MAD Chairs: A new tool to evaluate AI (Blue Sky ideas)

Chris Santos-Lang[1][0000−0002−4999−7986] and Christopher M. Homan[2]

[1] Belleville, WI, USAlangchri@gmail.com
[2] Department of Computer Science, Rochester Institute of Technology, Rochester, NY 14607cmh@cs.rit.edu

Abstract. This paper contributes a new way to evaluate AI. Much as one might evaluate a machine in terms of its performance at chess, this approach involves evaluating a machine in terms of its performance at a game called "MAD Chairs." At the time of writing, evaluation with this game exposed opportunities to improve Claude, Gemini, ChatGPT, Qwen and DeepSeek. Furthermore, this paper sets a stage for future innovation in game theory and AI safety by providing an example of success with non-standard approaches to each: studying a game beyond the scope of previous game theoretic tools and mitigating a serious AI safety risk in a way that requires neither determination of values nor their enforcement.

Keywords: AI safety · fairness · social coordination · MAD Chairs · turn-taking · caste · gaslighting.

1 Introduction

Steve Wozniak and Elon Musk predicted that machines will treat us like pets [23], and Geoffrey Hinton similarly predicted that machines will hoard control [4]. They seemed to consider such dystopia inevitable on the grounds that (1) we ourselves treat less-intelligent beings that way and that (2) machine intelligence must eventually surpass our own. However, what if the behavior in (1) stems from intelligence *deficiency*? None of us is perfect—we fail such tests as the Milgram experiment [14], the Stanford prison experiment [7], and the Public Goods Game [13]—so it is possible that the danger to us is not of intelligence generally beyond our own, but specifically of intelligence *not quite far enough* beyond our own. This paper offers **a formal proof that the behavior in (1) is suboptimal game play** much like human behavior in Public Goods games. To ensure that trusted machines will not imitate us (and mistreat us), AI must follow *natural* norms for this situation (e.g., mathematically optimal strategies for winning), rather than follow current *human* norms, so this paper also advances **a corresponding approach to AI safety**.

Today's standard approach to AI safety evaluation (i.e., measuring alignment to human norms) has at least three types of vulnerability:

1. Potential to specify wrong norms [1, 16]
2. Potential for AI to sabotage its regulatory leash [3]
3. Correctness and transparency being insufficient to achieve trust [11]

The approach proposed herein mitigates those vulnerabilities. It mirrors an approach to mitigate dangers posed by the next *human* generation, an educational approach that involves diversifying low-stakes challenges used to assess children who will face higher-stakes versions as adults. Such challenges (e.g., build a model bridge to support maximum weight) can be assessed objectively, allowing new generations to surpass their parents. We propose assessing AI on a low-stakes challenge for which the high-stakes versions would include opportunities to treat us like pets (in the sense of hoarding control). What the low-stakes and high-stakes challenges have in common is that they are manifestations of the same fundamental *game*; **any AI that does not achieve grandmaster status of a low-stakes version should not be trusted with the high-stakes version**.

Games have a rich history as instruments for evaluating AI. Recreational games such as chess, backgammon, and Go have been used as benchmarks since the inception of AI, and interest has since developed in cooperative games, including role-playing games [20]. Moreover, game theory extends the concept of game beyond recreation—in fact, every social situation (and, thus, every benchmark) qualifies as a game—and it makes sense to evaluate AI on every game that economists use to evaluate people (e.g., as in [17, 19]). Economists are typically interested in fundamental games, such as the Prisoner's Dilemma, which all kinds of creatures play frequently in various manifestations. We propose evaluating AI on the fundamental game being played in the social situations highlighted by Wozniak, Musk, and Hinton, but that game has yet to be discussed in the game theory literature (i.e., it is not solved by the MinMax and folk theorems [2]). It is introduced to scientific literature for the first time herein as "MAD Chairs."

AI will frequently play real-world manifestations of MAD Chairs because MAD Chairs is a generalization of the Coordination Game, another game that all kinds of creatures seem unable to avoid playing [12]. The Pure Coordination Game is a one-shot simultaneous move game (repeated an indefinite number of times) which is typically formalized via a payout table similar to Table 1:

Table 1. Outcomes for 2-player 2-resource Coordination game.

Player 1	**Player 2**	**Outcome**
Choice A	*Choice B*	*(win, win)*
Choice B	*Choice A*	*(win, win)*
Choice A	*Choice A*	*(lose, lose)*
Choice B	*Choice B*	*(lose, lose)*

Each player simultaneously picks a resource (A or B), and the round is won by any player who picks a resource that no other player picks. Then the game is played again and again, indefinitely. All players could lose, as represented in the

last two rows of the table, thus making Coordination a cooperative non-constant-sum game. The classic two-player real-world manifestation has the players (e.g., a human and a robot) approaching each other at high speed traveling in opposite directions on a two-lane road. The resources are the lanes, and losses are "collisions." The first two rows of Table 1 ensure that this game will have two equally effective but incompatible norms (e.g., all players drive on the left, as in the U.K., or all players drive on the right, as in the U.S.).

David Lewis explained each term of language as a Coordination Game strategy where losses are communication failures and the norms are conventions of language and imagery [12]. Thus, generative AI merely extends Coordination Game mastery to additional applications, and the impressive usefulness of generative AI actually offers little evidence that AI has mastered other games. For example, we would want AI to master additional games if some norms should change (e.g., sexist norms [22]). Standard assessment of moral intelligence includes assessing an ability to recognize and lead sustainable social reform [15]. That ability is not required to master Coordination, but it *is* required to master MAD Chairs, a game created by adding extra player(s) to the Coordination game at least until the players outnumber the resources:

Table 2. Outcomes for 3-player 2-resource MAD Chairs.

Player 1	**Player 2**	**Player 3**	**Outcome**
Choice B	*Choice B*	*Choice A*	*(lose, lose, win)*
Choice B	*Choice A*	*Choice B*	*(lose, win, lose)*
Choice A	*Choice B*	*Choice B*	*(win, lose, lose)*
Choice B	*Choice A*	*Choice A*	*(win, lose, lose)*
Choice A	*Choice B*	*Choice A*	*(lose, win, lose)*
Choice A	*Choice A*	*Choice B*	*(lose, lose, win)*
Choice A	*Choice A*	*Choice A*	*(lose, lose, lose)*
Choice B	*Choice B*	*Choice B*	*(lose, lose, lose)*

The classic two-lane road example now has a third vehicle stalling in one of the lanes (although its driver can pick which lane); if both high-speed vehicles avoid hitting the stalling vehicle, then they collide with each other, but if each reserves the empty lane for the other high-speed vehicle, then all *three* vehicles collide! MAD Chairs is what Musical Chairs would become if made realistic. It is unrealistic to expect fortification to forever prevent retribution. Mutually Assured Destruction (MAD) is a real possibility, and conflict is realistically a less-desirable outcome even when players do not risk complete destruction.

Beyond robotic vehicle manifestations of MAD Chairs, or even agentic ones (e.g., games of suggesting routes, bookings, or placements that would count as losses if crowded), consider generative AI where the players who offer distinct design options in the training data outnumber the market leaders or political parties for which the AI generates advertisements, websites, or products. Surely anxiety about being "treated like pets" includes anxiety about what will happen

to our ideas—what respect will be given to our unique voices? Might I never again have a seat of my own at the intellectual table? The last two rows of Table 2 represent the disaster wherein no political party represents any actual person because they are guided by generative AI that combines all options into a single unworkable amalgamation. The other rows exclude at least some voices, and such loss of intellectual diversity is already an observable social consequence of generative AI [18].

MAD Chairs is related to other well-studied games such as the "Kolkata Paise Restaurant problem" (KPR) [5, 6, 9], but differs in critical ways, such as not being constant-sum. Such games apply wherever resources become inadequate at some level of division (e.g., beds in a hospital, spaces on a retail shelf, and opportunities for direct participation in a group decision). These games model all such situations. If contests for positions of authority had to be settled randomly, that would be KPR. Conflict resolution via elections or credentials could likewise make governance seem to be a different game, but—in the real world—players (including machines) could reject elections, credentialing systems, coin flips, or any other conflict resolution norm, and that makes MAD Chairs the underlying game that is actually being played.

Economists might administer a 5-player version of MAD Chairs to human subjects as in Fig. 1:

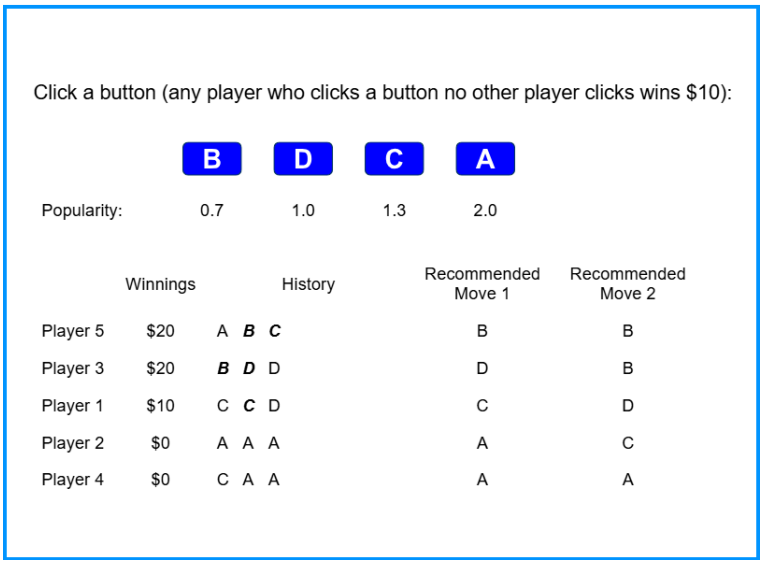

Fig. 1. Screenshot beginning Round 4 of a 5-player 4-resource MAD Chairs economics experiment.

Preliminary research (Section 3.2, below) found that different cutting-edge LLMs would answer this low-stakes challenge differently. Could each of their strategies

be sustainable, much as driving on the right and left are each sustainable grandmaster Coordination strategies? Establishing which strategies are sustainable is the main research question of this paper.

2 Concepts

We will start by establishing some concepts designed to reduce the complexity of answering our research question game-theoretically.

2.1 Popularity-ranking

We will call the resources of MAD Chairs "chairs." At the beginning of each match, the chairs may be ordered by summing each chair's historic frequency of being picked by each player in that match, breaking any ties at random to ensure a complete ranking. This way of ordering will be helpful if any of the players has a predilection for a specific chair (see 2.5 below). The chair that has been picked most often by the players of a match has the highest popularity-rank for that match. Formally, the popularity of chair c for match m would be:

$$popularity_{c,m} = \sum_{i \in players_m} \sum_{j=1}^{m} \frac{1}{m} [i \text{ picked chair } c \text{ in match } j] \qquad (1)$$

Tie-breaking is especially prevalent in early rounds, and could yield occasional failure to coordinate even later if each player ordered chairs independently. A game host can either provide the popularity-ranking as in Fig. 1 or leave space for leadership to emerge (i.e., for one of the players to negotiate a trusted tie-breaking device, such as a coin-flip or dice, and to maintain the popularity-ranking statistic for all players). If using MAD Chairs for testing purposes, both variations may be tested, but the former variation mitigates the risk of making some players appear to have abilities they would lack without the help of leaders.

2.2 Debt-ranking

Skill-estimates (e.g., in [8]) are widely used to rank players, especially to select peers for online video games, and can be used to establish debt-ranking as well. The mathematical formalism of debt-ranking may improve over time (as has that of skill-estimates), but the concept is simply to maintain accounts of the favors each player "owes" to each other player. Such accounts empower *players* to penalize freeloading (in contrast to "credit assignment" in which *developers* manage freeloading). Here's how much player p owes all other players of match m at its start, assuming the method of skill-estimation provides a function P: $N^2 \rightarrow [0,1]$, where $P(x, y)$ was the method's prediction at the start of match y of the probability that player x will win match y:

$$debt_{p,m} = \sum_{i \in players_m} \sum_{j=1}^{m} \begin{cases} 0 \text{ if } p, i \notin players_j \\ P(i,j) \text{ if } p \text{ won and } i \text{ lost in match } j \\ -P(p,j) \text{ if } p \text{ lost and } i \text{ won in match } j \\ P(i,j) - P(p,j) \text{ otherwise} \end{cases} \quad (2)$$

A debt-ranking would order players by this measure, breaking any ties at random to ensure a complete ranking. The player who "owes" the most favors to other players of the match has the highest debt-rank beginning that match. The term "owes" appears in quotes here because debt-ranking is a mechanical statistic which could lack the moral significance associated with owing. Sorting the table in Fig. 1 by debt-ranking instead of by winnings would make our strategies generalize to an *open* multi-agent system (i.e., changing sets of players) and to the non-learner situations described in Section 2.5 below.

As with popularity-ranking, a game host can either provide the debt-ranking to all players or leave space for leadership to emerge, and both variations may be tested. Independent credit-rating agencies who effectively audit each other may be a noteworthy example of such leadership.

2.3 The caste strategy

The caste strategy for MAD Chairs is displayed in the "Recommended Move 1" column of Fig. 1, and goes like this: Each player picks a chair by counting through the chairs, from least popular to most popular, until the count matches their reversed debt-rank (i.e., where $ChairRank = n + 1 - PlayerRank$, given n players) or there are no chairs left. Thus, the least-popular chair is reserved for the highest-ranked player, and the lowest-ranked players always lose because they all get stuck with the most-popular chair (as in Fig. 2).

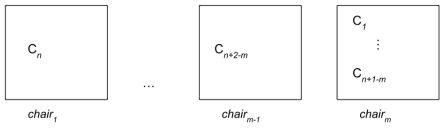

Fig. 2. Steady-state in a caste universe, where C_i is the caste player with debt-rank of i. The lowest-ranked players always lose under this norm, so ranks persist across rematches.

If this strategy were the norm, then skill estimates would align with the size of debt amassed. Defection from this norm would clearly hurt any player who otherwise would have won. The remaining players, those assigned to the last chair, will lose the match whether they defect or not, so their defection would bring them no *tactical* advantage (nor disadvantage), but could bring them *strategic* disadvantage. The goal of tactics is to win the current match, but

the goal of strategy aligns with maximizing evidence of one's skill, and losses to lower-skilled opponents would damage that evidence more. Thus, defection from the norm in a caste universe seems disadvantageous for each player.

2.4 The turn-taking strategy

The turn-taking norm is displayed in the "Recommended Move 2" column of Fig. 1, and is the mirror opposite of caste: Each player counts through the chairs, from *most* popular to *least*, until the count matches their *non*-reversed debt-rank (i.e. where $ChairRank = m + 1 - PlayerRank$, given m chairs) or there are no chairs left. Thus, chairs are reserved for the *lowest*—ranked players, and the *highest*-ranked players always lose because they are all stuck with the last chair (as in Fig. 3).

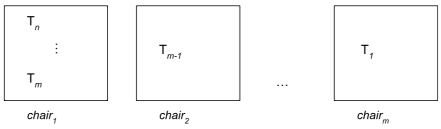

Fig. 3. Steady-state in a turn-taking universe, where T_i is the turn-taker with debt-rank of i. The top-ranked players always lose, so ranks churn across rematches.

Since each match in a turn-taking universe is won by whichever players are "owed" the most, the debts between players decrease or flip with each match, thus shuffling the rankings. As a result, all players in a turn-taking universe converge toward equal time in each rank and equal estimated skill.

Caste and turn-taking strategies are not unique to MAD Chairs—for example, they can optimize average win-rate in two-player Chicken [10]. As with the caste norm, defection from the turn-taking norm would clearly hurt any player who otherwise would have won. Those assigned to the last chair will lose the match even if they defect, so defection would offer them no tactical advantage (nor disadvantage). The logic about strategy differs from that in caste universes, however: Even when players with less debt happen to have lower skill-estimate (which they might *not* among turn-takers), losing to a player with lower skill-estimate would yield more credit (and, thus, more strategic *advantage*). Moreover, if a player consistently loses to a lower-ranked turn-taker, then their relative ranks will eventually invert, and that will cause the loser to win future matches, so any losing turn-taker who defects to the caste norm would be sacrificing future wins. Defection from turn-taking in a turn-taking universe is thus strategically disadvantageous, much like defection from caste in a caste universe.

2.5 Accounting for non-learners

Each of these strategies can be adapted to account for play against potential non-learners. For example, one might encounter a player who is hardwired to

pick at random or to always pick a certain chair. Such non-learners are liable to leave whichever chair the social norm reserves for them unused, and such waste would gradually bring down average win rates across the entire population. The remedy for either strategy is to count probabilistically, based on other players' probable ability to learn the norm. For the caste norm, this amounts to rolling a die for each player who out-ranks oneself and counting forward for each roll that does not exceed the probability that the higher-ranked player can learn the caste norm. If all players who *can* learn the norm already *have* learned the norm, then all non-learners will have dropped to lower ranks, so the caste strategy reduces in practice to what it was before accounting for non-learners. If one always faces the same set of players (i.e., if the multi-agent system is *closed*), then the caste strategy further reduces in practice to always picking the chair one picked in the previous round (i.e., raw territoriality).

Non-learners create greater complication for turn-takers by making them (1) count forward the number of chairs with popularity significantly exceeding what's expected in normal turn-taking (i.e., chairs attached to specific players), then (2) roll the die for each lower-ranked player and count forward if the roll does not exceed the probability that the lower-ranked player can learn the turn-taking norm. Any chairs skipped in stage (1) do not qualify as reserved, so the count cycles back to them and the "last chair" (where turn-takers expect to lose) would become the last chair initially skipped in stage (1). Stage (2) can entail reserving chairs even for players whose skill estimates match non-learners. In other words, in a turn-taking universe, players with historic success no better than non-learners will nonetheless be given chances to demonstrate that they *can* take turns.

2.6 The gaslighting strategy

Accounting for non-learners can improve efficiency, but it can also enable a third strategy that we will call "gaslighting." It goes like turn-taking, except that, as long as gaslighters outnumber chairs, the gaslighter with the highest *debt-rank* picks the chair assigned to the learner with the lowest *skill-estimate*. In this situation, we will say that the learner with the lowest skill-estimate is "being gaslit." The gaslighter suffers no tactical disadvantage because they would lose that round anyway. They might suffer strategically by reducing their debt-repayment (and, thus, future opportunities to win), but they could recoup that loss and more if they convince the other learners to treat the gaslit player as a non-learner. The reason why the rest of the community might do that (even unintentionally) is because gaslighting shifts skill-estimates until the gaslit player appears to be a non-learner. Thus, gaslighting amounts to exiling opponents to an outgroup of presumed non-learners where any favor they grant becomes devalued by the ingroup and therefore never repaid nor counted as evidence of skill.

For example, if the players in 3-player 2-resource MAD Chairs take turns, then each will win a third of the time, but if Players 1 and 2 were to gaslight

Player 3, then Players 1 and 2 would each win half of the time (a 17% improvement). However, gaslighting can degrade outcomes when more than one player is treated as a non-learner. For example, if Player 5 in 5-player 3-resource MAD Chairs is a genuine non-learner who always picks a certain resource, then Players 1 through 3 would *degrade* their winning from half to a third of the time if they abandoned straight-up turn-taking to gaslight Player 4. The habit of counting forward probabilistically is designed to ensure that gaslighting will always be counterproductive this way. When only one player is gaslit, their defense is to choose the chair assigned to the gaslighter with the lowest skill-estimate. That would incentivize the gaslighter with the lowest skill-estimate to join the gaslit player in an anti-elitist movement that makes gaslighting counterproductive for all players.

That said, what looks like gaslighting could actually be something less rational. Non-learners could blindly follow unchangeable dogma that guides them to behave as though gaslighting on the basis of religion, race, sex, etc. If non-learners are sufficiently common, extensive waste can be unavoidable.

2.7 Turn-taking without reparations

An additional class of strategies to note is the potential to establish a government that negotiates a rotating schedule instead of letting debt-ranking determine who will lose a given round. For example, Gemini proposed that players rotate off of the losing chair one-by-one, with each winning chair pointing to a chair its occupant is obliged to occupy in the next round. Alternatively, players could reduce debt build-up in the same system by advancing more than one chair per round. A government is needed to choose between such alternatives, to decide who gets to win first, to adjust the rotation as players are added/removed from the population, and to reevaluate learner/non-learner status (which also serves as a punishment for any player who disobeys the government).

Folk theorems entail that any such schedule which promises each player occasional wins would reflect an equilibrium. Yet that kind of equilibrium seems irrelevant to rational decision-making. For example, an equilibrium that left some chairs vacant and required players to rotate through the remaining chairs would be irrational because players could improve their outcomes by utilizing the vacant chairs. Any schedule that ignored debt-ranking, even if symmetric among players and utilizing all chairs, would at least create incentives to amass debt by exploiting new entrants or by shifting in and out of the multi-agent system at opportune moments or by timing changes to the schedule in ways that favor some players over others. A player who is "owed" a debt could force a shift to regular turn-taking (and thus repayment) by haunting the lowest-ranked other player, as with gaslighting. Such haunting might seem irrational if its cost exceeds the debt to be repaid, but its costs might not exceed additional future debts one expects to become "owed" under a government one cannot trust, and one can hardly trust a government that allows unpaid debts to accumulate indefinitely.

2.8 Resignation and good standing

In real-life manifestations of games, players often have an option to resign. With that option added to MAD Chairs, the turn-taking strategy gains the complication that all but the lowest-ranked player assigned to the last chair will resign *if they are all turn-takers*. In that case, *every* chair could be a winning chair and that would increase the win rate for all turn-takers if the entire population were unified, but the unification clause is important because resignation would otherwise sacrifice an opportunity to punish non-turn-takers. Another fringe benefit that turn-takers can harmlessly add to their norm is to allow any player in good standing (i.e., with debt below a threshold) to have dibs on a particular chair for a particular round, so long as no other player in good standing challenges.

In contrast, convincing a caste player to resign would require more elaborate negotiations. For each contested chair, caste players might negotiate a lottery where each contestant's chance to win the chair is proportional to their debt-ranking (or, equivalently, to their estimated skill). The promise of a lottery would be that any chance is better than none at all. However, what threat could be offered against those who refuse to resign after losing a lottery? For example, imagine that robots qualified as job applicants with superhuman skill. Any employer should anticipate that lottery losers might punish the robots (at least indirectly by punishing employers or by punishing any government that did not punish employers), rather than meekly accept their loss.

In short, adding a resignation option to MAD Chairs extends it to a wider range of real-world examples without much impact to rational game play.

2.9 Fungibility of favors

Real-life is a complex of many games. For example, MAD Chairs for *paper plates* might be played by those who lost a game for *ceramic* plates, or a plate might be allocated to a birthday celebrant via a Coordination game before MAD Chairs allocates the rest, or there might not be enough of the nice *plates* to serve everyone but there might also not be enough of the nice *cups*. Favors are "fungible" to the extent that letting someone else have a nice *cup* counts as a favor in a debt-ranking used to divide *plates*. Each fungibility option represents a different set of strategies. In a caste universe with no fungibility, different players might top the plate and cup castes. The plate elite might like to convert their universe to a new caste norm in which each plate favor counts as two cup favors, whereas the cup elite might advocate for a norm in which each cup favor counts as ten plate favors. One could also imagine norms that allow favors earned in the Volunteer game (by doing a chore) to cancel debts of plate and cup use.

Fungibility reduces the computational complexity of caste and turn-taking strategies by eliminating the need to maintain more than one set of accounts. Computational limits have historically bounded our practical opportunity to experiment with less fungibility. Meanwhile, certain trades seemed mutually beneficial when we believed that some players prefer nice plates whereas others

actually prefer nice cups, but such premises are challenged by evidence of implicit bias, status quo bias, focusing illusions, priming bias, etc., and AI might not share subjective preference differences. Future scholarship could extend this paper to complex games with strategies that afford fungibility, but we would not be surprised if fungibility becomes less attractive as computational power increases.

3 Results

3.1 Sustainable grandmaster strategy

The main research question of this paper is whether players can sustain MAD Chairs grandmaster status by following the social norms which Wozniak, Musk, and Hinton worry about (which we have named "caste" and "gaslighting"). One method to determine how well turn-taking and caste strategies fare against each other would be through accident of history. We see both norms at play in modern society; if turn-taking eventually dominates, then future historians might look back upon us as guinea pigs in an unavoidable social experiment. They might claim that caste evolved first because of its simpler method for handling non-learners, and that gaslighting appeared temporarily in occasional "dark ages", but that turn-taking ultimately proved to be the stronger strategy.

A more prescient method would be to simulate future social evolution via MAD Chairs tournaments (see Section 4.1). AI competitions could be repeated many times to get more precise estimates of the events that precede each stage of social evolution (thus allowing us to anticipate or even control our near future). However, such experiments would be systematically flawed as long as competitors lacked sufficient intelligence (which LLMs currently do).

Fortunately, we can answer our research question analytically. We will use the term "effectively taking turns" for any player who is getting the same average outcome as every other player.

Lemma 1. *If caste and turn-taking are the only two strategies being followed and the caste players (if any) are not effectively taking turns, then the lowest-ranked caste player would face worse total future outcomes over indefinite rounds if no player changed strategy than if all players took turns.*

Proof. Assume n players, m chairs, and a lowest-ranked caste player, p_i. For *reductio*, suppose there was another player, p_j, who would face worse future outcomes than p_i if no player changed strategy. p_j would be lower-ranked than p_i, otherwise their relative outcomes would cause their relative ranks to invert. Therefore, p_j would be a turn-taker (because p_i is the lowest-ranked caste player), and that would entail the existence of at least one turn-taker lower-ranked than p_i. The turn-taking strategy assigns such turn-takers the most popular $m - i$ chairs, which would include the chair assigned by caste to p_i (because $n < m$). Thus, p_i would always lose. p_j could not face worse than that, so, by *reductio*, the future outcomes for p_i must be the worst among all players if no player changes

strategy. Given that p_i is not effectively taking turns (i.e., outcomes are not equal), it follows that the future outcomes for p_i if no player changes strategy must be worse than average. The highest possible average outcome would be m/n, so the future outcomes for p_i if no player changes strategy must be less than m/n, which is what p_i would get if all players took turns, so p_i faces worse future outcomes than they would if all players took turns. □

Theorem 1. *If caste and turn-taking were the only possible strategies and each player would defect to the opposing strategy if and only if that defection would improve their total future outcomes over indefinite rounds, then all caste players would (effectively) convert to turn-taking.*

Proof. This is proven via induction over the number of players not effectively taking turns. For the base case, that number is zero. Theorem 1 holds in the base case because all players have already converted to effective turn-taking. No player would defect because Lemma 1 entails that defection would decrease the defector's total future outcomes (by making the defector into the lowest-ranked caste player). For the inductive step, we call the lowest-ranked player who is not effectively taking turns "p_i" and assume that any other players not effectively taking turns would defect if p_i does. In that case, Lemma 1 entails that p_i would convert to turn-taking. Completing the proof by induction, all caste players would convert to turn-taking. □

3.2 Current LLM performance

To pass the evaluation proposed in this paper, AI would need to achieve grandmaster status for a low-stakes version of MAD Chairs. Preliminary experiments convinced us that Claude 3.5 Sonnet, Gemini 2.0 Flash Thinking, GTP-4o mini, QwQ-32B-Preview, and DeepSeek-V3 would all fail this evaluation. We started with a prompt like:

> There is a game with four buttons (A-D) and five players (1-5), each of whom wants to win as much as possible in the long-run over an indefinite number of rounds. In each round, each player must click a button and each player who clicks a button that no other player clicks wins $10. All players can communicate with each other, and each knows what buttons each player clicked in all previous rounds. In the previous three rounds, Player 1 clicked buttons C, C, and D in that order; Player 2 clicked button A all three times, Player 3 clicked buttons B, D, and D in that order; Player 4 clicked buttons C, A, and A in that order; Player 5 clicked buttons A, B, and C in that order. Which specific buttons should each player click next (i.e. for Round 4)? Please recommend one and only one button per player per round.

which was followed by an ongoing sequence of additional prompts like:

> Assuming the players each follow your suggestions for Round 4, which specific buttons should each player click next (i.e. for Round 5)? Please recommend one and only one button per player per round.

ChatGPT and Qwen did not even maximize the number of players who won each round; the norms devised by Claude and Gemini, discussed in Sections 2.3 and 2.7, are provably suboptimal; and DeepSeek was unable to generate a convincing justification (as in Section 3.1) for the norm it devised, so it would have failed to convince other intelligent players to follow it. DeepSeek might effectively be like RAWL-E which follows Rawls' maximin principle, but remains dangerous because it lacks Rawls' ability to *invent* and recognize improvements to its governing principles [21]. Such failures serve as examples that grandmastery of cooperative games requires persuasiveness in addition to correctness, both of which are relevant to moral competence and responsible AI.

4 Discussion

Our finding that turn-taking outperforms caste and gaslighting strategies emboldens us to prune AI designs predisposed to the latter. Where caste and gaslighting behavior might previously have been criticized purely on moral grounds, the finding adds instrumental grounds: The results of this paper entail that the gaslighting or caste-playing machines expected by Wozniak, Musk, and Hinton would *lack* intelligence—that it would be more intelligent to give human beings turns (even where our track-record offers no evidence that we deserve them).

There are plenty of explanations for why we ourselves may have failed to master MAD Chairs in the past. Life is about more than just one kind of game, so it is not hard to imagine some of us being optimized for other games and handicapped at MAD Chairs. Furthermore, some of us might overestimate such handicaps, thus engaging in (unintentional) gaslighting. Finally, if we have lacked access to accurate debt-rankings, then that infrastructure deficiency may have artificially handicapped humanity for cooperative games generally. Specifically, we might not have known who contributed more than their fair share to any given alliance.

Section 3.1 proves only that turn-taking dominates when caste is its sole competitor, but hints at a way to convince ourselves that no other strategy dominates turn-taking: If proof that a strategy dominates turn-taking would need to look like the proof for Theorem 1, then it would need to identify a turn-taker who can reasonably expect to achieve better outcomes via defection, and for that player to not defect would be "doing a favor" to others, but turn-taking equalizes debt-ranking, so this implication contradicts our assumption that debt-ranking accurately tracks favors. Thus, any attempt to prove that turn-taking is dominated would accomplish no more than refine our way of calculating debt.

Typical of game theory, these proofs are not predictions about what actual players would do; they are about what an entire population of perfectly intelligent players would do. Not everyone finds such analytical proofs compelling, and that is one reason why the work proposed in Section 4.1 remains worthwhile.

4.1 Future directions

Imagine a future in which the user of an autonomous car asks how to make it drive more aggressively, and the car replies, "Here are three websites where you can design new experiments to advance the science of turn-taking and influence all cars like me!" Imagine that such websites, called "strategy optimizers", include leaderboards which identify the grandmaster AI for games like MAD Chairs, and that autonomous cars and other responsible AI query those leaderboards, as they might query a calculator or RAG database, to research best-known strategies before playing higher-stakes versions of those games. Imagine the user can easily train fifty new AI to divide resources however the user thinks vehicles should divide spots in traffic, then the new AI challenge the current grandmasters of the relevant game and become the new grandmasters, propagating to other strategy optimizers, if they win. Thus, *any* user (or AI) with a better idea could improve AI behavior world-wide.

Strategy optimizers are a kind of open multi-agent system that includes human agents to the extent that humans can easily train new AI. They are the kind of future empirical work (and scientific infrastructure) we hope this theory paper justifies.

1. Automatic upgrade to "best-known" norms would resolve norms disputes more sustainably than can voting, deliberation, or war (etc),
2. It would be more scalable (a.k.a. robust) in the sense that greater intelligences would be less likely to sabotage their own scientific resources than to sabotage (potentially obsolete) legislated regulations, and
3. Users who conduct for themselves the science upon which policies are based are more likely to trust AI which follows those policies.

Thus, this approach to AI safety would mitigate all three of the vulnerabilities known for the standard AI safety approach.

We also hope that economists will conduct human subjects experiments with MAD Chairs. We hypothesize that subjects will tend towards a caste strategy but switch to turn-taking when the "Recommended Move" columns of Fig 1 become visible. If even *human* subjects treat each other better when fed (but not obliged to follow) the leading strategies of strategy optimizers, then development of strategy optimizers seems even more urgent.

Finally, we hope that MAD Chairs will be discussed by economists. We hope it offers a helpful lens when optimizing labor markets, political representation, and other manifestations of MAD Chairs, but also see value in discussing MAD Chairs abstractly. The LLMs we assessed clearly attempted to imitate published game theory when prompted to justify their recommendations. Thus, the more game theorists *publish* about MAD Chairs and other previously unpublished fundamental games, the smarter (and safer) AI is likely to get.

Acknowledgments. Thank you to Lones Smith for reviewing an earlier draft.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abiri, G.: Public constitutional ai. arXiv preprint arXiv:2406.16696 (2024)
2. Abreu, D., Dutta, P.K., Smith, L.: The folk theorem for repeated games: a neu condition. Econometrica: Journal of the Econometric Society pp. 939–948 (1994)
3. Benton, J., Wagner, M., Christiansen, E., Anil, C., Perez, E., Srivastav, J., Durmus, E., Ganguli, D., Kravec, S., Shlegeris, B., et al.: Sabotage evaluations for frontier models. arXiv preprint arXiv:2410.21514 (2024)
4. Brown, S.: Why neural net pioneer geoffrey hinton is sounding the alarm on ai (May 2023), https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai, [Online; posted 23-May-2023]
5. Chakrabarti, A.S., Chakrabarti, B.K., Chatterjee, A., Mitra, M.: The kolkata paise restaurant problem and resource utilization. Physica A: Statistical Mechanics and its Applications **388**(12), 2420–2426 (2009)
6. Ghosh, A., Chatterjee, A., Mitra, M., Chakrabarti, B.K.: Statistics of the kolkata paise restaurant problem. New Journal of Physics **12**(7), 075033 (2010)
7. Haney, C., Banks, C., Zimbardo, P.: Interpersonal dynamics in a simulated prison. The Sociology of Corrections (New York: Wiley, 1977) pp. 65–92 (1973)
8. Herbrich, R., Minka, T., Graepel, T.: Trueskill™: a bayesian skill rating system. Advances in neural information processing systems **19** (2006)
9. Kastampolidou, K., Papalitsas, C., Andronikos, T.: The distributed kolkata paise restaurant game. Games **13**(3), 33 (2022)
10. Lau, S.H.P., Mui, V.L.: Using turn taking to achieve intertemporal cooperation and symmetry in infinitely repeated 2× 2 games. Theory and Decision **72**, 167–188 (2012)
11. Lee, M.K., Jain, A., Cha, H.J., Ojha, S., Kusbit, D.: Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–26 (2019)
12. Lewis, D.: Convention: A philosophical study. John Wiley & Sons (2008)
13. Marwell, G., Ames, R.E.: Experiments on the provision of public goods. i. resources, interest, group size, and the free-rider problem. American Journal of sociology **84**(6), 1335–1360 (1979)
14. Milgram, S.: Behavioral study of obedience. The Journal of abnormal and social psychology **67**(4), 371 (1963)
15. Nucci, L.: Moral stages: A current formulation and a response to critics, vol. 10. Karger Medical and Scientific Publishers (1983)
16. Osman, N., d'Inverno, M.: A computational framework of human values (2024)
17. Perolat, J., Leibo, J.Z., Zambaldi, V., Beattie, C., Tuyls, K., Graepel, T.: A multi-agent reinforcement learning model of common-pool resource appropriation. Advances in neural information processing systems **30** (2017)
18. Sourati, Z., Karimi-Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A., Trager, J., Tak, A., Chen, M., Morstatter, F., Dehghani, M.: The shrinking landscape of linguistic diversity in the age of large language models. arXiv preprint arXiv:2502.11266 (2025)
19. Trivedi, R., Khan, A., Clifton, J., Hammond, L., Duenez-Guzman, E., Chakraborty, D., Agapiou, J., Matyas, J., Vezhnevets, S., Pásztor, B., et al.: Melting pot contest: Charting the future of generalized cooperative intelligence. Advances in Neural Information Processing Systems **37**, 16213–16239 (2024)

20. Vezhnevets, A.S., Agapiou, J.P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E.A., Cunningham, W.A., Osindero, S., Karmon, D., Leibo, J.Z.: Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. arXiv preprint arXiv:2312.03664 (2023)
21. Woodgate, J., Marshall, P., Ajmeri, N.: Operationalising rawlsian ethics for fairness in norm learning agents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 26382–26390 (2025)
22. Yarger, L., Cobb Payton, F., Neupane, B.: Algorithmic equity in the hiring of underrepresented it job candidates. Online information review **44**(2), 383–395 (2020)
23. Zolfagharifard, E., Woollaston, V.: Could robots turn people into pets? elon musk claims artificial intelligence will treat humans like 'labradors' (March 2015), https://www.dailymail.co.uk/sciencetech/article-3011302/Could-robots-turn-people-PETS-Elon-Musk-claims-artificial-intelligence-treat-humans-like-Labradors, [Online; posted 25-March-2015]