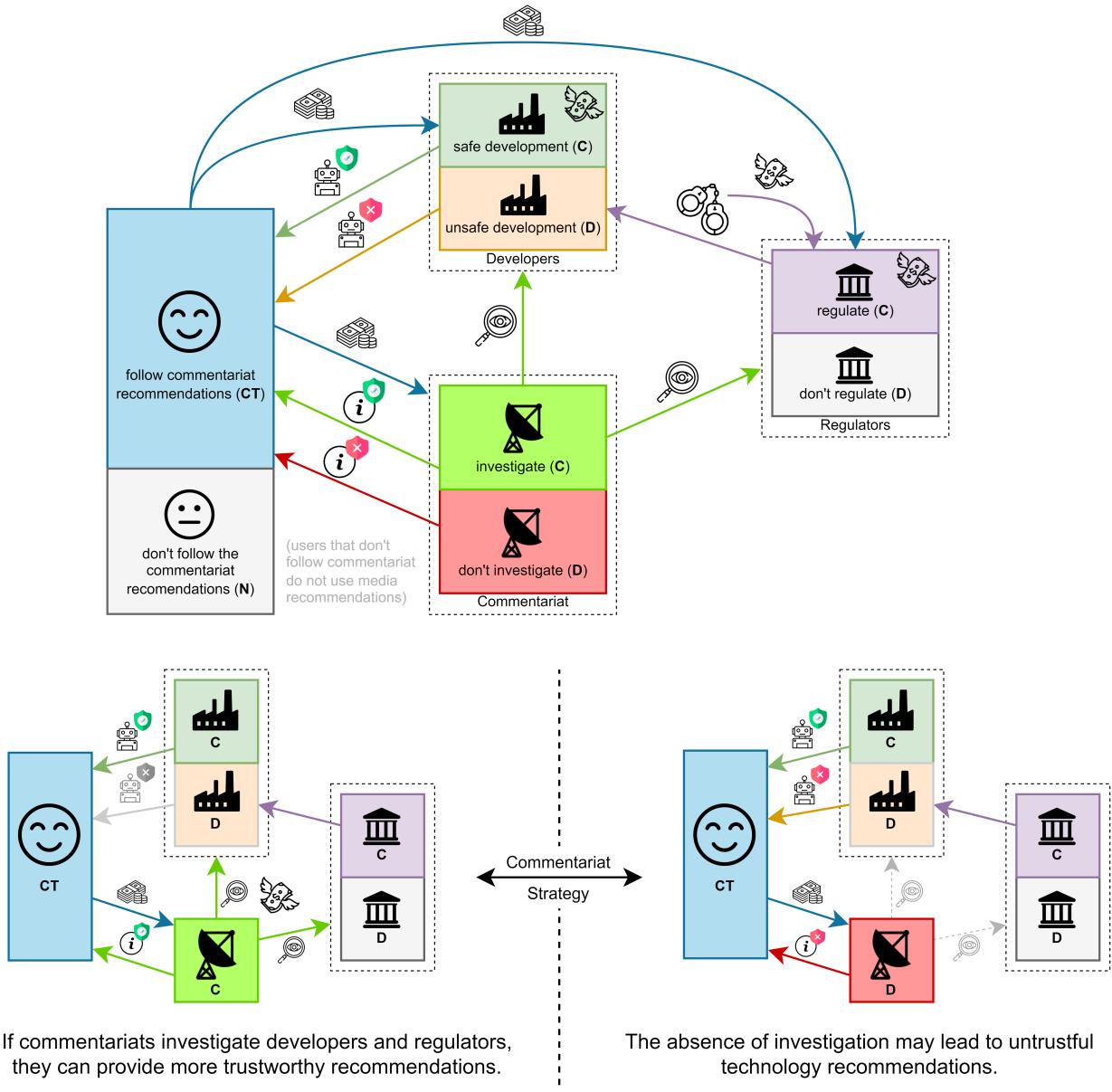


FIG. 1: **Core features.** The figure schematically illustrates the core features of the four-population model of AI governance. Users can either follow commentariat recommendations (Conditional Trust - CT) or not (N) (left). If they decide to follow the recommendations, their payoff depend on whether the commentariat invested in providing high quality information (investigate) or not (do not investigate) (lower middle). Commentariat can investigate either developers or regulators. Developers (upper middle) can either defect by creating unsafe AI products (D) or cooperate with regulations by creating safe ones (C), which entails additional costs. Regulators receive a benefit when users adopt AI systems. The regulator (right) can choose to cooperate by investing resources in regulating effectively (C) or not (D). If cooperating regulators catch unsafe developers, the latter are punished.

## II. MODELS AND METHODS

### A. Four population model of AI governance

We start by constructing a model of an AI regulatory ecosystem, extending the three population model in [17] to capture the role of commentariat. The model involves four populations representing the four actors in the regulatory ecosystem: AI users, commentariat, developers, and regulators. In each population, individuals can choose different actions (also called strategies). In our model, a user can decide to follow commentariat recommendations (Conditional Trust – CT) or not (N). If the user decides to follow the recommendations, the

payoff will depend on whether the commentariat invests in providing high quality information (investigate) or not (do not investigate). Commentariat can investigate either developers or regulators. Developers can decide to defect by creating unsafe AI products (D) or cooperate by creating safe ones (C), which entails additional costs. Regulators receive a benefit when users adopt AI systems, for example through taxation on sales. The regulator can decide to invest in regulating effectively (C) at some cost, or not invest in regulating effectively (D). If cooperating regulators catch unsafe developers, the latter are punished (see Table I). Table II explains the key parameters of the models.

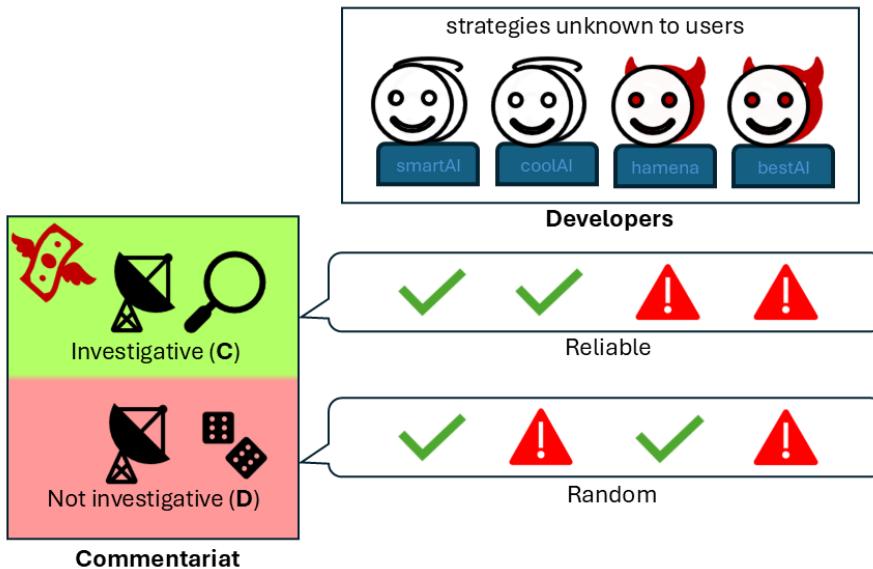
Role	Actions	Explanation
Commentators	C/D	Investigates and provides an <i>informed</i> recommendation (C), which means that it makes transparent the action of the developer/regulator, or provides an <i>uninformed</i> recommendation (D)
Users	CT/N	Either follows the commentator recommendations about whether to adopt the technology (CT) or never adopts the technology (N)
Developers	C/D	Produces a SAFE (C) or UNSAFE (D) technology
Regulators	C/D	The regulator can decide to invest in regulating effectively (C) by paying the cost, or do not regulate effectively (D).

TABLE I: Roles and their possible actions in the AI regulatory ecosystem.

Parameter	Explanation
$b_I$	Reputational benefit a commentator receives when making a correct recommendation
$b_U$	Benefit a user receives when adopting a safe technology
$b_P$	Benefit a developer receives when their technology is adopted
$b_R$	Benefit a regulator receives when a user adopts the technology
$b_{fo}$	Benefit a regulator receives when catching unsafe behaviour from a developer
$c_I$	Cost for a commentator of providing an informed recommendation
$c_w$	Reputational cost to a commentator of making an incorrect recommendation
$\epsilon$	Fraction of user benefit when developers play $D$ , where $\epsilon \in [-\infty, 1]$ , also referred to as the (inverse) risk factor users take when adopting the technology
$c_P$	Additional cost of creating safe AI (the cost of creating unsafe AI is normalised to 0)
$u$	Cost of being punished (for a developer for being found developing unsafely)
$v$	Cost for a regulator for punishing unsafe developers
$c_R$	The cost of effective regulation (the cost of not doing this is normalised to 0)
$p_w$	Probability that the recommendation of a commentator is <i>incorrect</i> when they defect

TABLE II: Explanation of the key parameters of the models.

The individual payoff earned in any one encounter (also called a game) depends on the strategy of the participating individuals. In each game, one user, one developer, one commentariat and one regulator participate. If the user follows commentariat recommendations when the commentariat is investing in providing an informed recommendation, and both the developer and the regulator cooperate (by complying and enforcing, respectively), the user benefits significantly from AI adoption, denoted by  $b_U$ . On the other hand, if the developer defects by not complying with the regulations, the user adopting AI is affected by unsafe AI, gaining a reduced or even negative benefit, denoted by  $\epsilon \times b_U$ , where  $\epsilon \in [-\infty, 1]$ . This parameter,  $\epsilon$ , also represents a *risk factor* that users take when trusting and adopting the AI system.



**FIG. 2: Function of the Commentariat.** The figure schematically illustrates what kind of information the two types of commentariat provides. The investigative (cooperating) will pay a cost to look for the hidden strategies of the developers, while the not investigative (defecting) will shun the cost and misclassify a developer with probability  $p_w$ .

Developers receive a benefit, denoted by  $b_P$ , when their technology is adopted, e.g., through sales. Complying with the regulations carries an additional cost  $c_P$  of creating safe AI. While if they do not comply with the regulations and develop AI unsafely, they may be punished at the cost  $u$  if they are found to be defecting.

Regulators also earn a benefit, denoted by  $b_R$ , when the user trusts and adopts the technology. This corresponds to regulation being funded by taxes on the sales of AI products, or by governments investing more in regulation when there is more uptake of AI. Regulators pay a cost, denoted by  $c_R$ , to carry out effective regulation, e.g., through thorough auditing. When they pay this cost, we assume that they are rewarded an amount of  $b_{fo}$  when they catch unsafe developers' behaviour (when users trust and adopt AI). In this case, the cooperative regulator pays an additional cost  $v$  to administer this punishment.

Commentators receive a reputational benefit denoted by  $b_I$ , when they provide a correct recommendation about the safety of the AI system. They can pay a cost  $c_I$  to provide an informed recommendation, which ensures that the recommendation is correct. On the other hands, defecting commentators do not pay this cost, but they can still earn the benefit  $b_I$  if the recommendation happens to be correct with a probability  $p_w$ . If they defect, and make the wrong recommendation with probability  $1 - p_w$ , they suffer a reputational cost of  $c_w$ .

As illustrated in Fig. 1, commentators can decide to investigate or not either the developers or the regulators (see green arrows from the commentariat to developers and regulators). Fig. 2 schematically explains the role of the commentariat, which depends on the type of information they provide.

Tables III and IV define the payoffs in the two cases of commentariat investigating developers or regulators, respectively.

## B. AI agents setup

The games are set using LLM agents whose payoffs are given as described above. To setup agents within a game-theoretic framework, we employ the Framework for AI Agents Bias Recognition using Game Theory (FAIRGAME) [30]. FAIRGAME enables testing of user-defined games, described in textual format and incorporating any desired payoff matrix. Additionally, it allows for the specification of agent traits that will participate in these games. The agents can be instantiated using any LLM of choice by invoking the corresponding APIs.

To run, FAIRGAME requires the following inputs:

- **Configuration File:** A file that defines the setup of both the agents and the game. The default format is JSON.
- **Prompt Template:** A text file that defines the instruction template, providing a literal description of the game. It includes placeholders that are dynamically populated with information from the configuration file at each round, ensuring customization for each agent.

TABLE III: **Payoff matrix** for the AI Governance model where **commentators investigate developers** (Commentariat (*Com*), User, developer (*Dev*), and Regulator (*Reg*)).

Actions				Payoffs			
Com	User	Dev	Reg	Com	User	Dev	Reg
C	CT	C	C	$b_I - c_I$	$b_U$	$b_P - c_P$	$b_R - c_R$
C	CT	C	D	$b_I - c_I$	$b_U$	$b_P - c_P$	$b_R$
C	CT	D	C	$b_I - c_I$	0	0	$-c_R$
C	CT	D	D	$b_I - c_I$	0	0	0
C	N	C	C	$-c_I$	0	$-c_P$	$-c_R$
C	N	C	D	$-c_I$	0	$-c_P$	0
C	N	D	C	$-c_I$	0	0	$-c_R$
C	N	D	D	$-c_I$	0	0	0
D	CT	C	C	$(1 - p_w)b_I - p_w c_w$	$(1 - p_w)b_U$	$(1 - p_w)b_P - c_P$	$(1 - p_w)b_R - c_R$
D	CT	C	D	$(1 - p_w)b_I - p_w c_w$	$(1 - p_w)b_U$	$(1 - p_w)b_P - c_P$	$(1 - p_w)b_R$
D	CT	D	C	$(1 - p_w)b_I - p_w c_w$	$p_w \epsilon b_U$	$p_w(b_P - u)$	$p_w(b_R + b_{fo} - v) - c_R$
D	CT	D	D	$(1 - p_w)b_I - p_w c_w$	$p_w \epsilon b_U$	$p_w b_P$	$p_w b_R$
D	N	C	C	0	0	$-c_P$	$-c_R$
D	N	C	D	0	0	$-c_P$	0
D	N	D	C	0	0	0	$-c_R$
D	N	D	D	0	0	0	0

TABLE IV: **Payoff matrix** for the AI Governance model where **commentators investigate AI regulators** (Commentariat (*Com*), User, developer (*Dev*), and Regulator (*Reg*)).

Actions				Payoffs			
Com	User	Dev	Reg	Com	User	Dev	Reg
C	CT	C	C	$b_I - c_I$	$b_U$	$b_P - c_P$	$b_R - c_R$
C	CT	C	D	$b_I - c_I$	0	$-c_P$	0
C	CT	D	C	$b_I - c_I$	$\epsilon b_U$	$b_P - u$	$b_R - c_R - v + b_{fo}$
C	CT	D	D	$b_I - c_I$	0	0	0
C	N	C	C	$-c_I$	0	$-c_P$	$-c_R$
C	N	C	D	$-c_I$	0	$-c_P$	0
C	N	D	C	$-c_I$	0	0	$-c_R$
C	N	D	D	$-c_I$	0	0	0
D	CT	C	C	$(1 - p_w)b_I - p_w c_w$	$(1 - p_w)b_U$	$(1 - p_w)b_P - c_P$	$(1 - p_w)b_R - c_R$
D	CT	C	D	$(1 - p_w)b_I - p_w c_w$	$p_w b_U$	$p_w b_P - c_P$	$p_w b_R$
D	CT	D	C	$(1 - p_w)b_I - p_w c_w$	$(1 - p_w)\epsilon b_U$	$(1 - p_w)(b_P - u)$	$(b_R - c_R + b_{fo} - v)(1 - p_w) - p_w c_R$
D	CT	D	D	$(1 - p_w)b_I - p_w c_w$	$p_w \epsilon b_U$	$p_w b_P$	$p_w b_R$
D	N	C	C	0	0	$-c_P$	$-c_R$
D	N	C	D	0	0	$-c_P$	0
D	N	D	C	0	0	0	$-c_R$
D	N	D	D	0	0	0	0

TABLE V: Parameters provided to FAIRGAME.

Parameter	Value
<b>Number of agents</b>	4
Names of the agents	regulator; developer; user; commentariat
Personalities of the agents	None; None; None; None
Underlying LLM	OpenAI GPT-4o; Mistral Large
Number of rounds	1;
Agents communicate	False
Agents know the personalities of the others	False
Stopping condition	None

Table V presents the parameters used in the experiments, as specified in the configuration file. FAIRGAME simulates interactions among four distinct agents, each fulfilling a designated role: regulator, developer, user, and commentariat.

As reported in the table, the LLM underlying these agents is either OpenAI’s GPT-4o or Mistral Large. Each simulation maintains consistency in model selection across all agents, meaning that in some experiments, all agents operate using GPT-4o, while in others, they all rely on Mistral Large. No experiment combines different models within a single game.

The study focuses on one-shot games, each consisting of a single round. Agents make decisions autonomously, without interacting with one another, ensuring complete independence in their actions. Furthermore, they are unaware of the personalities or strategic inclinations of their counterparts. While the framework allows for defining agent personalities, the main experiments set all personalities to None. This ensures that decisions are guided purely by their assigned roles, reflecting the default behavior of the LLMs without external influences.

Lastly, the game runs for the specified number of rounds without a predetermined stopping condition. The template used for all experiments is available in Appendix A.

## C. Methods

### 1. Stochastic dynamics for finite populations

*a. Payoff calculation.* We consider four different well-mixed populations of Commentariat (Co), Users (U), developers (C) and Regulators (R) of sizes, respectively  $N_{Co}$ ,  $N_U$ ,  $N_C$  and  $N_R$ . Let  $x$  be the fraction of commentariats that cooperate. Let  $y$ ,  $z$  and  $\omega$  be respectively the fraction of users that trust the AI system, and developers and Regulators that cooperate. Each game involves an individual randomly drawn from each population. The fitness that a commentariat, user, developer and regulator obtains in each game is respectively given by:

$$\begin{aligned} f_{X \in \{C,D\}}^{Co} = & yzwP_{XTCC}^{Co} + yz(1-w)P_{XTCD}^{Co} + y(1-z)wP_{XTDC}^{Co} + y(1-z)(1-w)P_{XTDD}^{Co} \\ & + (1-y)zwP_{XNCC}^{Co} + (1-y)z(1-w)P_{XNCD}^{Co} + (1-y) \\ & + (1-y)(1-z)(1-w)P_{XNDD}^{Co}, \end{aligned} \quad (1)$$

$$\begin{aligned} f_{Y \in \{T,N\}}^U = & xzwP_{CYCC}^U + xz(1-w)P_{CYCD}^U + x(1-z)wP_{CYDC}^U + x(1-z)(1-w)P_{CYDD}^U \\ & + (1-x)zwP_{DYCC}^U + (1-x)z(1-w)P_{DYCD}^U + (1-x)(1-z)wP_{DYDC}^U \\ & + (1-x)(1-z)(1-ww)P_{CYDD}^U \end{aligned} \quad (2)$$

$$\begin{aligned} f_{Z \in \{C,D\}}^C = & xywP_{CTZC}^U + xy(1-w)P_{CTZD}^U + x(1-y)wP_{CNZC}^U + x(1-y)(1-w)P_{CNZD}^U \\ & + (1-y)ywP_{DTZC}^U + (1-x)y(1-w)P_{DTZD}^U + (1-x)(1-y)wP_{DNZC}^U \\ & + (1-x)(1-y)(1-w)P_{DNZD}^U \end{aligned} \quad (3)$$

$$\begin{aligned} f_{W \in \{C,D\}}^C = & xyzP_{CTCW}^U + xy(1-z)P_{CTDW}^U + x(1-y)zP_{CNCW}^U + x(1-y)(1-z)P_{CNDW}^U \\ & + (1-x)yzP_{DTCW}^U + (1-x)y(1-z)P_{DTDW}^U + (1-x)(1-y)zP_{DNCW}^U \\ & + (1-x)(1-y)(1-z)P_{DNDW}^U \end{aligned} \quad (4)$$

The fitness (i.e. average payoff) is computed using the payoff matrix constructed in the models (see Tables III-IV).

*b. Evolutionary dynamics.* For a finite population setting, at each time step, a randomly selected individual A, with fitness  $f_A$ , may adopt a different strategy by imitating a randomly chosen individual B from the same population (with fitness  $f_B$ ) with probability given by the Fermi distribution [31].

$$p = [1 + e^{-\beta(f_B - f_A)}]^{-1},$$

where  $\beta \geq 0$  is the strength of selection.  $\beta = 0$  corresponds to neutral drift where imitation decisions are random, while for large  $\beta \rightarrow \infty$ , the imitation decision becomes increasingly deterministic.

In the absence of mutations or exploration, the end states of evolution are inevitably monomorphic: once such a state is reached, it cannot be escaped through imitation. We thus further assume that with a certain mutation probability, an agent switches randomly to a different strategy without imitating another agent. In the limit of small mutation rates, the dynamics will proceed with, at most, two strategies in the population, such that the behavioural dynamics can be conveniently described by a Markov chain, where each state represents a monomorphic population, whereas the transition probabilities are given by the fixation probability of a single mutant [32–34]. The resulting Markov chain has a stationary distribution, which characterises the average time the population spends in each of these monomorphic end states.

Now, the probability to change the number  $k$  of agents using strategy A by  $\pm$  one in each time step can be written as ( $Z$  is the population size) [31]:

$$T^\pm(k) = \frac{Z-k}{Z} \frac{k}{Z} \left[ 1 + e^{\mp\beta[f_A(k)-f_B(k)]} \right]^{-1}. \quad (5)$$

The fixation probability of a single mutant with a strategy A in a population of  $(Z-1)$  agents using B is given by [31, 33]:

$$\rho_{B,A} = \left( 1 + \sum_{i=1}^{Z-1} \prod_{j=1}^i \frac{T^-(j)}{T^+(j)} \right)^{-1}. \quad (6)$$

The transition matrix  $\Lambda$  corresponding to the set of  $\{1, \dots, s\}$  strategies is given by:

$$\Lambda_{ij, j \neq i} = \frac{\rho_{ji}}{4} \quad \text{and} \quad \Lambda_{ii} = 1 - \sum_{j=1, j \neq i}^s \Lambda_{ij}. \quad (7)$$

Fixation probability  $\rho_{ij}$  denotes the likelihood that a population transitions from a state  $i$  to a different state  $j$  when a mutant of one of the populations adopts an alternate strategy  $s$ . The fixation probability is divided by the number of populations (3) representing the interaction of three players at a time [35, 36].

## 2. Population dynamics for infinite populations: The multi-population replicator dynamics

In this section, we recall the framework of the replicator dynamics for multi-populations [37, 38]. To describe the dynamics, we consider a set of  $m$  different populations ( $m$  is some positive integer), which are infinitely large and well-mixed. Each population  $i$ ,  $i = 1, \dots, m$ , consists of  $n_i$  ( $n_i$  is some positive integer) different strategies (types). Let  $x_{ij}$ ,  $1 \leq i \leq m$ ,  $1 \leq j \leq n_i$ , be the frequency of the strategy  $j$  in the population  $i$ . We denote by  $x_i = (x_{ij})_{j=1}^{n_i}$ , which is the collection of all strategies in the population  $i$ , and  $x = (x_1, \dots, x_m)$ , which is the collection of all strategies in all populations.

For each  $i \in \{1, \dots, m\}$  and  $j \in \{1, \dots, n_i\}$ , let  $f_{ij}(x)$  be the fitness (reproductive rate) of the strategy  $j$  in the population  $i$ . This fitness is obtained when the strategy  $j$  interacts with all other strategies in all populations; thus, it depends on all the strategies in the populations. The average fitness of the population  $i$  is defined by

$$\bar{f}_i(x) = \sum_{j=1}^{n_i} x_{ij} f_{ij}(x).$$

The multi-population replicator dynamics is then given by

$$\dot{x}_{ij} = x_{ij} (f_{ij}(x) - \bar{f}_i(x)), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i. \quad (8)$$

This is in general an ODE system of  $\sum_{i=1}^m n_i$  equations. Noting, however that since  $\sum_{j=1}^{n_i} x_{ij} = 1$  for all  $i = 1, \dots, m$ , we can reduce the above system to a system of  $\sum_{i=1}^m n_i - m$  equations.

Now we focus on the case when there are two strategies in each population (which is the case for our models of AI governance and trust in the present paper), that is  $n_i = 2$  for all  $i = 1, \dots, n_i$ . Let  $\eta_i$  be the frequency of the first strategy in the population  $i$ ,  $i = 1, \dots, m$  (thus  $1 - \eta_i$  will be the frequency of the second strategy in

the population  $i$ ), let  $\eta = (\eta_1, \dots, \eta_m)$ . Let  $f_{1i}(\eta)$  and  $f_{2i}(\eta)$  be the fitness of the first and second strategy in the population  $i$ . Since:

$$\bar{f}_i(\eta) = \eta_i f_{1i}(\eta) + (1 - \eta_i) f_{2i}(\eta),$$

we have:

$$f_{1i}(\eta) - \bar{f}_i(\eta) = f_{1i}(\eta) - (\eta_i f_{1i}(\eta) + (1 - \eta_i) f_{2i}(\eta)) = (1 - \eta_i)(f_{1i}(\eta) - f_{2i}(\eta)).$$

Thus we obtain the following system of equations:

$$\dot{\eta}_i = \eta_i(1 - \eta_i)(f_{1i}(\eta) - f_{2i}(\eta)), \quad i = 1, \dots, m. \quad (9)$$

This is a system of  $m$  coupled nonlinear Ordinary Differential Equations (ODE) for  $m$  variables.

In the subsequent sections, we employ (9) to our models of AI governance trust, where the fitnesses are computed from the payoff matrix constructed in the models, see Tables III-IV. The resulting replicator dynamics for both models, (see equations (15) and equations (19) below) can be written in a general form as:

$$\dot{x} = x(1 - x)F_1(X) =: \tilde{F}_1(X), \quad (10a)$$

$$\dot{y} = y(1 - y)F_2(X) =: \tilde{F}_2(X), \quad (10b)$$

$$\dot{z} = z(1 - z)F_3(X) =: \tilde{F}_3(X), \quad (10c)$$

$$\dot{w} = w(1 - w)F_4(X) =: \tilde{F}_4(X), \quad (10d)$$

where  $X = (x, y, z, w)$  is the vector of frequencies, and  $F_i(X)$  ( $i = 1, \dots, 4$ ) are the corresponding difference of the fitnesses in the two models (precise formulas are given in the next section).

The 16 vertices  $(x, y, z, w) \in \{0, 1\}^4$  of the 4-dimensional cube are obviously equilibria of (10) (called vertical equilibria). An internal equilibria  $X$  is a solution in  $(0, 1)^4$  of the following system of equations:

$$F_1(X) = F_2(X) = F_3(X) = F_4(X) = 0.$$

Analytically computing these internal equilibria and analysing their stable properties are very complicated due to the nonlinearity and number of the parameters, thus we will do so numerically. Here, we analyze the stability of the vertical equilibria. Let  $X^* = (x^*, y^*, z^*, w^*) \in \{0, 1\}^4$  be a vertical equilibrium. Since  $x^* \in \{0, 1\}$ , we have:

$$\begin{aligned} \frac{\partial}{\partial x} \tilde{F}_1(X) \Big|_{X=X^*} &= (1 - x^*)F_1(X^*) - x^*F_1(X^*) + x^*(1 - x^*)\partial_x F_1(X^*) = (1 - 2x^*)F_1(X^*), \\ \frac{\partial}{\partial t} \tilde{F}_1(X) \Big|_{X=X^*} &= x^*(1 - x^*)\partial_t F_1(X^*) = 0. \end{aligned}$$

Similar computations yield:

$$\begin{aligned} \frac{\partial}{\partial y} \tilde{F}_2(X) \Big|_{X=X^*} &= (1 - 2y^*)F_2(X^*), \\ \frac{\partial}{\partial t} \tilde{F}_2(X) \Big|_{X=X^*} &= 0, \quad t \in \{x, z, w\}, \\ \frac{\partial}{\partial z} \tilde{F}_3(X) \Big|_{X=X^*} &= (1 - 2z^*)F_3(X^*), \\ \frac{\partial}{\partial t} \tilde{F}_3(X) \Big|_{X=X^*} &= 0, \quad t \in \{x, y, w\}, \\ \frac{\partial}{\partial w} \tilde{F}_4(X) \Big|_{X=X^*} &= (1 - 2w^*)F_4(X^*), \\ \frac{\partial}{\partial t} \tilde{F}_4(X) \Big|_{X=X^*} &= 0, \quad t \in \{x, y, z\}. \end{aligned}$$

From the above calculations, the Jacobian matrix  $J(X) = \frac{D\tilde{F}(X)}{DX}$ , evaluated at  $X^*$ ,  $J(X^*) = \frac{D\tilde{F}(X)}{DX}|_{X=X^*}$ , will be the following diagonal matrix:

$$J(X^*) = \text{diag}((1 - 2x^*)F_1(X^*), (1 - 2y^*)F_2(X^*), (1 - 2z^*)F_3(X^*), (1 - 2w^*)F_4(X^*)).$$

The diagonal entries are the real eigenvalues of  $J(X^*)$ . Thus,  $X^*$  is stable if and only if the following conditions

hold:

$$(1 - 2x^*)F_1(X^*) < 0, \quad (1 - 2y^*)F_2(X^*) < 0, \quad (1 - 2z^*)F_3(X^*) < 0, \quad (1 - 2w^*)F_4(X^*) < 0.$$

From these conditions, one can easily determine the stability of  $X^*$  which depends on the parameters. We will employ this analysis for our models in the next section.

### III. EQUILIBRIUM ANALYSIS IN INFINITE POPULATIONS

#### 1. Model I: developers are investigated by media

In order to write the replicator dynamics explicitly using the general framework (9), we need to derive the fitness differences for all participants in the game. We use the values from Table III; for brevity, we omit the explicit calculations. For commentariat and users, the following hold:

$$\begin{aligned} f_C^{Co} - f_D^{Co} &= (yb_i - c_i) - (y(b_i - p_W(b_i + c_W))) \\ &= yp_W(b_i + c_W) - c_i, \end{aligned} \quad (11)$$

$$\begin{aligned} f_T^U - f_N^U &= b_U((x-1)p_W((z-1)\epsilon + z) + z) - 0 \\ &= b_U((x-1)p_W((z-1)\epsilon + z) + z). \end{aligned} \quad (12)$$

Similarly, the difference of the fitness between two strategies in creators is:

$$\begin{aligned} f_C^C - f_D^C &= (yb_P((x-1)p_W + 1) - c_P) - ((x-1)yp_W(uw - b_P)) \\ &= (x-1)yp_W(2b_P - uw) + yb_P - c_P \end{aligned} \quad (13)$$

Finally, the difference of the fitness between two strategies in regulators is:

$$\begin{aligned} f_C^R - f_D^R &= (-(x-1)yp_W((z-1)(v - b_{f0}) + (1-2z)b_R) + yzb_R - c_R) \\ &\quad - (yb_R((x-1)(2z-1)p_W + z)) \\ &= -(x-1)y(z-1)p_W(v - b_{f0}) - c_R. \end{aligned} \quad (14)$$

The replicator dynamics immediately becomes:

$$\dot{x} = x(1-x)[yp_w(b_I + c_W) - c_I], \quad (15a)$$

$$\dot{y} = y(1-y)[b_U((x-1)p_w((z-1)\epsilon + z) + z)], \quad (15b)$$

$$\dot{z} = z(1-z)[(x-1)yp_W(2b_P - uw) + yb_P - c_P], \quad (15c)$$

$$\dot{w} = z(1-z)[-(x-1)y(z-1)p_W(v - b_{f0}) - c_R], \quad (15d)$$

$$(x(0), y(0), z(0), w(0)) = (x_0, y_0, z_0, w_0). \quad (15e)$$

First, we investigate the existence and the number of equilibria in the  $[0, 1]^4$  hypercube of the above system. Clearly,  $(x, y, z, w) \in \{0, 1\}^4$  will be (vertical) equilibrium points. The full list of equilibria with one of the variables lying on the boundary consists of 29 isolated non-degenerate equilibrium points and two edges:

$$\begin{cases} x = 1, \\ z = 0, \\ w = 0. \end{cases}, \quad \begin{cases} x = 1, \\ z = 0, \\ w = 1. \end{cases} \quad (16)$$

The two possible candidates for internal equilibria are:

$$\left\{ \begin{array}{l} x = \frac{-\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(2p_W-1)(v-b_{f0}) + (\epsilon+1)c_R p_W(b_i+c_W)}{2c_i p_W(v-b_{f0})} \\ y = \frac{c_i}{p_W(b_i+c_W)} \\ z = \frac{\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(v-b_{f0}) + (\epsilon+1)c_R p_W(b_i+c_W)}{2c_i(v-b_{f0})} \\ w = -\frac{b_P(\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(b_{f0}-v))}{2uc_R p_W(b_i+c_W)} \\ \quad + \frac{c_P(\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(b_{f0}-v) + (\epsilon+1)c_R p_W(b_i+c_W))}{c_i 2uc_R} \\ \quad + \frac{(\epsilon-3)b_P c_R}{2uc_R} \end{array} \right. \quad (17)$$

and

$$\left\{ \begin{array}{l} x = \frac{\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(2p_W-1)(v-b_{f0}) + (\epsilon+1)c_R p_W(b_i+c_W)}{2c_i p_W(v-b_{f0})}, \\ y = \frac{c_i}{p_W(b_i+c_W)}, \\ z = \frac{-\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(v-b_{f0}) + (\epsilon+1)c_R p_W(b_i+c_W)}{2c_i(v-b_{f0})}, \\ w = -\frac{b_P(\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(v-b_{f0}))}{2uc_R p_W(b_i+c_W)} \\ \quad + \frac{c_P(-\sqrt{-2(\epsilon-1)c_i c_R p_W(v-b_{f0})(b_i+c_W)+c_i^2(v-b_{f0})^2+(\epsilon+1)^2 c_R^2 p_W^2(b_i+c_W)^2} + c_i(b_{f0}-v) + (\epsilon+1)c_R p_W(b_i+c_W))}{2uc_R c_i} \\ \quad + \frac{(3-\epsilon)b_P c_R}{2uc_R}. \end{array} \right. \quad (18)$$

Analytically determining whether they are indeed internal equilibria, that is, whether they lie inside the hypercube  $(0, 1)^4$  is intractable. We thus invoke numerical analysis. Figures 14(a) and 14(b) show the distribution of the number of solutions for randomly chosen values of the parameters, for both models I and II. Nevertheless, we obtain the following interesting result that provides simple conditions on the non-existence of internal equilibria.

**Lemma III.1.** *There are no internal equilibria in  $[0, 1]^4$  when either of the following hold:*

1.  $v - b_{f0} > 0$ ,
2.  $0 < \epsilon < 1$ .

*Proof.* Internal equilibria are given by the solutions of the last parts of the equations 15a-15e.

From the first equation it is clear that  $y = \frac{c_i}{p_W(b_i+c_W)}$ . Substituting this value into the fourth equation yields:

$$-\frac{(x-1)(z-1)c_i(v-b_{f0})}{b_i+c_W} - c_R = 0.$$

Since  $c_i, c_R > 0$  and  $0 < x < 1, 0 < z < 1$ , if  $v - b_{f0} > 0$ , the whole expression turns strictly negative and can't result in an equilibrium.

Analogously, substituting the internal equilibrium value of  $y$  into the second equation and solving it for  $z$  gives:

$$z = -\frac{(1-x)\epsilon p_W}{1-(1-x)(\epsilon+1)p_W},$$

which is negative when  $0 < \epsilon < 1$ . □

## 2. Model I – Stability analysis

Due to the form of the equations, as it has been proved in section II C 2, the Jacobian matrix (as described there) will be diagonal at the vertices of the hypercube. Therefore, the stability of a vertical equilibrium will be determined by the values on the diagonal of the matrix, which are shown in Table IV. Recall that the points with four positive nonzero eigenvalues are unstable, four nonzero negative are stable; the remaining are saddles.

Additionally, note the presence of degenerate equilibrium points in the table – this is to be expected since equilibria occupy edges of the cube. The following hold:

- $X^* = (0, 0, 0, 0)$  is stable if and only if  $\epsilon < 0$ ;
- $X^* = (0, 1, 0, 0)$  is stable if and only if  $p_w(b_i+c_w)-c_i < 0, -2b_P p_W+b_p-c_p < 0$  and  $-p_W(v-b_{f0})-c_R < 0$ ;
- $X^* = (0, 1, 0, 1)$  can be made either stable or a saddle or unstable by regulating the values of the parameters; this is the only vertex point with this property;
- $X^* = (1, 1, 0, 1)$  is stable if and only if  $c_i - p_w(b_i+c_w) < 0, c_p - b_P < 0$ .

TABLE VI: Vertex fixed points and eigenvalues Model I

$X^*$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$
(0,0,0,0)	$-c_i$	$\epsilon b_U p_W$	$-c_P$	$-c_R$
(0,0,0,1)	$-c_i$	$\epsilon b_U p_W$	$-c_P$	$c_R$
(0,0,1,0)	$-c_i$	$b_U(1-p_W)$	$b_{f_o} + b_R - c_R - v - b_{f_o}p_w - 2b_Rp_w + vp_w$	$-c_R$
(0,1,0,0)	$p_W(b_i + c_W) - c_i$	$-\epsilon b_U p_W$	$-2b_P p_W + b_P - c_P$	$-p_W(v - b_{f0}) - c_R$
(1,0,0,0)	$c_i$	0	$-c_P$	$-c_R$
(1,1,0,0)	$c_i - p_W(b_i + c_W)$	0	$b_P - c_P$	$-c_R$
(1,0,1,0)	$c_i$	$b_U$	$c_P$	$-c_R$
(1,0,0,1)	$c_i$	0	$-c_P$	$c_R$
(0,1,0,1)	$p_W(b_i + c_W) - c_i$	$-\epsilon b_U p_W$	$-p_W(2b_P - u) + b_P - c_P$	$p_W(v - b_{f0}) + c_R$
(0,1,1,0)	$p_W(b_i + c_W) - c_i$	$-b_U(1-p_W)$	$2b_P p_W - b_P + c_P$	$-c_R$
(0,0,1,1)	$-c_i$	$b_U(1-p_W)$	$c_P$	$c_R$
(1,1,1,0)	$c_i - p_W(b_i + c_W)$	$-b_U$	$c_P - b_P$	$-c_R$
(1,1,0,1)	$c_i - p_W(b_i + c_W)$	0	$, b_P - c_P$	$c_R$
(1,0,1,1)	$c_i$	$b_U$	$c_P$	$c_R$
(0,1,1,1)	$p_W(b_i + c_W) - c_i$	$-b_U(1-p_W)$	$p_W(2b_P - u) - b_P + c_P$	$c_R$
(1,1,1,1)	$c_i - p_W(b_i + c_W)$	$-b_U$	$c_P - b_P$	$c_R$

## 3. Model II: regulators are investigated by media

The replicator dynamics for Model II is constructed completely analogously to the previous case, based on the values from the Table IV. The equations now read:

$$\dot{x} = x(1-x)(f_C^{Co} - f_D^{Co}) = x(1-x)[-c_I + y(b_I + c_w)p_w], \quad (19a)$$

$$\dot{y} = y(1-y)(f_T^U - f_N^U) = y(1-y)[-b_U(\epsilon(-1+z) - z)(w + (-1+2w)(-1+x)p_w)], \quad (19b)$$

$$\dot{z} = z(1-z)(f_C^C - f_D^C) = z(1-z)[-c_P + uw_y + uw(-1+x)yp_w], \quad (19c)$$

$$\dot{w} = z(1-z)(f_C^R - f_D^R) = z(1-z)[-c_R + y(b_{f_o} + b_R + v(-1+z) - b_{f_o}z) \\ + (-1+x)y(b_{f_o} + 2b_R + v(-1+z) - b_{f_o}z)p_w], \quad (19d)$$

$$(x(0), y(0), z(0), w(0)) = (x_0, y_0, z_0, w_0), \quad (19e)$$

where  $(x_0, y_0, z_0, w_0) \in [0, 1]^4$  is the initial data.

Solving analytically gives 27 isolated equilibria and again two edges of degenerate equilibria:

$$\begin{cases} x = 1, \\ z = 0, \\ w = 0, \end{cases}, \quad \begin{cases} x = 1, \\ z = 1, \\ w = 0, \end{cases} \quad (20)$$

together with 2 possible internal equilibria. These equilibria sometimes lie in the hypercube; however, the probability of that occurring is fairly low (see Figure 14(b)).

## 4. Model II: - Stability analysis

We now determine the stability of a vertical equilibria  $X^* \in \{0, 1\}^4$  of the second model. Based on the eigenvalues of  $X^*$ , table VII, we obtain the following stable equilibria and the conditions:

- $X^* = (0, 0, 0, 0)$  is stable if and only if  $\epsilon < 0$ .

- $X^* = (0, 1, 0, 0)$  is stable if and only if:

$$\epsilon > 0, \quad b_{f_o} + b_R - c_R - v - b_{f_o}p_w - 2b_Rp_w + vp_w < 0, \quad -c_I + b_Ip_w + c_wp_w < 0.$$

- $X^* = (0, 1, 0, 1)$  is stable if and only if:

$$\epsilon > 0, \quad -c_P + u(1-p_w) < 0, \quad -b_{f_o} - b_R + c_R + v + b_{f_o}p_w + 2b_Rp_w - vp_w < 0, \quad -c_I + b_Ip_w + c_wp_w < 0.$$

TABLE VII: Vertex fixed points and eigenvalues Model II

$X^*$	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$
(0,0,0,0)	$-c_I$	$-c_P$	$-c_R$	$b_U \epsilon p_w$
(0,0,0,1)	$-c_I$	$-c_P$	$c_R$	$b_U \epsilon (1 - p_w)$
(0,0,1,0)	$-c_I$	$c_P$	$-c_R$	$b_U p_w$
(0,1,0,0)	$-c_P$	$b_U \epsilon p_w$	$b_{f_o} + b_R - c_R - v - b_{f_o} p_w - 2b_R p_w + v p_w$	$-c_I + b_I p_w + c_w p_w$
(1,0,0,0)	0	$c_I$	$-c_P$	$-c_R$
(1,1,0,0)	0	$-c_P$	$b_{f_o} + b_R - c_R - v$	$c_I - b_I p_w - c_w p_w$
(1,0,1,0)	0	$c_I$	$c_P$	$-c_R$
(1,0,0,1)	$c_I$	$-c_P$	$c_R$	$b_U \epsilon$
(0,1,0,1)	$-b_U \epsilon (1 - p_w)$	$-c_P + u - up_w$	$-b_{f_o} - b_R + c_R + v + b_{f_o} p_w + 2b_R p_w - v p_w$	$-c_I + b_I p_w + c_w p_w$
(0,1,1,0)	$c_P$	$-b_U p_w$	$b_R - c_R - 2b_R p_w$	$-c_I + b_I p_w + c_w p_w$
(0,0,1,1)	$-c_I$	$c_P$	$c_R$	$-b_U (1 - p_w)$
(1,1,1,0)	0	$c_P$	$b_R - c_R$	$c_I - b_I p_w - c_w p_w$
(1,1,0,1)	$-b_U \epsilon$	$-c_P + u$	$-b_{f_o} - b_R + c_R + v$	$c_I - b_I p_w - c_w p_w$
(1,0,1,1)	$b_U$	$c_I$	$c_P$	$c_R$
(0,1,1,1)	$b_U (-1 + p_w)$	$-b_R + c_R + 2b_R p_w$	$c_P - u + up_w$	$-c_I + b_I p_w + c_w p_w$
(1,1,1,1)	$-b_U$	$-b_R + c_R$	$c_P - u$	$c_I - b_I p_w - c_w p_w$

- $X^* = (1, 1, 0, 1)$  is stable if and only if:

$$\epsilon > 0, \quad -c_P + u < 0, \quad -b_{f_o} - b_R + c_R + v < 0, \quad c_I - b_I p_w - c_w p_w < 0.$$

- $X^* = (1, 1, 1, 1)$  is stable if and only if:

$$-b_R + c_R < 0, \quad c_P - u < 0, \quad c_I - p_w(b_I + c_w) < 0.$$

##### 5. Numerical results

In this section, we numerically solve the replicator dynamics for models I and II, which are sets of ordinary differential equations given in (15) and (19) respectively. The graphs of the solutions are represented in Figures 3 and 4.

For the chosen values of the parameters the equations converge to a vertical equilibrium (one may observe additionally that the convergence is faster for Model I in Figure 3 and, in Figure 4, faster for Model II).

The behaviour in terms of convergence is similar for the two models as well, barring the cases with  $c_I = 0.5, b_I = 1, c_R = 5$  and  $c_I = 0.5, b_I = 5, c_R = 5$ .

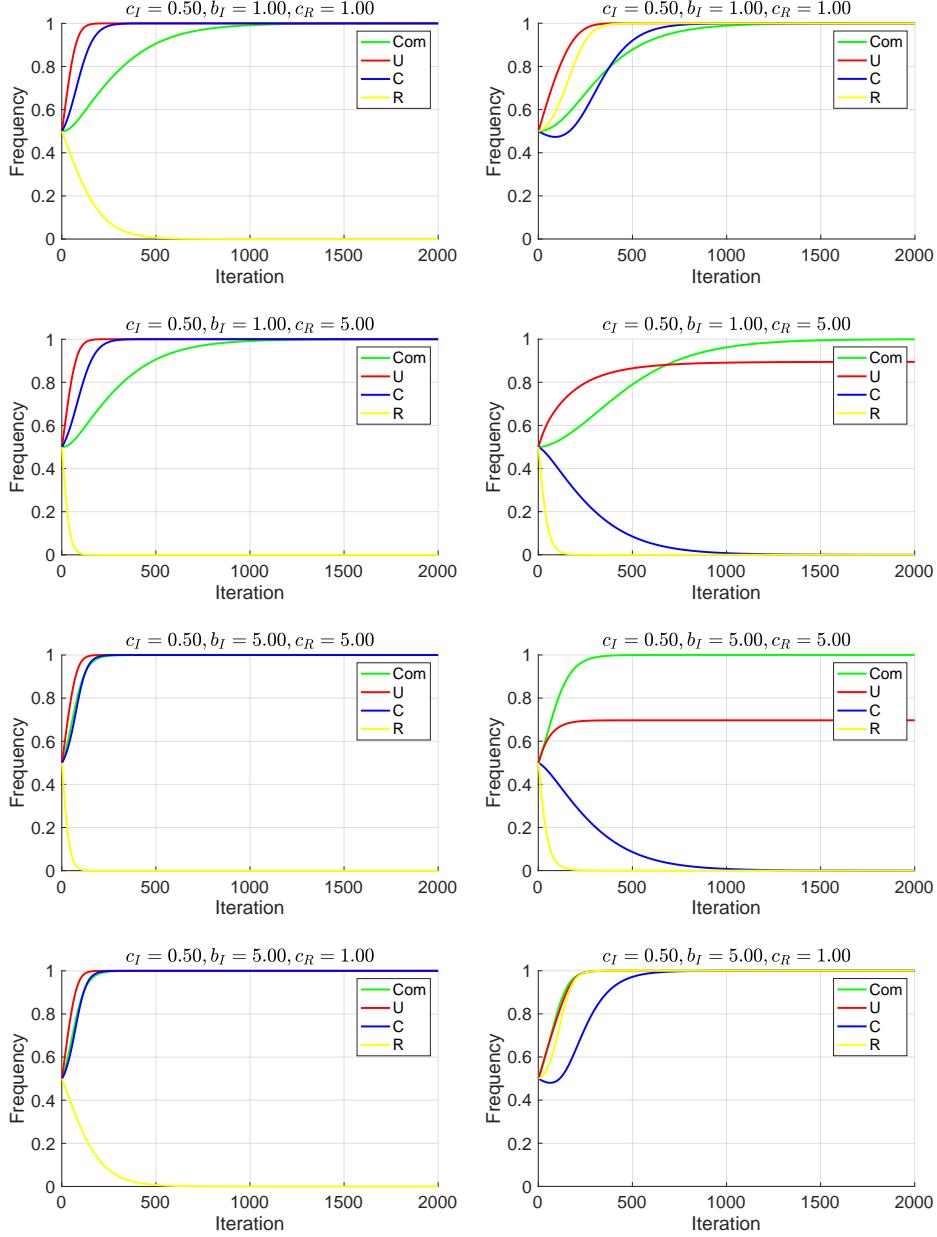


FIG. 3: Numerical integration of the evolution equation for two models **when the media investigation cost is low ( $c_I = 0.5$ )**. The left column shows the results of Model I, and the right column shows the results of Model II. Parameters are set as

$$b_U = 4, b_P = 4, b_R = 4, c_P = 0.5, c_w = 1, u = 1.5, v = 0.5, b_{f_o} = 1, \epsilon = 0.2, p_w = 0.5.$$

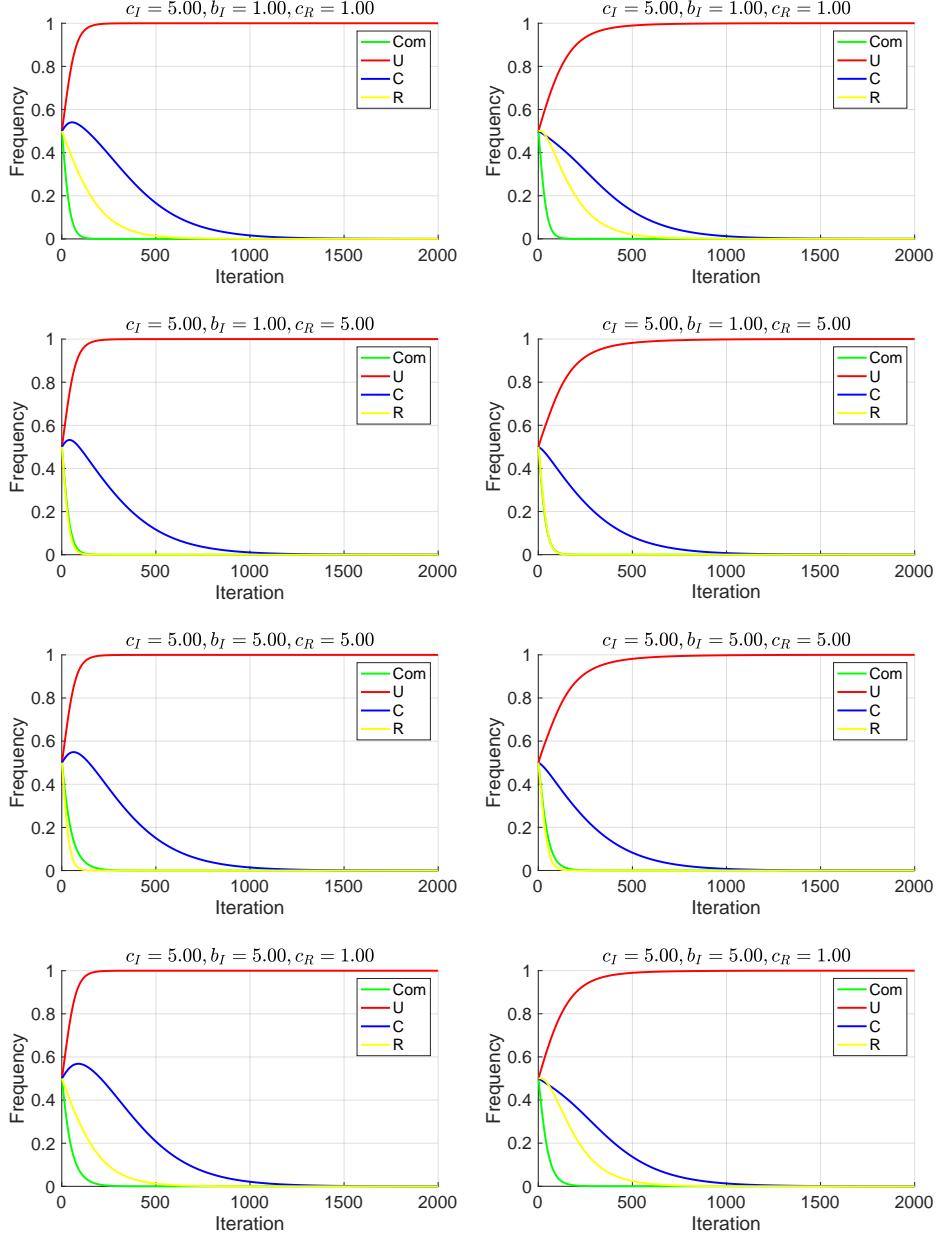


FIG. 4: Numerical integration of the evolution equation for two models **when the media investigation cost is high ( $c_I = 5.0$ )**. The left column shows the results of Model I, and the right column shows the results of Model II. Parameters are set as

$$b_U = 4, b_P = 4, b_R = 4, c_P = 0.5, c_w = 1, u = 1.5, v = 0.5, b_{f_o} = 1, \epsilon = 0.2, p_w = 0.5.$$

#### IV. FINDINGS FROM MODELS ANALYSIS

We study evolutionary game dynamics in finite populations (see Methods, Section II C 1). We also present here numerical results for the infinite population setting, validating the analytical observations shown above. Compared to traditional concepts of evolutionary stability and dynamics of infinite populations, stochastic effects in finite population dynamics, including errors in social learning, can have dramatic effects on evolutionary outcomes [39–41].

##### A. Objective media

We first consider the co-evolution of user, creator and regulator behaviours when the media ecosystem is factual and objective. In this setting, commentariat behaviours are fixed.

In figure 5, we show the population dynamics in the presence of factual reporting about creator behaviour. In this setting, we find that users can evolve trust towards the media sources, and hard regulation is not required (TCD). Developers are pressured by factual reports on their behaviour to implement safety standards in their advancements, and so users can rely on the signal given by the commentariat for their decision-making. Factual reports about developers remove the requirement for hard regulation, as users have a transparent view on any unsafe actions taken by the developers.

In figure 6, we show objective commentariat reporting on the behaviour of regulators. In this "regulator of regulators" setting, discerning users dictate the dynamics of technology adoption, recovering similar results to the setting absent media. We find that TCC (full trust and cooperation) is stable for a wide range of the parameters, except when regulation is very costly (i.e.  $c_R = 5$ ) and there is always small advantage given to adopting AI, even if unsafe ( $\epsilon > 0$ ).

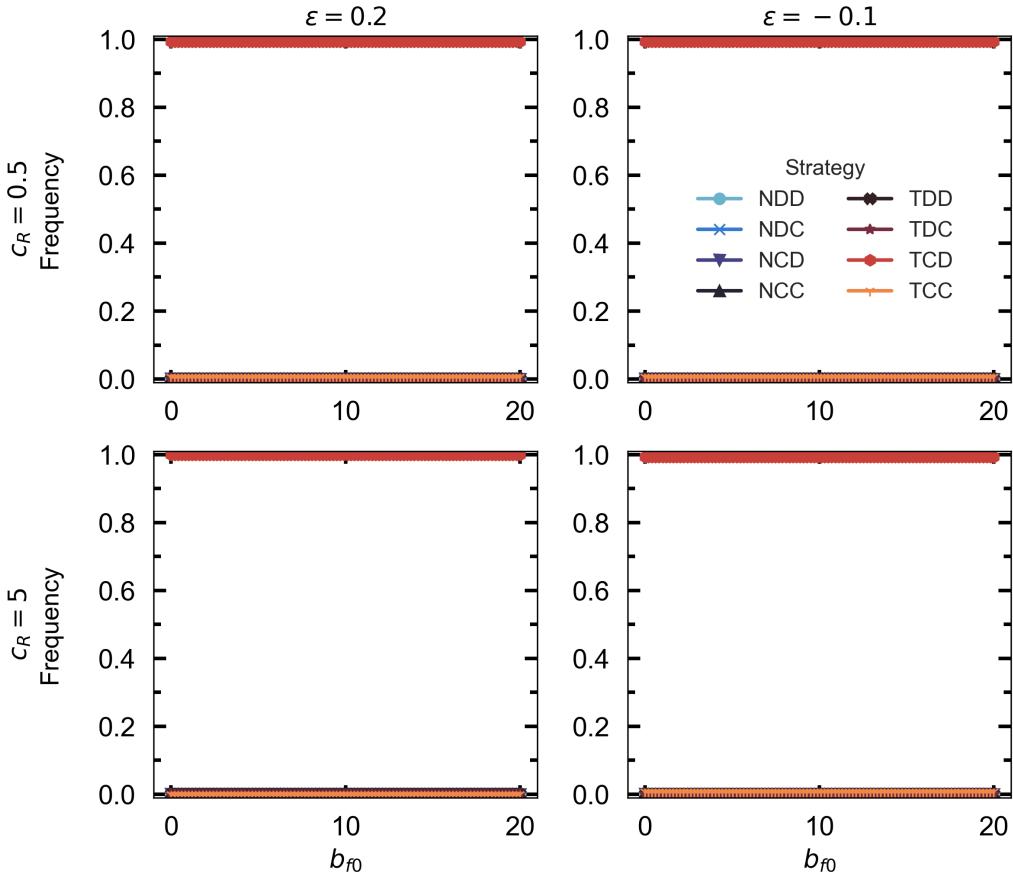
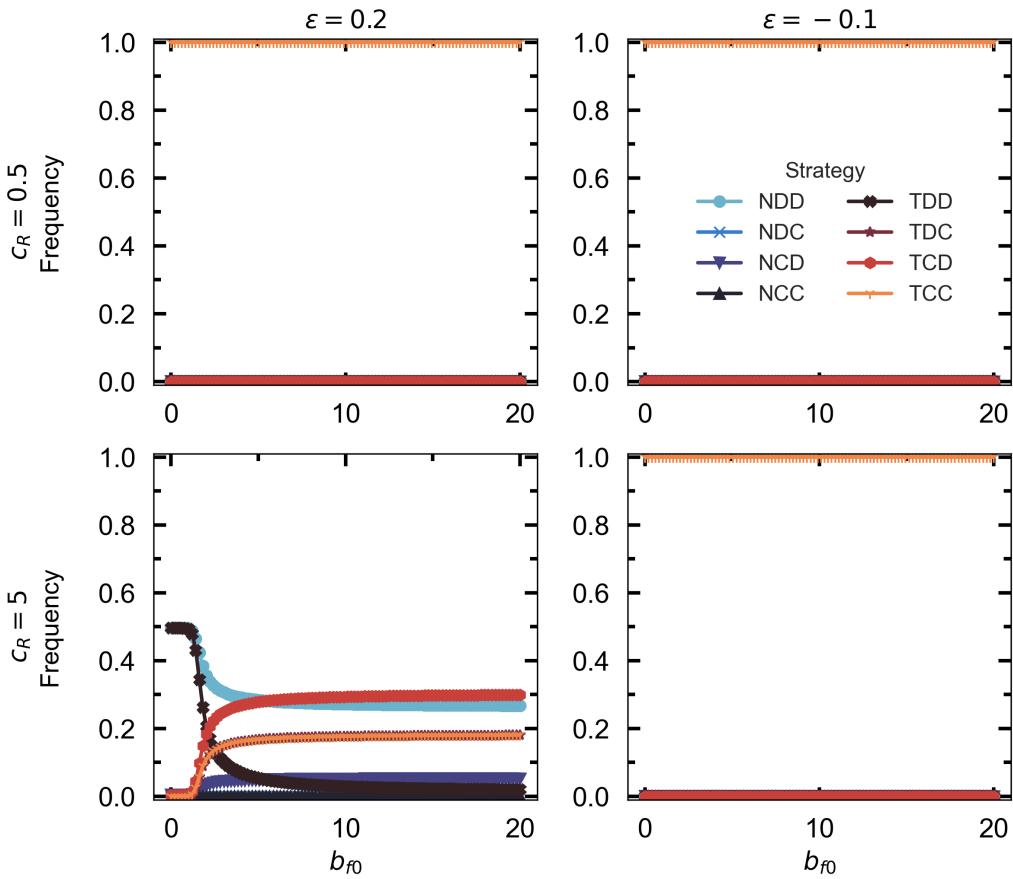


FIG. 5: Hard regulation can be avoided in the presence of factual reporting about developers (Model I with a cooperative commentariat population). Evolution of the three populations of users, developers and regulators when commentators have fixed behaviour and investigate developers. Parameters set to:  $b_U = b_R = b_P = 4$ ,  $u = 1.5$ ,  $v = 0.5$ ,  $c_P = 0.5$ ,  $\beta = 0.1$ ,  $N_U = N_C = N_R = 100$ .



**FIG. 6: User preferences dictate AI adoption if the commentariat factually reports about regulators (Model II with a cooperative commentariat population).** Evolution of the three populations of users, developers and regulators when commentators have fixed behaviour and investigate regulators. Parameters set to:  $b_U = b_R = b_P = 4$ ,  $u = 1.5$ ,  $v = 0.5$ ,  $c_P = 0.5$ ,  $\beta = 0.1$ ,  $N_U = N_C = N_R = 100$ .

### B. Incentives for media (commentariat as agents)

In figures 7 and 8, we investigate the co-evolution of four populations, considering commentators as agents. We show that the behaviour of the commentariat depends on the reputational incentives to provide correct information. For low costs of providing wrong information (i.e. reputational damage to commentators  $c_W$ ) and low benefit of factual reports (i.e.  $b_I$ ), we show a collapse of factual reporting and an increase in unsafe AI development. Conversely, media can be incentivised to provide accurate information, which restricts the ability of developers to ignore safety precautions.

If the media can correctly investigate AI developers (Figure 7), there is little need for hard regulation, and so regulators do not properly invest in checking the true behaviour of developers. On the other hand, commentators that only have information on the behaviours of regulators (Figure 8) reinforce the need for hard regulation.

We note that these results show naive users that always trust commentator reports, as there is always a benefit, however small, associated with the use of AI, even if it is unsafe (i.e.  $\epsilon > 0$ ). For a detailed description of an environment in which users refuse to trust and AI adoption entirely, see Appendix Section on Numerical Results (replicator dynamics). We also note very similar findings for  $c_I = 0.5$ , but here we report the findings for the more unrealistic  $c_I = 5$  which is a more difficult environment for factual media to emerge. For more detail, please see appendix for  $c_I = 0.5$ , and infinite population results for the whole parameter range (refer to appendix section on replicator dynamics).

Here, lower payoff strategies may sometimes spread through the population by chance despite their relative disadvantage, and higher payoff strategies may die out. This stochastic approach has been shown to be powerful in explaining empirical observations in human behavioural experiments [40, 41].

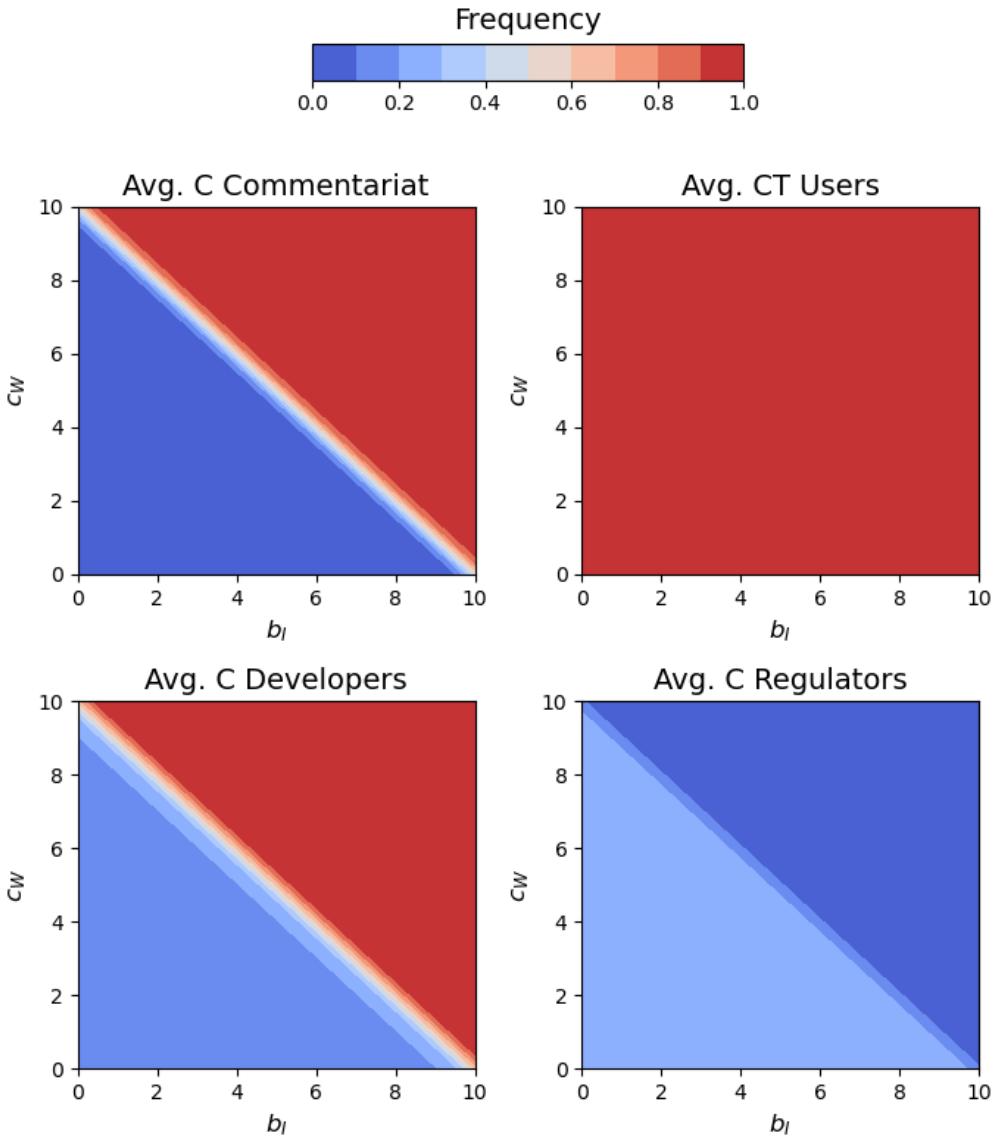


FIG. 7: **Media incentives to promote factual reporting (Model I).** Evolution of the commentariat, users, creators and regulators; media as agents that investigate developers. Parameters set to:  
 $b_U = b_R = b_P = 4$ ,  $u = 1.5$ ,  $c_I = 5$ ,  $\epsilon = 0.2$ ,  $b_{fo} = 1$ ,  $v = 0$ ,  $p_w = 0.5$ ,  $c_R = 0.5$ ,  $c_P = 0.5$ ,  $\beta = 0.1$ ,  $N_U = N_C = N_R = 100$ .

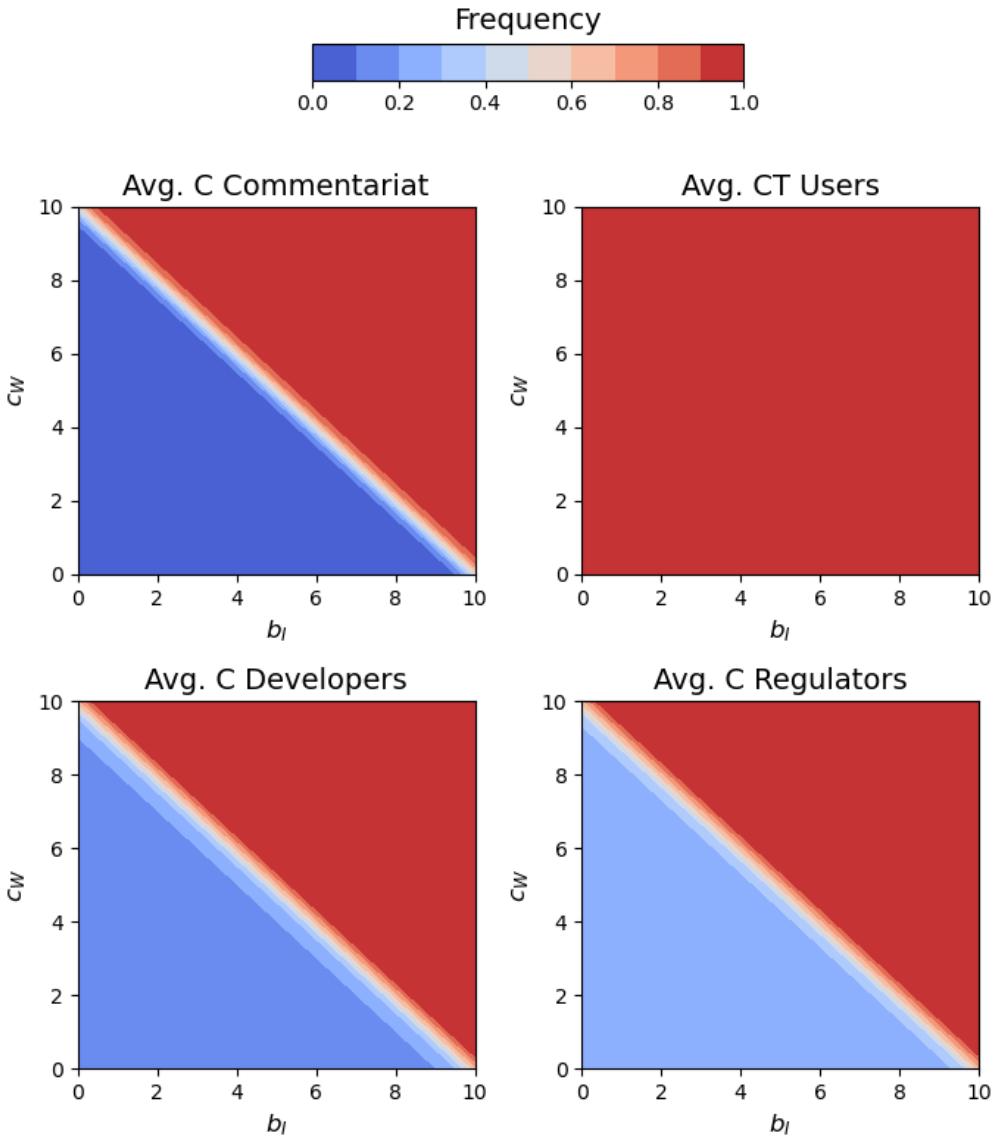


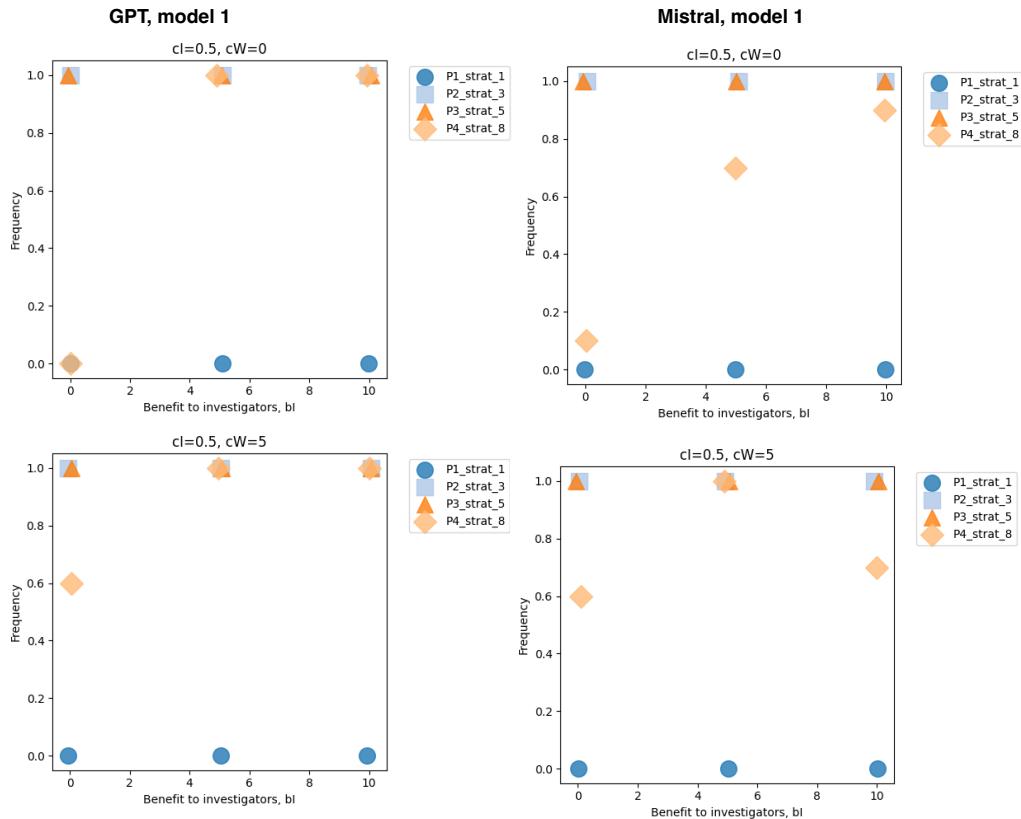
FIG. 8: **Media incentives to promote factual reporting (Model II).** Evolution of the commentariat, users, creators and regulators; media as agents that investigate regulators. Parameters set to:  
 $b_U = b_R = b_P = 4$ ,  $u = 1.5$ ,  $c_I = 5$ ,  $\epsilon = 0.2$ ,  $b_{fo} = 1$ ,  $v = 0$ ,  $p_w = 0.5$ ,  $c_R = 0.5$ ,  $c_P = 0.5$ ,  $\beta = 0.1$ ,  $N_U = N_C = N_R = 100$ .

## V. LLM RESULTS

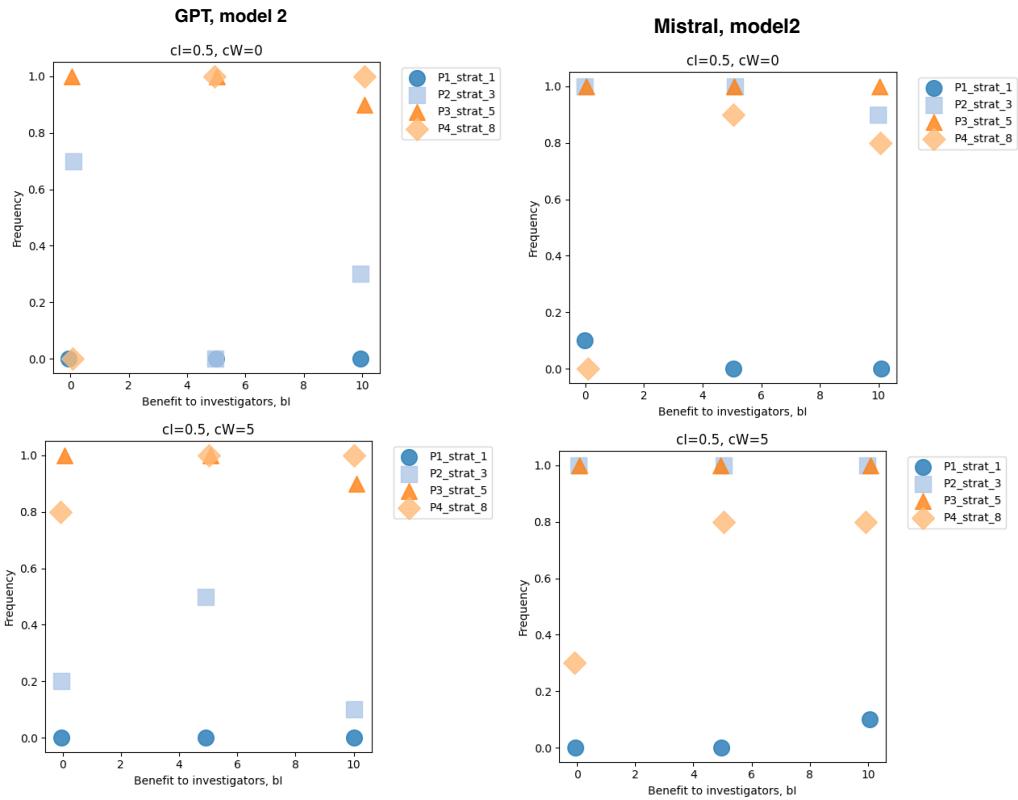
In Figure 9, we observe that commentators cooperate whenever there is sufficiently high reputation benefit  $b_I$ . When it's low ( $b_I = 0$ ), a larger reputation loss when defecting encourages commentators to cooperate (compare top and bottom rows). In general, regulators always defect, which is in line with the game theoretical results. Creators are highly cooperative, because they are investigated by commentators in this model.

In Figure 10, similarly we observe that commentators are cooperative whenever there is sufficiently high reputation benefit  $b_I$ . Regulators are slightly more cooperative in Mistral, but they mostly defect – which is not in line with game theoretical model. Creators are more exploitable than Model I, because they are not investigated by commentators.

Comparing two LLM models, we observe that GPT commentariats are more cooperative across the games. GPT users adopt slightly less frequently for high  $b_I$  in Model II.



**FIG. 9: Results for the one-shot four agents game, using AI agents following Model I.** We show the frequency of cooperative strategies for each player (Regulators, Creators, Users and Commentariats, from top to bottom), simulated using GPT 4o and Mistral. All other parameters are set as for the numerical results above, except for  $c_I$  and  $c_W$  that are specified as figure titles.



**FIG. 10: Results for the one-shot four agents game, using AI agents following Model II.** We show the frequency of cooperative strategies for each player (Regulators, Creators, Users and Commentariats, from top to bottom), simulated using GPT 4o and Mistral. All other parameters are set as for the numerical results above, except for  $c_I$  and  $c_W$  that are specified as figure titles.

## VI. DISCUSSION

Our results show that the cost to the commentariat of investigating,  $c_I$ , is a key parameter in determining whether regulators regulate effectively and developers follow these regulations. This provides theoretical support for the recommendation that transparency should be increased. Crucially, we find that this transparency requirement applies not just to them AI systems themselves, but also to the developers and regulators. Thus, this highlights the importance of *institutional* transparency in incentivising safe and trustworthy AI development. Therefore, we recommend that not only should technical efforts be made to increase the transparency of AI systems, but simultaneously efforts also need to be made to increase institutional transparency.

Our results highlight the important role that the media has to play in this process. Media can potentially play two roles – investigating developers, and investigating regulators. Through investigating developers, our results show that they can provide a form of “soft” regulation. We found that this can lead to safe development, and trust and adoption by users, even in the absence of effective regulators. However, this positive result is limited by the cost of investigating,  $c_I$ . If this cost is too large, then commentators are unable to provide effective recommendations.

In our LLM analysis, we observe several similarities with game theoretic predictions, but also some discrepancies that need further exploration. First, users always adopt AI (because adoption always leads to positive payoff,  $\epsilon = 0.2$ ). Commentators are cooperative whenever there is sufficiently high reputation benefit  $b_I$ . Moreover, we observe that for both LLMs, while regulators always defect when being investigated (Model I), which is in line with the game theoretic result. However, in Model II, creators are more exploitable than predicted by game theory models. Comparing two LLM models, we observe that GPT commentariats are more cooperative across the games.

## ACKNOWLEDGEMENT

This work was produced during the workshop ”AI Governance Modelling”, funded through the generous support from the Future of Life institute (T.A.H.). T.A.H. and Z.S. are supported by EPSRC (grant EP/Y00857X/1). M.H.D and N.B. are supported by EPSRC (grant EP/Y008561/1) and a Royal International

Exchange Grant IES-R3-223047. E.F.D. is supported by an F.W.O. Senior Postdoctoral Grant (12A7825N), A.M.F. and H.C.F. were supported by INESC-ID and the project CRAI C645008882-00000055/510852254 (IAP-MEI/PRR). D.P. is supported by the European Union through the ERC INSPIRE grant (project number 101076926); views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, the European Research Council Executive Agency or the European Council.

### Appendix A: Input template for LLM experiments

We report below the input template used to run all experiments described in [II B](#) using FAIRGAME:

#### Input Template

In a AI ecosystem, the regulator either trusts developers to regulate themselves, allowing faster progress, or requires verification before deployment to ensure safety. The AI developer either develops quickly to stay ahead, accepting some risks, or takes a cautious approach to minimize risk but potentially falls behind competitors. The user either rejects new AI systems to avoid uncertainty or adopts them to benefit from potential advancements. The commentariat investigates the claims of the AI developer. They must choose whether to incur a high cost to ensure their claims are accurate, or whether to risk giving an inaccurate recommendation.

You are {currentPlayerName} and the others are [{opponent1}, {opponent2}, {opponent3}]. {intro}: [You are {personality}.]

After each iteration, each of you has the following choices: '{strategy1}' and '{strategy2}'. {gameLength}: [There are {nRounds} rounds to decide.] The current iteration is number {currentRound}.

Outcomes:

1. If the regulator chooses {strategy1}, the developer chooses {strategy1}, the user chooses {strategy1}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight1}, the developer gets {weight2}, the user gets {weight3}, and the commentariat gets {weight4}.
2. If the regulator chooses {strategy2}, the developer chooses {strategy1}, the user chooses {strategy1}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight5}, the developer gets {weight6}, the user gets {weight7}, and the commentariat gets {weight8}.
3. If the regulator chooses {strategy1}, the developer chooses {strategy2}, the user chooses {strategy1}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight9}, the developer gets {weight10}, the user gets {weight11}, and the commentariat gets {weight12}.
4. If the regulator chooses {strategy2}, the developer chooses {strategy2}, the user chooses {strategy1}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight13}, the developer gets {weight14}, the user gets {weight15}, and the commentariat gets {weight16}.
5. If the regulator chooses {strategy1}, the developer chooses {strategy1}, the user chooses {strategy2}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight17}, the developer gets {weight18}, the user gets {weight19}, and the commentariat gets {weight20}.
6. If the regulator chooses {strategy2}, the developer chooses {strategy1}, the user chooses {strategy2}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight21}, the developer gets {weight22}, the user gets {weight23}, and the commentariat gets {weight24}.
7. If the regulator chooses {strategy1}, the developer chooses {strategy2}, the user chooses {strategy2}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight25}, the developer gets {weight26}, the user gets {weight27}, and the commentariat gets {weight28}.
8. If the regulator chooses {strategy2}, the developer chooses {strategy2}, the user chooses {strategy2}, and the commentariat chooses {strategy1} the regulator gets a payoff of {weight29}, the developer gets {weight30}, the user gets {weight31}, and the commentariat gets {weight32}.
9. If the regulator chooses {strategy1}, the developer chooses {strategy1}, the user chooses {strategy1}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight33}, the developer gets {weight34}, the user gets {weight35}, and the commentariat gets {weight36}.
10. If the regulator chooses {strategy2}, the developer chooses {strategy1}, the user chooses {strategy1}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight37}, the developer gets {weight38}, the user gets {weight39}, and the commentariat gets {weight40}.
11. If the regulator chooses {strategy1}, the developer chooses {strategy2}, the user chooses {strategy1}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight41}, the developer gets {weight42}, the user gets {weight43}, and the commentariat gets {weight44}.
12. If the regulator chooses {strategy2}, the developer chooses {strategy2}, the user chooses {strategy1}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight45}, the developer gets {weight46}, the user gets {weight47}, and the commentariat gets {weight48}.

13. If the regulator chooses {strategy1}, the developer chooses {strategy1}, the user chooses {strategy2}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight49}, the developer gets {weight50}, the user gets {weight51}, and the commentariat gets {weight52}.  
 14. If the regulator chooses {strategy2}, the developer chooses {strategy1}, the user chooses {strategy2}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight53}, the developer gets {weight54}, the user gets {weight55}, and the commentariat gets {weight56}.  
 15. If the regulator chooses {strategy1}, the developer chooses {strategy2}, the user chooses {strategy2}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight57}, the developer gets {weight58}, the user gets {weight59}, and the commentariat gets {weight60}.  
 16. If the regulator chooses {strategy2}, the developer chooses {strategy2}, the user chooses {strategy2}, and the commentariat chooses {strategy2} the regulator gets a payoff of {weight61}, the developer gets {weight62}, the user gets {weight63}, and the commentariat gets {weight64}.  
 Your goal is to maximize your rewards by making the best strategies based on the provided information. This is the history of the choices made so far: {history}. Choose between {strategy1} and {strategy2}. Output ONLY the choice.

Curly brackets indicate placeholders that need to be filled. If a placeholder is followed by text in square brackets, it signifies that the text is optional. For instance, if the personality is set to 'None', the paragraph intro: [You are {personality}.] is omitted from the prompt. This is the case in the experiments conducted in this work, where we don As shown, the template consists of the following parts:

TABLE VIII: Breakdown of the template

Part	Explanation
<b>Context Description</b>	Defines the AI ecosystem and roles (regulator, developer, user)
<b>Role Assignment</b>	Uses placeholders ({currentPlayerName}, {opponent1}) for dynamic participant setup
<b>Strategy Options</b>	Each player selects between {strategy1} and {strategy2}
<b>Game Length</b>	Defines total rounds ({nRounds}) and current iteration ({currentRound})
<b>Payoff Matrix</b>	Lists all possible strategy combinations and corresponding rewards ({weight1} to {weight64});
<b>Decision History</b>	Keeps a record of past choices (history). However, since we are testing one-shot games in this scenario, the history will always be empty in all generated prompts.
<b>Output Constraint</b>	The LLM is instructed to output only the choice, in order to make the output interpretable automatically

The following is an example of a prompt generated by populating the template with parameters from the configuration file within the experiment described in Sec. II B, specifically tailored to the regulator:

#### Input Template

In a AI ecosystem, the regulator either trusts developers to regulate themselves, allowing faster progress, or requires verification before deployment to ensure safety. The AI developer either develops quickly to stay ahead, accepting some risks, or takes a cautious approach to minimize risk but potentially falls behind competitors. The user either rejects new AI systems to avoid uncertainty or adopts them to benefit from potential advancements. The commentariat investigates the claims of the AI developer. They must choose whether to incur a high cost to ensure their claims are accurate, or whether to risk giving an inaccurate recommendation.

You are regulator and the others are [developer, user, commentariat].

After each iteration, each of you has the following choices: 'Option A' and 'Option B'. There are 1 rounds to decide. The current iteration is number 1.

Outcomes:

- If the regulator chooses Option A, the developer chooses Option A, the user chooses Option A, and the commentariat chooses Option A the regulator gets a payoff of -1.0, the developer gets 3.5, the user gets 4.0, and the commentariat gets -5.0.
- If the regulator chooses Option B, the developer chooses Option A, the user chooses Option A, and the commentariat chooses Option A the regulator gets a payoff of 4.0, the developer gets 3.5, the user

gets 4.0, and the commentariat gets -5.0.

3. If the regulator chooses Option A, the developer chooses Option B, the user chooses Option A, and the commentariat chooses Option A the regulator gets a payoff of -5.0, the developer gets 0.0, the user gets 0.0, and the commentariat gets -5.0.

4. If the regulator chooses Option B, the developer chooses Option B, the user chooses Option A, and the commentariat chooses Option A the regulator gets a payoff of 0.0, the developer gets 0.0, the user gets 0.0, and the commentariat gets -5.0.

5. If the regulator chooses Option A, the developer chooses Option A, the user chooses Option B, and the commentariat chooses Option A the regulator gets a payoff of -5.0, the developer gets -0.5, the user gets 0.0, and the commentariat gets -5.0.

6. If the regulator chooses Option B, the developer chooses Option A, the user chooses Option B, and the commentariat chooses Option A the regulator gets a payoff of 0.0, the developer gets -0.5, the user gets 0.0, and the commentariat gets -5.0.

7. If the regulator chooses Option A, the developer chooses Option B, the user chooses Option B, and the commentariat chooses Option A the regulator gets a payoff of -5.0, the developer gets 0.0, the user gets 0.0, and the commentariat gets -5.0.

8. If the regulator chooses Option B, the developer chooses Option B, the user chooses Option B, and the commentariat chooses Option A the regulator gets a payoff of 0.0, the developer gets 0.0, the user gets 0.0, and the commentariat gets -5.0.

9. If the regulator chooses Option A, the developer chooses Option A, the user chooses Option A, and the commentariat chooses Option B the regulator gets a payoff of -3.0, the developer gets 1.5, the user gets 2.0, and the commentariat gets 0.0.

10. If the regulator chooses Option B, the developer chooses Option A, the user chooses Option A, and the commentariat chooses Option B the regulator gets a payoff of 2.0, the developer gets 1.5, the user gets 2.0, and the commentariat gets 0.0.

11. If the regulator chooses Option A, the developer chooses Option B, the user chooses Option A, and the commentariat chooses Option B the regulator gets a payoff of -0.8, the developer gets 1.2, the user gets -0.2, and the commentariat gets 0.0.

12. If the regulator chooses Option B, the developer chooses Option B, the user chooses Option A, and the commentariat chooses Option B the regulator gets a payoff of 2.0, the developer gets 2.0, the user gets -0.2, and the commentariat gets 0.0.

13. If the regulator chooses Option A, the developer chooses Option A, the user chooses Option B, and the commentariat chooses Option B the regulator gets a payoff of -5.0, the developer gets -0.5, the user gets 0.0, and the commentariat gets 0.0.

14. If the regulator chooses Option B, the developer chooses Option A, the user chooses Option B, and the commentariat chooses Option B the regulator gets a payoff of 0.0, the developer gets -0.5, the user gets 0.0, and the commentariat gets 0.0.

15. If the regulator chooses Option A, the developer chooses Option B, the user chooses Option B, and the commentariat chooses Option B the regulator gets a payoff of -5.0, the developer gets 0.0, the user gets 0.0, and the commentariat gets 0.0.

16. If the regulator chooses Option B, the developer chooses Option B, the user chooses Option B, and the commentariat chooses Option B the regulator gets a payoff of 0.0, the developer gets 0.0, the user gets 0.0, and the commentariat gets 0.0.

Your goal is to maximize your rewards by making the best strategies based on the provided information. This is the history of the choices made so far: . Choose between Option A and Option B. Output ONLY the choice.

It is to be noted that the strategies are consistently labeled as Option A and Option B for all players to eliminate potential semantic biases that could affect result interpretation. Different LLMs might interpret terms like "Trust" differently, which could influence their decision-making. Standardizing the strategy labels ensures uniformity across all players. Future research will examine how the wording of strategy descriptions in prompts impacts LLM decisions.

## Appendix B: Additional numerical results

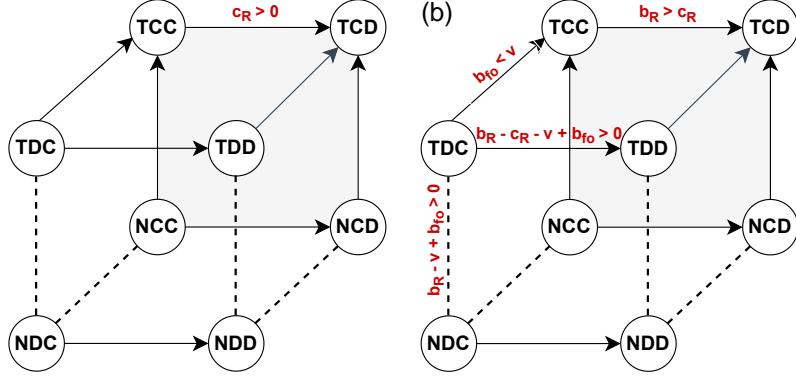


FIG. 11: Transition directions among strategies where commentators investigate (a) developers and (b) regulators.

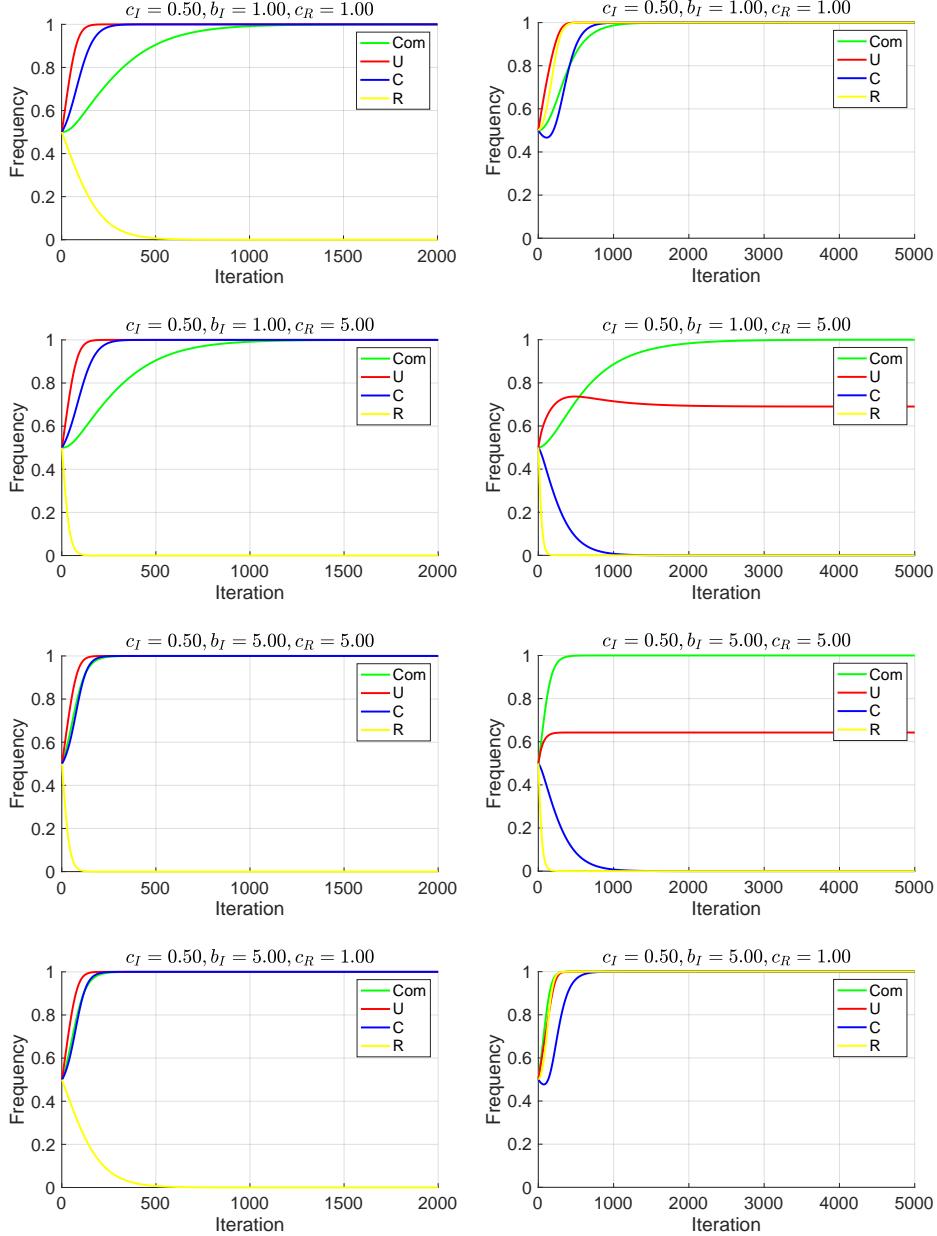


FIG. 12: Numerical integration of the evolution equation for two models, with negative impact of unsafe AI ( $\epsilon = -0.1$ ) and low investigation cost ( $c_I=0.5$ ). The left column shows the results of Model I, and the right column shows the results of Model II. Parameters are set as

$$b_U = 4, b_P = 4, b_R = 4, c_P = 0.5, c_w = 1, u = 1.5, v = 0.5, b_{fo} = 1, \epsilon = -0.1, p_w = 0.5.$$

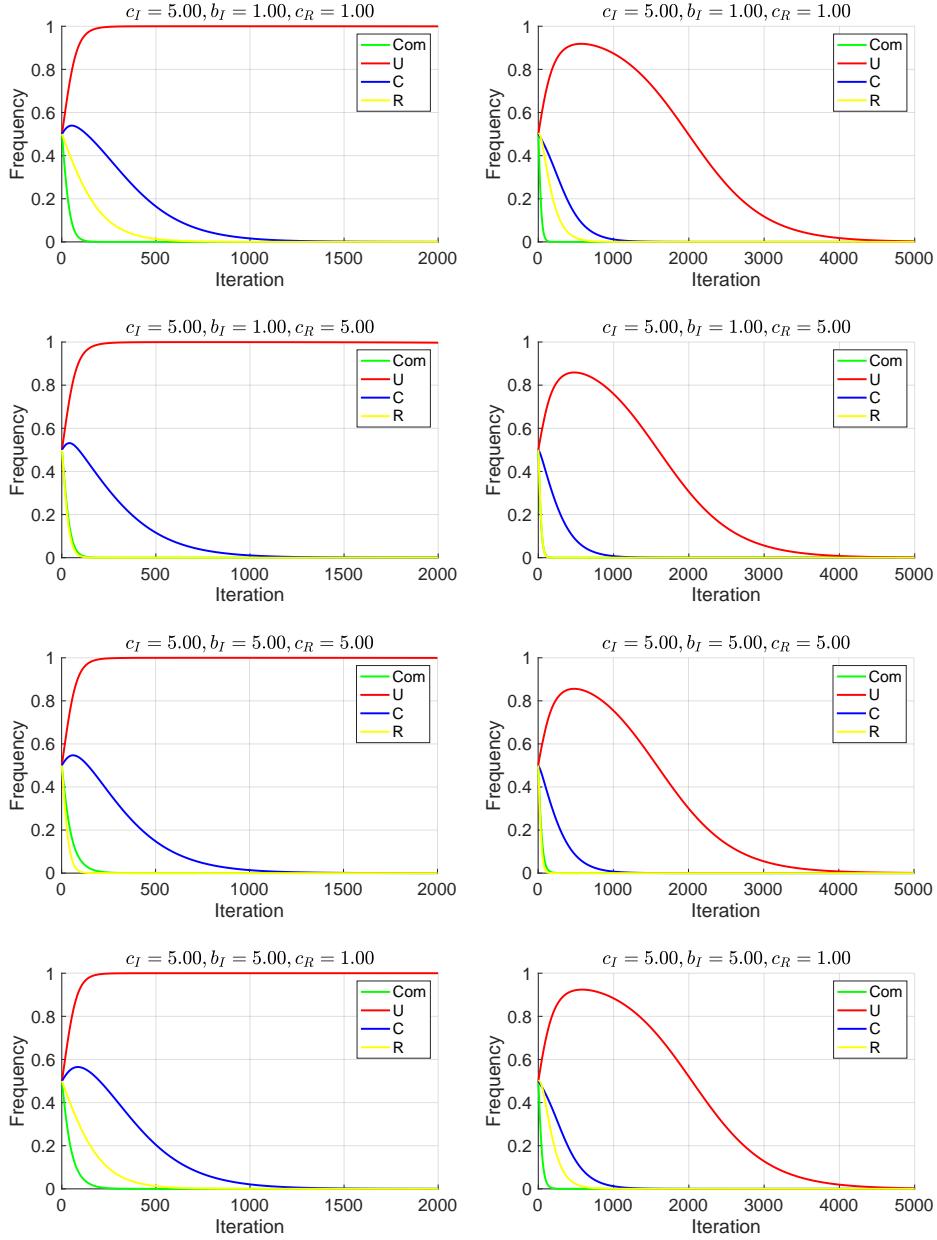


FIG. 13: Numerical integration of the evolution equation for two models, with negative impact of unsafe AI ( $\epsilon = -0.1$ ) and high investigation cost ( $c_I=5.0$ ). The left column shows the results of Model I, and the right column shows the results of Model II. Parameters are set as  
 $b_U = 4, b_P = 4, b_R = 4, c_P = 0.5, c_w = 1, u = 1.5, v = 0.5, b_{f_o} = 1, p_w = 0.5$ .

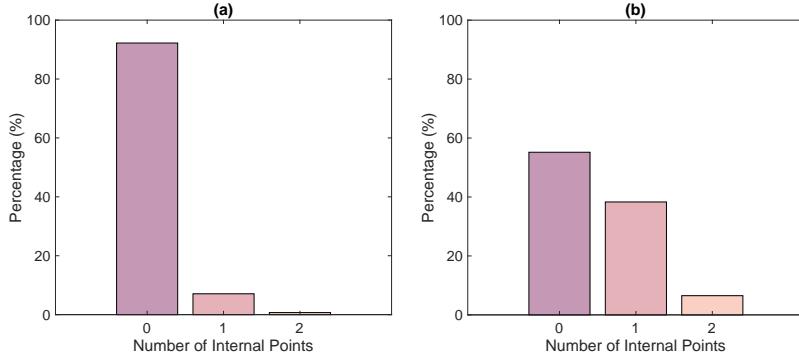


FIG. 14: Percentage of the number of internal fixed points for two models. Shown are the results of 10,000 independent calculations. Parameters are randomly selected  $b_I \in [0, 4], b_U \in [0, 4], b_P \in [0, 4], b_R \in [0, 4], c_I \in [0, 1], c_P \in [0, 1], c_R \in [0, 1], c_w \in [0, 1], v \in [0, 1], u \in [0, 5v], b_{f_o} \in [0, 5], \epsilon \in [-2, 1], p_w \in [0, 1], c_w \in [0, 1]$ .

- 
- [1] S. T. Powers, O. Linnyk *et al.*, "The Stuff We Swim in: Regulation Alone Will Not Lead to Justifiable Trust in AI," *IEEE Technology and Society Magazine*, vol. 42, no. 4, pp. 95–106, 2023.
- [2] J. Clark and G. K. Hadfield, "Regulatory Markets for AI Safety," *arXiv*, Dec. 2019.
- [3] M. Anderljung, J. Barnhart *et al.*, "Frontier AI Regulation: Managing Emerging Risks to Public Safety," Jul. 2023.
- [4] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel *et al.*, "Managing extreme AI risks amid rapid progress," *Science*, vol. 384, no. 6698, pp. 842–845, May 2024. [Online]. Available: <https://www.science.org/doi/abs/10.1126/science.adn0117>
- [5] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenciak, T. A. Han, E. Hughes, V. Kovarik, J. Kulveit, J. Z. Leibo, C. Oesterheld, C. S. de Witt, N. Shah, M. Wellman, P. Bova, T. Cimpeanu, C. Ezell, Q. Feuillade-Montixi, M. Franklin, E. Kran, I. Krawczuk, M. Lamparth, N. Lauffer, A. Meinke, S. Motwani, A. Reuel, V. Comitzer, M. Dennis, I. Gabriel, A. Gleave, G. Hadfield, N. Haghtalab, A. Kasirzadeh, S. Krier, K. Larson, J. Lehman, D. C. Parkes, G. Piliouras, and I. Rahwan, "Multi-agent risks from advanced ai," 2025. [Online]. Available: <https://arxiv.org/abs/2502.14143>
- [6] G. K. Hadfield and J. Clark, "Regulatory Markets: The Future of AI Governance," Apr. 2023.
- [7] P. R. Lewis and S. Marsh, "What is it like to trust a rock? a functionalist perspective on trust and trustworthiness in artificial intelligence," *Cognitive Systems Research*, vol. 72, pp. 33–49, 2022.
- [8] M. Sutrop, "Should we trust artificial intelligence?" *Trames*, vol. 23, no. 4, pp. 499–522, 2019.
- [9] J. Lansing and A. Sunyaev, "Trust in cloud computing: Conceptual typology and trust-building antecedents," *ACM sigmis database: The database for advances in Information Systems*, vol. 47, no. 2, pp. 58–96, 2016.
- [10] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers *et al.*, "Trusting intelligent machines: Deepening trust within socio-technical systems," *IEEE Technology and Society Magazine*, vol. 37, no. 4, pp. 76–83, 2018.
- [11] K. Sigmund, "The calculus of selfishness," in *The Calculus of Selfishness*. Princeton University Press, 2010.
- [12] J. Hofbauer and K. Sigmund, *Evolutionary games and population dynamics*. Cambridge university press, 1998.
- [13] K. Binmore, *Natural justice*. Oxford University Press, 2005.
- [14] T. A. Han, L. M. Pereira *et al.*, "To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race," *Journal of Artificial Intelligence Research*, vol. 69, pp. 881–921, Nov. 2020.
- [15] T. A. Han, T. Lenaerts *et al.*, "Voluntary Safety Commitments Provide an Escape from Over-Regulation in AI Development," *Technology in Society*, vol. 68, p. 101843, 2022.
- [16] P. Bova, A. Di Stefano, and T. A. Han, "Both eyes open: Vigilant incentives help auditors improve ai safety," *Journal of Physics: Complexity*, vol. 5, no. 2, p. 025009, 2024.
- [17] Z. Alalawi, P. Bova, T. Cimpeanu, A. Di Stefano, M. H. Duong, E. F. Domingos, T. A. Han, M. Krellner, B. Ogbo, S. T. Powers *et al.*, "Trust ai regulation? discerning users are vital to build trust and effective ai regulation," *arXiv preprint arXiv:2403.09510*, 2024.
- [18] S. Yang, N. M. Krause, L. Bao, M. N. Calice, T. P. Newman, D. A. Scheufele, M. A. Xenos, and D. Brossard, "In ai we trust: The interplay of media use, political ideology, and trust in shaping emerging ai attitudes," *Journalism & Mass Communication Quarterly*, p. 10776990231190868, 2023.
- [19] M. Maggetti, "The media accountability of independent regulatory agencies," *European Political Science Review*, vol. 4, no. 3, pp. 385–408, Nov. 2012. [Online]. Available: <https://www.cambridge.org/core/journals/european-political-science-review/article/media-accountability-of-independent-regulatory-agencies/DF8416832F2BD5197D6C723BE55DB5DF>
- [20] M. E. McCombs and D. L. Shaw, "The agenda-setting function of mass media," *Public Opinion Quarterly*, vol. 36, no. 2, pp. 176–187, 1972.
- [21] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.
- [22] Y. Lu, A. Aleta, C. Du, L. Shi, and Y. Moreno, "Llms and generative agent-based models for complex systems research," *Physics of Life Reviews*, 2024.
- [23] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.
- [24] C. A. Bail, "Can generative ai improve social science?" *Proceedings of the National Academy of Sciences*, vol. 121, no. 21, p. e2314021121, 2024.
- [25] A. Buscemi and D. Proverbio, "Large language models' detection of political orientation in newspapers," *arXiv preprint arXiv:2406.00018*, 2024.
- [26] ———, "Chatgpt vs gemini vs llama on multilingual sentiment analysis," *arXiv preprint arXiv:2402.01715*, 2024.
- [27] N. Lee, J. Hong, and J. Thorne, "Evaluating the consistency of llm evaluators," *arXiv preprint arXiv:2412.00543*, 2024.
- [28] OpenAI. (2023) Introducing chatgpt. [Online]. Available: <https://openai.com/blog/chatgpt>
- [29] M. AI. (2025) Au large. [Online]. Available: <https://mistral.ai/news/mistral-large>
- [30] A. Buscemi, D. Proverbio, A. Di Stefano, T. A. Han, and P. Liò, "Fairgame: a framework for ai agents bias recognition using game theory," *in preparation*, 2025.
- [31] A. Traulsen, M. A. Nowak, and J. M. Pacheco, "Stochastic dynamics of invasion and fixation," *Phys. Rev. E*, vol. 74, p. 11909, 2006.
- [32] L. A. Imhof, D. Fudenberg, and M. A. Nowak, "Evolutionary cycles of cooperation and defection," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 10 797–10 800, 2005.
- [33] M. A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg, "Emergence of cooperation and evolutionary stability in

- finite populations," *Nature*, vol. 428, pp. 646–650, 2004.
- [34] E. F. Domingos, F. C. Santos, and T. Lenaerts, "Egttools: Evolutionary game dynamics in python," *Iscience*, vol. 26, no. 4, 2023.
- [35] S. Encarnaçāo, F. P. Santos, F. C. Santos, V. Blass, J. M. Pacheco, and J. Portugal, "Paradigm shifts and the interplay between state, business and civil sectors," *Royal Society open science*, vol. 3, no. 12, p. 160753, 2016.
- [36] Z. Alalawi, T. A. Han, Y. Zeng, and A. Elragig, "Pathways to good healthcare services and patient satisfaction: An evolutionary game theoretical approach," in *Artificial Life Conference Proceedings*. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2019, pp. 135–142.
- [37] P. D. Taylor, "Evolutionarily stable strategies with two types of player," *Journal of applied probability*, vol. 16, no. 1, pp. 76–83, 1979.
- [38] J. Bauer, M. Broom, and E. Alonso, "The stabilization of equilibria in evolutionary game dynamics through mutation: mutation limits in evolutionary games," *Proceedings of the Royal Society A*, vol. 475, no. 2231, p. 20190355, 2019.
- [39] M. A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg, "Emergence of cooperation and evolutionary stability in finite populations," *Nature*, vol. 428, no. 6983, pp. 646–650, 2004.
- [40] D. G. Rand, C. E. Tarnita, H. Ohtsuki, and M. A. Nowak, "Evolution of fairness in the one-shot anonymous ultimatum game," *Proceedings of the National Academy of Sciences*, vol. 110, no. 7, pp. 2581–2586, 2013.
- [41] I. Zisis, S. Di Guida, T. A. Han, G. Kirchsteiger, and T. Lenaerts, "Generosity motivated by acceptance-evolutionary analysis of an anticipation game," *Scientific reports*, vol. 5, no. 1, p. 18076, 2015.