Fairness and Stability in Allocation-Induced Hedonic Games

Ayşe Mutlu Derya
TOBB University of Economics and Technology
aysederya@etu.edu.tr

Abstract

Understanding when fair allocation mechanisms lead to stable coalition structures is fundamental for designing robust multi-agent systems. We investigate the compatibility between stable allocations in transferable utility (TU) games and the stability of coalition structures in induced hedonic games, where agents' preferences over coalitions are derived from payoffs assigned by a fixed allocation rule applied to all subgames of the TU game. We analyze FX-FE strong Nash stability (SNS) in induced hedonic games, which captures a strong form of robustness under free-exit and free-entry conditions. We show that any efficient allocation rule ensuring core membership for the grand coalition induces a hedonic game that guarantees the existence of an FX-FE strong Nash stable partition. Examining the Shapley value, we further show that when it lies outside the core, the resulting hedonic game may or may not possess FX-FE strongly Nash stable partitions, highlighting a sensitive interaction between Shapley-based fairness and coalition-level stability. Our framework bridges SHAP methodology from explainable AI with hedonic coalition theory, providing theoretical foundations for understanding when fair allocation mechanisms shape coalition-level incentives in a way that ensures strategic stability in team formation and multi-agent systems.

1 Introduction

Game-theoretic concepts such as fairness and stability are receiving growing attention in the design of intelligent, cooperative systems. In domains ranging from multi-agent coordination to explainable AI (XAI), systems are expected to allocate value fairly while ensuring that agents have no incentive to deviate from their assigned coalitions. In cooperative game theory, fairness is often formalized via axiomatically characterized allocation rules such as the Shapley value, while stability is captured by solution concepts like the core or stable partitions in hedonic coalition formation games. These concepts, though historically distinct, now intersect in modern AI settings, particularly in multi-agent systems (MAS), where allocation mechanisms must satisfy both equity and strategic robustness.

Hedonic coalition formation games provide a foundational framework for modeling group formation and are widely applied in MAS [Chevaleyre *et al.*, 2006; Aziz *et al.*, 2013; Aziz and Savani, 2016; Jang *et al.*, 2018; Caskurlu and Eser, 2025]. In these games, agents evaluate coalitions solely based on their members, independently of how others are grouped, a property termed hedonic preferences [Drèze and Greenberg, 1980]. Formal foundations were developed in [Farrell and Scotchmer, 1988; Banerjee *et al.*, 2001; Bogomolnaia and Jackson, 2002]; for surveys, see [Hajduková, 2006; Sung and Dimitrov, 2007; Aziz and Savani, 2016]. This framework is particularly suitable for studying how allocation rules induce coalition preferences.

We study a structured class of hedonic games in which agents derive preferences over coalitions from payoffs assigned by a fixed allocation rule applied to an underlying transferable utility (TU) game. In such settings, allocation rules play a structural role by shaping agents' incentives to deviate or form coalitions, and stability emerges endogenously from the induced preferences.

We focus on FX-FE strong Nash stability [Karakaya, 2011] (FX = free exit, FE = free entry; also referred to as strong Nash stability (SNS) of [Aziz and Brandl, 2012]), a stability concept requiring that no group of agents can jointly deviate—by moving to existing coalitions, to the empty set, or by freely reassigning themselves in any arrangement (including splitting into subgroups that join different coalitions or the empty set)—to strictly improve their payoffs under free-exit and free-entry membership rights, without needing permission from any agent.

Our contributions are threefold. First, we bridge the XAI literature—particularly SHAP as an allocation-based explanation method—and the MAS literature on coalition stability via a unified framework of allocation-induced hedonic games. Second, we prove that efficient allocation rules ensuring core membership (defined formally in Section 1.2) for the grand coalition induce FX-FE strongly Nash stable partitions, even if subcoalition allocations lie outside their cores. Third, we show that when Shapley lies outside the core, stability may fail, revealing a sensitive fairness-stability interaction.

These results suggest that fairness and stability are neither inherently conflicting nor automatically aligned; rather, their relationship depends on how fairness notions interact with the underlying game structure and induced coalition dynamics. This leads to the central question of the paper: If fairness is not aligned with stability, should we still want it when designing multi-agent systems? Our results suggest that the answer naturally depends on whether fairness is embedded via allocation rules that preserve coalition-level incentives.

1.1 Cooperative Games with Transferable Utility (TU Games)

Let $N = \{1, 2, \ldots, n\}$ be a finite set of players with $n \geq 2$. Any non-empty subset $S \subseteq N$ is called a *coalition*. A *transferable utility game* (or *TU game*) is a pair (N, v), where $v : 2^N \to \mathbb{R}$ is a *characteristic function* that assigns a real-valued worth $v(S)$ to each coalition S, with the standard assumption that $v(\emptyset) = 0$. The value $v(S)$ represents the total utility the members of coalition S can achieve by cooperating [Moulin, 1988; Moulin, 2003].

For any coalition $S \subseteq N$, the *subgame* on S is the pair (S, v^S), where $v^S : 2^S \to \mathbb{R}$ is defined by $v^S(T) = v(T)$ for all $T \subseteq S$. Note that for $S = N$, $(N, v^N) \equiv (N, v)$.

A *payoff vector* for a coalition $S \subseteq N$ is a vector $x_S \in \mathbb{R}^{|S|}$ assigning a payoff to each player $i \in S$. A coalition $T \subseteq N$ is said to *block* a payoff vector $x_N \in \mathbb{R}^n$ if its members can obtain a better outcome on their own, i.e., if $v(T) > \sum_{i \in T} (x_N)_i$.

Definition 1 (Core of a Game)**.** *The core is a solution concept applicable to any TU game [Gillies, 1953]. It identifies the set of payoff vectors that are both efficient and stable against blocking by any subcoalition. Formally, the core of a TU game (S, v^S), denoted by $C(S, v^S)$, is the set:*

$$C(S, v^S) = \left\{ x_S \in \mathbb{R}^{|S|} \;\middle|\; \begin{array}{l} \underbrace{\sum_{i \in S} (x_S)_i = v(S)}_{\textit{Efficiency}} \textit{ and} \\ \underbrace{\sum_{i \in T} (x_S)_i \geq v(T)\,, \forall T \subseteq S}_{\textit{Stability}} \end{array} \right\}$$

The first condition states that the full value of coalition S is to be distributed, which is known as *efficiency*. The second condition ensures that no subcoalition can profitably secede and is known as *coalition rationality* or *stability*. The core of a game or any of its subgames may be empty.

1.2 Allocation Rules and Core Selection

While the core defines a set of stable allocations, an *allocation rule (or value)* provides a more decisive solution by selecting a single payoff vector for any given game. Formally, an *allocation rule* is a function Φ that maps a game (S, v^S) to a payoff vector $\Phi(S, v^S) \in \mathbb{R}^{|S|}$.

An allocation rule $\Phi(S, v^S)$ is *efficient* if $\sum_{i \in S} \Phi_i(S, v^S) = v(S)$ for all (S, v^S). Efficiency is a standard property required of most prominent allocation rules in the literature.

A key question is how a rule relates to the core. This relationship can be characterized by a strong, universal property or a weaker, local one.

Definition 2 (Core-Selective Rule)**.** *An allocation rule Φ is core-selective if, for any game (S, v^S) with a non-empty core, the assigned payoff vector is an element of the core. That is:*

$$C(S, v^S) \neq \emptyset \implies \Phi(S, v^S) \in C(S, v^S).$$

Being core-selective is a powerful, global property, as it must hold for all games. Many prominent rules, like the Shapley value, are not core-selective. This motivates a weaker, localized notion.

Definition 3 (Core Membership)**.** *An allocation rule Φ provides a core-member for a specific game (S, v^S) if its allocation for that game is an element of the core:*

$$\Phi(S, v^S) \in C(S, v^S).$$

The Shapley value, for example, is not core-selective in general, but it does provide a core-member in every convex game (S, v^S).

This distinction is critical. This highlights the difference between a rule being core-selective (holding across all games) and providing a core-member (for a single game). Our analysis relies on this weaker, game-specific property, and for brevity, we say that an allocation "lies in the core" of a specific game.

1.3 The Shapley Value and Convex Games

A central focus of our analysis is the *Shapley value* [Shapley, 1953], a prominent allocation rule uniquely characterized by four axioms: efficiency, symmetry, additivity and the null player property. It distributes the total worth of a coalition based on each player's average marginal contribution. Formally, for each player $i \in S$:

$$Sh_i(S, v^S) = \sum_{T \subseteq S \setminus \{i\}} \frac{|T|!(|S| - |T| - 1)!}{|S|!} [v(T \cup \{i\}) - v(T)].$$

Despite its strong fairness properties, the Shapley value is not core-selective, but it does provide a core-member for each convex game (S, v^S).

Definition 4 (Convex TU Game)**.** *A TU game (S, v^S) is convex if it exhibits increasing marginal returns, meaning a player's incentive to join a coalition grows as the coalition itself grows. An equivalent condition, known as supermodularity, requires that for any two coalitions $M, N \subseteq S$: $v(M) + v(N) \leq v(M \cup N) + v(M \cap N)$.*

A seminal result by [Shapley, 1971] connects these concepts, establishing that the Shapley value of a convex game always lies in the core. This, in turn, proves that convexity is a sufficient condition for a TU game to have a non-empty core.

Subsequently, the notion of convexity was generalized to *average convexity* by [Inarra and Usategui, 1993]. While we omit the technical details here, average convexity provides a weaker condition than convexity that still ensures the Shapley value is a core-member, thereby characterizing a broader class of games for which the core-membership property still holds.

1.4 Hedonic Games and Stability

The final component of our framework is the hedonic game, which models how agents form coalitions based on their own preferences. In a hedonic game, agents are concerned only with the members of their own coalition and do not consider how other agents form groups.

Formally, a *hedonic coalition formation game* (or simply a *hedonic game*) is a pair $(N, \succeq)$, where N is a finite set of players and $\succeq = (\succeq_i)_{i \in N}$ is a profile of preferences. For each player $i \in N$, the preference relation $\succeq_i$ is a complete and transitive ordering over $\Sigma_i = \{S \subseteq N \mid i \in S\}$, the set of all coalitions that contain the player i. As usual, for each player i, strict preferences are denoted by $\succ_i$ and indifference by $\sim_i$, both derived from the weak preference relation $\succeq_i$. Formally, for each player i, the strict preference relation $\succ_i$ is defined by $x \succ_i y$ if and only if $x \succeq_i y$ and not $y \succeq_i x$, while indifference $\sim_i$ holds when $x \succeq_i y$ and $y \succeq_i x$.

A central objective in hedonic games is to identify stable partitions. The stability of a partition is evaluated based on its resilience to deviations. While many such stability concepts exist, we focus on three foundational and related notions: Nash stability, core stability, and the more general FX-FE strong Nash stability.

Individual Deviations and Nash Stability

The most basic deviation is a movement by a single player. When a player i leaves her current coalition $\pi(i)$ and joins another existing coalition $H \in (\pi \cup \{\emptyset\})$ (or forms a new coalition by herself by joining the empty set), she induces a new partition π'. We say that π' *is reachable from* π *via player* i, denoted $\pi \xrightarrow{i} \pi'$. A player $i \in N$ *Nash blocks* a partition π if there exists a coalition $H \in (\pi \cup \{\emptyset\})$ such that $H \cup \{i\} \succ_i \pi(i)$.

Definition 5 ((FX-FE) Nash Stability). *A partition π is FX-FE Nash stable if there does not exist a player who Nash blocks it. This corresponds to the standard definition of Nash stability in the literature, where a player can freely exit her coalition and freely enter another without asking permission.*

Group Deviations and Stronger Stability Concepts

More powerful stability notions allow coordinated deviations by groups of players. A classical example is core stability, which permits a group to deviate only by jointly forming a new stand-alone coalition, i.e., each agent in the group leaves their current coalition and then they form a single new coalition among themselves.

Definition 6 (Core Stability). *A partition π is core stable if there does not exist a coalition $T \subseteq N$ such that for all $i \in T$, $T \succ_i \pi(i)$. If such a coalition T exists, then we say that T blocks π.*

Although core stability is a powerful concept, it imposes a restriction: any blocking group must form a new stand-alone coalition and is not permitted to join existing ones or exchange their current coalitions. Our work instead focuses on a more general and demanding stability notion introduced by [Karakaya, 2011].

To formalize this, we define a *movement* by a group $H \subseteq N$, which generalizes the notion of a single-player movement. A partition π' $(\pi' \neq \pi)$ is *reachable from π by movements of the players in H*, denoted by $\pi \xrightarrow{H} \pi'$, if, for all $i, j \in (N \setminus H)$ with $i \neq j$, $\pi(i) = \pi(j) \Leftrightarrow \pi'(i) = \pi'(j)$. That is, $(\pi \xrightarrow{H} \pi')$ if each player not in H (a non-deviating agent) stays grouped with all their previous partners who are also non-deviators, while the players in H are free to leave their current coalitions and freely reorganize in any way they want. Based on this, we define our central group stability concept.

Definition 7 (FX-FE Strong Nash Stability). *A partition π is FX-FE strongly Nash stable if there does not exist a pair (π', H), where $\pi' \neq \pi$ and $\emptyset \neq H \subseteq N$, such that:*

(i) $\pi \xrightarrow{H} \pi'$ (π' is reachable from π by movements of H), and

(ii) for all $i \in H$, $\pi'(i) \succ_i \pi(i)$ (all agents in the deviating coalition H be strictly better off.)

If such a pair (π', H) exists, then we say that H FX-FE strongly Nash blocks π.

This stability concept is particularly strong, which we refer to as *FX-FE strong Nash stability*. It is equivalent to the strong Nash equilibrium of a non-cooperative game induced by the hedonic game. Moreover, any FX-FE strongly Nash stable partition is both core and Nash stable. However, a partition that is either core or Nash stable may not be FX-FE strongly Nash stable (see Example 1).

1.5 Model: Allocation-Induced Hedonic Games

We consider a setting that connects transferable utility (TU) games with hedonic games by deriving each player's preferences over coalitions not abstractly, but endogenously from a fixed allocation rule. Specifically, we focus on a class of hedonic games in which each player's preferences over coalitions are induced by the payoff they would receive from applying an allocation rule to the relevant subgame.

Given a TU game (N, v) and an allocation rule Φ, player $i \in N$ compares any two coalitions $S, S' \in \Sigma_i$ based on the payoff they would receive from Φ in each corresponding subgame (similar to [Magaña and Carreras, 2018]).

Formally, a TU game (N, v) and an allocation rule Φ induce a preference profile $\succeq^{(v,\Phi)}$ as follows: for each player $i \in N$ and any two coalitions $S, S' \in \Sigma_i$:

$$S \succeq_i^{(v,\Phi)} S' \quad \text{if and only if} \quad \Phi_i(S, v^S) \geq \Phi_i(S', v^{S'}).$$

For simplicity, we henceforth write $\succeq_i$ to denote the induced preference relation. All previously introduced stability concepts—namely, Nash stability, core stability, and FX-FE strong Nash stability—are also well-defined within the context of the allocation-induced hedonic games.

Example 1. ***(A motivating example)*** *The following example can be interpreted as a heterogeneous multi-agent system in which agents differ in capabilities or roles, with two pairs of agents sharing similar characteristics. Such structures arise naturally in team formation and collaborative multi-agent systems where partnerships differ in productivity. Consider the following 4-player TU game with $N = \{1, 2, 3, 4\}$: $v(\{1, 3\}) = 8$, $v(\{1, 4\}) = 10$, $v(\{2, 3\}) = 14$, $v(\{2, 4\}) = 12$, and $v(S) = 0$ otherwise.*

To determine the induced hedonic game via Shapley value and analyze its stability, we compute Shapley values of each player at each possible subgame to drive preferences. Shapley values for all non-singleton, non-trivial subgames are shown in Table 1.

S	$Sh_i(S, v^S)$			
	$i=1$	$i=2$	$i=3$	$i=4$
$\{1,2,3,4\}$	$-2/3$	$2/3$	0	0
$\{1,2,3\}$	$-10/3$	$-1/3$	$11/3$	–
$\{1,2,4\}$	$-7/3$	$-4/3$	–	$11/3$
$\{1,3,4\}$	3	–	-2	-1
$\{2,3,4\}$	–	$13/3$	$-5/3$	$-8/3$
$\{1,3\}$	4	–	4	–
$\{1,4\}$	5	–	–	5
$\{2,4\}$	–	6	–	6
$\{2,3\}$	–	7	7	–

Table 1: Shapley values $Sh_i(S, v^S)$ of subgames v^S.

These payoffs induce the following preference orderings:

$\succeq_1$: $\{1,4\} \succ_1 \{1,3\} \succ_1 \{1,3,4\} \succ_1 \{1\} \sim_1 \{1,2\} \succ_1 \{1,2,3,4\} \succ_1 \{1,2,4\} \succ_1 \{1,2,3\}$

$\succeq_2$: $\{2,3\} \succ_2 \{2,4\} \succ_2 \{2,3,4\} \succ_2 \{1,2,3,4\} \succ_2 \{2\} \sim_2 \{1,2\} \succ_2 \{1,2,3\} \succ_2 \{1,2,4\}$

$\succeq_3$: $\{2,3\} \succ_3 \{1,3\} \succ_3 \{1,2,3\} \succ_3 \{3\} \sim_3 \{3,4\} \sim_3 \{1,2,3,4\} \succ_3 \{2,3,4\} \succ_3 \{1,3,4\}$

$\succeq_4$: $\{2,4\} \succ_4 \{1,4\} \succ_4 \{1,2,4\} \succ_4 \{4\} \sim_4 \{3,4\} \sim_4 \{1,2,3,4\} \succ_4 \{1,3,4\} \succ_4 \{2,3,4\}$

With the preference profile above:

(i) $\pi_1 = \{\{1,3\},\{2,4\}\}$ *is FX-FE Nash stable, but neither FX-FE strongly Nash stable nor core stable: No single agent can move to an existing coalition or to a singleton and strictly improve, so FX-FE Nash stability holds. However, it fails FX-FE strong Nash and core stability since players* 2 *and* 3 *both strictly prefer to form* $\{2,3\}$.

(ii) $\pi_2 = \{\{2,3\},\{1,4\}\}$ *is FX-FE strongly Nash stable, hence also FX-FE Nash and core stable: Except player* 4*, every other player prefers their top coalitions. Player* 4 *may prefer to form* $\{2,4\}$*, but player* 2 *strictly prefers her current coalition to* $\{2,4\}$*, thus refuses. Since no deviating coalition is unanimously preferred by its members,* π_2 *is FX-FE strongly Nash stable, and thus FX-FE Nash and core stable.*

(iii) Every other partition fails FX-FE Nash stability, and therefore also fails core stability and FX-FE strong Nash stability, as FX-FE Nash stability is weaker than both: In any other partition, at least one lone player has a strictly preferred existing coalition and can deviate unilaterally, violating FX-FE Nash stability and consequently also the stronger notions.

2 Guaranteed Stability from Core-membership

We emphasize the precise requirements placed on the allocation rule Φ. Our main condition requires that the allocation for the grand game, $\Phi(N, v)$ lies in the core $C(N, v)$. However, for any proper subgame (S, v^S), we do not require the corresponding allocation $\Phi(S, v^S)$ to lie in the subgame's core $C(S, v^S)$. We only impose the standard condition of efficiency on all subgames. This distinction is key, as it shows that hedonic stability can be achieved even if the allocation rule does not provide core-membership within smaller, deviating coalitions. The following theorem identifies conditions under which allocation-induced fairness notions are compatible with FX-FE strong Nash stability induced by allocation-based preferences.

Theorem 1. *Let* Φ *be an allocation rule that is efficient on all subgames of a game* (N, v) *and be a core-member for* (N, v)*. In the induced hedonic game* $(N, \succeq)$ *derived from* (N, v) *and* Φ*, the partition consisting of the grand coalition,* $\pi = \{N\}$*, is an FX-FE strongly Nash stable partition.*

Proof. We prove by contradiction. Assume that under the conditions of the theorem, the grand coalition $\pi = \{N\}$ is not an FX-FE strongly Nash stable partition in the induced hedonic game $(N, \succeq)$.

By the definition of FX-FE strong Nash stability, this assumption implies that there exists a deviating coalition $H \subsetneq N$ and a partition π' ($\pi' \neq \pi$) such that $\pi \xrightarrow{H} \pi'$ and for all $i \in H$, $\pi'(i) \succ_i \pi(i)$. The only possible movements for H in $\pi = \{N\}$ are for agents in H to leave N and form their own coalition H, or possibly form a partition among themselves—which includes the former case as well—, i.e., $\pi' = \{H_1, H_2, \dots H_k, N \setminus H\}$ where $\{H_1, H_2, \dots, H_k\}$ is a partition of H for some $k \geq 1$. Thus, by definition, for each $j \in \{1, \dots, k\}$ and for each $i \in H_j$,

$$\Phi_i(H_j, v^{H_j}) > \Phi_i(N, v). \tag{1}$$

For each $j \in \{1, \dots, k\}$, summing this inequality over all players i within a single subcoalition H_j, we get:

$$\sum_{i \in H_j} \Phi_i(H_j, v^{H_j}) > \sum_{i \in H_j} \Phi_i(N, v). \tag{2}$$

By the hypothesis of the theorem, the allocation rule Φ is efficient on all subgames. Applying this to the subgame (H_j, v^{H_j}), the left-hand side of (2) simplifies to:

$$\sum_{i \in H_j} \Phi_i(H_j, v^{H_j}) = v^{H_j}(H_j) = v(H_j). \tag{3}$$

Furthermore, we have that $\Phi(N, v)$ lies in the core $C(N, v)$. By the definition of the core, this allocation must satisfy the coalitional rationality constraint for all coalitions, including any H_j:

$$\sum_{i \in H_j} \Phi_i(N, v) \geq v(H_j). \tag{4}$$

Now, substituting (3) into (2) yields:

$$v(H_j) > \sum_{i \in H_j} \Phi_i(N, v).$$

This result directly contradicts the core condition in (4).

Therefore, the initial assumption must be false. No such deviating coalition H can exist. Thus, $\pi = \{N\}$ is an FX-FE strongly Nash stable partition. $\square$

Notably, core membership is required only for the grand coalition. Since FX-FE strong Nash stability is stronger than both Nash stability and core stability, our result also implies these weaker stability notions.[1] Theorem 1 confirms that core-membership criteria can translate into FX-FE strong Nash stability under our model.

By Theorem 1, we obtain the following corollaries. These results collectively show that the grand coalition partition, which is often considered to form from the perspective of cooperative game theory, also exhibits strong stability properties when evaluated through the lens of hedonic games derived from a fixed allocation rule.

Corollary 1. *Let Φ be an efficient core-selective allocation rule. Then for any TU game (N, v) with a non-empty core, $\pi = \{N\}$ is an FX-FE strongly Nash stable partition in the hedonic game induced by (N, v) and Φ.*

This result implies that any core-selective rule ensures strong stability of the grand coalition whenever the core is non-empty. Examples include the nucleolus [Schmeidler, 1969] and the per-capita nucleolus [Grotte, 1970; Moulin, 1988], both inducing FX-FE strongly Nash stable outcomes.

Corollary 2. *Let (N, v) be a TU game. If the Shapley value of (N, v) lies in the core of v, then in the hedonic game induced by (N, v) and the Shapley value, $\pi = \{N\}$ is an FX-FE strongly Nash stable partition.*

The Shapley value ensures stability whenever it lies in the core, as in average convex games [Inarra and Usategui, 1993]. Since it always lies in the core of convex games, we also have:

Corollary 3. *If (N, v) is a convex game, then in the hedonic game induced by (N, v) and the Shapley value, $\pi = \{N\}$ is an FX-FE strongly Nash stable partition.*

3 The Shapley Value and the Nuance of Stability

The preceding results establish a sufficient condition for stability: in particular, when the Shapley value lies in the core, the grand coalition is guaranteed to be FX-FE strongly Nash stable. We now examine cases where this condition fails, either because the Shapley value is outside the core or because the core is empty. As this section demonstrates, the stability guarantee breaks down and the outcome becomes highly sensitive to the specific structure of the game.

We first consider the case where the core is non-empty but the Shapley value lies outside it. As the next result shows, stability need not be lost.

[1]Our result also applies to strictly strong Nash stability (SSNS) as defined by [Aziz and Brandl, 2012], which allows group deviations even if some agents are indifferent. Since our contradiction arises from a violation of core constraints by any subcoalition containing at least one strictly better-off agent, it also rules out such weakly blocking deviations. Furthermore, any partition that is stable under FX-FE cannot be blocked under weaker membership rights, including FX-AE (free exit–approved entry), AX-FE (approved exit–free entry), and AX-AE (approved exit–approved entry), as defined by [Karakaya, 2011]. Hence, our result extends to these variants of strong Nash stability notions as well.

Theorem 2. *There exists a TU game (N, v) such that:*

(i) The core $C(N, v)$ is non-empty.

(ii) The Shapley value $Sh(N, v)$ is not a core-member for (N, v) .

(iii) The hedonic game induced by (N, v) and the Shapley value has an FX-FE strongly Nash stable partition.

Proof. We prove this constructively with the classic 3-player glove game. Let $N = \{1, 2, 3\}$, where players 1 and 2 each has a right-hand glove and player 3 has a left-hand glove. A coalition achieves a worth of 1 if it contains a pair of gloves, and 0 otherwise. The characteristic function is therefore: $v(\{1, 3\}) = v(\{2, 3\}) = v(N) = 1$, and $v(S) = 0$ for all other coalitions S.

The core is non-empty, $C(N, v) = \{(0, 0, 1)\}$. The Shapley value for the grand game is $Sh(N, v) = (\frac{1}{6}, \frac{1}{6}, \frac{2}{3})$ which is clearly not a core-member. For subgame $S = \{1, 3\}$ or $S = \{2, 3\}$, $Sh(S, v^S) = (\frac{1}{2}, \frac{1}{2})$.

These payoffs induce the following preference orderings:

$\succeq_1$: $\{1, 3\} \succ_1 \{1, 2, 3\} \succ_1 \{1, 2\} \sim_1 \{1\}$

$\succeq_2$: $\{2, 3\} \succ_2 \{1, 2, 3\} \succ_2 \{1, 2\} \sim_2 \{2\}$

$\succeq_3$: $\{1, 2, 3\} \succ_3 \{1, 3\} \sim_3 \{2, 3\} \succ_3 \{3\}$

We now test the stability of the grand coalition: For a coalition T to deviate, all its members must strictly prefer T to N. However, while player 1 would prefer to form $\{1, 3\}$, player 3 would not. Similarly, while player 2 would prefer to form $\{2, 3\}$, player 3 would again refuse. Since no deviating coalition is unanimously preferred by its members, $\{\{1, 2, 3\}\}$ is an FX-FE strongly Nash stable partition. □

This result is crucial: it shows that the core-membership condition is sufficient for stability, but not necessary. However, the existence of such a game does not imply a general guarantee. The following complementary result demonstrates that, in a similar setting, the opposite outcome is also possible.

Theorem 3. *There exists a TU game (N, v) such that:*

(i) The core $C(N, v)$ is non-empty.

(ii) The Shapley value $Sh(N, v)$ is not a core-member for (N, v).

(iii) The hedonic game induced by (N, v) and the Shapley value does not have any FX-FE strongly Nash stable partition.

Proof. For this, we use an example adapted from [Magaña and Carreras, 2018].[2] Consider: $v(\{i\}) = 0$, $v(\{1, 3\}) = v(\{2, 3\}) = 1$, and $v(\{1, 2\}) = v(\{1, 2, 3\}) = 6$.

(i) The core is non-empty; e.g., $(3, 3, 0) \in C(N, v)$.

(ii) The Shapley value, $Sh(N, v) = (\frac{17}{6}, \frac{17}{6}, \frac{2}{6})$, is not in the core, because player 3's allocation is positive, whereas any core allocation requires $x_1 + x_2 = v(\{1, 2\}) = 6$, implying $x_3 = 0$.

[2]The hedonic stability notion of [Magaña and Carreras, 2018] differs from ours; in this example, the partition $\{\{1, 2\}, \{3\}\}$ is regarded as stable under their framework.

(iii) To analyze the induced hedonic game and its stability, we compute Shapley values for all non-singleton subgames to derive preferences, as shown in Table 2. These payoffs induce

S	$Sh_i(S, v^S)$		
	$i=1$	$i=2$	$i=3$
$\{1,2,3\}$	$17/6$	$17/6$	$1/3$
$\{1,2\}$	3	3	–
$\{1,3\}$	$1/2$	–	$1/2$
$\{2,3\}$	–	$1/2$	$1/2$

Table 2: Non-trivial Shapley values-to induce preferences.

the following preference orderings:

$\succeq_1$: $\{1,2\} \succ_1 \{1,2,3\} \succ_1 \{1,3\} \succ_1 \{1\}$

$\succeq_2$: $\{1,2\} \succ_2 \{1,2,3\} \succ_2 \{2,3\} \succ_2 \{2\}$

$\succeq_3$: $\{1,3\} \sim_3 \{2,3\} \succ_3 \{1,2,3\} \succ_3 \{3\}$

An analysis of these preferences shows that no partition is stable: The coalition $\{1,2\}$ is the top choice for both players 1 and 2, meaning any stable partition must contain it. This leaves only $\pi' = \{\{1,2\},\{3\}\}$. However, π' is itself unstable, since coalition $\{3\}$ would be better off by joining $\{1,2\}$. Thus, there is no FX-FE strongly Nash stable partition. □

Together, Theorems 2 and 3 establish a fundamental indeterminacy: when the Shapley value is not a core-member for (N, v), stability is possible but not assured. To complete our analysis, we now examine the most extreme case of instability in the underlying TU game: when the core is empty.

Theorem 4. *There exists a TU game (N, v) such that:*

(i) The core $C(N, v)$ is empty.

(ii) The hedonic game induced by (N, v) and the Shapley value has an FX-FE strongly Nash stable partition.

Proof. Consider the following 3-player game:
$v(\{1\}) = 1$, $v(\{2\}) = 2$, $v(\{3\}) = 0$, $v(\{1,2\}) = 5$, $v(\{1,3\}) = 3$, $v(\{2,3\}) = 2$, and $v(\{1,2,3\}) = \alpha$, where α is a real number such that $4 < \alpha < 5$.

(i) Core: The core $C(N, v)$ is empty. For instance, the balanced coalition collection $\mathcal{B} = \{\{1,2\},\{3\}\}$ with weights $\delta_{\{1,2\}} = 1, \delta_{\{3\}} = 1$ violates the Bondareva-Shapley [Bondareva, 1962; Shapley, 1967] condition, as $x_1 + x_2 + x_3 \geq \delta_{\{1,2\}} v(\{1,2\}) + \delta_{\{3\}} v(\{3\}) = 5 > \alpha = v(N)$.

(ii) Hedonic Game and Stability:

S	$Sh_i(S, v^S)$		
	$i=1$	$i=2$	$i=3$
$\{1,2,3\}$	$(\alpha+2)/3$	$(\alpha+2)/3$	$(\alpha-4)/3$
$\{1,2\}$	2	3	–
$\{1,3\}$	2	–	1
$\{2,3\}$	–	2	0
$\{1\}$	1	–	–
$\{2\}$	–	2	–
$\{3\}$	–	–	0

Table 3: Shapley values for each subgame for any α.

This induces the following preferences:

$\succeq_1$: $\{1,2,3\} \succ_1 \{1,2\} \sim_1 \{1,3\} \succ_1 \{1\}$,

$\succeq_2$: $\{1,2\} \succ_2 \{1,2,3\} \succ_2 \{2,3\} \sim_2 \{2\}$,

$\succeq_3$: $\{1,3\} \succ_3 \{1,2,3\} \succ_3 \{2,3\} \sim_3 \{3\}$.

We now test the stability of the grand coalition: For a coalition T to deviate, all its members must strictly prefer T to N. However, while player 2 would prefer to form $\{1,2\}$, player 1 would not. Similarly, while player 3 would prefer to form $\{1,3\}$, player 1 would again refuse. Since no deviating coalition is unanimously preferred by its members, $\{\{1,2,3\}\}$ is an FX-FE strongly Nash stable partition. □

Remark 1. *The stability outcome of this game is highly sensitive to the value of α. In all cases where $\alpha < 5$, the core of the underlying TU game is empty. In particular:*

(i) When $\alpha = 4$, the induced preferences differ from those in Theorem 4, yet $\{\{1,2,3\}\}$ remains to be an FX-FE strongly Nash stable partition.

(ii) When $1 < \alpha < 4$, the preferences differ from both case (i) and Theorem 4. In this range, the only FX-FE strongly Nash stable partitions are $\{\{1,2\},\{3\}\}$ and $\{\{1,3\},\{2\}\}$.

(iii) When $\alpha < 1$, the induced preferences again differ from all previous cases. Nevertheless, the only FX-FE strongly Nash stable partitions are again $\{\{1,2\},\{3\}\}$ and $\{\{1,3\},\{2\}\}$.

Such diverse outcomes in an empty-core setting demonstrate that Shapley's fairness insights can produce unexpectedly varied stability patterns, even when the core provides no solutions. However, as before, this possibility is not a certainty.

Theorem 5. *There exists a TU game (N, v) such that:*

(i) The core $C(N, v)$ is empty.

(ii) The hedonic game induced by (N, v) and the Shapley value does not have any FX-FE strongly Nash stable partition.

Proof. For this final case, we use the following 3-player TU game from [Aumann and Drèze, 1974], which has an empty core. The game is defined by: $v(\{i\}) = 0$ for all $i \in N$, $v(\{1,2\}) = v(\{1,3\}) = v(\{2,3\}) = 8$, and $v(N) = 9$.

The Shapley value for the grand coalition is $(3,3,3)$, while for any two-player subgame, it is $(4,4)$. This induces the following preferences:

$\succeq_1$: $\{1,2\} \sim_1 \{1,3\} \succ_1 \{1,2,3\} \succ_1 \{1\}$,

$\succeq_2$: $\{1,2\} \sim_2 \{2,3\} \succ_2 \{1,2,3\} \succ_2 \{2\}$,

$\succeq_3$: $\{1,3\} \sim_3 \{2,3\} \succ_3 \{1,2,3\} \succ_3 \{3\}$.

From these preferences, it is clear that no partition can be stable. Any partition with a lone player (i.e. any $\{\{i,j\},\{k\}\}$ or $\{\{1\},\{2\},\{3\}\}$) is unstable because that player would rather join another. The grand coalition is also unstable because any pair of players would rather leave to form their own, more valuable coalition. Thus, no FX-FE strongly Nash stable partition exists. □

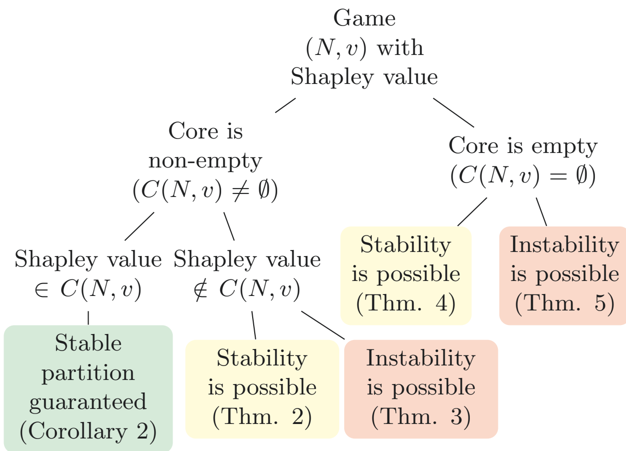

Figure 1: Summary of FX-FE strong Nash stability outcomes for hedonic games induced by the Shapley value.

The findings in this section reveal a clear dichotomy. Figure 1 provides a brief summary of the section, illustrating the conditions under which stable partitions are guaranteed, merely possible, or may not exist at all.

4 Related Work

Our research contributes to a growing body of work at the intersection of cooperative game theory and explainable or trustworthy AI. The Shapley value plays a central role in this discourse, serving as both a normative standard for fair resource allocation in MAS [Heuillet *et al.*, 2022; Gray *et al.*, 2023; Li *et al.*, 2024] and a foundation for interpretability via SHAP [Lundberg and Lee, 2017; Nizri *et al.*, 2022; Salih *et al.*, 2025]. While a large body of recent literature has explored the axiomatic foundations and computational challenges of approximating the Shapley value [Zhou *et al.*, 2023; Li *et al.*, 2024], its downstream effects on agent behavior and system stability remain a critical, underexplored area.

Our approach is partially motivated by recent critiques of the Shapley value in AI contexts. For instance, [Li *et al.*, 2024] highlight the model sensitivity of Shapley, i.e., small perturbations in data or models can induce significant instability in computed values. Similarly, [Yan and Procaccia, 2021] propose replacing Shapley with alternative solution concepts to better align fairness with stability. Acknowledging the value of these critiques while taking a complementary direction, we retain the Shapley rule and investigate the conditions under which it continues to support stability. This leads to a different type of contribution, namely not by discarding Shapley's fairness principle, but by understanding its structural limits and the precise conditions under which it ensures strategic robustness. In doing so, we develop a framework that bridges core stability (in TU games) and hedonic coalition stability.

This positions our work firmly within the literature on coalition formation in MAS, and more specifically, within the theory of hedonic games [Burani and Zwicker, 2003; Aziz and Brandl, 2012; Suksompong, 2015; Çaşkurlu *et al.*, 2023], which are extensively studied in artificial intelligence settings, including MAS and swarm robotics [Aziz *et al.*, 2013; Lang *et al.*, 2015; Bilò *et al.*, 2018; Jang *et al.*, 2018; Czarnecki and Dutta, 2021], as well as more generalized models [Darmann *et al.*, 2012; Darmann *et al.*, 2018]. These works often assume externally defined agent preferences or impose structural restrictions on the domain. In contrast, our contribution is to study how the choice of allocation rule itself—specifically, the Shapley value—can endogenously induce preferences and thereby shape coalition stability.

5 Conclusion

We established a formal bridge between two fundamental stability paradigms: the centralized, cooperative stability guaranteed by the core, and the decentralized, non-cooperative stability of FX-FE strong Nash partitions in the hedonic games they induce. Our findings show that while core-membership of an efficient allocation is a sufficient condition for stability, its absence—a common case for the Shapley value—leads to a structural indeterminacy that the induced hedonic game may or may not have an FX-FE strongly stable partition. These results highlight a fundamental trade-off, or limitation in multi-agent systems: fairness guarantees alone do not ensure strategic robustness across diverse domains and applications. The choice of an allocation rule is not merely a normative decision but a critical design parameter that directly shapes the structure of agents' strategic interactions.

6 Outlook and Open Questions

Our findings open several avenues for future research at the intersection of cooperative game theory and AI. Below are a few illustrative open questions.

- **Beyond Sufficient Conditions:** Core-membership is a sufficient condition for FX-FE strong Nash stability, but establishing necessary conditions or a sharper characterization remains an open problem.
- **Dynamic Stability:** Following work in evolving coalition structures [Bullinger and Romen, 2024; Boehmer *et al.*, 2025; Fanelli *et al.*, 2025], whether agents under non-core Shapley allocations converge to stable outcomes or cycle remains open. Our Theorems 2–5 show both possibilities exist, rather than guaranteeing convergence.
- **Computational Tractability:** Determining whether a specific TU game induces a stable hedonic game under Shapley-induced preferences appears to be computationally hard. Following [Peters, 2016], identifying tractable subclasses or effective approximation heuristics remains crucial for practical deployment.
- **Constrained Systems:** Inspired by [Deligkas *et al.*, 2025], a future direction is extending our results to constrained systems (networked/spatial) where agents only form coalitions within their local neighborhood.

We hope these directions stimulate further exploration into the essential interplay of fairness and stability in the design of autonomous, cooperative AI systems.

Acknowledgments

The author thanks Mehmet Karakaya and Buğra Çaşkurlu for helpful discussions on hedonic games.

References

[Aumann and Drèze, 1974] Robert J. Aumann and Jacques H. Drèze. Cooperative games with coalition structures. *International Journal of Game Theory*, 3(4):217–237, December 1974.

[Aziz and Brandl, 2012] Haris Aziz and Florian Brandl. Existence of stability in hedonic coalition formation games. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, volume 2 of *AAMAS '12*, page 763–770, 2012.

[Aziz and Savani, 2016] Haris Aziz and Rahul Savani. Hedonic games, chapter 15. *Handbook of Computational Social Choice, F. Brandt, V. Conitzer, J. Lang U. Endriss, and AD Procaccia (Eds.). Cambridge University Press, Cambridge*, 2016.

[Aziz *et al.*, 2013] Haris Aziz, Felix Brandt, and Hans Georg Seedig. Computing desirable partitions in additively separable hedonic games. *Artificial Intelligence*, 195:316–334, 2013.

[Banerjee *et al.*, 2001] Suryapratim Banerjee, Hideo Konishi, and Tayfun Sönmez. Core in a simple coalition formation game. *Social Choice and Welfare*, 18(1):135–153, 2001.

[Bilò *et al.*, 2018] Vittorio Bilò, Angelo Fanelli, Michele Flammini, Gianpiero Monaco, and Luca Moscardelli. Nash stable outcomes in fractional hedonic games: existence, efficiency and computation. *Journal of Artificial Intelligence Research*, 62(1):315–371, 2018.

[Boehmer *et al.*, 2025] Niclas Boehmer, Martin Bullinger, and Anna Maria Kerkmann. Causes of stability in dynamic coalition formation. *ACM Transactions on Economics and Computation*, 13(2):1–45, 2025.

[Bogomolnaia and Jackson, 2002] Anna Bogomolnaia and Matthew O. Jackson. The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2):201–230, 2002.

[Bondareva, 1962] Olga N. Bondareva. The theory of the core in an n-person game. *Vestnik Leningradskogo Universiteta*, 13:141–142, 1962. In Russian.

[Bullinger and Romen, 2024] Martin Bullinger and René Romen. Stability in online coalition formation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):9537–9545, 2024.

[Burani and Zwicker, 2003] Nadia Burani and William S. Zwicker. Coalition formation games with separable preferences. *Mathematical Social Sciences*, 45(1):27–52, 2003.

[Caskurlu and Eser, 2025] Bugra Caskurlu and Ali Eser. (Submodular) hedonic games with common ranking property. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 2449–2451, 2025.

[Çaşkurlu *et al.*, 2023] Buğra Çaşkurlu, Fatih Erdem Kizilkaya, and Berkehan Özen. Hedonic expertise games. *Annals of Mathematics and Artificial Intelligence*, 2023.

[Chevaleyre *et al.*, 2006] Yann Chevaleyre, Paul Dunne, Endriss Ulle, Lang Jérôme, Lemaître Michel, Nicolas Maudet, Julian Padget, Steve Phelps, Juan Rodríguez-Aguilar, and Paulo Sousa. Issues in multiagent resource allocation. *Informatica*, 30:3–31, 2006.

[Czarnecki and Dutta, 2021] Emily Czarnecki and Ayan Dutta. Scalable hedonic coalition formation for task allocation with heterogeneous robots. *Intelligent Service Robotics*, 14(3):501–517, 2021.

[Darmann *et al.*, 2012] Andreas Darmann, Edith Elkind, Sascha Kurz, Jérôme Lang, Joachim Schauer, and Gerhard Woeginger. Group activity selection problem. In *Proceedings of the 8th International Conference on Internet and Network Economics*, WINE'12, page 156–169, 2012.

[Darmann *et al.*, 2018] Andreas Darmann, Edith Elkind, Sascha Kurz, Jérôme Lang, Joachim Schauer, and Gerhard Woeginger. Group activity selection problem with approval preferences. *International Journal of Game Theory*, 47(3):767–796, 2018.

[Deligkas *et al.*, 2025] Argyrios Deligkas, Eduard Eiben, Stavros D. Ioannidis, Dusan Knop, and Simon Schierreich. Balanced and fair partitioning of friends. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):13754–13762, 2025.

[Drèze and Greenberg, 1980] Jacques H. Drèze and Joseph Greenberg. Hedonic coalitions: Optimality and stability. *Econometrica*, 48(4):987–1003, 1980.

[Fanelli *et al.*, 2025] Angelo Fanelli, Laurent Gourvès, Ayumi Igarashi, and Luca Moscardelli. Individually stable dynamics in coalition formation over graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):13831–13838, 2025.

[Farrell and Scotchmer, 1988] Joseph Farrell and Suzanne Scotchmer. Partnerships. *The Quarterly Journal of Economics*, 103(2):279–297, 1988.

[Gillies, 1953] Donald B. Gillies. *Some Theorems on n-Person Games*. PhD thesis, Department of Mathematics, Princeton University, 1953.

[Gray *et al.*, 2023] Robert C. Gray, Jichen Zhu, and Santiago Ontañón. Distribution fairness in multiplayer AI using Shapley constraints. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 19(1):233–243, 2023.

[Grotte, 1970] Jeffrey H. Grotte. Computation of and observations on the nucleolus, the normalized nucleolus, and the central games. Master's thesis, Department of Applied Mathematics, Cornell University, Ithaca, NY, USA, 1970.

[Hajduková, 2006] Jana Hajduková. Coalition formation games: A survey. *International Game Theory Review*, 8(4):613–641, 2006.

[Heuillet *et al.*, 2022] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with Shapley values. *IEEE Computational Intelligence Magazine*, 17(1):59–71, 2022.

[Inarra and Usategui, 1993] Elena Inarra and Jose Usategui. The Shapley value and average convex games. *International Journal of Game Theory*, 22:13–29, 02 1993.

[Jang *et al.*, 2018] Inmo Jang, Hyo-Sang Shin, and Antonios Tsourdos. Anonymous hedonic game for task allocation in a large-scale multiple agent system. *IEEE Transactions on Robotics*, 34(6):1534–1548, 2018.

[Karakaya, 2011] Mehmet Karakaya. Hedonic coalition formation games: A new stability notion. *Mathematical Social Sciences*, 61(3):157–165, 2011.

[Lang *et al.*, 2015] Jérôme Lang, Anja Rey, Jörg Rothe, Hilmar Schadrack, and Lena Schend. Representing and solving hedonic games with ordinal preferences and thresholds. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, page 1229–1237, 2015.

[Li *et al.*, 2024] Meng Li, Hengyang Sun, Yanjun Huang, and Hong Chen. Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, 4(1), 2024.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, 2017.

[Magaña and Carreras, 2018] Antonio Magaña and Francesc Carreras. Coalition formation and stability. *Group Decision and Negotiation*, 27(3):467–502, 2018.

[Moulin, 1988] Hervé Moulin. *Axioms of cooperative decision making*. Cambridge University Press, Cambridge, 1988.

[Moulin, 2003] Hervé Moulin. *Fair Division and Collective Welfare*. The MIT Press, Cambridge, 2003.

[Nizri *et al.*, 2022] Meir Nizri, Noam Hazon, and Amos Azaria. Explainable Shapley-based allocation (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):13023–13024, 2022.

[Peters, 2016] Dominik Peters. Complexity of hedonic games with dichotomous preferences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016.

[Salih *et al.*, 2025] Ahmed M Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Karim Lekadir, and Gloria Menegaz. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1):2400304, 2025.

[Schmeidler, 1969] David Schmeidler. The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics*, 17(6):1163–1170, 1969.

[Shapley, 1953] Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pages 307–317. Princeton University Press, Princeton, NJ, 1953.

[Shapley, 1967] Lloyd S. Shapley. On balanced sets and cores. *Naval Research Logistics Quarterly*, 14(4):453–460, 1967.

[Shapley, 1971] Lloyd S. Shapley. Cores of convex games. *International Journal of Game Theory*, 1(1):11–26, 1971.

[Suksompong, 2015] Warut Suksompong. Individual and group stability in neutral restrictions of hedonic games. *Mathematical Social Sciences*, 78:1–5, 2015.

[Sung and Dimitrov, 2007] Shao Chin Sung and Dinko Dimitrov. On myopic stability concepts for hedonic games. *Theory and Decision*, 62(1):31–45, 2007.

[Yan and Procaccia, 2021] Tom Yan and Ariel D. Procaccia. If you like Shapley then you'll love the core. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5751–5759, 2021.

[Zhou *et al.*, 2023] Zijian Zhou, Xinyi Xu, Rachael Hwee Ling Sim, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Probably approximate Shapley fairness with applications in machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5910–5918, 2023.