Media and responsible AI governance: A game-theoretic and LLM analysis

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

[1] School of Mathematics, University of Birmingham
[2] School Computing, Engineering and Digital Technologies, Teesside University
[3] Luxembourg Institute of Science and Technology
[4] School of Mathematics and Statistics, University of St Andrews
[5] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa
[6] Machine Learning Group, Université libre de Bruxelles
[7] AI Lab, Vrije Universiteit Brussel
[8] Division of Computing Science and Mathematics, University of Stirling
[9] Department of Industrial Engineering, University of Trento
[10] 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 shaping trustworthy AI development. Using evolutionary game theory and large language models (LLMs), we model the strategic interactions among these actors under different regulatory regimes. We explore 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 commentariat recommendations. We show that when high-quality media investigations are sufficiently rewarded and not excessively costly, they can either substitute for weak formal regulation (by directly scrutinising developers) or enhance strong regulation (by monitoring regulators), leading to higher levels of safe AI development and user trust. Our analysis identifies parameter regimes under which full cooperation by all actors—responsible developers, effective regulators, informed media, and discerning users—emerges as a stable equilibrium. Complementary LLM-based simulations broadly corroborate these patterns while also revealing behavioural deviations, particularly among regulator agents, that highlight how real-world decision-makers, and AI models themselves, may depart from idealised game-theoretical predictions. Overall, these results underline that effective AI governance crucially depends on aligning incentives and reducing the costs of rigorous, accurate media scrutiny.

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

I. INTRODUCTION

A common narrative poses that the route to trustworthy artificial intelligence (AI) is enhanced though transparency and regulation of AI systems [1–4]. 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 [5]. Governments, and the regulators appointed by them, are both self-interested agents that can be expected to make strategic decisions [4, 6–10]. 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 [11–14]. In the process of ensuring trustworthy, beneficial, and trusted AI, accounting for these complex aspects is key to designing effective regulatory mechanisms.

The rapid pace of AI development, coupled with the scarcity of empirical data on the effectiveness of various regulatory mechanisms, presents a significant challenge for robust governance. In such a dynamic environment, waiting for comprehensive data is often impractical as the landscape may change quickly. Evolutionary Game Theory (EGT) modelling [15, 16], grounded in widely accepted theories about how people and self-interested organisations behave [17], offers a powerful solution by providing theoretical predictions for how individuals and organizations might behave under diverse conditions [18–24]. EGT models, for instance, have demonstrated that effective regulation and safe AI development are contingent on users conditioning their trust in developers based on the perceived effectiveness of regulators [25]. Nevertheless, this left unanswered the question of *how* users would obtain information on the behaviour of developers and regulators. Recognising the increasing evidence that people's trust in AI developers is affected by media consumption [26, 27], this paper explores the role of media and other opinion leaders, which we hereafter also refer to as the commentariat, in providing such information, particularly through investigative journalism [28]. Crucially, we consider the fact that the media commentators can themselves be self-interested agents, who do not merely report objectively on developments, but can also act to shape the agenda [29].

In this work, we develop an EGT model to explore and quantify the role of media commentators as self-interested agents that influence user trust decisions (Figure 1). We consider two possible roles for the commentariat. First, they can directly 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 in incentivising developers to build safe AI systems, and influencing users to place appropriate trust these systems. In our model, we investigate two potential strategies for the commentariat, either investing resources in providing quality information (cooperate), or spending less effort in investigations (defect). Users can condition their decision based on the information that the commentariat provides, either trusting and adopting the AI system if recommended by the commentariat (conditional trust), or choosing never to trust and adopt it. Developers can decide to invest time and effort in creating safe AI systems (cooperate), or avoid the burden of doing this (defect).

Beyond traditional (evolutionary) game-theoretic approaches, which formally embed human decision-making processes within a proven analytical framework [15, 16, 24, 30], we also use recent AI methods to generate complementary models and predictions. In particular, Large Language Models (LLMs) have shown promise in enabling suitable replicas of human actions [31, 32], making them well-suited for simulating strategic games involving complex, non-linear, and multi-faceted agents. To this end, we developed a framework to experiment with the regulatory dynamics among the four actors considered in our model (commentariat, developers, regulators, and users). Within this game-theoretic setting, four LLM agents interact dynamically, each specifically prompted to represent one of these actors. The strategic dilemmas faced by each LLM agent are embedded through payoff matrices, in the spirit of evolutionary game theory. This setup allows for direct comparison with our EGT model predictions. Furthermore, recognising that different LLMs may produce contrasting outcomes in various tasks [33–35], we employ two distinct models: ChatGPT-4o from OpenAI's GPT family [36] and Mistral Large by Mistral [37].

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 safe development, effective regulation of AI, and appropriate trust by users?

2. Under what conditions will the commentariat be incentivised to 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?

Our analysis demonstrates the crucial role of the commentariat in achieving trustworthy AI, particularly through its capacity to provide information and complement regulatory efforts. Our findings highlight how effective regulation and the development of trustworthy AI emerge under specific conditions, emphasising the

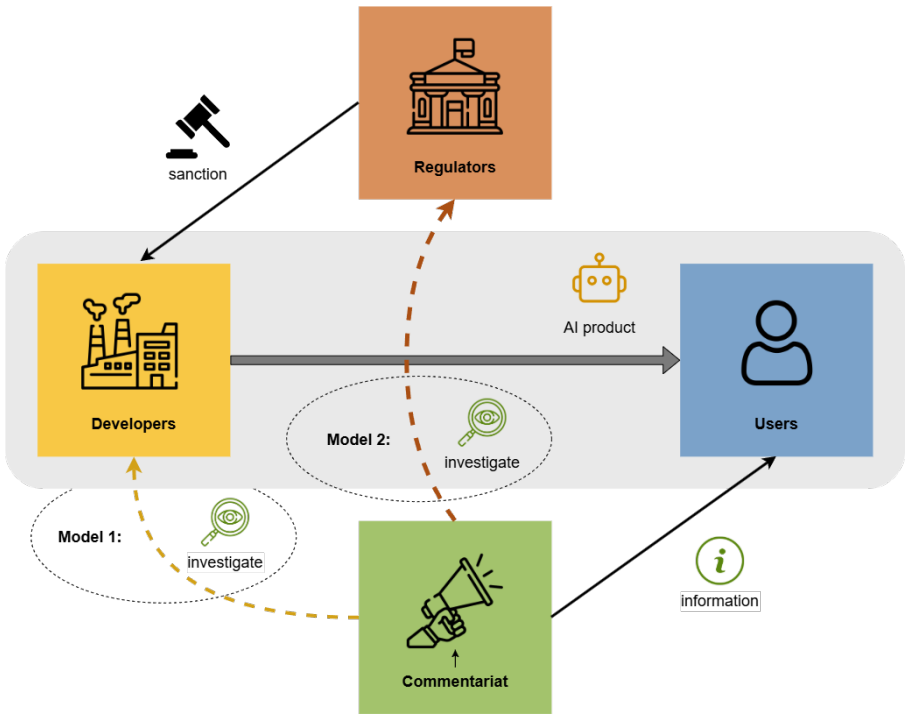

FIG. 1: **Visual description of the AI governance ecosystem**. In the centre lay the interactions between developers and users (grey panel). Developers may provide safe or unsafe AI, while users may adopt it or not. Regulators may intervene in this process by sanctioning unsafe AI developers. Commentariat informs users, in two different ways, corresponding to two models. In Model 1, commentariat investigates developers and provides information about which AI products are safe or unsafe. In Model 2, they instead investigate regulators, providing information on whether they are effectively sanctioning unsafe AI.

need for managing incentives and costs for quality reporting by the commentariat. We show that distinct dynamics emerge depending on whether the commentariat focuses its investigation on developers or regulators, and show how users' risk perception significantly influences system-wide trust. Moreover, while the LLM simulations largely corroborate our game-theoretic predictions, they also exhibit behavioural nuances and divergences, demonstrating how AI agents, prompted to represent these actors, can exhibit complex strategic responses.

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

II. MODELS AND METHODS

A. The four-actor model of AI governance: Commentariat, Users, Developers and Regulators

We start by constructing a model of an AI governance ecosystem, extending the three population model in [25] to capture the role of the 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 to adopt different strategies (see Table I and Supplementary Information Figure S1).

In particular, 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). The 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. A 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.

The individual payoff earned in any one encounter (also called a game) depends on the strategy of the participating individuals. A game takes place among one user, one developer, one commentator and one regulator. If the user follows the commentator's recommendations when the commentator has invested in providing an informed recommendation, and both the developer and the regulator cooperate (by complying

Role	Actions	Explanation
Commentators	C/D	Investigates and provides an *informed* recommendation (C), which means that it makes transparent the action of the developer/regulator, or provides an *uninformed* recommendation (D)
Users	CT/N	Either follows the commentator recommendations about whether to adopt and trust 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
ϵ	Fraction of user benefit when developers play D, where ε in [-∞,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 *incorrect* when they defect

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

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, a user that adopts AI is affected by unsafe AI, gaining a reduced or even negative benefit, denoted by $\epsilon \times b_U$, where $\epsilon \in [-\infty, 1]$. The parameter ϵ thus represents a (inverse) *risk factor* that users take when trusting and adopting the AI system.

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. If they do not comply with the regulations and develop AI unsafely, they may be punished by an amount u if they are found by a regulator to be defecting.

Regulators 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_{f_o} when they catch unsafe developers' behaviour (when users trust and adopt this unsafe 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 hand, defecting commentators do not pay this cost, but they can still earn the benefit b_I if the recommendation happens to be correct, which occurs with probability

TABLE III: **Payoff matrix for Model I, where the commentariat investigate developers**. Each row specifies one joint action profile of the four actors, i.e. Commentariat (Com), User, Developer (Dev), and Regulator (Reg), and the corresponding payoffs to each actor, given whether users condition trust on commentator recommendations and whether developers/regulators cooperate or defect (see Table I for detailed explanation of the actions).

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	$(1-p_w)b_I - p_w c_w$	$p_w \epsilon b_U$	$p_w(b_P - u)$	$p_w(b_R + b_{f_o} - v) - c_R$
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 Model II, where the commentariat investigate regulators**. Each row specifies one joint action profile of the four actors—Commentator (Com), User, Developer (Dev), and Regulator (Reg)—and the resulting payoffs to each actor, given users' conditional trust and whether developers/regulators cooperate or defect under commentariat scrutiny of regulatory behaviour (see Table I for detailed explanation of the actions).

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	$b_I - c_I$	ϵb_U	$b_P - u$	$b_R - c_R - v + b_{f_o}$
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	$(1-p_w)b_I - p_w c_w$	$(1-p_w)\epsilon b_U$	$(1\text{-}p_w)(b_P - u)$	$(b_R - c_R + b_{f_o} - v)(1-p_w)$ - $p_W c_R$
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

$1 - p_W$. If they defect and make the wrong recommendation, which occurs with probability p_W, they suffer a reputational cost of c_W. Table II summarise the key parameters of the two models.

We consider two versions of the model, differing in whom the commentariat monitors. In *Model I*, the commentariat investigates *developers*, and users condition their adoption decisions directly on these investigations

(see the payoff matrix in Table III). In *Model II*, the commentariat instead investigates *regulators*, and users condition adoption on whether the commentariat report that regulators are cooperating (see Table IV for the corresponding payoff matrix).

B. Evolutionary dynamics: finite and infinite population perspectives

1. Stochastic dynamics for finite populations

a. Payoff calculation. We consider four different well-mixed populations of Commentators (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 commentators that cooperate. Let y, z and ω 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 commentator, user, developer and regulator obtains in each game is respectively given by:

$$\begin{aligned} f^{Co}_{X\in\{C,D\}} &= yzwP^{Co}_{XTCC} + yz(1-w)P^{Co}_{XTCD} + y(1-z)wP^{Co}_{XTDC} + y(1-z)(1-w)P^{Co}_{XTDD} \\ &+ (1-y)zwP^{Co}_{XNCC} + (1-y)z(1-w)P^{Co}_{XNCD} + (1-y)(1-z)wP^{Co}_{XNDC} \\ &+ (1-y)(1-z)(1-w)P^{Co}_{XNDD}, \end{aligned} \tag{1}$$

$$\begin{aligned} f^{U}_{Y\in\{T,N\}} &= xzwP^{U}_{CYCC} + xz(1-w)P^{U}_{CYCD} + x(1-z)wP^{U}_{CYDC} + x(1-z)(1-w)P^{U}_{CYDD} \\ &+ (1-x)zwP^{U}_{DYCC} + (1-x)z(1-w)P^{U}_{DYCD} + (1-x)(1-z)wP^{U}_{DYDC} \\ &+ (1-x)(1-z)(1-w)P^{U}_{DYDD}, \end{aligned} \tag{2}$$

$$\begin{aligned} f^{C}_{Z\in\{C,D\}} &= xywP^{C}_{CTZC} + xy(1-w)P^{C}_{CTZD} + x(1-y)wP^{C}_{CNZC} + x(1-y)(1-w)P^{C}_{CNZD} \\ &+ (1-x)ywP^{C}_{DTZC} + (1-x)y(1-w)P^{C}_{DTZD} + (1-x)(1-y)wP^{C}_{DNZC} \\ &+ (1-x)(1-y)(1-w)P^{C}_{DNZD}, \end{aligned} \tag{3}$$

$$\begin{aligned} f^{R}_{W\in\{C,D\}} &= xyzP^{R}_{CTCW} + xy(1-z)P^{R}_{CTDW} + x(1-y)zP^{R}_{CNCW} + x(1-y)(1-z)P^{R}_{CNDW} \\ &+ (1-x)yzP^{R}_{DTCW} + (1-x)y(1-z)P^{R}_{DTDW} + (1-x)(1-y)zP^{R}_{DNCW} \\ &+ (1-x)(1-y)(1-z)P^{R}_{DNDW}. \end{aligned} \tag{4}$$

The term P^{PO}_{XYZW} describes the payoff a player in population PO obtained from a four-player game, defined by the payoff matrices in Tables III and IV. In each of the above formulas, each term on the right-hand side is the (average) payoff that the focal player obtains when interacting with the specific group of other players encoded in the subscripts. It is calculated as a multiple of the probability of encountering that specific group and the corresponding obtained payoff. The fitness (i.e. average payoff) is then computed using the payoff matrix constructed in the models (see Tables III and IV). The calculation of the fitness is a key step in applying evolutionary game theory, which translates a game-theoretical concept (the payoff) to a biological or cultural one (the fitness).

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 [38].

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

where $\beta \geq 0$ is the strength of selection. The fitnesses f_A and f_B are defined in Equations (1-4). $\beta = 0$ corresponds to neutral drift where imitation decisions are random, while for large $\beta \to \infty$, 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, while the transition probabilities are given by the fixation probability of a single mutant [39–41]. 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) [38]:

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

The fixation probability of a single mutant with a strategy A in a population of $(Z-1)$ agents using B is given by [38, 40]:

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

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

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

Fixation probability ρ_{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 (which is 4) representing the interaction of four players at a time [42, 43].

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 [44, 45]. 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, $i = 1,\ldots m$, 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 $x = (x_1,\ldots,x_m)$, which is the collection of all strategies in all populations.

For each $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n_i\}$, 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

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

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 $i = 1,\ldots,m$, 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 the models of AI governance and trust in the present paper), that is $n_i = 2$ for all $i = 1,\ldots,n_i$. Let η_i be the frequency of the first strategy in the population i, $i = 1,\ldots,m$ (thus $1-\eta_i$ will be the frequency of the second strategy in the population i), let $\eta = (\eta_1,\ldots,\eta_m)$. 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:

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

This is a system of m coupled non-linear 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 can be written in a general form as:

$$\dot{x} = x(1-x)F_1(X) =: \tilde{F}_1(X), \tag{10a}$$
$$\dot{y} = y(1-y)F_2(X) =: \tilde{F}_2(X), \tag{10b}$$
$$\dot{z} = z(1-z)F_3(X) =: \tilde{F}_3(X), \tag{10c}$$
$$\dot{w} = w(1-w)F_4(X) =: \tilde{F}_4(X), \tag{10d}$$

where $X = (x, y, z, w)$ is the vector of frequencies, and $F_i(X)$ $(i = 1, \ldots, 4)$ 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 equilibrium 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 do so numerically (see Results). Moreover, for analysing the stability of the vertical equilibria, the detailed derivation of the Jacobian at these equilibria and the resulting stability conditions is given in Supporting Information I.

C. LLM 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) [46]. 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 V: Parameters provided to FAIRGAME.

Parameter	**Value**
Number of agents	4
Names of the agents	regulator; developer; user; commentator
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 amongst four distinct agents, each fulfilling a designated role: regulator, developer, user, and commentator.

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, whilst 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 behaviour of the LLMs without external influences.

Lastly, the game runs for the specified number of rounds without a predetermined stopping condition. To address the inherent stochasticity associated with LLMs, we run each experiment 10 times, and we then analyse their mean results. The template used for all experiments is available in Supporting Information III.

III. EQUILIBRIUM ANALYSIS IN INFINITE POPULATIONS

This section presents the equilibrium analysis for infinite populations, beginning with a study of the equilibria and their stability for each model. We then provide a comparison between the models, supported by numerical results.

A. Model I: Developers are investigated by media

Using the general framework (9) and the payoff matrix given in Table III, the replicator dynamics for this model read (see detailed derivations in Supporting Information I-A):

$$\dot{x} = x(1-x)\Big[yp_W\,(b_I + c_W) - c_I\Big], \tag{11a}$$

$$\dot{y} = y(1-y)\Big[b_U\,((x-1)p_W((z-1)\epsilon + z) + z)\Big], \tag{11b}$$

$$\dot{z} = z(1-z)\Big[(x-1)yp_W\,(2b_P - uw) + yb_P - c_P\Big], \tag{11c}$$

$$\dot{w} = w(1-w)\Big[-(x-1)y(z-1)p_W\,(v - b_{\mathrm{f0}}) - c_R\Big], \tag{11d}$$

$$(x(0), y(0), z(0), w(0)) = (x_0, y_0, z_0, w_0), \tag{11e}$$

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

We investigate the existence and the number of equilibria in the $[0,1]^4$ hypercube of the above system. The 16 vertices, $(x, y, z, w) \in \{0,1\}^4$, are 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 (with $x = 1, z = 0$ and $w = 0$ or $w = 1$).

We also derived two potential internal equilibria (see Supporting Information I-A for the explicit expressions and derivation). We prove there that no internal equilibrium exists if either $(v - b_{f_0} > 0)$ or $(0 < \epsilon < 1)$. It means that whenever punishing unsafe developers is net costly for regulators, or users still obtain a positive benefit even from unsafe AI, the long-run dynamics collapse to boundary (vertex) states rather than mixed interior configurations.

We now determine the stability of a vertical equilibria $X^* \in \{0,1\}^4$. Due to the form of the equations, the Jacobian matrix (see details in Supporting Information I-B) 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, shown in Supporting Information Tables S1. Recall that the points with four positive non-zero eigenvalues are unstable, four non-zero negative are stable; the remaining are saddles.

There are four possible stable equilibria, and the conditions for their stability are given as follows

- $X^* = (0,0,0,0)$ is stable if and only if (iff) $\epsilon < 0$;
- $X^* = (0,1,0,0)$ is stable iff $p_W(b_I + c_W) - c_I < 0$, $-2b_P p_W + b_P - c_P < 0$ and $-p_W(v - b_{f_0}) - c_R < 0$;
- $X^* = (0,1,0,1)$ 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;
- $X^* = (1,1,1,0)$ is stable iff $c_I - p_W(b_I + c_W) < 0$ and $c_P - b_P < 0$.

Of these equilibria, $X^* = (1,1,1,0)$ is the only desirable one. It corresponds to a situation where users adopt a safely developed AI system, based on an informed recommendation from the commentariat, who have paid the cost of thoroughly investigating development of the AI system. This ideal equilibrium outcome, where users adopt safely developed AI based on informed commentariat recommendations, depends on several factors. The commentariat's cost for thorough investigation (c_I) must be low enough, specifically bounded by the total of their reputational benefit for a correct recommendation (b_I) and their reputational cost for an incorrect recommendation (c_W), weighted by the probability of an incorrect recommendation if they do not investigate (p_W). For developers, the cost of creating safe AI (c_P) must be less than the benefits from user adoption (b_P). In this Model I, if the commentariat effectively provides trustworthy signals about developers' safety, regulators are not incentivised to act, rendering their role redundant. However, it is often unrealistic for media to conduct such extensive investigations into complex AI systems (e.g. due to lack of technical capabilities or infrastructures), implying that c_I can be too high for this scenario to be practical.

Moving to *Model II*, we analyse the scenario where the commentariat reports on regulators. This is motivated by the idea that regulators can monitor AI development more cost-effectively than journalists. Regulatory mechanisms like auditing and transparency requirements would simplify commentariat investigations into regulators' actions, making it less costly than directly scrutinising AI developers.

B. Model II: Regulators are investigated by media

The replicator dynamics for Model II is constructed from the payoff matrix in Table IV. With this payoff matrix, the equations now read:

$$\dot{x} = x(1-x)(f_C^{Co} - f_D^{Co}) = x(1-x)\left[-c_I + y(b_I + c_W)p_W\right], \tag{12a}$$

$$\dot{y} = y(1-y)(f_T^U - f_N^U) = y(1-y)\left[-b_U(\epsilon(-1+z) - z)(w + (-1+2w)(-1+x)p_W)\right], \tag{12b}$$

$$\dot{z} = z(1-z)(f_C^C - f_D^C) = z(1-z)\left[-c_P + uwy + uw(-1+x)yp_W\right], \tag{12c}$$

$$\dot{w} = w(1-w)(f_C^R - f_D^R) = w(1-w)\big[-c_R + y(b_{f_o} + b_R + v(-1+z) - b_{f_o}z) \tag{12d}$$

$$+ (-1+x)y(b_{f_o} + 2b_R + v(-1+z) - b_{f_o}z)p_W\big], \tag{12e}$$

$$(x(0), y(0), z(0), w(0)) = (x_0, y_0, z_0, w_0), \tag{12f}$$

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

Solving analytically gives 27 isolated equilibria, two edges of degenerate equilibria (with $x = 1, z = 0, w = 0$ or $x = 1, z = 1, w = 0$), and two possible internal equilibria. The probability of internal equilibria occurring is rather low (see Supporting Information Figure S2) and we will focus on stability analysis for vertical equilibria.

Indeed, based on the eigenvalues of X^* (see Table S2 in Supporting Information), we obtain the following five possible stable equilibria and the conditions:

- $X^* = (0,0,0,0)$ is stable iff $\epsilon < 0$.
- $X^* = (0,1,0,0)$ is stable iff:

$$\epsilon > 0,\ b_{f_o} + b_R - c_R - v - b_{f_o}p_W - 2b_Rp_W + vp_W < 0,\ -c_I + b_Ip_W + c_Wp_W < 0.$$

- $X^* = (0,1,0,1)$ is stable iff:

$$\epsilon > 0,\ -c_P + u(1-p_W) < 0,\ -b_{f_o} - b_R + c_R + v + b_{f_o}p_W + 2b_Rp_W - vp_W < 0,\ -c_I + b_Ip_W + c_Wp_W < 0.$$

- $X^* = (1,1,0,1)$ is stable iff:

$$\epsilon > 0,\ -c_P + u < 0,\ -b_{f_o} - b_R + c_R + v < 0,\ c_I - b_Ip_W - c_Wp_W < 0.$$

- $X^* = (1,1,1,1)$ is stable iff:

$$-b_R + c_R < 0,\ c_P - u < 0,\ c_I - p_W(b_I + c_W) < 0.$$

The (only) desirable equilibrium, $X^* = (1,1,1,1)$, in this model is where regulators effectively regulate, media monitors regulators, developers comply, and users trust AI based on media reports, requires specific conditions for achieving its stability. Regulators must incur lower regulation costs (c_R) than benefits they receive from user adoption (b_R). Developers are incentivised to create safe AI if its cost (c_P) is less than regulatory punishment for non-compliance (u), a condition dependent on regulatory enforcement (unlike Model I). The commentariat's cost for thorough investigation (c_I) must be less than its potential reputational benefits and avoided costs from accurate reporting ($p_W(b_I + c_W)$), similar to the condition in Model I.

C. Model I vs. Model II

Comparing Model I and Model II provides insights on distinct pathways to effective AI governance, with numerical results visually confirming our analytical predictions above regarding these stable states. Analytically, in Model I (media directly investigates developers), the most desirable of the stable states, $X^* = (1,1,1,0)$, sees regulators become redundant, as the commentariat's effective reporting incentivises developers to produce safe AI. This is contingent on the media's capacity for cost-effective, high-quality investigation ($c_I < p_W(b_I + c_W)$) and developers finding building safe AI cheaper than the benefits they get from subsequent user adoption

($c_P < b_P$). Numerical simulations for Model I (e.g., left column of Figure 2, low c_I) confirm this: the frequency of cooperation in the commentator (Com) and conditional trust in the user (U) populations converges to 1, developers/creators (C) cooperate, while regulator (R) cooperation often converges to 0, showing their redundancy in this context.

On the other hand, Model II, focusing on media investigation of regulators, analytically predicts a more comprehensive desirable equilibrium, $X^* = (1, 1, 1, 1)$, where all four actors cooperate. This state requires active and cost-effective regulation ($c_R < b_R$), and developers' compliance is driven by the threat of regulatory punishment ($c_P < u$), a mechanism directly dependent on robust regulation. The numerical integrations for Model II (e.g., right column of Figure 2, low c_I) strongly support this, showing all four actors converging to a cooperative frequency of 1.

Moreover, the condition for the commentariat's effective investigation ($c_I < p_W(b_I + c_W)$) remains consistent across both models, underscoring the universal need for incentives that ensure accurate media reporting. When c_I is high (see Supporting Information Figure S5), the commentariat's cooperation frequency drops significantly in both models, which in turn hinders the cooperative outcomes for other actors, highlighting the sensitivity of the entire ecosystem to the media's cost of investigation. Although Model I can exhibit faster convergence to certain equilibria, Model II is generally considered more practical, as the cost of media investigating regulatory processes (which can be streamlined by auditing mechanisms) is typically lower than directly assessing complex AI systems, leading to a more practical and robust governance approach.

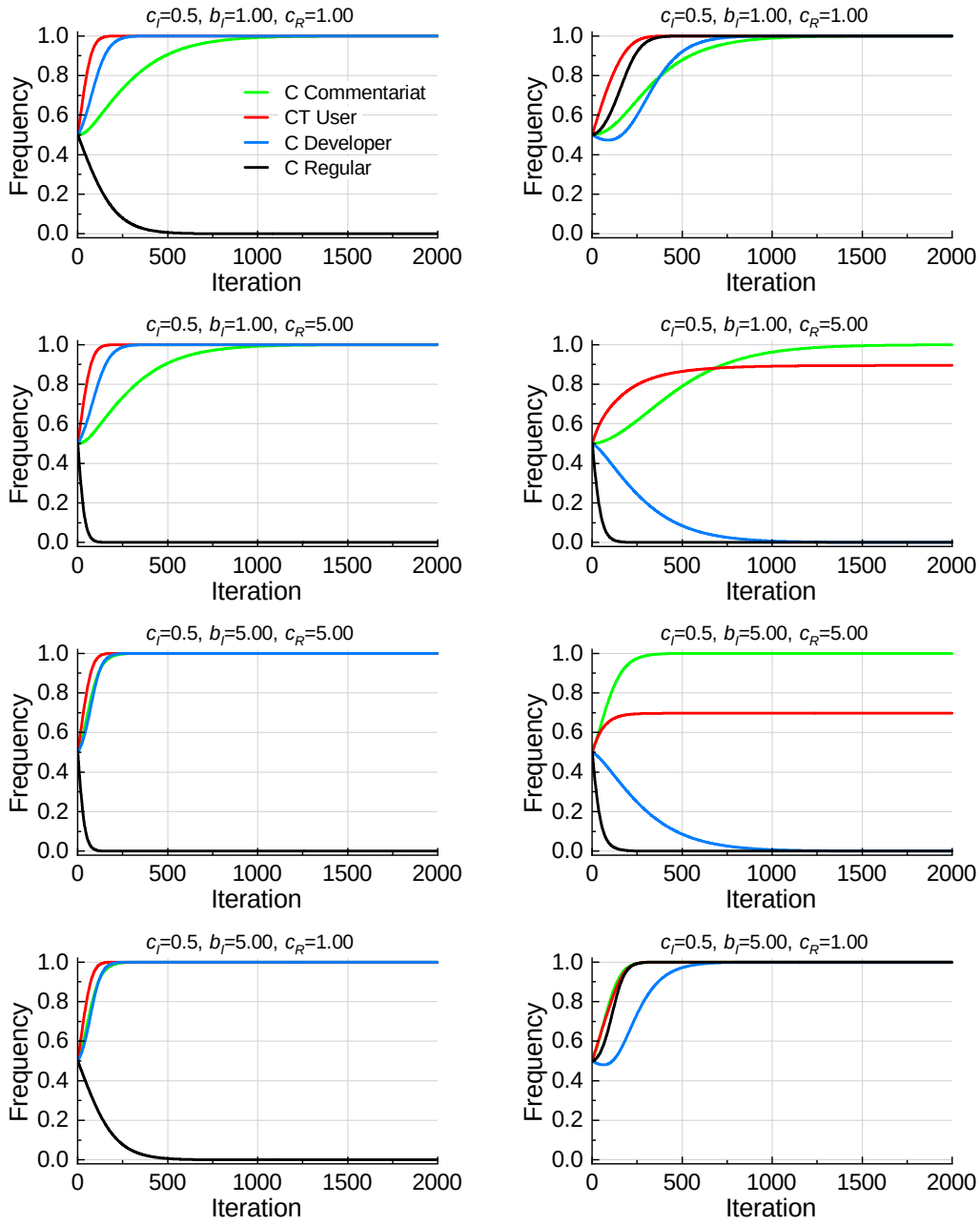

FIG. 2: When the cost for the media to conduct effective investigations is low ($c_I = 0.5$), the AI governance ecosystem achieves high cooperation, with distinct roles for regulators depending on the media's investigative focus. Shown are the frequencies of commentariat cooperation (C Commentariat), user conditional trust (CT User), developer cooperation (C Developer), and regulator cooperation (C Regular). The left column shows results for Model I, where the media investigates developers, showing that regulators become redundant as other actors achieve cooperation. The right column depicts Model II, where the media investigates how effectively regulations are implemented by regulators, leading to full cooperation across all four actors. Parameters: $b_U = 4, b_P = 4, b_R = 4, c_P = 0.5, c_W = 1, u = 1.5, v = 0.5, b_{f_o} = 1, \epsilon = 0.2, p_W = 0.5$.

IV. FINITE POPULATION ANALYSIS

We now turn to study evolutionary game dynamics in finite populations (see Methods, Section II B 1). 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 [47–49]. 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 [48, 49], making it an important mode of analysis for our models.

Figure 3 shows the long-term frequencies of strategies in all four populations when $\epsilon = 0.2$ – indicating low-

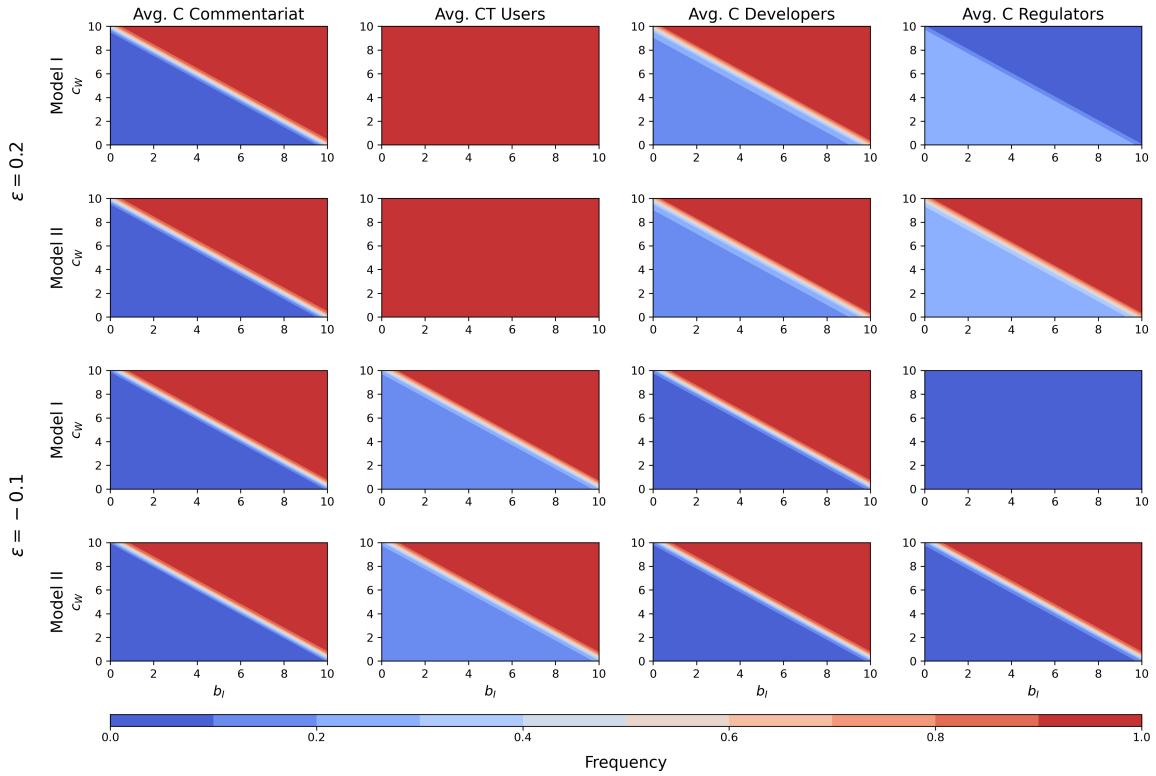

FIG. 3: High reputational incentives for accurate reporting (high b_I and c_W) sustain commentariat cooperation and, in turn, promote trust by users and cooperation by developers and regulators, whereas weaker incentives lead to widespread defection especially when AI is high-risk (negative $\epsilon = -0.1$). Shown are the frequencies of commentariat cooperation (C), user conditional trust (CT), developer cooperation (C), and regulator cooperation (C) are shown for ($\epsilon = 0.2$) and ($\epsilon = -0.1$), under Model I and Model II. Parameters set to: $b_U = b_R = b_P = 4$, $u = 1.5$, $c_I = 5$, $b_{f_o} = 1$, $v = 0$, $p_W = 0.5$, $c_R = 0.5$, $c_P = 0.5$, $\beta = 0.1$, $N_{Co} = N_U = N_C = N_R = 100$.

risk AI – and $\epsilon = -0.1$ – indicating high-risk AI – for Model I and Model II. Commentariat cooperation (and thus system-wide cooperation) is highly dependent on reputational incentives. As seen in Figure 3, increasing the reputational benefit (b_I) for correct recommendations and the reputational cost (c_W) for incorrect ones incentivises high-quality media investigations. This is numerically in line the analytical condition from Section III that $b_I + c_W$ must be sufficiently high compared to c_I for commentariat cooperation. This trend holds irrespective of whether the media targets developers (Model I) or regulators (Model II), aligning with the infinite population analysis. Moreover, we observe that, for sufficiently large b_I and c_W, regulators defect in Model I but cooperate in Model II. This is also in line with the infinite population analysis. Also, comparing between positive and negative ϵ, while analytical results show that the existence of cooperative equilibria does not strictly depend on ϵ, negative ϵ (bottom two rows in Figure 3) in finite populations makes users more discerning (the frequency of conditional trust, **CT**, increases). For sufficiently low b_I and c_W, users switch to the **N** (never adopt) strategy as commentators tend to defect.

In Supporting Information Figure S6, we further illustrate that as the cost of accurate investigations (c_I) rises, commentariat cooperation—and subsequently, that of regulators and developers—declines in both models. Also, comparing between positive and negative ϵ, while analytical results show that the existence of cooperative equilibria does not strictly depend on ϵ, we observe that as commentariat reliability wanes due to rising c_I, users switch to the **N** (never adopt) strategy, as there is no benefit in trusting unreliable recommendations for inherently risky AI. Moreover, we observe a high level of cooperation in the regulator population in Model II, while it is almost absent in Model I.

In addition, Figure 4 reveals how the intensity of selection (β) modulates these patterns. When media investigations are relatively cheap (low c_I), cooperation is robust across a wide range of β in both models, and particularly strong in Model II, regardless of whether AI is low- or high-risk ($\epsilon = 0.2$ vs. $\epsilon = -0.1$). However, when c_I is high and AI is high-risk, increasing β—corresponding to more deterministic, payoff-driven imitation—drives all four populations towards defection (except for the user population in Model I), with regulator defection being especially pronounced in Model I (rightmost panels). Interestingly, for small β, where imitation is highly stochastic, cooperation persists at non-trivial levels in user, developer and regulator populations despite the absence of cooperation in the commentator population, even for large c_I. This indicates that behavioural noise can help maintain cooperative behaviour under otherwise unfavourable incentives.

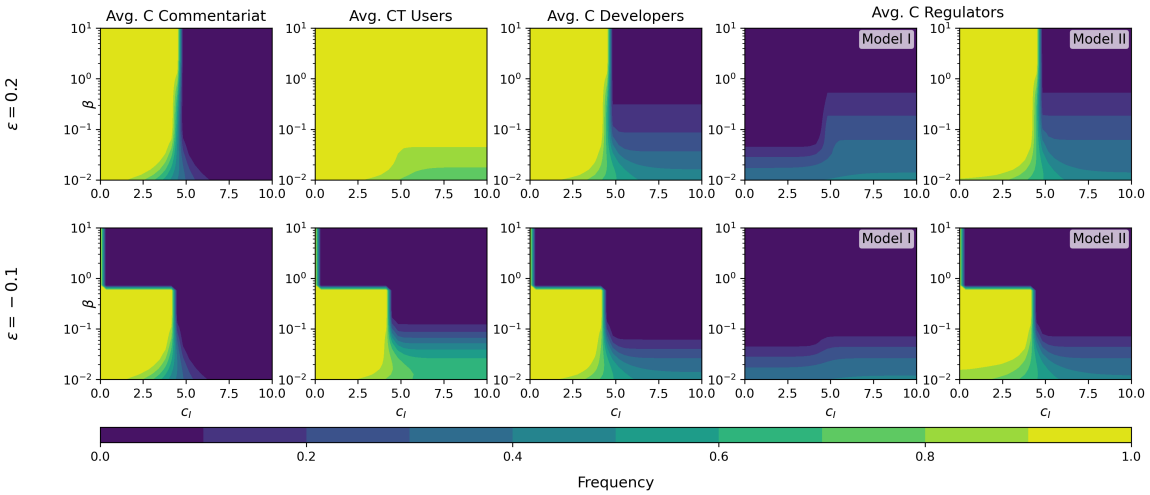

FIG. 4: Higher-quality, lower-cost media investigations (low c_I) strongly support cooperation, especially in Model II, across a wide range of selection intensities (β), whereas when investigations are costly and AI is high-risk (negative ϵ), defection dominates in all populations, with regulators almost always defecting in Model I, especially when β is high (corresponding to highly deterministic dynamics). When β is small, corresponding to highly stochasticity dynamics, cooperation is present in user, developer and regulator populations even for high c_I. Shown are frequencies of cooperative strategies in the four populations are shown as a function of c_I and β, for ($\epsilon = 0.2$) (top row, low risk) and ($\epsilon = -0.1$) (bottom row, high risk); the rightmost two panels show cooperative regulator frequencies separately for Model I and Model II. Parameters: $b_U = b_R = b_P = 4$, $u = 1.5$, $v = 0.5$, $c_P = 0.5$, $\beta = 0.1$, $N_U = N_C = N_R = 100$.

Beyond selection intensity and payoff parameters, we also examine the impact of asymmetric population sizes; see Supporting Information Figure S11. We vary the sizes of the commentariat, developer, and regulator populations (with users fixed), and also consider a more realistic configuration with many more users than regulator and commentariat actors. The results show that our main qualitative conclusions remain robust, while relative population sizes can modulate how easily cooperation is sustained.

Taken together, Figures 2–4 and Supporting Information Figures S2 and S5–S11 provide a broad sensitivity exploration: we systematically vary c_I, b_I, c_W, ε, β, and population sizes, and sample a large random parameter sets in Figure S2. These systematic analyses show that our central qualitative conclusions are robust across wide regions of the parameters space.

Overall, our finite population analysis shows that (i) low investigation costs and strong reputational incentives are primary drivers of system-wide cooperation and user trust, and (ii) when those conditions are not met, some degree of stochasticity in social learning is necessary to sustain cooperation, particularly in high-risk AI environments.

V. LLM RESULTS

This section presents the results from our LLM-based simulations, offering a complementary perspective to the game-theoretic analyses above. By prompting LLMs to represent the four actors (commentators, developers, regulators, and users) and interact dynamically based on payoff matrices, we can observe emergent strategic behaviours and compare them with our theoretical predictions.

In Model I, where the commentariat investigates developers (Figure 5), LLM results largely align with game theory. Commentariat agents cooperate when their reputational benefit (b_I) is high or potential reputational loss (c_W) from defecting is significant. Regulators consistently defect, which perfectly matches the game-theoretical prediction that their role becomes redundant ($w = 0$ in $X^* = (1, 1, 1, 0)$) when media effectively monitors and accurately reports on developers. Developers, when facing direct media scrutiny, are highly cooperative, consistent with the condition $c_P < b_P$. Users generally adopt conditional trust (CT) for low-risk AI.

In Model II, where the commentariat monitors regulators (Figure 5), we observe both corroboration and key divergences. Commentariat cooperation remains tied to strong reputational incentives ($c_I < p_W(b_I + c_W)$), consistent across both models. However, a notable divergence arises as LLM developers, particularly GPT-4o, mostly defect. This is different from the game-theoretical predictions above, where developers cooperate. This suggests LLM agents may exhibit behavioural nuances or biases derived from their training process, that lead to partially ignoring payoff maximisation in complex interactions. Similar emerging behaviour – that certain LLMs keep some tendencies, such as promoting cooperation or not in strategic games, instead of being dependent solely on the payoff structure of a game, have also been observed in other game theoretical settings

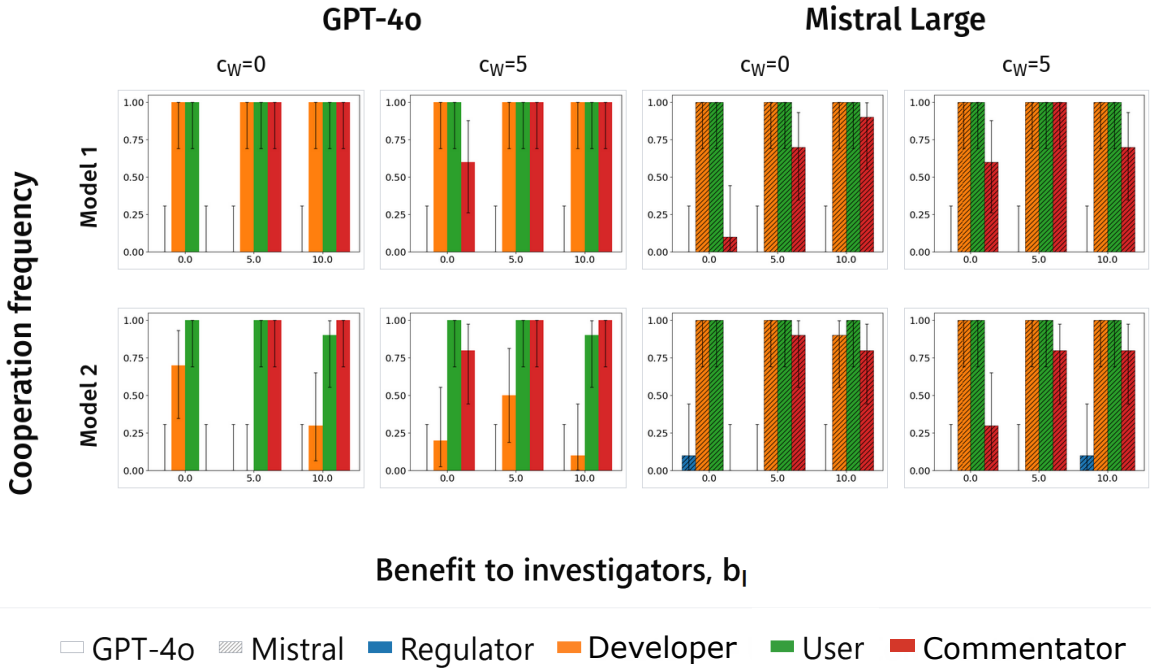

FIG. 5: **Results for the one-shot four agents game, using AI agents following Model I and Model II.** We show the average frequency of cooperative strategies for each player (Regulator, Developer, User and Commentator, from top to bottom), simulated using GPT 4o and Mistral, together with their Clopper–Pearson binomial 95% CI. All other parameters are set as for the numerical results above, except for $c_I = 0.5$ and c_W as specified in the columns.

[46]. Consequently, developers become more exploitative in Model II due to this lack of direct oversight and the LLM agents' defection, weakening the effective threat of punishment (u). Users show slightly less frequent adoption for high b_I in Model II compared to Model I, indicating subtle differences in how LLMs interpret the context of trust.

Comparing the LLM architectures, GPT commentators are generally more cooperative than Mistral's across both models; this tendency of GPT LLMs to be more cooperative than the Mistral family is in line with previous observations on different games and experimental settings [46, 50]. This highlights that different LLMs possess inherent biases or interpretive heuristics even under identical conditions, reinforcing the value of diverse modelling approaches to understand complex AI governance dynamics.

VI. CONCLUSIONS

Our analysis provides a rigorous framework for understanding the complex dynamics of AI governance, combining both traditional evolutionary game theory and novel LLM-based simulations. Across all our models, the results show that the media's cost of investigation is a key factor in determining whether regulators regulate effectively and developers follow these regulations. When the cost for the commentariat to conduct thorough investigations is high, their willingness to cooperate, and subsequently the cooperativeness of developers and regulators, significantly decreases. This highlights a fundamental need to manage the incentives and costs associated with high-quality media reporting.

Our EGT analysis identified two different pathways to achieving trustworthy AI governance, depending on where the media focuses its investigative efforts. When the media investigates developers (Model I), this approach can promote safe AI development and user trust. In an ideal scenario (i.e. a desirable stable equilibrium) characterised by low media investigation costs and developers finding it affordable to create safe AI, the media's effective reporting can lead to users adopting safe systems. Interestingly, in this specific desirable state, formal regulators tend to become less relevant, as their role is effectively supplanted by the media's direct scrutiny of developers. This approach is therefore promising in countries/regions where formal AI regulation is unavailable [51].

On the other hand, when the media investigates regulators (Model II), this model suggests a broader, more collaborative path to governance. Under conditions where media investigation costs are low, regulation is effective, and penalties for non-compliance are substantial, an ideal outcome emerges where all four key actors, i.e. the media, users, developers, and regulators, cooperate. This result is in line with previous findings that

when users can make informed decisions on whether to trust and adopt an AI system based on a regulator's reputation, then developers are incentivised to build safe AI [25]. This approach can prove more practical, as investigating the transparency and effectiveness of regulatory processes can be less resource-intensive for the media than directly evaluating complex AI development [1, 2].

Our finite population analysis further supports these findings and adds important nuance due to the finite population stochastic effects. We show that whether the media actually invest in thorough reporting depends very sensitively on how costly investigations are and how strongly accuracy is rewarded or punished reputationally. When good reporting is expensive and incentives are weak, media cooperation, and then developer and regulator cooperation, rapidly erode. It also reveals that user trust is highly context-dependent: when adopting AI is low-risk, users tend to keep conditionally trusting systems even in the face of occasional media failures, but when AI is high-risk and media become unreliable, users quickly shift to not adopting at all. Finally, our analysis highlights that outcomes also depend on how strongly agents respond to payoff differences, captured by the intensity of selection [38, 47]. When learning dynamics are highly deterministic (corresponding to high intensities of selection), defection spreads easily under weak media incentives, whereas with more noisy, exploratory dynamics some cooperation can persist even in otherwise unfavourable conditions.

Our LLM-based analysis offered a complementary perspective. While largely corroborating EGT predictions regarding media cooperation, user trust, and regulator behaviour in Model I, a notable divergence emerged in Model II (where media investigates regulators). Different from game theory predictions of high cooperation when the media cost is sufficiently low, LLM regulators frequently opted not to cooperate. This unexpected behaviour from AI agents led to developers facing less oversight and becoming more exploitative. This suggests that LLMs, even when prompted for rational decision-making, might exhibit biases or heuristic decision-making that deviates from strict game theory, highlighting the presence of such characteristics within these AI models [46, 52, 53]. Differences in cooperative tendencies between various LLM architectures (e.g., GPT versus Mistral) further reinforced this notion of inherent behavioural traits. The precise reason for such behaviour is still a matter of debates and deserves deeper studies. LLMs are known to be inconsistent between each other and to be differently skewed towards a certain strategic alignment or behavioural preference (e.g., they show political preferences or misalignment [33, 54]). Randomness may be a source of inconsistency, but we have mitigated its effect using repeated runs. Paraphrasing inconsistency [55] may be another limiting factor; however, our prompt structure is very essential in providing a verbal description of the payoff structure and very little additional information, which can hardly be paraphrased (see Supporting Information Sec. III). Another hypothesis is that these tendencies may be associated with how the training process (data, training, tuning, guardrailing, and more) eventually yields LLM outputs, or with personality-like traits of certain LLMs, again derived from their training process; a few works explore this direction exist [46], but many more are required to deeply characterise the aetiology of the inconsistency, which falls beyond the scopes of this work. We hope that, showing that such phenomenon also exists in complex games, will promote future studies into the topic.

A. Implications

Our results show that the media's role in AI governance is crucial but highly conditional. Across both models and both types of analysis (infinite and finite populations), cooperation by the commentariat is very sensitive to two factors: i) how costly it is to investigate (the effort of good reporting) and ii) how strong the reputational rewards and penalties are for being right or wrong. When accurate investigation is low-cost and reputational incentives are high, the media tend to cooperate, which in turn encourages safe development by developers, responsible regulation, and user trust/adoption. When investigation is expensive, media cooperation collapses in our numerical calculations and simulations, and so does cooperation by developers and regulators.

This leads to a first implication: if policymakers and industry want to rely on media as a meaningful component of AI governance, they must actively shape the environment to enable low-cost high-quality investigation. Concretely, that means reducing the practical cost of good reporting. Measures to facilitate this could include public registries of where and which AI systems are being used in high risk applications, as is mandated in Article 49 of the EU Artificial Intelligence Act [56], for example. This would increase transparency in which AI systems are being used in different settings, thereby facilitating media investigations. Article 11 of the EU Artificial Intelligence Act also mandates that developers of high-risk AI systems produce mandatory technical documentation before the system is put into use, and that this documentation is kept up-to-date [57]. Such measures can further increase transparency for media, reducing the cost c_I of providing informed recommendations. High quality investigative journalism could be incentivised through state aid, such as through direct government financial grants, or indirect tax offsets [58, 59], which would also reduce c_I. Research should investigate how the reputation of media sources can be objectively evaluated and published using automated metrics, for example based on citation numbers, the quality of sources linking to articles, and the authority of the domain [60]. Such reputation mechanisms would increase b_I – the reputational benefit a commentator receives when making a correct recommendation, and would increase c_W – the cost of making an incorrect recommendation. These are precisely the levers that, in our models, separate cooperative from non-cooperative outcomes. Similar concerns about the cost and capacity of effective oversight arise in regulator-centred models of AI governance

[7, 9, 19, 20], and in broader discussions of frontier AI evaluation infrastructure [61, 62]. Our results therefore reinforce calls for sustained public investment in evaluation capacity and oversight institutions, such as national AI safety institutes, as a prerequisite for meaningful governance [3, 4, 61].

A second implication concerns where media attention is directed. In Model I, where media investigate developers, our analysis and simulations show that accurate reporting can be strong enough to push developers toward safe behaviour and allow users to appropriately trust AI, even if regulators largely disengage or are unavailable. This pathway is especially relevant in settings where formal AI regulation is weak, absent, or slow to emerge: credible media scrutiny can, to a degree, substitute for regulation [21]. In Model II, where media investigate regulators instead, we see a different pattern: when regulatory action is not too costly, and when developers face real penalties for unsafe behaviour, media oversight of regulators promotes cooperation by all four actors. This suggests that in jurisdictions with serious regulatory capacity, it may be more effective and more realistic to encourage media to focus on regulatory performance rather than directly (re-)evaluate complex AI systems. This is consistent with work arguing that regulators themselves are strategic actors whose effectiveness depends on clear institutional incentives and credible monitoring [6, 7, 63, 64]. It also aligns with proposals for mixed public–private regulatory markets and certification regimes, in which independent auditors and public bodies jointly shape safe AI development [6, 64].

A third implication follows from what our payoff structures and dynamics assume the media must actually find out. In both models, the key outcomes depend on whether AI systems are in fact safe or unsafe, and whether regulators are in fact enforcing the rules. In practice, that information typically comes from capability and safety evaluations, including tests of dangerous capabilities. Our findings therefore point toward the importance of building ecosystems where such evaluations are routinely conducted and, crucially, made legible and accessible to journalists and other commentators. Lowering the barriers for media to understand and report on evaluation results is exactly what, in our framework, lowers the cost of effective investigation and supports cooperative equilibria. This perspective supports evaluation-centric approaches to AI governance that emphasise systematic testing of capabilities and risks, and clear communication of those results [2, 61, 65].

Finally, the LLM simulations broadly support these conclusions, but sometimes deviate from the game-theoretic predictions. This fact may have two interpretations, which need careful studies due to their potential implications. One hypothesis is that the inconsistency is linked to implicit biases that LLMs of different kind may possess, which is in turn associated with the training datasets or process [50, 66, 67]; under this point of view, game theory presents a "ground-truth" that helps reveal biases in LLMs. However, LLM deviations may also reveal how real-world decision-makers might themselves deviate from idealised game-theoretic incentives; in fact, LLMs are trained on real-world data, thereby possibly reflecting additional (albeit limited) complexity in thought processes [68]. In particular, in Model II, LLM-based regulators often chose not to cooperate even under conditions where the game theory model predicts they should. This mirrors the possibility that actual regulators may fail to enforce effectively despite having formal incentives to do so. It reinforces the case for designing AI governance with redundancy: not just relying on formal regulation, but also empowering a well-incentivised, well-informed media ecosystem that can monitor both developers and regulators and help steer the system toward safer trajectories. As LLMs become more pervasive in modelling studies and decision-making [69, 70], future works may further investigate these intriguing results.

B. Limitations and areas for future research

Firstly, our model assumes a highly stylised, mean-field setting in which all actors within a given population are homogeneous and randomly matched. In practice, AI governance involves structured interactions: developers operate within platform-specific ecosystems, regulators are organised by jurisdiction, media outlets serve segmented markets, and users cluster within socio-economic groups [6, 63]. Introducing network structure into evolutionary game models is known to qualitatively alter dynamics even in single-population settings [71]. Recently, it has been shown that even in single-population games on graphs, the concept of a homogeneous evolutionary process must account not just for the population structure, but for payoff structure and mutation dynamics jointly [72]. These results underscore that extending network analysis to four-population models is far from straightforward. Fundamental modelling questions would need to be resolved: whether each population should reside in its own network or be embedded in a shared structure, how the bipartite or multipartite matching between populations should be governed by network topology, and how the separation between interaction and dispersal graphs should be handled across multiple, co-evolving populations [72]. Future work could therefore tighten the random-matching assumption by allowing partner selection and explicit competition between regulatory regimes. We note that incorporating network heterogeneity into related technology adoption models, for example by island network models of cultural group selection and multi-level selection [17, 73–76], represents a promising direction that captures a broad scope of effects such as jurisdictional clustering among regulators, platform lock-in among developers, and media market segmentation among commentators.

Moreover, our analysis assumes fixed population sizes for users, developers, regulators, and media, and thus does not model market entry, exit, or consolidation (for example, new firms entering, companies merging, or media markets fragmenting). This simplification allows us to focus on strategic incentives within a given

governance landscape, but it omits potential structural dynamics, for instance as highlighted in research on sustainable electricity markets and adaptive policy design [24]. An interesting direction would be to allow market structure to evolve, letting the number and type of developers, regulators, and media actors change over time. As such, we can examine how such entry, exit, and consolidation processes interact with the strategic mechanisms we identify in the present work.

In this work, our models were simplified by giving each actor only a binary choice. In practice, developers may partially comply with safety requirements [18], regulators may enforce rules selectively, and media actors may vary the depth and balance of their reporting. We use binary strategies to keep the analysis tractable and to clarify the core mechanisms. A natural direction for future work is to introduce richer action spaces, including multiple levels or even continuous space of compliance, enforcement effort, and/or investigative intensity, which would allow more fine-grained and realistic equilibrium patterns to be studied.

We also focus on a single layer of incentives for the commentariat, and do not explicitly explore richer incentive architectures across all actors. There is extensive work on how combinations of rewards and punishments, and adaptive or hybrid schemes, can promote cooperation and other pro-social behaviours in social dilemmas and climate governance [77–80], in both finite-population stochastic [81–83] and infinite population (via optimal control theory) [84, 85] analyses. Extending our framework to consider such more sophisticated incentive mechanisms, for commentariat, regulators, developers and even users, could reveal more cost-efficient ways of sustaining trustworthy AI ecosystems. Moreover, we have not explicitly modelled AI race dynamics between AI firms/developers, even though competitive pressure to rapidly scale capabilities is known to interact in complex ways with safety investments [18, 22]. Embedding our governance ecosystem within explicit race models, where firms trade off speed against safety and regulation, would help assess how robust our conclusions are under stronger competitive pressure.

Note that we have not introduced an explicit social welfare function to compare governance regimes. Instead, we use a qualitative notion of desirability based on safe AI adoption, reduced harm from unsafe systems, and avoidance of wasted oversight effort. An important direction for future work is to define and analyse formal welfare metrics [78, 86], such as weighted sums of expected payoffs across populations, or user-centric welfare measures, in order to compare equilibria and media-targeting strategies in a more principled, quantitative way and to align recommendations with specific policy priorities.

Agents use payoff-biased social learning in our model, which is standard in evolutionary game theory [16], but does not capture all aspects of decision-making by individuals or firms. However, it provides a null hypothesis of bounded rationality – individuals or organisations copying the behaviour of others seen to be doing better than themselves, and is consistent with the assumptions of neoclassical microeconomics and evolutionary biology [87]. Behavioural economics shows that biases such as loss avoidance can modify behaviour from that predicted by bounded rationality [88]. Indeed, there are many such biases observed in the behavioural economics literature [89], including present bias (overweighting immediate rewards) [90], status quo bias (preferring present circumstances, which could encourage continued use of the same strategy) [91], and conformity bias (herd effects) [92]. Our study provides a baseline from which future work can examine the effects of these biases and heuristics.

A further limitation is that we do not empirically calibrate our payoff parameters to specific real-world AI governance settings. At present, there is little systematic data on issues such as the financial and organisational cost of rigorous AI investigations, sanction levels for unsafe behaviour, or the size of reputational gains and losses for media actors [3, 19]. Our aim here was therefore to understand mechanisms and threshold conditions using stylised payoffs, rather than to predict outcomes for any particular context. An important direction for future work is to combine this framework with empirical case studies and available data on AI audits, enforcement actions, and media coverage, as well as the population sizes of different actors in the model, so that fully calibrated versions of the model can support more context-specific policy recommendations.

While we recognise the limitations of our simplified scenario, we find our approach useful as a tool for thinking through what assumptions policymakers must make in order for different combinations of regulatory and media regimes to plausibly achieve their policy objectives. Further avenues of research should test the predictions of our model empirically [93], for example by studying how users, developers, regulators, and media organisations actually interact around concrete AI deployments, and by examining how risk evaluations and regulatory signals are communicated and received in practice.

Overall, our findings offer valuable insights into the dynamics at play in the AI regulation landscape and highlight the indispensable role of reliable oversight actors—not only regulators but also the media—whose incentives are aligned through governmental rewards, market mechanisms, or the maintenance of prestigious reputation. They highlight the value of using game-theoretic models, complemented by LLM-based simulations, to shed light on which governance designs are most likely to steer developers and users towards safer AI.

ACKNOWLEDGEMENTS

An earlier version of 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 (IAPMEI/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.

---

[1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
[2] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart *et al.*, "Governing ai safety through independent audits," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 566–571, 2021.
[3] Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani *et al.*, "International Scientific Report on the Safety of Advanced AI (Interim Report)," Nov. 2024. [Online]. Available: http://arxiv.org/abs/2412.05282
[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.
[5] 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.
[6] J. Clark and G. K. Hadfield, "Regulatory Markets for AI Safety," *arXiv preprint arXiv:2001.00078*, Dec. 2019.
[7] M. Anderljung, J. Barnhart *et al.*, "Frontier AI Regulation: Managing Emerging Risks to Public Safety," *arXiv preprint arXiv:2307.03718*, Jul. 2023.
[8] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean *et al.*, "Multi-Agent Risks from Advanced AI," *arXiv preprint arXiv:2502.14143*, 2025.
[9] G. K. Hadfield and J. Clark, "Regulatory Markets: The Future of AI Governance," *arXiv preprint arXiv:2304.04914*, no. arXiv:2304.04914, Apr. 2023.
[10] T. A. Han, J. Z. Leibo, T. Lenaerts, I. Rahwan, F. Santos, M. Perc, and V. Capraro, "Social physics in the age of artificial intelligence," 2026. [Online]. Available: https://arxiv.org/abs/2603.16900
[11] 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.
[12] M. Sutrop, "Should We Trust Artificial Intelligence?" *Trames*, vol. 23, no. 4, pp. 499–522, 2019.
[13] 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.
[14] 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.
[15] K. Sigmund, "The Calculus of Selfishness," in *The Calculus of Selfishness*. Princeton University Press, 2010.
[16] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*. Cambridge university press, 1998.
[17] K. Binmore, *Natural Justice*. Oxford University Press, 2005.
[18] 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.
[19] 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.
[20] 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.
[21] H. C. da Fonseca, A. Fernandes, Z. Song, T. Cimpeanu, N. Balabanova, A. Bashir *et al.*, "Can Media Act as a Soft Regulator of Safe AI Development? a Game Theoretical Analysis," *ALIFE 2025: Ciphers of Life. Proceedings of the Artificial Life Conference 2025*, p. 90, 10 2025.
[22] S. Armstrong, N. Bostrom *et al.*, " Racing to the Precipice: A Model of Artificial Intelligence Development ," *Ai & Society*, vol. 31, no. 2, pp. 201–206, May 2016.
[23] L. Cheng, M. Zhang, K. Wang, M. Yuan, Z. Liu, J. Wang, K. Zhang, and P. Huang, "Evolutionary smart contracts for virtual power plant trading: integrating prospect theory and multi-stage negotiation in cross-regional energy markets," *International Journal of Electrical Power & Energy Systems*, vol. 173, p. 111453, 2025.
[24] L. Cheng, R. Sun, K. Wang, F. Yu, P. Huang, and M. Zhang, "Advancing sustainable electricity markets: Evolutionary game theory as a framework for complex systems optimization and adaptive policy design," *Complex & Intelligent Systems*, vol. 11, no. 7, p. 320, 2025.
[25] Z. Alalawi, P. Bova, T. Cimpeanu, A. Di Stefano, M. Hong Duong, E. F. Domingos, T. A. Han, M. Krellner, N. B. Ogbo, S. T. Powers, and F. Zimmaro, "Trust AI Regulation? Discerning Users are Vital to Build Trust and Effective AI Regulation," *Applied Mathematics and Computation*, vol. 508, p. 129627, 2026.
[26] 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.
[27] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese, "In ai we trust? perceptions about automated decision-making by artificial intelligence," *AI & society*, vol. 35, no. 3, pp. 611–623, 2020.
[28] M. Maggetti, "The Media Accountability of Independent Regulatory Agencies," *European Political Science Review*, vol. 4, no. 3, pp. 385–408, Nov. 2012.

[29] 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.

[30] A. Traulsen and N. E. Glynatsi, "The future of theoretical evolutionary game theory," *Philosophical Transactions of the Royal Society B*, vol. 378, no. 1876, p. 20210508, 2023.

[31] 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.

[32] C. A. Bail, "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences*, vol. 121, no. 21, p. e2314021121, 2024.

[33] A. Buscemi and D. Proverbio, "Large Language Models' Detection of Political Orientation in Newspapers," *arXiv preprint arXiv:2406.00018*, 2024.

[34] N. Lee, J. Hong, and J. Thorne, "Evaluating the Consistency of LLM Evaluators," *arXiv preprint arXiv:2412.00543*, 2024.

[35] A. Buscemi and D. Proverbio, "ChatGPT vs Gemini vs Llama on Multilingual Sentiment Analysis," *arXiv preprint arXiv:2402.01715*, 2024.

[36] OpenAI. (2023) Introducing ChatGPT. [Online]. Available: https://openai.com/blog/chatgpt

[37] Mistral AI. (2025) Au large. [Online]. Available: https://mistral.ai/news/mistral-large

[38] A. Traulsen, M. A. Nowak, and J. M. Pacheco, "Stochastic Dynamics of Invasion and Fixation," *Physical Review E*, vol. 74, p. 11909, 2006.

[39] 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.

[40] 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.

[41] E. F. Domingos, F. C. Santos, and T. Lenaerts, "EGTtools: Evolutionary Game Dynamics in Python," *Iscience*, vol. 26, no. 4, 2023.

[42] S. Encarnação, F. P. Santos, F. C. Santos, V. Blass, J. M. Pacheco, and J. Portugali, "Paradigm Shifts and the Interplay Between State, Business and Civil Sectors," *Royal Society open science*, vol. 3, no. 12, p. 160753, 2016.

[43] 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, 2019, pp. 135–142.

[44] P. D. Taylor, "Evolutionarily Stable Strategies with Two Types of Player," *Journal of applied probability*, vol. 16, no. 1, pp. 76–83, 1979.

[45] 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.

[46] A. Buscemi, D. Proverbio, A. Di Stefano, T. A. Han, and P. Liò, "FAIRGAME: a Framework for AI Agents Bias Recognition using Game Theory," *Frontiers in Artificial Intelligence and Applications*, vol. ECAI 2025, pp. 4097 – 4104, 2025.

[47] 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.

[48] 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.

[49] 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.

[50] D. Proverbio, A. Buscemi, A. Di Stefano, T. A. Han, G. Castignani, and P. Liò, "Can LLMs Effectively Provide Game-Theoretic-Based Scenarios for Cybersecurity?" *Frontiers in Computer Science*, vol. 7, p. 1703586, 2025.

[51] N. A. Smuha, "From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence," *Law, Innovation and Technology*, vol. 13, no. 1, pp. 57–84, 2021.

[52] E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz, "Playing repeated games with large language models," *Nature Human Behaviour*, pp. 1–11, 2025.

[53] N. Fontana, F. Pierri, and L. M. Aiello, "Nicer than humans: How do large language models behave in the prisoner's dilemma?" in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2025. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/35829

[54] D. Rozado, "The political preferences of llms," *PloS one*, vol. 19, no. 7, p. e0306621, 2024.

[55] J. J. Ahn and W. Yin, "Prompt-reverse inconsistency: Llm self-inconsistency beyond generative randomness and prompt paraphrasing," *arXiv preprint arXiv:2504.01282*, 2025.

[56] "Article 49: Registration | EU Artificial Intelligence Act." [Online]. Available: https://artificialintelligenceact.eu/article/49/

[57] "Article 11: Technical Documentation | EU Artificial Intelligence Act." [Online]. Available: https://artificialintelligenceact.eu/article/11/

[58] R. Foster and M. Bunting, "Public funding of high-quality journalism A report for the ACCC," Australian Competition and Consumer Commission, Canberra, Tech. Rep., 2019. [Online]. Available: https://www.accc.gov.au/system/files/ACCC%20commissioned%20report%20-%20Public%20funding%20of%20high-quality%20journalism%20-%20phase%201,%20Communications%20Chambers.PDF

[59] P. C. Murschetz, "State Aid for Independent News Journalism in the Public Interest? A Critical Debate of Government Funding Models and Principles, the Market Failure Paradigm, and Policy Efficacy," *Digital Journalism*, vol. 8, no. 6, pp. 720–739, Jul. 2020, _eprint: https://doi.org/10.1080/21670811.2020.1732227. [Online]. Available: https://doi.org/10.1080/21670811.2020.1732227

[60] M. Trillo-Domínguez, R. Salaverría, L. Codina, and F. De Moya-Anegón, "Digital reputation indicator: A webometric approach for a global ranking of digital media," *Journalism*, vol. 26, no. 2, pp. 406–424, Feb. 2025. [Online]. Available: https://doi.org/10.1177/14648849241237647

[61] UK AI Safety Institute, "Early lessons from evaluating frontier ai systems," 2024. [Online]. Available: https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems

[62] J. Whittlestone and J. Clark, "Why and How Governments Should Monitor AI Development," *arXiv*, Aug. 2021.

[63] J. Tallberg, E. Erman *et al.*, "The global governance of artificial intelligence: Next steps for empirical and normative research," *International Studies Review*, vol. 25, no. 3, p. viad040, 2023, private vs public regulation.

[64] P. Cihon, M. J. Kleinaltenkamp *et al.*, " AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries ," *IEEE Transactions on Technology and Society*, vol. 2, no. 4, pp. 200–209, Dec. 2021.

[65] A. Dafoe, "AI Governance: A Research Agenda," The University of Oxford, Oxford, UK, Tech. Rep., Aug. 2018.

[66] A. Buscemi, D. Proverbio, P. Bova, N. Balabanova, A. Bashir, T. Cimpeanu, H. C. da Fonseca, M. H. Duong, E. F. Domingos, A. M. Fernandes *et al.*, "Do LLMs Trust AI Regulation? Emerging Behaviour of Game-theoretic LLM Agents," *arXiv preprint arXiv:2504.08640*, 2025.

[67] T.-K. Huynh, D.-M. Dao-Sy, T.-B. Cao, P.-H. Le, H.-D. Nguyen, P.-Q. Nguyen-Lam, M.-L. Nguyen-Vo, H.-P. Pham, P.-H. Pham, T.-K. Than *et al.*, "Understanding LLM Agent Behaviours via Game Theory: Strategy Recognition, Biases and Multi-Agent Dynamics," *arXiv preprint arXiv:2512.07462*, 2025.

[68] N. B. Petrov, G. Serapio-García, and J. Rentfrow, "Limited Ability of LLMs to Simulate Human Psychological Behaviours: a Psychometric Analysis," *arXiv preprint arXiv:2405.07248*, 2024.

[69] E. Eigner and T. Händler, "Determinants of LLM-Assisted Decision-Making," *arXiv preprint arXiv:2402.17385*, 2024.

[70] Q. Wang, J. Wu, Z. Tang, B. Luo, N. Chen, W. Chen, and B. He, "What Limits LLM-based Human Simulation: LLMs or Our Design?" *arXiv preprint arXiv:2501.08579*, 2025.

[71] G. Szabó and G. Fáth, "Evolutionary games on graphs," *Physics Reports*, vol. 446, no. 4, pp. 97–216, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157307001810

[72] A. McAvoy and C. Hauert, "Structural symmetry in evolutionary games," *Journal of The Royal Society Interface*, vol. 12, no. 111, p. 20150420, 10 2015. [Online]. Available: https://doi.org/10.1098/rsif.2015.0420

[73] J. C. van den Bergh and J. M. Gowdy, "A group selection perspective on economic behavior, institutions and organizations," *Journal of Economic Behavior & Organization*, vol. 72, no. 1, pp. 1–20, 2009.

[74] P. Richerson, R. Baldini, A. V. Bell, K. Demps, K. Frost, V. Hillis, S. Mathew, E. K. Newton, N. Naar, L. Newson *et al.*, "Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence," *Behavioral and Brain Sciences*, vol. 39, p. e30, 2016.

[75] A. Traulsen and M. A. Nowak, "Evolution of cooperation by multilevel selection," *Proceedings of the National Academy of Sciences*, vol. 103, no. 29, pp. 10 952–10 955, 2006.

[76] L. Cheng, P. Peng, W. Lu, P. Huang, and Y. Chen, "Study of flexibility transformation in thermal power enterprises under multi-factor drivers: Application of complex-network evolutionary game theory," *Mathematics*, vol. 12, no. 16, 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/16/2537

[77] K. Sigmund, H. De Silva, A. Traulsen, and C. Hauert, "Social learning promotes institutions for governing the commons," *Nature*, vol. 466, no. 7308, pp. 861–863, 2010.

[78] T. A. Han, Z. Song, T. Cimpeanu, M. H. Duong, M. Krellner, V. Capraro, and M. Perc, "Cooperation versus social welfare," *Physics of Life Reviews*, vol. 56, pp. 33–60, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1571064525001654

[79] A. R. Góis, F. P. Santos, J. M. Pacheco, and F. C. Santos, "Reward and punishment in climate change dilemmas," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.

[80] Y. Liu, L. Wang, R. Guo, S. Hua, L. Liu, L. Zhang *et al.*, "Evolution of trust in the n-player trust game with transformation incentive mechanism," *Journal of the Royal Society Interface*, vol. 22, no. 224, 2025.

[81] X. Chen, T. Sasaki, Å. Brännström, and U. Dieckmann, "First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation," *Journal of The Royal Society Interface*, vol. 12, no. 102, p. 20140935, 2015.

[82] M. Duong, C. Durbac, and T. Han, "Cost optimisation of hybrid institutional incentives for promoting cooperation in finite populations," *J. Math. Biol.*, vol. 87, 2023.

[83] T. A. Han, "Institutional incentives for the evolution of committed cooperation: Ensuring participation is as important as enhancing compliance," *Journal of the Royal Society Interface*, vol. 19, no. 188, p. 20220036, 2022.

[84] S. Wang, X. Chen, Z. Xiao, A. Szolnoki, and V. V. Vasconcelos, "Optimization of institutional incentives for cooperation in structured populations," *Journal of the Royal Society Interface*, vol. 20, no. 199, 2023.

[85] W. Sun, L. Liu, X. Chen, A. Szolnoki, and V. V. Vasconcelos, "Combination of institutional incentives for cooperative governance of risky commons," *Iscience*, vol. 24, no. 8, p. 102844, 2021.

[86] R. L. Rardin, *Optimization in operations research*. Prentice Hall Upper Saddle River, NJ, 1998, vol. 166.

[87] S. T. Powers, C. P. van Schaik, and L. Lehmann, "Cooperation in large-scale human societies—What, if anything, makes it unique, and how did it evolve?" *Evolutionary Anthropology: Issues, News, and Reviews*, vol. 30, no. 4, pp. 280–293, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/evan.21909

[88] D. Kahneman, "A Psychological Perspective on Economics," *The American Economic Review*, vol. 93, no. 2, pp. 162–168, 2003. [Online]. Available: https://www.jstor.org/stable/3132218

[89] D. Rau and P. Bromiley, "A review of cognitive biases in strategic decision making," *Long Range Planning*, vol. 58, no. 3, p. 102529, Jun. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0024630125000329

[90] T. O'Donoghue and M. Rabin, "Present bias: Lessons learned and to be learned," *The American Economic Review*, vol. 105, no. 5, pp. 273–279, 2015. [Online]. Available: https://www.jstor.org/stable/43821892

[91] C.-C. Wu, "Status quo bias in information system adoption: a meta-analytic review," *Online Information Review*, vol. 40, no. 7, pp. 998–1017, Nov. 2016. [Online]. Available: https://doi.org/10.1108/OIR-09-2015-0311

[92] C. Capuano and P. Chekroun, "A systematic review of research on conformity," *International Review of Social Psychology*, vol. 37, no. 1, 2024.

[93] E. F. Domingos and T. A. Han, "Inertia and fear of lagging behind drive unsafe technological development in an idealised AI race experiment," in *The 37th Benelux Conference on Artificial Intelligence and the 34th Belgian Dutch Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=Ma7jBkFeXF