Modélisation de valeurs humaines : le cas des vertus dans les jeux hédoniques

Thibaut Vallée [1], **Grégory Bonnet** [2], **Thibault de Swarte** [3]

1. *Normandie Université, UNICAEN, GREYC, CNRS UMR 6072*
   thibaut.vallee@unicaen.fr
2. *Normandie Université, UNICAEN, GREYC, CNRS UMR 6072*
   gregory.bonnet@unicaen.fr
3. *IMT Atlantique, Idea Lab LASCO-IMT, UBL, F-35576 Cesson-Sévigné Cedex*
   thibault.deswarte@imt-atlantique.fr

RÉSUMÉ. *De nombreux domaines applicatifs mettent en présence plusieurs agents qui doivent interagir, décider conjointement et coopérer. Dans ce contexte, un agent doit non seulement tenir compte de critères éthiques au regard de ses objectifs mais aussi sur la manière dont il coopère, et en particulier sur la manière dont il tient compte des objectifs des autres agents. En partant de ce constat, nous nous intéressons dans cet article à la modélisation d'une éthique des vertus, c'est-à-dire respectant une valeur cardinale, dans le cadre de la formation de collectifs d'agents. Pour ce faire, nous proposons un nouveau modèle de formation de coalitions, appelé jeu hédonique de déviation, où chaque agent décide des collectifs à former au regard de règles de comportements atomiques. Nous montrons comment ces règles peuvent être composées pour représenter une pluralité de valeurs cardinales, en particulier des valeurs de liberté, altruisme et hédonisme, et nous permettent de caractériser de nouveaux concepts de solution.*

ABSTRACT. *In many applicative contexts, several autonomous artificial agents must interact, make decision and cooperate in a collective way. In those contexts, an agent must not only take into account ethical criterion to reach its goals but also need to decide how it will cooperate and how it will take the other agents' goals into account. Thus, in this article, we are interested by modeling a virtue ethics, ie. which aims at supporting a cardinal moral value, for collective formation. To this end, we propose a new model of coalition formation, called deviation hedonic games, where each agent decides which coalition it will form according to a set of atomic behavioral rules. Then we show how those rules can be combined in order to represent a plurality of values. As examples, we model three values – liberty, altruism and hedonism – which allow us to characterize new solution concepts.*

MOTS-CLÉS : *éthique des vertus, coalitions, systèmes multi-agents, valeurs humaines.*

KEYWORDS: *coalitions, human values, multi-agent systems, virtue ethics.*

DOI:10.3166/RIA.32.519-546

1. Introduction

L'introduction croissante d'agents autonomes artificiels dans certains domaines applicatifs soulève des questionnements éthiques dans le sens où les utilisateurs de ces systèmes peuvent avoir des attentes distinctes des problématiques d'optimalité ou de conformité légale du comportement des agents. Par exemple, dans le domaine de la justice prédictive, il est désirable que les agents autonomes artificiels soient exempts de biais. Dans le domaine de l'aide à la décision médicale, il est désirable que ces agents respectent le code de déontologie médicale. De manière plus prospective, il pourrait être désirable de disposer de robots domestiques capables de politesse ou de voitures autonomes faisant preuve de civilité. Ainsi, se pose la question de concevoir des agents autonomes pouvant faire preuve de comportements qui pourraient être qualifiés d'éthiques, sous-tendus par des valeurs humaines et des principes éthiques et moraux. Ceci implique de pouvoir caractériser, évaluer, raisonner et décider en fonction de ces valeurs et principes.

De nombreux travaux se sont intéressés à la modélisation de la moralité et de l'éthique ; la plupart s'appuient sur la caractérisation d'éléments fondamentaux qui peuvent être la causalité (Berreby *et al.*, 2017 ; Halpern, 2015), la responsabilité individuelle ou collective (Lorini, 2012), des droits et des devoirs spécifiques à des applications (Arkin, 2009) ou des principes éthiques plus généraux (Ganascia, 2007 ; Pereira, Saptawijaya, 2007 ; Bringsjord, Taylors, 2012). Ainsi, ces travaux portent essentiellement sur l'éthique du comportement individuel de l'agent dans l'accomplissement de ses objectifs. Toutefois, dans de nombreux domaines applicatifs, plusieurs agents sont mis en présence et doivent interagir, décider conjointement ou coopérer. Dans ce contexte, un agent doit non seulement tenir compte de critères éthiques au regard de ses objectifs mais aussi sur la manière dont il coopère, et en particulier de la manière dont il tient compte des objectifs des autres agents car il ne saurait y avoir de coopération sans prise en compte de l'intérêt d'autrui.

En partant de ces constats, nous nous intéressons dans cet article à la modélisation d'une éthique des vertus – respectant une valeur humaine [1] cardinale – dans le cadre de la formation de collectifs d'agents, c'est-à-dire de groupes d'agents qui vont devoir par la suite collaborer pour résoudre une tâche. Au-delà de la collaboration en soi, nous nous intéressons plus précisément au processus lui-même de formation de ces collectifs au regard de valeurs que les agents désirent respecter.

Afin d'illustrer notre propos, considérons l'exemple suivant. Quatre agents – Alice (A), Bob (B), Charles (C) et David (D) – ont pour objectif de faire un trajet commun et cherchent une solution de co-voiturage. Chaque agent dispose de préférences, représentant ses intérêts, vis-à-vis des passagers avec qui ils peuvent partager une voiture. Supposons qu'Alice et David ne s'apprécient pas et refusent de partager la même voiture, c'est-à-dire préfèrent être seuls qu'avec l'autre. En revanche, tous deux préfèrent

1. En raison de la polysémie du terme « valeur », ce dernier signifiera toujours « valeur humaine » lorsque nous l'emploierons dans la suite de cet article.

être avec Bob et Charles en même temps plutôt qu'être seulement avec l'un d'entre eux. Bob préfère partager sa voiture avec Charles qui, lui, préfère faire deux voitures de deux en fonction de ses affinités pour Alice, Bob et enfin David. Ces préférences peuvent être représentées par les relations d'ordre suivantes :

Alice : $ABC \succ AB \succ AC \succ A \succ ABCD \succ ABD \succ ACD \succ AD$
Bob : $BC \succ ABC \succ BCD \succ AB \succ BD \succ ABCD \succ ABD \succ B$
Charles : $AC \succ BC \succ CD \succ ABC \succ BCD \succ ABCD \succ ACD \succ C$
David : $BCD \succ BD \succ CD \succ D \succ ABCD \succ ABD \succ ACD \succ AD$

Ici, $ABC \succ AB$ représente le fait que Alice préfère partager la voiture avec Bob et Charles plutot qu'avec Bob seulement. Ici, la question de la coopération est celle de la répartition des agents, sachant que chacun peut alors décider individuellement ou collectivement de changer de voiture. Par exemple, si la répartition consiste à faire deux voitures, Alice seule et Bob, Charles et David ensemble, Charles peut décider de changer de voiture pour rejoindre Alice. Ce choix s'opère alors en fonction des intérêts de Charles, sa préférence pour Alice, mais aussi en fonction de la manière dont il va tenir compte des préférences de Bob et David, représenté par le respect d'une valeur. Supposons que ces valeurs sont respectivement pour Alice, Bob, Charles et David, la liberté, l'hédonisme, l'égoïsme et l'altruisme. Ces dernières pourraient être caractérisées par les règles de comportement suivantes :

Libertaire : Alice rejoint une voiture tant que cela ne se fait pas détriment des autres agents,

Hédoniste : Bob ne rejoint une voiture que s'il préfère voyager avec les passagers de cette dernière et que les passagers des voitures qu'il quitte et qu'il rejoint le préfèrent aussi,

Égoïste : Charles rejoint une voiture s'il préfère voyager avec les passagers de cette dernière,

Altruiste : David ne rejoint une autre voiture que si cela est préféré par tous les autres agents, sans considération pour ses propres préférences.

Supposons alors la répartition des passagers en deux voitures avec Alice et Charles dans l'une et David et Bob dans l'autre. Cette répartition satisfait les valeurs des agents car aucun ne désire alors changer de voiture. Elle fait alors consensus au sens qu'elle respecte au mieux les intérêts de chacun et les manières hétérogènes dont chacun tient compte des intérêts des autres. Ceci a d'autant plus d'intêret dans le cas d'applications impliquant à la fois des agents artificiels et des acteurs humains, où cela peut avoir du sens de concevoir des agents artificiels qui ne sont pas d'égoïstes dans leur manière de former des collectifs.

Ainsi dans cet article, nous proposons de modéliser ce type de problème en nous fondant sur un modèle de jeux de coalitions hédoniques qui répondent au problème du partitionnement des agents au regard des préférences de chaque agent vis-à-vis des

groupes auxquels il peut appartenir (section 2). Afin de représenter les valeurs cardinales des agents, nous enrichissons ce modèle en proposant des *jeux de déviations* où chaque agent décide de changer de groupe au regard de règles de comportements – appelées *concept de déviations* – qui lui sont propres. Une solution fait consensus lorsqu'aucun agent ne désire changer de coalition (section 3). Nous montrons ensuite comment ces règles de déviations peuvent être composées pour représenter une pluralité de valeurs cardinales, en particulier des valeurs de liberté, altruisme et hédonisme, amenant ainsi les agents à suivre une éthique de la vertu dans leur processus de formation de coalitions (section 4).

2. État de l'art

2.1. Éthique des vertus et valeurs

Proposée par l'École d'Athènes, l'éthique des vertus vise à penser les valeurs – comme la sagesse, le courage, la justice – qui peuvent guider le comportement humain (Platon, 1966). Afin de distinguer les choix éthiques de ceux qui ne le sont pas, il convient alors d'identifier ces valeurs et de comprendre comment elles peuvent soutenir ou non une prise de décision (Hursthouse, 2013). D'un point de vue général, les valeurs sont des qualités abstraites d'un état du monde que des acteurs perçoivent comme bon ou idéal (Brey, 2014 ; 2015) comme par exemple la liberté, la justice, la sagesse, l'honnêteté, l'efficacité, la beauté, la sérénité, l'amitié ou la bien-portance. Si les valeurs sont universellement présentes à des degrés divers dans toutes les cultures (S. Schwartz, Bilsky, 1990 ; Bardi *et al*., 2009), elles évoluent dans leur définition et leur usage au cours du temps et dans l'espace (Boltanski, Thévenot, 2006 ; Dambra, 2005). De ce fait, elles peuvent être associées à des normes sociales et deviennent des *valeurs morales* lorsqu'elles concernent la conduite envers autrui.

Dans le domaine des sciences sociales, de nombreuses études ont proposé des définitions et une analyse de l'usage des valeurs (Rokeach, 1973 ; S. Schwartz, Bilsky, 1990 ; Valette-Florence *et al*., 1996 ; S. H. Schwartz, 2012). Par exemple, une des études les plus complètes a été réalisée par Shalom Schwartz qui a cartographié 56 valeurs groupées en 10 catégories (S. Schwartz, Bilsky, 1990 ; S. H. Schwartz, 2012). De manière générale, les valeurs semblent exister en nombre fini, être relativement stables et présenter des relations de conflits ou de compatibilités entre elles (Bardi *et al*., 2009) sous forme de *systèmes de valeurs*. Toutefois, les valeurs sont aussi dynamiques dans le temps et l'espace : différentes cultures peuvent avoir différent systèmes de valeurs qui évoluent dans le temps (Boltanski, Thévenot, 2006). Quoi qu'il en soit, les valeurs et les systèmes de valeurs forment un langage. D'un point de vue relativiste (supposant que les valeurs différent selon les cultures) ou absolutiste (supposant que les différences observées ne sont dues qu'à des différences dans l'expression des valeurs), les valeurs permettent de discuter ou caractériser une attitude ou une décision et de les comparer. En ce sens, il nous semble pertinent de modéliser et de caractériser des valeurs dans le cadre de systèmes d'agents autonomes devant traiter de questions éthiques.

2.2. Implémentation d'éthiques dans la décision individuelle

Dans le domaine de l'Intellligence Artificielle, de nombreux travaux se sont intéressés à la modélisation de l'éthique et de la morale pour les agents autonomes. Au-delà des travaux consistant à implémenter des comportements contraints par des considérations éthiques (Arkin, 2009 ; Wiegel, van den Berg, 2009 ; Dennis *et al*., 2015), il s'agit généralement d'approches logiques visant à formaliser et implémenter des principes éthiques (Berreby *et al*., 2017 ; Bringsjord, Taylors, 2012 ; Ganascia, 2007 ; Lorini, 2012 ; Saptawijaya, Pereira, 2014). D'autres approches plus générales proposent des architectures d'agents permettant un raisonnement éthique et moral dans un cadre plus large de décision, coopération et jugement (Berreby *et al*., 2017 ; Cointe *et al*., 2016 ; Lorini, 2012).

Toutefois, dans ces approches, les notions de valeurs sont souvent représentées comme des propriétés abstraites de haut niveau dont la définition exacte est laissée à la discrétion de l'utilisateur, à l'instar des systèmes d'argumentation valués, Bench-Capon (2002) considère les valeurs comme des étiquettes associées à des arguments qui peuvent être comparées par une relation d'ordre (non nécessairement transitive).

Une dernière catégorie de travaux d'ordre plus méthodologique ont proposé des définitions pour des valeurs spécifiques aux agents artificiels. Par exemple, Gigerenzer (2010) les voit comme des heuristiques désirables (imitation des pairs, partage égalitaire de ressources, coopération a priori, respect des normes). Dans la même idée, Coleman (2001) a proposé une taxonomie très complète de valeurs (appelée vertus) pour agents artificiels : des vertus agentives (adaptabilité, autonomie, autopoièse par exemple), des vertus sociales (comme la disposition à dire la vérité), des vertus environnementales (comme l'utilisation parcimonieuse des ressources) et des valeurs morales (comme la disposition à se mettre dans un état de vulnérabilité).

Toutefois, la question qui nous intéresse est l'usage de l'éthique des vertus dans le cadre de la formation de collectifs d'agents (ou d'un collectif d'agents), question qui n'a pas été traitée dans la littérature à notre connaissance.

2.3. Formation de coalitions

Le problème de la formation de collectifs a été étudié dans le cadre des jeux à n-joueurs, aussi appelés jeux de coalitions. Étant donné une partition des agents, il s'agit de calculer si chaque agent est disposé à rester dans son sous-ensemble actuel, appelé coalition, ou s'il dévie pour en rejoindre ou en former un autre. Si les approches traditionnelles des jeux de coalitions sont des approches utilitaristes qui se fondent sur l'existence d'une fonction caractéristique représentant l'utilité à former une coalition donnée (Morgenstern, Von Neumann, 1953 ; Nash, 1950), nous considérons le modèle qualitatif des jeux de coalitions hédoniques où chaque agent exprime des préférences vis-à-vis des coalitions auxquelles il peut appartenir (Dreze, Greenberg, 1980). Nous introduisons dans la suite la définition formelle de ces jeux et les notations que nous employons dans le reste de cet article.

DÉFINITION 1 (Jeu hédonique). — *Un* jeu hédonique *est un tuple* $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *où* $N = \{a_1, \ldots, a_n\}$ *désigne l'ensemble des agents et* $\succeq_i$ *désigne les préférences de l'agent* a_i *représentées par un pré-ordre total sur* $\mathcal{N}_i = \{C \subseteq N \ : \ a_i \in C\}$.

Le problème associé aux jeux de coalitions hédoniques est donc de calculer une partition Π de l'ensemble N qui satisfait aux mieux les préférences de chaque agent. Dans la suite de cet article, nous dénotons par $\mathcal{P}_N$ l'ensemble des partitions possibles pour les agents de N et, étant donnée une partition $\Pi \in \mathcal{P}_N$, $C_i(\Pi)$ désigne la coalition de l'agent a_i dans Π.

Pour caractériser une partition qui satisfait « au mieux » les préférences des agents, de nombreux *concepts de solution* ont été proposés dans la littérature. Les concepts de solution définissent les propriétés que doit satisfaire une partition pour être considérée comme *stable*, c'est-à-dire une partition où aucun agent ne désire changer de coalition. Le tableau 1 présente les concepts de solution classiquement considérés dans la littérature (Greenberg, 1994 ; Bogomolnaia, Jackson, 2002 ; Ballester, 2004 ; Elkind, Wooldridge, 2009 ; Aziz *et al*., 2011 ; Aziz, Brandt, Seedig, 2013 ; Aziz, Brandt, Harrenstein, 2013 ; Brandl *et al*., 2015 ; Peters, Elkind, 2015).

Tableau 1. Principaux concepts de solution

Concept de solution	**Propriété**
Stabilité de Nash (NS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} \ : \ C \cup \{a_i\} \succ_i C_i(\Pi)$
Stabilité Individuelle (IS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} \ : \ C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$
Stabilité Individuelle Contractuelle (ICS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} \ : \ C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
Stabilité Contractuelle de Nash (CNS)	$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} \ : \ C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
Stabilité du Cœur (CS)	$\forall a_i \in N, \nexists C \in \mathcal{N}_i \ : \ C \succ_i C_i(\Pi)$ $\wedge \forall a_j \in C, C \succeq_j C_j(\Pi)$
Optimalité (O)	$\forall a_i \in N, \nexists C \in \mathcal{N}_i \ : \ C \succ_i C_i(\Pi)$
Pareto-Optimalité (PO)	$\nexists \Pi_2 \ : \ \forall a_i \in N, C_i(\Pi_2) \succeq_i C_i(\Pi)$ $\wedge \exists a_j \in N, C_j(\Pi_2) \succ_j C_j(\Pi)$

Chacun de ces concepts de solution représente un comportement spécifique que doivent suivre les agents dans le processus de formation de coalitions. À titre d'exemple, la stabilité au sens de Nash représente des partitions où aucun agent ne désire individuellement rejoindre une autre coalition déjà présente dans cette partition. De plus, tous les concepts de solution canoniques reposent sur une hypothèse forte : tous les agents présentent le même comportement et cherchent à satisfaire le même concept de solution.

Nous avons proposé dans (Vallée, Bonnet, 2017) une généralisation des jeux de coalitions hédoniques – appelée jeux à concepts de solutions multiples – où chaque agent dispose de son propre concept de solution. Cette généralisation permet ainsi de représenter des agents hétérogènes quant aux concepts qu'ils désirent satisfaire.

De plus, dans les modèles avec des agents homogènes, Sung et Dimitrov (2007) ont proposé de représenter explicitement les comportements en redéfinissant les concepts de solution non pas à partir d'une caractérisation globale mais d'une conjonction de cinq ensembles de déviations, chacun représentant une propriété sur le fait qu'une déviation soit autorisée ou non. Une partition est alors considérée comme stable lorsque aucun agent ne désire dévier.

Le modèle que nous proposons dans la suite de cet article intègre alors ces deux aspects, une généralisation de Sung et Dimitrov (2007) pour représenter des comportements individuels sous forme de déviations et, comme nous l'avions proposé dans (Vallée, Bonnet, 2017), la prise en compte d'une hétérogénéité des comportements, afin de modéliser une éthique des vertus dans le cadre de la formation de coalitions.

3. Des jeux de déviations

Contrairement aux jeux hédoniques classiquement considérés dans la littérature, nous proposons ici un nouveau modèle où la notion de stabilité est définie par l'absence de déviation au regard de conditions propres à chaque agent.

3.1. Modéliser les déviations

Dans un jeu hédonique classique, l'absence de déviation est caractérisée par la satisfaction de propriétés globales. Nous proposons ici de considérer explicitement une déviation comme tout changement de coalition de la part d'un sous-ensemble d'agents pour en former ensemble une nouvelle.

DÉFINITION 2 (Déviation). — *Soit* $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *un jeu hédonique et* $\Pi \in \mathcal{P}_N$ *une partition. Une* déviation *est une coalition* $D \subseteq N, D \notin \Pi, D \neq \emptyset$ *telle que l'ensemble des agents de* D *quittent leurs coalitions courantes dans* Π *pour former la nouvelle coalition* D.

Nous distinguons deux types de déviations : les déviations *individuelles* et les déviations *collectives*. Une déviation individuelle est une déviation qui nécessite qu'un seul agent a_i quitte sa coalition courante pour rejoindre les autres membres de D, c'est-à-dire $D \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$. À l'inverse, les déviations collectives nécessitent qu'aux moins deux agents distincts quittent leurs coalitions courantes.

Dans la suite, nous désignons par $[D \to \Pi]$ l'application de la déviation D sur la partition Π.

DÉFINITION 3 (Application d'une déviation). — *La partition* Π' *résultant de* $[D \to \Pi]$ *est telle que :*

– $\forall a_i \in D, C_i(\Pi') = D$
– $\forall a_j \in N \ : \ \exists a_i \in D, C_j(\Pi) = C_i(\Pi), C_j(\Pi') = C_j(\Pi) \setminus D$
– $\forall a_k \in N \ : \ \nexists a_i \in D, C_k(\Pi) = C_i(\Pi), C_k(\Pi') = C_k(\Pi)$

Étant donnée une partition Π, nous désignons par $AllD_i(\Pi) = \{D \subseteq N, D \notin \Pi \ : \ a_i \in D\}$ l'ensemble des déviations qui impliquent l'agent a_i.

Plaçons nous maintenant du point de vue d'un agent $a_i \in N$ et considérons une partition $\Pi \in \mathcal{P}_N$. Nous modélisons les déviations que l'agent a_i désirerait voir se réaliser au regard de ses préférences et d'autres critères qui lui sont propres à l'aide de *conditions* (au sens large) qui doivent être satisfaites.

DÉFINITION 4 (Condition de déviation). — *Soit $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ un jeu hédonique, $a_i \in N$ un agent, $\Pi \in \mathcal{P}_N$ une partition et $D \in AllD_i(\Pi)$ une déviation. Une* conditition de déviation Δ_X *désigne une propriété que doit satisfaire la déviation D au regard de l'agent a_i, de la partition courante Π et du profil de préférence pour que D soit désirable pour l'agent a_i.*

Dans la suite $\Delta_X(a_i, D, \Pi, HG)$ désigne la fonction booléenne vérifiant si une déviation D satisfait la condition Δ_X du point de vue de l'agent a_i étant donnés la partition Π et le jeu HG.

Afin d'illustrer notre propos, nous nous limitons dans cet article aux conditions ci-dessous. Leur choix est dicté par leur sémantique et par leurs liens avec les concepts de solution classiquement considérés dans la littérature. Nous montrons plus spécifiquement ce lien dans la section 3.3.

Condition de Rationalité : $\Delta_R := D \succ_i C_i(\Pi)$ – la déviation D est *rationnelle* du point de vue l'agent a_i s'il préfére (strictement) la déviation à sa coalition courante.

Condition d'Acceptation : $\Delta_A := \forall a_j \in D \setminus \{a_i\}, D \succ_j C_j(\Pi)$ – la déviation D est *acceptable* si tous les membres de D préfèrent (strictement) la déviation à leur coalition courante.

Condition de Défection : $\Delta_D := \forall a_k \in N \setminus D \ : \ \exists a_j \in D, C_k(\Pi) = C_j(\Pi), C_k(\Pi) \setminus D \succ_k C_k(\Pi)$ – la déviation D est une *défection* si le départ des agents de D est (strictement) préférable du point de vue des autres membres de leurs coalitions initiales.

Condition d'Optimalité : $\Delta_O := \nexists C \subseteq N \ : \ C \succ_i D$ – la déviation D est *optimale* du point de vue de l'agent a_i si elle fait partie de ses coalitions préférées.

Condition de Pareto : $\Delta_{PO} := \exists \Pi' \in \mathcal{P}_N, D \in \Pi' \ : \ \forall a_j \in N, C_j(\Pi') \succ_j C_j(\Pi)$ – la déviation D est *Pareto-compatible* [2] s'il existe une partition Π'

2. Il est important de remarquer qu'il ne s'agit pas ici d'une propriété de Pareto-dominance. En effet, la Pareto-dominance classique sur laquelle s'appuie la Pareto-optimalité est caractérisée par le fait qu'au

contenant D, où toutes les coalitions de Π' sont (strictement) préférées à celles de Π par tous les agents.

Condition d'Individualité : $\Delta_I := \forall a_j \in D \setminus \{a_i\}, C_j(\Pi) = D \setminus \{a_i\}$ – la déviation D est *individuelle* si tous les autres agents de D forment déjà la coalition, c'est-à-dire si l'agent a_i est le seul agent qui change de coalition lors de la déviation.

Condition de Collectivité : $\Delta_C := \exists a_j \in D \setminus \{a_i\} : C_j(\Pi) \neq D \setminus \{a_i\}$ – la déviation D est *collective* si au moins un autre agent que a_i n'appartenaient pas à D avant de la rejoindre.

Remarquons que nous avons ici deux familles de conditions. D'un côté, les conditions Δ_R, Δ_A, Δ_D, Δ_O et Δ_{PO} portent sur la satisfaction des préférences des agents, tandis que les conditions Δ_I et Δ_C portent sur l'identité des agent déviants.

Les conditions que nous présentons ici sont fortement complémentaires. Concernant les conditions d'identité, nous avons nécessairement :

– $\Delta_I(a_i, D, \Pi, HG) \vee \Delta_C(a_i, D, \Pi, HG)$ est une tautologie,

– $\Delta_I(a_i, D, \Pi, HG) \wedge \Delta_C(a_i, D, \Pi, HG) = \emptyset$.

Concernant les conditions de préférences, elles portent toutes sur la satisfaction des préférences des différents agents :

– Δ_R et Δ_O portent uniquement sur l'agent a_i,

– Δ_A porte uniquement sur les autres agents de la déviation,

– Δ_D porte sur les agents impactés par la déviation, c'est-à-dire ceux n'effectuant pas la déviation D mais dont aux moins l'un des membres de leur coalition dévie,

– Δ_{PO} porte sur la satisfaction de tous les agents.

Remarquons que ces conditions présentent une forme d'anonymat car nous ne distinguons pas les agents au sein de l'ensemble auxquels ils appartiennent. Du point de vue d'un agent déviant, les agents a_i et a_j impactés par la déviation ont autant d'importance l'un que l'autre. Comme dans les jeux hédoniques classiques, il n'existe pas de relations privilégiées entre les agents en dehors de celles exprimées par les profils de préférences. Remarquons également qu'une déviation optimale pour l'agent a_i est nécessairement rationnelle pour cet agent.

$$\Delta_O(a_i, D, \Pi, HG) \implies \Delta_R(a_i, D, \Pi, HG)$$

Nous n'avons présenté ici que des versions *fortes* des conditions sur les préférences dans le sens où les préférences considérées sont strictes. Nous notons par Δ_X^- les équivalents *affaiblis* usant de préférences non strictes. Par exemple, une déviation D

---

moins un agent préfère strictement une issue à une autre et que les autres agents la préfère ou sont indifférents. La Pareto-optimalité se retrouve si tous les agents considèrent la condition de Pareto affaiblie en conjonction avec la condition de rationalité, comme indiqué dans le tableau 2 de la section 3.3.

qui satisfait Δ_A^- signifie que les agents de D autres que a_i peuvent également être indifférents au changement de coalition de a_i :

$$\Delta_A^- := \forall a_j \in D \setminus \{a_i\}, D \succeq_j C_j(\Pi)$$

La condition de Pareto diffère des autres conditions de déviation. En effet, cette condition ne compare pas uniquement les coalitions de la partition Π avec les coalitions appartenant à la partition Π' résultant de la déviation $[D \to \Pi]$. Elle compare les coalitions de Π avec toutes les coalitions appartenant à toutes les partitions contenant D. Cette spécificité nous permet de considérer non plus uniquement la déviation D, mais une succession de déviations.

EXEMPLE 5. — Considérons la partition $\Pi = \{\{a_1, a_3\}, \{a_2, a_4\}\}$ dans le jeu :

$$\begin{aligned}
N =& \{a_1, a_2, a_3, a_4\} \\
\succeq_1 =& \{a_1, a_2\} \succ_1 \{a_1, a_3\} \succ_1 \{a_1\} \\
\succeq_2 =& \{a_1, a_2\} \succ_2 \{a_2, a_4\} \succ_2 \{a_2\} \\
\succeq_3 =& \{a_3, a_4\} \succ_3 \{a_1, a_3\} \succ_3 \{a_3\} \\
\succeq_4 =& \{a_3, a_4\} \succ_4 \{a_2, a_4\} \succ_4 \{a_4\}
\end{aligned}$$

Du point de vue de a_1, $\forall D \in AllD_1(\Pi)$, il existe au moins un agent $a_j \in N$ tel que, pour Π' résultant de $[D \to \Pi]$, $C_j(\Pi) \succ_j C_j(\Pi')$. Le même raisonnement se tient pour les autres agents. Ainsi, quel que soit l'agent ou le groupe d'agents qui effectue une déviation, elle se fait au détriment d'au moins un agent. Par exemple, en considérant la déviation $D = \{a_1, a_2\}$, nous avons $\Pi' = \{\{a_1, a_2\}, \{a_3\}, \{a_4\}\}$ où $C_3(\Pi) \succ_3 C_3(\Pi')$. Cependant, en effectuant cette déviation qui est désavantageuse pour a_3 et pour a_4, ces derniers peuvent désormais eux-même envisager la déviation $D_2 = \{a_3, a_4\}$ avec $\Pi'' = \{\{a_1, a_2\}, \{a_3\, a_4\}\}$ qui, elle, satisfait $\forall a_i \in N, C_i(\Pi'') \succ_i C_i(\Pi)$. □

3.2. Concepts de déviation

Les conditions de déviation permettent à un agent de définir les règles individuelles permettant de caractériser les déviations qu'il désire réaliser. Trivialement, un agent a_i peut vouloir satisfaire simultanément plusieurs conditions, ou encore qu'au moins l'une soit satisfaite. Par exemple, un agent peut exprimer avec la proposition $\Delta_R \wedge \Delta_A$ le fait de désirer une déviation D si et seulement si D est préférable à la fois pour lui-même mais aussi pour tous les autres agents de D. Ainsi, une agrégation de conditions de déviation est appelée le *concept de déviation* de l'agent a_i.

DÉFINITION 6 (Concept de déviation). — *Soit $a_i \in N$. Le* concept de déviation $\mathbb{D}_i$ *de l'agent a_i est une formule propositionnelle portant sur un ensemble $\{\Delta_1, \ldots, \Delta_k\}$ de*

conditions de déviation. Toute déviation $D \in AllD_i(\Pi)$ qui satisfait $\mathbb{D}_i$ (noté $D \vDash \mathbb{D}_i$) est considérée comme désirable pour l'agent a_i.

Dans la suite, étant donnés l'agent $a_i \in N$, la partition $\Pi \in \mathcal{P}_N$ et le jeu HG, nous désignons par $\mathbb{D}_i(\Pi, HG)$ l'ensemble des déviations désirables pour l'agent a_i :

$$\mathbb{D}_i(\Pi, HG) = \{D \in AllD_i(\Pi) | D \vDash \mathbb{D}_i\}$$

EXEMPLE 7. — Considérons l'agent Charles de l'introduction. Ce dernier ne rejoint une voiture que s'il préfère voyager avec le(s) passager(s) de cette dernière. Ceci correspond au concept de déviation $\mathbb{D}_{\text{Charles}} = \Delta_R \wedge \Delta_I$, signifiant qu'il recherche les déviations individuelles strictement préférées à sa coalition courante. Considérons maintenant Bob pour qui est désirable toute déviation (individuelle ou collective) telle qu'elle soit strictement préférée par lui-même, par les autres agents déviants et par les agents impactés par la déviation. Ce concept de déviation peut être formalisé comme suit $\mathbb{D}_{\text{Bob}} = (\Delta_I \vee \Delta_C) \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$ qui peut être réduit à $\mathbb{D}_{\text{Bob}}$ à : $\Delta_R \wedge \Delta_A \wedge \Delta_D$ car $(\Delta_I \vee \Delta_C)$ est une tautologie. □

Comme nous le montre l'exemple 7, plusieurs agents peuvent suivre des concepts de déviation différents, représentant des agents hétérogènes dans leurs processus de décision vis-à-vis des déviations. Nous pouvons donc définir un nouveau modèle de jeu hédonique – les *jeux hédoniques de déviation* – où chaque agent exprime ses désirs de déviation au regard de son propre concept de déviation.

DÉFINITION 8 (Jeu hédonique de déviation). — *Un* jeu hédonique de déviation *est un triplet $HGD = \langle N, (\succeq_i)_{a_i \in N}, (\mathbb{D}_i)_{a_i \in N} \rangle$ où $N = \{a_1, \ldots, a_n\}$ est l'ensemble des agents, $\succeq_i$ les préférences de l'agent a_i vis-à-vis des coalitions et $\mathbb{D}_i$ le concept de déviation de l'agent a_i.*

Le problème de partitionnement de ce modèle reste le problème classique des jeux de coalitions hédoniques : trouver une partition $\Pi \in \mathcal{P}_N$ telle qu'aucun agent ne désire dévier. Cependant, contrairement aux jeux de coalitions hédoniques canoniques, cette recherche de stabilité ne passe non pas par la satisfaction de propriétés globales pour tous les agents mais par l'absence de déviation désirable du point de vue d'un agent. Ainsi, du point de vue d'un agent a_i, une partition est *localement stable* lorsqu'il n'existe pas de déviation qui satisfasse son concept de déviation, c'est-à-dire lorsque $\mathbb{D}_i(\Pi, HGD) = \emptyset$. Nous avons donc les deux notions de stabilité suivantes :

DÉFINITION 9 (Stabilité). — *Soit HGD un jeu hédonique de déviation et $\Pi \in \mathcal{P}_N$ une partition.* Π *est* localement stable *du point de vue l'agent $a_i \in N$ si $\mathbb{D}_i(\Pi, HGD) = \emptyset$, et* Π *est* collectivement stable *si $\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$.*

3.3. Liens avec les concepts de solutions canoniques

Comme énoncé précédemment, les conditions de déviation que nous considérons ici ont un fort lien avec les concepts de solutions classiquement utilités dans les jeux de coalitions hédoniques. En effet, Sung et Dimitrov (2007) ont déjà utilisé la notion de

déviation pour caractériser les concepts de solution canoniques mais n'a pas introduit la notion de condition de déviation. Afin de montrer les liens entre les concepts de solution canoniques et les concepts de déviation que nous proposons, nous allons ici considérer une hypothèse d'homogénéité : tous les agents expriment le même concept de déviation.

Nous prouvons ici le lien entre stabilité au sens de Nash et un concept de déviation associé. Les preuves pour les autres concepts canoniques sont semblables. Nous présentons ensuite dans le tableau 2 les correspondances entre les concepts de solution canoniques et les concepts de déviation.

PROPRIÉTÉ 10. — *Soit HGD un jeu hédonique de déviation et $\Pi \in \mathcal{P}_N$ une partition. Si $\forall a_i \in N, \mathbb{D}_i := \Delta_I \wedge \Delta_R$, alors l'équivalence suivante est vraie :*

$$\Pi \in NS \Longleftrightarrow \forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$$

PREUVE 11. — Fixons un jeu de déviation HGD et une partition $\Pi \in \mathcal{P}_N$.

Par définition, $\Pi \in NS$ si :

$$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} \ : \ C \cup \{a_i\} \succ_i C_i(\Pi) \tag{1}$$

Cette formulation de l'équilibre de Nash est équivalente à la suivante :

$$\forall a_i \in N, \nexists C \subseteq N, a_i \in C \ : \ C \setminus \{a_i\} \in \Pi \cup \{\emptyset\} \wedge C \succ_i C_i(\Pi) \tag{2}$$

Distinguons trois parties dans la formule :

1. $\nexists C \subseteq N, a_i \in C$, est ici équivalent à $\nexists C \in AllD_i(\Pi)$ puisque C doit nécessairement être différente de $C_i(\Pi)$,

2. $C \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$ est ici équivalent par définition à $\Delta_I(a_i, C, \Pi, HGD)$, c'est-à-dire que C satisfait la condition de déviation individuelle,

3. $C \succ_i C_i(\Pi)$ est ici équivalent par définition à $\Delta_R(a_i, C, \Pi, HGD)$, c'est-à-dire que C satisfait la condition de rationalité.

Ainsi, une partition est stable au sens de Nash si, pour aucun agent, il n'existe pas de déviation individuelle vers une coalition déjà existante dans Π qui soit rationnelle. Nous pouvons alors réécrire la formule 2 par :

$$\forall a_i \in N, \nexists D \in AllD_i(\Pi) \ : \ \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD) \tag{3}$$

Par hypothèse, $\forall a_i, \mathbb{D}_i := \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD)$. La formule 3 est alors équivalente à :

$$\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset \tag{4}$$

Ainsi, par définition, une partition Π est stable au sens de Nash si le concept de déviation $\mathbb{D}_i := \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD)$ est vide pour tout les agents. ■

Le tableau 2 indique les concepts de déviation $\mathbb{D}_i$ correspondant aux différents concepts de solution canoniques, c'est-à-dire tels que si tous les agents expriment $\mathbb{D}_i$ alors $\Pi \in SC \iff \forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$.

Tableau 2. Association entre concepts de solution et concepts de déviation

Concept de solution	**Concept de déviation**
Stabilité au sens de Nash	$\Delta_I \wedge \Delta_R$
Stabilité individuelle	$\Delta_I \wedge \Delta_R \wedge \Delta_A^-$
Stabilité contractuelle de Nash	$\Delta_I \wedge \Delta_R \wedge \Delta_D^-$
Stabilité individuelle contractuelle	$\Delta_I \wedge \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$
Stabilité au sens du cœur forte	$\Delta_R \wedge \Delta_A$
Stabilité au sens du cœur faible	$\Delta_R \wedge \Delta_A^-$
Optimalité	Δ_O
Pareto-optimalité	$\Delta_R \wedge \Delta_{PO}^-$

Ces concepts de déviation se présentent tous sous la forme d'une conjonction de deux clauses :

Clause sur l'identité : la clause sur l'identité définit si les déviations désirables par l'agent doivent être individuelles (Δ_I) ou peuvent également être collectives ($\Delta_I \vee \Delta_C$). Du fait de la tautologie, ce second cas est implicite. De part leur complémentarité, il n'existe pas non plus de déviations qui puisse satisfaire $\Delta_I \wedge \Delta_C$. De plus, il est intéressant de constater qu'aucun concept de solution canonique ne considère uniquement les déviations collectives alors qu'une telle condition pourrait pourtant être envisageable. Un tel cas représenterait un agent qui n'accepte de dévier que si au moins un autre agent désire dévier avec lui.

Clause sur les préférences : l'autre clause représente la prise en compte des préférences des autres agents. Il est important de constater que pour les concepts de solution canoniques, la condition de rationalité est toujours présente car il s'agit d'une hypothèse classique en théorie des jeux. De plus, dans de nombreux cas, ce sont les formes affaiblies des conditions qui sont considérées. Enfin, la clause est toujours une conjonction de conditions qui doivent être positivement satisfaites.

Cette représentation des concepts de déviation sous forme normale conjonctive amène à un second lien entre les concepts de solution canoniques et les concepts de déviation. Les concepts de solution canoniques présentent en effet des relations d'inclusion et ces relations se retrouvent entre les concepts de déviation. Par exemple, trivialement, pour une partition Π, s'il existe une déviation $D \in All D_i(\Pi)$ telle que $D \vDash \Delta_I \wedge \Delta_R \wedge \Delta_A$ alors D satisfait également le concept de déviation $\Delta_I \wedge \Delta_R$. Ainsi, une telle partition Π ne satisfait pas les concepts de déviation associés à la stabilité individuelle et à la stabilité au sens de Nash. Nous retrouvons alors le fait que toute partition qui n'est pas individuellement stable ne peut pas être stable au sens de Nash, représenté classiquement par l'inclusion $NS \subseteq IS$.

De manière générique, l'inclusion d'un concept de déviation $\mathbb{D}_i^1$ dans un autre concept de déviation $\mathbb{D}_i^2$ – noté $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$ – correspond au fait que que toute déviation autorisée par $\mathbb{D}_i^1$ satisfait également les conditions de $\mathbb{D}_i^2$.

DÉFINITION 12 (Concept de déviation inclus). — *Un concept de déviation $\mathbb{D}_i^1$ est* inclus *dans un concept $\mathbb{D}_i^2$ si, pour tout jeu hédonique de déviation et pour toute partition $\Pi \in \mathcal{P}_N$, $D \in \mathbb{D}_i^1(\Pi, HGD) \implies D \in \mathbb{D}_i^2(\Pi, HGD)$.*

Nous pouvons alors facilement déduire certaines de ces relations d'inclusion.

PROPRIÉTÉ 13. — *Soit $\mathbb{D}_i^1$ et $\mathbb{D}_i^2$ deux concepts de déviation. Soit A (resp. B) l'ensemble des conditions de déviation qui définissent $\mathbb{D}_i^1$ (resp. $\mathbb{D}_i^2$). Si $B \subseteq A$, le concept de déviation $\mathbb{D}_i^1$ est inclus dans $\mathbb{D}_i^2$.*

PREUVE 14. — Fixons HGD un jeu quelconque et une partition $\Pi \in \mathcal{P}_N$. Soit $\mathbb{D}_i^1$ et $\mathbb{D}_i^2$ deux concepts de déviation. Soit A (resp. B) l'ensemble des conditions de déviation qui définissent $\mathbb{D}_i^1$ (resp. $\mathbb{D}_i^2$). Suposons que $B \subseteq A$ et montrons que nous avons nécessairement l'inclusion $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$.

Les concepts de déviations $\mathbb{D}_i^1$ et $\mathbb{D}_i^2$ peuvent être définis par les formes normales conjonctives :

$$\mathbb{D}_i^1 := \bigwedge_{\Delta_X \in A} \Delta_X \text{ et } \mathbb{D}_i^2 := \bigwedge_{\Delta_X \in B} \Delta_X$$

Comme $B \subseteq A$, nous pouvons réécrire $\mathbb{D}_i^1$ sous la forme suivante :

$$(\bigwedge_{\Delta_{X_1} \in B} \Delta_{X_1}) \wedge (\bigwedge_{\Delta_{X_2} \in A \setminus B} \Delta_{X_2})$$

Ainsi,

$$\forall D \in AllD_i(\Pi) \ : \ D \vDash (\bigwedge_{\Delta_{X_1} \in B} \Delta_{X_1}) \wedge (\bigwedge_{\Delta_{X_2} \in A \setminus B} \Delta_{X_2}) \implies D \vDash \bigwedge_{\Delta_{X_1} \in B} \Delta_{X_1}$$

En conséquence,

$$\forall D \in AllD_i(\Pi), D \in \mathbb{D}_i^1(\Pi, HGD) \implies D \in \mathbb{D}_i^2(\Pi, HGD)$$

Nous avons donc nécessairement l'inclusion $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$. ■

Pour illuster cette propriété, considérons les quatre concepts de déviation suivants :

1. $\mathbb{D}_i^1 := \Delta_I \wedge \Delta_R$ (Nash stabilité)
2. $\mathbb{D}_i^2 := \Delta_I \wedge \Delta_R \wedge \Delta_A$ (Stabilité Individuelle)

3. $\mathbb{D}_i^3 := \Delta_I \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$ (Stabilité Individuelle contractuelle)
4. $\mathbb{D}_i^4 := \Delta_R \wedge \Delta_A$ (Stabilité du Cœur)

Ici, nous obtenons les relations d'inclusion suivantes : $\mathbb{D}_i^3 \subseteq \mathbb{D}_i^2 \subseteq \mathbb{D}_i^1$ et $\mathbb{D}_i^3 \subseteq \mathbb{D}_i^4$. Nous retrouvons alors les relations d'inclusion entre les concepts de solution canoniques : $NS \subseteq IS \subseteq ICS$ et $CS \subseteq IS \subseteq ICS$.

3.4. Vers de nouveaux concepts

Le lien avec les concepts de solution classiquement utilisés dans les jeux de coalitions hédoniques nous amène à un constat important. Même en nous limitant à 7 conditions de déviations (5 portant sur la satisfaction des préférences et 2 sur l'identité des agents), de nombreux cas ne sont pas couverts par les concepts classiques. Le tableau 3 met en avant certain de ces manques dans la littérature. Pour des raisons de lisibilité, nous ne présentons ici qu'un sous-ensemble des concepts de solution manquants. Les colonnes donnent les clauses portant sur les conditions d'identité et les lignes les clauses sur les conditions de préférence. Les « ? » représentent des concepts de solution ne correspondant à notre connaissance à aucun concept de solution canonique.

Tableau 3. Des concepts de déviation non couvert

	Δ_I	Δ_C	$\Delta_I \vee \Delta_C$
Δ_R	NS	?	?
$\Delta_R \wedge \Delta_A$	IS	?	CS
$\Delta_R \wedge \Delta_D$	CNS	?	?
$\Delta_R \wedge \Delta_A \wedge \Delta_D$	ICS	?	?
$\Delta_R \wedge \Delta_{\overline{PO}}$	?	?	PO
Δ_O	?	?	O
Δ_A	?	?	?
Δ_D	?	?	?
$\Delta_A \wedge \Delta_D$	?	?	?

Comme nous l'avons fait remarquer précédement, l'un des principaux manques vient du fait que tous les concepts de solution considérent la condition de rationalité. Cependant, il est possible de considérer des agents qui cherchent à maximiser le bien-être social et ce même si la déviation est à leur détriment personnel. L'autre principal manque est l'absence de concepts de solution n'incluant que des déviations collectives. De tel concepts peuvent cependant représenter un agent ne désirant pas être le seul responsable de l'instabilité d'une partition.

4. Modéliser les valeurs humaines

Dans le processus de formation des coalitions, le choix des agents de rester ou de dévier peut être guidé par une éthique des vertues, représentée une valeur cardinale

personnelle. Nous proposons alors de montrer dans cette section comment modéliser des valeurs à l'aide d'un concept de déviation et comment ces valeurs viennent compléter le tableau 3. De manière générale, nous proposons de définir pour une valeur v et un agent a_i un concept de déviation $\mathbb{D}_i^v$ tel que toute déviation D qui satisfait $\mathbb{D}_i^v$ est une déviation qui respecte la valeur v. Une partition stable Π représente une répartition des agents telle qu'aucun d'entre eux ne peut changer de coalition sans trahir ses valeurs.

Afin d'illustrer notre propos, nous proposons d'instancier dans cet article trois valeurs en nous fondant sur leur définition dans la littérature : la *liberté*, l'*altruisme* et l'*hédonisme*. Nous proposons ici des concepts minimaux dans le sens où ces concepts sont des conjonctions des conditions qui doivent être minimalement satisfaites. Cependant, nous ne considérons pas que cette association comme absolue. En effet, tout autre concept de déviation qui satisfait au moins ces conditions satisfait également la valeur correspondante. Ainsi, pour une même valeur, des agents hétérogènes peuvent y associer des concepts de déviation différents.

4.1. Modélisation de la liberté

La liberté est une valeur qui a été grandement étudiée dans la littérature philosophique et politique. Considérons les quatre définitions (non exhaustives) suivantes :

La Liberté selon John Stuart Mill : dans (Mill, 1869), deux formes de libertés sont considérées : la « liberté de pensée » et la « liberté d'action ». La *liberté de pensée* représente le fait que tout homme doit pouvoir former son opinion et l'exprimer sans réserve. Mill indique que satisfaire cette liberté est un impératif pour l'intelligence et la nature morale de l'Homme. La *liberté d'action* désigne, elle, le fait que « les hommes soient libres d'agir selon leurs opinions, c'est-à-dire libres de les appliquer à leur vie sans que leurs semblables les en empêchent physiquement ou moralement, tant que leur liberté ne s'exerce qu'à leurs seuls risques et périls. »

La Liberté dans la Constitution : selon l'article 4 de la Déclaration des Droits de l'Homme et du Citoyen de 1789 (DDHC, 1789), « la liberté consiste à pouvoir faire tout ce qui ne nuit pas à autrui : ainsi, l'exercice des droits naturels de chaque homme n'a de bornes que celles qui assurent aux autres Membres de la Société la jouissance de ces mêmes droits. Ces bornes ne peuvent être déterminées que par la Loi. »

La Liberté selon Montesquieu : « Il est vrai que dans les démocraties le peuple parait faire ce qu'il veux ; mais la liberté politique ne consiste point à faire ce que l'on veut. Dans un État, c'est-à-dire dans une société où il y a des lois, la liberté ne peut consister qu'à pouvoir faire ce que l'on doit vouloir, et à n'être point contraint de faire ce que l'on ne doit pas vouloir. Il faut se mettre dans l'esprit ce que c'est que l'indépendance, et ce que c'est que la liberté. La liberté est le droit de faire tout ce que les lois permettent ; et si un citoyen pouvait faire

ce qu'elles défendent, il n'aurait plus de liberté, parce que les autres auraient tout de même ce pouvoir. » (Montesquieu, 1867) (livre XI, Chapitre III)

La Liberté selon Durkheim : « La vraie liberté individuelle ne consiste donc pas dans la suppression de toute réglementation, mais est le produit d'une réglementation ; car cette égalité n'est pas dans la nature. » (Durkheim, 1893) (Chapitre II)

Dans ces quatre définitions, une même contrainte apparaît clairement : l'absence d'atteinte aux autres. Cette contrainte est illustrée par la maxime populaire : « La liberté des uns s'arrête là où commence celle des autres ». En effet, dans le cadre des jeux de déviation, la liberté de pensée telle que définie par Mill correspond au fait que chaque agent est libre d'exprimer des préférences vis-à-vis des coalitions. Il reste donc à satisfaire la liberté d'action qui consiste à ne pas changer de coalition si cela nuit à un autre agent. Ainsi, un agent est libre de dévier de sa coalition courante si :

1. il ne nuit pas à ceux qu'il rejoint,
2. il ne nuit pas à ceux qu'il quitte.

Ces deux points correspondent respectivement aux formes affaiblies des conditions d'Acceptation (Δ_A^-) et de Défection (Δ_D^-). Remarquons que, bien que Durkheim met en avant une notion de liberté au niveau individuel, la liberté s'applique à tous les agents et il n'y a donc pas de conditions d'identité. De plus, la liberté, tant qu'elle ne nuit pas à autrui, peut nuire à l'agent déviant. Il n'y a donc pas non plus de condition de rationalité. Ainsi, nous pouvons donc définir la liberté par le concept de déviation :

$$\mathbb{D}_i \;:=\; \Delta_A^- \wedge \Delta_D^-$$

4.2. Modélisation de l'altruisme

Considérons maintenant la valeur d'altruisme. S'il y a débat sur l'existence d'actes purement altruistes en s'appuyant sur le fait que tout acte peut être motivé par une forme ou une autre de compensation égoïste (Batson, 2014), cette considération prend sens dans un contexte dynamique où les actions de l'agent à un instant donné influent sur les actions et les croyances des autres agents dans le futur. Ainsi, l'altruisme a été régulièrement étudié dans le cadre de jeux de négociation comme les *gift exchange games* et les *dictator games* (Akerlof, 1984 ; Hoffman *et al*., 1996 ; Eckel, Grossman, 1996 ; Bardsley, 2008). Par exemple, Nongaillard et Mathieu (2011) ont montré que des stratégies altruistes où des agents acceptent des offres qui leur sont désavantageuses permet d'atteindre plus tard une solution optimale.

Quoi qu'il en soit, comme les jeux de déviation que nous considérons sont statiques, la question des motivations liées à la mise en oeuvre d'un comportement altruiste est hors de notre cadre d'étude. Afin de définir des déviations altruistes, nous considérons les deux définitions suivantes :

L'Altruisme selon Rand : dans (Rand, 1964), l'altruisme est vu comme la réponse à l'égoïsme : « The ethics of altruism has created the image of the brute,

as its answer, in order to make men accept two inhuman tenets: (a) that any concern with one's own interests is evil, regardless of what these interests might be, and (b) that the brute's activities are in fact to one's own interest (which altruism enjoins man to renounce for the sake of his neighbors) ». Plus récemment, Rand (2005) a redéfini l'altruisme comme le fait de chercher à satisfaire en premier lieux le bien-être des autres avant son propre intêret : « altruism is the doctrine which demands that man lives for others and places others above self ».

L'Altruisme selon Comte : l'altruisme est le fait de « vivre pour autrui » (Comte, 1966).

Il est important de noter que Rand comme Comte définissent l'altruisme en opposition à l'égoïsme. Si nous considérons l'égoïsme d'un agent comme le fait de satisfaire uniquement ses préférences, alors la stabilité au sens de Nash modélise l'égoïsme. Comme le fait de vouloir satisfaire prioritairement les préférences des autres agents n'est, à notre connaissance, représenté par aucun concept de solution canonique, nous proposons de définir une nouvelle condition de déviation consistant à améliorer la satisfaction des préférences d'au moins un autre agent.

DÉFINITION 15 (Condition d'altruisme). — *Soit* $\Pi \in \mathcal{P}_N$ *et* $D \in AllD_i(\Pi)$ *une déviation. D satisfait la condition d'altruisme (notée* Δ_{alt}*), si pour* $\Pi' = [D \to \Pi]$,

$$\exists a_j \in N \setminus \{a_i\} \ : C_j(\Pi') \succ_j C_j(\Pi)$$
$$\wedge \forall a_k \in N \setminus \{a_i\} \ : C_k(\Pi') \succeq_k C_k(\Pi)$$

La première partie de la condition implique que la déviation doit être profitable pour au moins un agent, la seconde qu'elle ne doit pas être au désavantage d'un tiers quel qu'il soit. Remarquons que, comme notre modèle n'exprime pas de relation privilégiées entre les agents, il ne pourrait être considéré comme altruiste une déviation profitable à un agent si cela se fait au désavantage d'un tiers. En effet, sans un modèle de relation privilégiée, dévier pour améliorer la satisfaction d'un tiers au détriment d'un autre ne permet pas l'existence d'une partition stable. Pour preuve, si a_i dévie pour améliorer a_j au détriment de a_k, il existera toujours pour a_i une déviation altruiste dans la nouvelle partition formée : dévier pour améliorer a_k au détriment a_j. Nous avons là un cycle qui empêche la stabilité.

De plus, la définition de l'altruisme insiste sur le fait qu'il s'agit avant tout d'un acte personnel que nous pouvons représenter par la condition d'individualité Δ_I. Enfin, un acte altruiste peut être soit à l'avantage, soit au désavantage de l'agent qui l'effectue. Durkheim appelait ce dernier cas un *suicide altruiste* (Durkheim, 1897) : un agent commet un suicide altruiste lorsqu'il effectue une déviation qui lui est défavorable pour le bien d'un autre. Cela nous permet alors de définir deux concepts de déviation associés à l'altruisme :

Altruisme : $\mathbb{D}_i := \Delta_I \wedge \Delta_{alt}$

Suicide altruiste : $\mathbb{D}_i := \Delta_I \wedge \Delta_{alt} \wedge \neg\Delta_R$

Comme dit précédemment, Rand (1964) oppose l'altruisme à l'égoïsme. Si nous considérons un égoïsme modélisé par une stabilité au sens de Nash, cette opposition se retrouve bel et bien lorsque nous considérons le suicide altruiste. En effet, le suicide altruiste implique nécessairement des déviations irrationnelles [3] De plus, une étude d'Endriss (2006) propose une taxonomie des règles de concession dans le contexte de la négotiation et le concept de *concession égocentrique* où un agent fait volontairement une proposition en sa défaveur pour améliorer l'utilité des autres est à rapprocher de cette notion de suicide altruiste.

4.3. Modélisation de l'hédonisme

L'hédonisme est une valeur morale fondée sur la satisfaction des plaisirs personnels. Si la question de la recherche du plaisir a été fortement discutée, en particulier par les philosophes cyrénaïques et épicuriens, Épicure indiquait que « le plaisir excessif actuel doit être évité s'il conduit à une douleur future ». Plus récemment, Mill (1889) discutait ainsi que la satisfaction des plaisirs : « pleasure, and freedom from pain, are the only things desirable as ends; and that all desirable things are desirable either for the pleasure inherent in themselves, or as means to the promotion of pleasure and the prevention of pain ».

Dans le deux cas, la satisfaction des plaisirs ne prend de sens que dans l'évitement des douleurs. Ainsi, nous nous fonderons sur la définition de l'hédonisme donnée par Nicolas de Chamfort (Chamfort, 1857 ; Onfray, 2011) : « Jouis et fais jouir, *sans faire de mal ni à toi, ni à personne*, voilà je crois, toute la morale ».

D'un côté, un agent hédonique doit chercher à satisfaire ses propres préférences. De l'autre côté, l'agent doit aussi satisfaire les préférences des autres. Ces deux aspects se traduisent respectivement par la satisfaction des conditions de rationalité (Δ_R), d'acceptation (Δ_A) et de défection (Δ_D). Ainsi, à partir de cette définition, l'hédonisme peut être associé au concept de déviation :

$$\mathbb{D}_i \ := \ \Delta_R \wedge \Delta_A \wedge \Delta_D$$

En termes de concept de solution, cet hédonisme est équivalent à un concept de *stabilité du cœur contractuelle*, concept de solution qui n'existe pas dans la littérature classique. Notons que comme pour la stabilité du cœur, le concept d'hédonisme peut être affaibli en considérant des préférences non strictes. Cet hédonisme faible (que nous définissons par $\mathbb{D}_i \ := \ \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$) signifie que l'agent a_i va chercher à satisfaire ses préférences, sans aller à l'encontre des préférences des autres.

3. Trivialement, le suicide altruiste inclut par définition $\neg\Delta_R$.

4.4. Propriétés de ces nouveaux concepts de solutions

Modéliser les trois valeurs précédentes à l'aide de concepts de déviation nous permet de définir des solutions à un jeu de coalitions qui ne sont pas couvertes par les concepts de solution classiquement utilisé dans la littérature. En faisant l'hypothèse que les agents désirent respecter les mêmes valeurs, nous pouvons définir les nouveaux concepts de solution suivants :

Stabilité au sens de la liberté : $\Pi \in \mathcal{P}_N$ est *stable au sens de la liberté* (noté $\Pi \in LS$) si et seulement si :

$$\forall a_i \in N, \forall C \in N_i \ : \exists a_j \in N \setminus \{a_i\} \ : \ C_j(\Pi) \succ_j C_j([C \to \Pi])$$

Stabilité altruiste : $\Pi \in \mathcal{P}_N$ est *altruistement stable* (noté $\Pi \in AS$) si et seulement si :

$$\begin{aligned} &\forall a_i \in N, \nexists C \in N_i \ : \exists a_j \in N \setminus \{a_i\} \ : \ C_j([C \to \Pi]) \succ_j C_j(\Pi) \\ &\quad \wedge \forall a_k \in N \setminus \{a_i\}, C_k([C \to \Pi]) \succeq_j C_j(\Pi) \end{aligned}$$

Stabilité hédonique : $\Pi \in \mathcal{P}_N$ est *hédoniquement stable* (noté $\Pi \in HS$) si et seulement si :

$$\begin{aligned} &\forall a_i \in N, \nexists C \in N_i \ : \ C \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \succ_j C_j(\Pi) \\ &\wedge \forall a_k \in N \setminus C \ : \ (\exists a_j \in C, C_k(\Pi) = C_j(\Pi)), C_k(\Pi) \setminus C \succ_k C_k(\Pi) \end{aligned}$$

Le tableau 4 positionne ces trois nouveaux concepts de solution en fonction des concepts de déviation qui leur sont associés.

Tableau 4. Concepts de solution en fonction des concepts de déviation

	Δ_I	$\Delta_I \vee \Delta_C$
Δ_R	Nash-stabilité	?
$\Delta_R \wedge \Delta_A$	Stabilité Individuelle	Stabilité du Cœur
$\Delta_R \wedge \Delta_D$	Stabilité Contractuelle de Nash	?
$\Delta_R \wedge \Delta_A \wedge \Delta_D$	Stabilité Individuelle Contractuelle	**Hédonisme**
$\Delta_R \wedge \Delta_{PO}^-$	?	Pareto-Optimalité
Δ_O	?	Optimalité
Δ_{alt}	**Altruisme**	?
$\neg\Delta_R \wedge \Delta_{alt}$	**Suicide altruiste**	?
$\Delta_A^- \wedge \Delta_D^-$	?	**Liberté**

Il s'agit ici d'un complément du tableau 3 où nos trois concepts de solution correspondent à des situations qui ne sont pas représentées par les concepts de solution canoniques. Étudions quelques propriétés de ces nouveaux concepts de solution en considérant d'un côté l'existence d'une solution qui les satisfait, et de l'autre leurs relations d'inclusion vis-à-vis des concepts de solution canoniques.

4.4.1. Existence des partitions stables au sens de la liberté

La stabilité au sens de la liberté est un concept de solution où il n'existe pas nécessairement de solution stable.

PROPRIÉTÉ 16. — *Il existe des jeux hédoniques HG tel que $LS = \emptyset$.*

Intuitivement, la stabilité au sens de la liberté est un concept de solution pouvant être vide car les agents peuvent désirer réaliser des déviations irrationnelles tant que celles-ci ne mécontentent pas les autres agents.

PREUVE 17 (Par l'exemple). — Considérons le jeu de coalitions hédonique HG suivant :

- $N = \{a_1, a_2\}$
- $\{a_1, a_2\} \succ_1 \{a_1\}$
- $\{a_2\} \succ_2 \{a_1, a_2\}$

Dans ce jeu, nous avons deux partitions possibles : $\Pi_1 = \{\{a_1\}, \{a_2\}\}$ et $\Pi_2 = \{\{a_1, a_2\}\}$. Π_1 n'est pas stable au sens de la liberté puisque a_2 peut réaliser la déviation $D_1 = \{a_1, a_2\}$. Cette déviation est cependant irrationnelle en termes de satisfaction des préférences pour a_2. De même, Π_2 n'est pas stable au sens de la liberté puisque a_1 peut réaliser la déviation $D_2 = \{a_1\}$. Ainsi, ce jeu HG ne possède pas de partition stable au sens de la liberté. ■

4.4.2. Existence des partitions hédoniquement stables

La stabilité hédonique est un concept de solution non vide.

PROPRIÉTÉ 18. — *Soit $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ un jeu hédonique. Il existe nécessairement au moins une partition $\Pi \in \mathcal{P}_N$ tel que $\Pi \in HS$.*

PREUVE 19. — Nous allons prouver l'existence d'une partition hédoniquement stable par construction d'un jeu hédonique de déviation où tous les agents considèrent comme concept de déviation $\mathbb{D}_i := \Delta_R \wedge \Delta_A \wedge \Delta_D$. Nous montrons qu'il existe nécessairement au moins une partition $\Pi \in \mathcal{P}_N$ tel que $\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$, cette partition étant hédoniquement stable.

Soit un jeu de déviation $HGD = \langle N, (\succeq_i)_{a_i \in N}, (\mathbb{D}_i)_{a_i \in N} \rangle$ avec n agents et une première partition $\Pi_1 = \{\{a_1\}, \ldots, \{a_n\}\}$.

Considérons dans un premier temps l'agent a_1. $\mathbb{D}_1(\Pi_1, HGD) = \emptyset$ signifie que quelles que soient les déviations que a_1 propose, cela est au désavantage d'au moins un autre agent. Notons alors $\Pi_2 = \Pi_1$.

Supposons maintenant que $D \in \mathbb{D}_1(\Pi_1, HGD) \neq \emptyset$. Soit $D^* \in \mathbb{D}_1(\Pi_1, HGD)$ telle que $\forall D \in \mathbb{D}_1(\Pi_1, HGD), D^* \succ_1 D$. Soit $\Pi_2 = [D^* \to \Pi_1]$. Par choix de D^*, nous avons alors nécessairement $\mathbb{D}_1(\Pi_2, HGD) = \emptyset$.

Considérons maintenant l'agent a_2 à partir de la partition Π_2. Par construction de Π_2, s'il existe $D \in \mathbb{D}_2(\Pi_2, HGD)$, alors nécessairement $a_1 \notin D$.

L'agent a_2 peut ainsi effectuer la déviation $D^{*_2} \in \mathbb{D}_2(\Pi_2, HGD)$ telle que $\forall D \in \mathbb{D}_2(\Pi_2, HGD), D^{*_2} \succ_2 D$ pour passer dans une partition $\Pi_3 = [D^{*_2} \to \Pi_2]$.

Ainsi en appliquant successivement pour chaque agent les déviations appartenant à $D^{*_i} \in \mathbb{D}_i(\Pi_i, HGD)$, nous obtenons nécessairement une partition Π_n telle que $\forall a_i \in N, \mathbb{D}_i(\Pi_n, HGD) = \emptyset$, c'est-à-dire que une partition hédoniquement stable. ■

4.4.3. Existence des partitions altruistement stables

Il n'existe pas nécessairement de partition altruistement stable.

PROPRIÉTÉ 20. — *Il existe des jeux hédoniques HG tel que $AS = \emptyset$.*

PREUVE 21 (Par l'exemple). — Reprenons l'exemple de la preuve 17. Π_1 n'est pas altruistement stable puisque pour satisfaire les préférences de a_1, a_2 désire la déviation $D_1 = \{a_1, a_2\}$. $\Pi_2 = \{\{a_1, a_2\}\}$ n'est elle non plus pas altruistement stable puisque l'agent a_1 désire la déviation $\{a_1\}$ pour satisfaire les préférences de a_2. Ainsi, ce jeu ne possède pas de partition altruistement stable. ■

De manière intéressante, ce cas illustre des situations où par « politesse » deux personnes se laissent mutuellement la priorité, conduisant à des situations d'interblocage.

4.4.4. Relations d'inclusion de les nouveaux concepts

Certaines propriétés intéressantes des concepts de solution sont leurs relations d'inclusion. Pour cela, nous nous fondons sur la propriété 13. Par lisibilité, nous dénotons dans la suite par $\mathbb{D}_{SC}$ le concept de déviation associé au concept de solution SC. Par exemple, $\mathbb{D}_{LS} := \Delta_A^- \wedge \Delta_D^-$. En fin de section, la figure 1 résume l'ensemble des relations d'inclusion entre les concepts de solution.

Considérons dans un premier temps le cas de la stabilité au sens de la liberté et de la stabilité hédonique.

PROPRIÉTÉ 22. — *Toute partition $\Pi \in \mathcal{P}_N$ stable au sens de la liberté est nécessairement hédoniquement stable.*

PREUVE 23. — Nous avons les deux concepts de déviation : $\mathbb{D}_{LS} := \Delta_A^- \wedge \Delta_D^-$ et $\mathbb{D}_{HS} := \Delta_R \wedge \Delta_A \wedge \Delta_D$.

Par définition des formes affaiblies des conditions de déviation, toute déviation D qui satisfait la condition Δ_A (resp. Δ_D) satisfait nécessairement sa forme affaiblie Δ_A^- (resp Δ_D^-). De par la propriété 13, nous avons la relation d'inclusion $\mathbb{D}_{HS} \subseteq \mathbb{D}_{LS}$. Cette relation d'inclusion entre les concepts de déviation se traduit par la relation d'inclusion $LS \subseteq HS$. ■

Considérons maintenant le cas de la stabilité hédonique et de la stabilité individuelle contractuelle.

PROPRIÉTÉ 24. — *Toute partition $\Pi \in \mathcal{P}_N$ hédoniquement stable est nécessairement individuellement contractuellement stable.*

PREUVE 25. — Nous avons les deux concepts de déviation : $\mathbb{D}_{HS} := \Delta_R \wedge \Delta_A \wedge \Delta_D$ et $\mathbb{D}_{ICS} := \Delta_I \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$. De par la propriété 13, nous avons la relation d'inclusion $\mathbb{D}_{ICS} \subseteq \mathbb{D}_{HS}$. Cette relation d'inclusion entre les concepts de déviation se traduit par la relation d'inclusion $HS \subseteq ICS$. ■

Remarquons que comme $LS \subseteq HS$, nous avons également la relation d'inclusion $LS \subseteq ICS$. De la même manière, la stabilité hédonique est un concept de solution inclus dans les concepts de stabilité individuelle, de stabilité au sens de Nash et de stabilité au sens du cœur (la preuve suit le même principe que précédemment). Enfin, une partition Pareto-optimale est nécessairement hédoniquement stable.

PROPRIÉTÉ 26. — *La stabilité hédonique satisfait les relations d'inclusion suivantes : $NS \subseteq IS \subseteq HS$, $CS \subseteq IS \subseteq HS$ et $PO \subseteq HS$.*

Nous allons montrer ici uniquement la relation d'inclusion $PO \subseteq HS$.

PREUVE 27. — Toute partition hédoniquement stable n'est pas nécessairement Pareto-optimale car la Pareto-optimalité considère des successions de déviations, ce que ne fait pas la stabilité hédonique. Nous montrons dans la suite que toute partition Pareto-optimale est nécessairement hédoniquement stable.

Considérons une partition $\Pi \in PO$ et supposons que $\Pi \notin HS$. Par définition de la stabilité hédonique, il existe une déviation D telle que, pour Π' la partition résultante de $[D \rightarrow \Pi]$, nous avons $\forall a_i \in N, C_i(\Pi') \subset C_i(\Pi)$. Cela va à l'encontre de la définition de la Pareto-optimalité et donc de notre hypothèse de $\Pi \in PO$. Nous avons donc une contradiction. ■

La stabilité au sens de la liberté ne présente pas de relation d'inclusion avec le concept de stabilité individuelle (et par extention avec la stabilité au sens du cœur et la stabilité au sens de Nash).

PROPRIÉTÉ 28. — *La stabilité au sens de la liberté satisfait la relation $IS \not\subseteq LS$.*

PREUVE 29 (Par l'exemple). — Considérons dans un premier temps le jeu HG_1 avec $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ avec :

- $N = \{a_1, a_2, a_3\}$
- $\{a_1, a_3\} \succ_1 \{a_1, a_2\} \succ_1 \{a_1\}$
- $\{a_1, a_2\} \succ_2 \{a_2\}$
- $\{a_1, a_3\} \succ_3 \{a_3\}$

Soit la partition $\Pi = \{\{a_1, a_2\}, \{a_3\}\}$. Cette partition n'est pas individuellement stable puisque l'agent a_1 peut effectuer la déviation $D = \{a_1, a_3\}$ qui satisfait les conditions Δ_I, Δ_R et Δ_A. Par contre, elle n'est pas stable au sens de la liberté puisqu'il n'existe pas de déviation qui satisfait la condition Δ_D. Nous avons ainsi une partition $\Pi \in \mathcal{P}_N$ telle que $\Pi \in IS$ et $\Pi \notin LS$.

Considérons maintenant le jeu HG_2 avec $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ avec :

- $N = \{a_1, a_2, a_3\}$

– $\{a_1, a_2, a_3\} \succ_1 \{a_1\}$
– $\{a_2, a_3\} \succ_2 \{a_1, a_2, a_3\} \succ_2 \{a_2\}$
– $\{a_2, a_3\} \succ_3 \{a_1, a_2, a_3\} \succ_3 \{a_3\}$

Soit la partition $\Pi = \{\{a_1, a_2, a_3\}\}$. Cette partition est individuellement stable. Par contre, l'agent a_1 peut réaliser la déviation $D = \{a_1\}$ puisque celle-ci ne s'effectue qu'à ses propres dépends. Nous avons ainsi une partition $\Pi \in \mathcal{P}_N$ telle que $\Pi \notin IS$ et $\Pi \in LS$. ■

Considérons enfin le cas de l'altruisme.

PROPRIÉTÉ 30. — *La stabilité altruiste satisfait le relation $LS \subseteq AS$.*

PREUVE 31. — Rappelons que la définition de la condition d'altruisme Δ_{alt} implique nécessairement la satisfaction des deux conditions Δ_A^- et Δ_D^-. Ainsi, nous pouvons écrire $\mathbb{D}_{alt} := \Delta_I \wedge \Delta_{alt} \wedge \Delta_A^- \wedge \Delta_D^-$. En conséquence par la propriété 13, nous avons nécessairement la relation d'inclusion $\mathbb{D}_{AS} \subseteq \mathbb{D}_{LS}$. En termes de concepts de solution, nous avons donc $LS \subseteq AS$. ■

Remarquons que comme l'altruisme permet (voir oblige pour le cas du suicide altruiste) les déviations irrationnelles, il n'existe pas de relation d'inclusion entre la stabilité altruiste et les concepts de solution canoniques (dont la Pareto-optimalité), ni entre la stabilité altruiste et la stabilité hédonique.

Enfin, la figure 1 résume l'ensemble des relations d'inclusion entre les concepts de solution canoniques et les concepts que nous proposons.

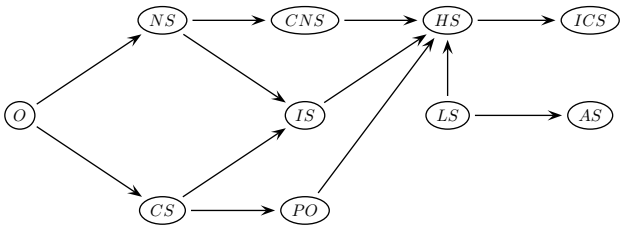

Figure 1. Relations d'inclusions entre les concepts de solution

5. Conclusion et perspectives

Dans cet article, nous nous sommes intéressés au problème de la formation de collectifs d'agents guidés par une éthique des vertus. Au-delà du problème de la coopération entre les agents une fois les collectifs formés, nous nous sommes plus spécifiquement intéressés au processus de formation de coalitions au regard de valeurs humaines que les agents désirent respecter.

Pour ce faire, nous avons proposé un nouveau modèle de jeux de coalitions – les *jeux de déviation hédoniques* – où chaque agent exprime des conditions qui lui sont propres pour caractériser la manière dont il tient compte des objectifs des autres

agents. Nous avons reformulé les concepts de solution classiquement utilisés dans la littérature sous forme de formules appliquées à des conditions de déviation. Nous avons montré que certaines compositions de condition permettent de retrouver les concepts classiques de la littérature. Nous avons ensuite modélisé trois valeurs humaines – la liberté, l'altruisme et l'hédonisme – en nous appuyant sur ces mêmes conditions afin de définir de nouveaux concepts de solution.

Nous nous sommes limités dans cet article à des conjonctions de six conditions de déviations portant sur les préférences (sous leurs formes fortes et affaiblies) et deux conditions sur l'identité des agents impactés par la déviation. Cependant, nous pouvons constater qu'il existe visiblement des concepts de solution qui n'ont pas encore été étudiés. Tout comme pour la liberté, l'altruisme et l'hédonisme, ces manques peuvent peut-être être associés à des valeurs. Une perspective à court terme sera alors de définir ces nouveaux concepts de solutions. Ceci pourrait se faire soit par une approche descendante comme nous l'avons réalisé – partir de définitions en philosophie et sciences sociales puis caractériser des conditions de déviation – soit par une approche ascendante – partir des conditions de déviation que nous avons caractérisées et identifier des valeurs selon leurs compositions. Par exemple, il pourrait être intéressant de penser d'autres formes de liberté telles que considérer que toute déviation est acceptable tant qu'elle n'empêche pas les autres agents de dévier à leur tour.

Dans des travaux précédents, nous avons montré que les jeux hédoniques avec des concepts de solutions propres à chaque agent que les problèmes d'existence d'une solution sont dans Σ_2^p tandis que les problèmes d'appartenance sont coNP-complet (Boissier *et al*., 2018). Il serait alors intéressant d'étendre ces résultats au cas des jeux de déviation.

Enfin, pour les perspectives à long terme, deux points nous paraissent intéressants. Dans un premier temps, il s'agirait de permettre aux agents d'exprimer une éthique des vertus non pas sur une unique valeur cardinale mais sur un ensemble de valeurs humaines afin de modéliser la notion de *système de valeurs* en nous appuyant sur des travaux en sociologie. Ces systèmes pourraient soit correspondre à des concepts de déviation satisfaisant plusieurs valeurs simultanément, soit à une relation de préférence entre plusieurs concepts de déviation. Dans un second temps, il serait intéressant d'étudier les protocoles permettant de calculer une solution pour jeux de déviation hédonique, voire de les implémenter dans une architecture d'agent éthique comme, par exemple, celle de Cointe *et al*. (2016).

Remerciements

Ce travail a été réalisé dans le cadre du projet ANR ETHICAA (ANR-13-CORD-0006).

Bibliographie

Akerlof G. A. (1984). Gift exchange and efficiency-wage theory: Four views. *The American Economic Review*, vol. 74, n° 2, p. 79–83.

Arkin R. (2009). *Governing lethal behavior in autonomous robots*. Chapman and Hall.

Aziz H., Brandt F., Harrenstein P. (2013). Pareto optimality in coalition formation. *Games and Economic Behavior*, vol. 82, p. 562 - 581.

Aziz H., Brandt F., Seedig H. G. (2011). Stable partitions in additively separable hedonic games. In *10th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 183–190.

Aziz H., Brandt F., Seedig H. G. (2013). Computing desirable partitions in additively separable hedonic games. *Artificial Intelligence*, vol. 195, p. 316–334.

Ballester C. (2004). NP-completeness in hedonic games. *Games and Economic Behavior*, vol. 49, n° 1, p. 1–30.

Bardi A., Lee J., Hofmann-Towfigh N., Soutar G. (2009). The structure of intraindividual value change. *Journal of Personality and Social Psychology*, vol. 97, n° 5, p. 913-929.

Bardsley N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, vol. 11, n° 2, p. 122–133.

Batson C. D. (2014). *The altruism question: Toward a social-psychological answer*. Psychology Press.

Bench-Capon T. (2002). Agreeing to differ: modelling persuasive dialogue between parties with different values. *Informal Logic*, vol. 22, n° 3, p. 231-245.

Berreby F., Bourgne G., Ganascia J.-G. (2017). A declarative modular framework for representing and applying ethical principles. In *16th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 96-104.

Bogomolnaia A., Jackson M. O. (2002). The stability of hedonic coalition structures. *Games and Economic Behavior*, vol. 38, n° 2, p. 201–230.

Boissier O., Bonnet G., Cointe N., de Swarte T., Vallée T. (2018). *Building ethical collectives*. Rapport technique. ANR ETHICAA.

Boltanski L., Thévenot L. (2006). *On justification: Economies of worth*. Princeton University Press.

Brandl F., Brandt F., Strobel M. (2015). Fractional hedonic games: Individual and group stability. In *14th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 1219–1227.

Brey P. (2014). From moral agents to moral factors: the structural ethics approach. In P. Kroes, P.-P. Verbeek (Eds.), *The moral status of technical artefacts*, p. 125-142. Springer Netherlands.

Brey P. (2015). *International differences in ethical standards and in the interpretation of legal frameworks*. Rapport technique. SATORI.

Bringsjord S., Taylors J. (2012). Introducing divine-command robot ethics. In P. Lin, G. Bekey, K. Abney (Eds.), *Robot ethics: the ethical and social implication of robotics*, p. 85-108. MIT Press.

Chamfort N. (1857). *Maximes, pensées, anecdotes, caractères et dialogues*. Durr.

Cointe N., Bonnet G., Boissier O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *15th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 1106–1114.

Coleman K. (2001). Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, vol. 3, n° 4, p. 247-265.

Comte A. (1966). *Catéchisme positiviste*. Arnaud.

Dambra S. (2005). Durkheim et la notion de morale. *Revue Interrogation*, vol. 1.

DDHC. (1789). *Déclaration des Droits de l'Homme et du Citoyen de 1789 - Article 4.*

Dennis L., Fisher M., Winfield A. (2015). Towards verifiably ethical robot behaviour. In *1st International Workshop on AI and Ethics*, p. 45–52.

Dreze J. H., Greenberg J. (1980). Hedonic coalitions: Optimality and stability. *Econometrica*, vol. 48, n° 4, p. 987–1003.

Durkheim E. (1893). *De la division du travail social: étude sur l'organisation des sociétés supérieures*. Alcan.

Durkheim E. (1897). *Le suicide: étude de sociologie*. Alcan.

Eckel C. C., Grossman P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, vol. 16, n° 2, p. 181–191.

Elkind E., Wooldridge M. (2009). Hedonic coalition nets. In *8th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 417–424.

Endriss U. (2006). Monotonic concession protocols for multilateral negotiation. In *5th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 392–399.

Ganascia J. (2007). Modeling ethical rules of lying with answer set programming. *Ethics and Information Technology*, vol. 9, p. 39-47.

Gigerenzer G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, vol. 2, n° 3, p. 528-554.

Greenberg J. (1994). Coalition structures. In R. Aumann, S. Hart (Eds.), *Handbook of game theory with economic applications (volume 2)*, p. 1305–1337. Elsevier.

Halpern J. Y. (2015). Cause, responsibility, and blame: a structural-model approach. *Law, Probability, and Risk*, vol. 14, n° 2, p. 91–118.

Hoffman E., McCabe K., Smith V. L. (1996). Social distance and other-regarding behavior in dictator games. *The American Economic Review*, vol. 86, n° 3, p. 653–660.

Hursthouse R. (2013). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.*

Lorini E. (2012). On the logical foundations of moral agency. In *11th International Conference on Deontic Logic in Computer Science*, p. 108-122.

Mill J. S. (1869). *On liberty*. Longmans, Green, Reader, and Dyer.

Mill J. S. (1889). *L'utilitarisme*. Alcan.

Montesquieu C.-L. d. S. de. (1867). *Esprit des lois*. Librairie de Firmon Didot Frères.

Morgenstern O., Von Neumann J. (1953). *Theory of games and economic behavior*. Princeton University Press.

Nash J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, n° 1, p. 48–49.

Nongaillard A., Mathieu P. (2011). Reallocation problems in agent societies: a local mechanism to maximize social welfare. *Journal of Artificial Societies and Social Simulation*, vol. 14, n° 3, p. 5.

Onfray M. (2011). *Manifeste hédoniste*. Autrement.

Pereira L., Saptawijaya A. (2007). Modelling morality with prospective logic. *Lecture Notes in Artificial Intelligence*, vol. 4874, p. 99-111.

Peters D., Elkind E. (2015). Simple causes of complexity in hedonic games. In *14th International Conference on Autonomous Agents and Multi-Agent Systems*, p. 617–623.

Platon. (1966). *La République*. Garnier Flammarion Paris (traduction G. Leroux).

Rand A. (1964). *The virtue of selfishness*. Penguin.

Rand A. (2005). *The fountainhead*. Penguin.

Rokeach M. (1973). *The nature of human values*. New York Free Press.

Saptawijaya A., Pereira L. (2014). Towards modeling morality computationally with logic programming. In *16th International Symposium on Practical Aspects of Declarative Languages*, p. 104-119.

Schwartz S., Bilsky W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross cultural replications. *Journal of Personality and Social Psychology*, vol. 58, p. 878-891.

Schwartz S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, vol. 2, n° 1, p. 11.

Sung S. C., Dimitrov D. (2007). On myopic stability concepts for hedonic games. *Theory and Decision*, vol. 62, n° 1, p. 31–45.

Valette-Florence P., Odin Y., Vinais J. (1996). *Analyse confirmatoire des domaines motivationnels de schwartz : Une application au domaine des media*. Rapport technique. CERAG – Université Grenoble 2.

Vallée T., Bonnet G. (2017). Jeux de coalitions hédoniques à concepts de solution multiples. In *25es Journées Francophones sur les Systèmes Multi-Agents*, p. 53–62.

Wiegel V., van den Berg J. (2009). Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, vol. 1, n° 3, p. 233-242.