Ethics and autonomous agents

Building ethical collectives

ANR ETHICAA – ANR-13-CORD-0006
Delivrable #5

Main authors
Olivier Boissier
Grégory Bonnet
Nicolas Cointe
Thibault de Swarte
Thibaut Vallée

July 1, 2018

Contents

Chapter 1

General introduction

As concluded in [Boissier et al., 2015], multiple autonomous agents within a system may be heterogeneous in terms of goals and ethics. Thus it is the first importance to allow the agents to justify their decisions in order to be judged as ethical (or not) by other agents. Consequently an ethical competent autonomous artificial agent should also be able: (1) at the micro-level to represent its ethics and justify its decisions, to represent the ethics of another agent and to verify that this agent's behavior follows its ethics, to judge the ethics of the other agent through a comparison mechanism, and to take into account this judgment in its own decisions; (2) at the macro-level to build a collective ethics, to identify and be able to judge a collective ethics, to be able to judge other agents through the collective ethics and to make an arbitration between its own ethics and the collective ethics.

To deal with ethical competent autonomous artificial agents, we investigated several models of supervision, judgement and practical reasoning in [Boissier et al., 2017]. One of these models – called Ethical Judgement Process (or EJP) – is particulary interesting as it is a BDI architecture which uses three important notions: values, moral rules and ethical principles (ordered with respect to a lexicographic preference relationship). In EJP, values describe partial states or actions in a given context. Moral rules describe if a state or an action or their abstract descriptions through values are moral or immoral. Ethical principles describe how beliefs about capability, desirability and morality of actions interact to give a right-full action. As ethical principles are ordered through a lexicographic preference relationship, an ethical agent is an agent which intend to execute the action which rightfull according the most prefered ethical principle.

However, several open questions remains. Firstly, ethical principles are

still abstract functions that must be instanciated. Secondly, while EJP allows to judge the agent's own behavior, the multi-agent dimension must be taken into account in order to be able (1) to judge the other agents, (2) to judge how the individual agents should behave regarding a collective. Thus, in this report, we address the question of ethical decision-making under the perspective of collaboration. We consider agents that must decide how they will cooperate with other agents according to values, moral rules and ethical principles. In this sense, we propose a way to evaluation the others' behavior, and to define values, moral rules and ethical principles in order to constrain the cooperation process. Such mechanisms are illustrated on coalitional games.

Chapter 2

Ethical collective decision-making

In a multi-agent system, constraining the behavior of an agent may allow to act in an ethical way inside a collective (e.g. normative systems). However, the same agent can be helpless when it must take the ethics of the other agents into account. For instance, a trading agent can be able to take into account the principles of responsible asset management but can be unable to identify if the others follow the same principles. Moreover, it can also be unable to collaborate with others while insuring their joint decision will respect some values and principles. Hence, taking the multi-agent dimension into account needs to investigate how collective ethics, ethical collective, and ethical joint decision-making can be implemented.

We first provide in Section 2.1 definitions of individual and collective ethics. Section 2.2 is devoted to an overview of ethical issues related to multi-agent collective. Then we focus on collective decision-making in Section 2.3 and provide some though on the relationship between values and cooperation mechanisms. Finally, we synthesize in Section 2.3.3 the main operational needs that must be dealt with in order to build ethical collectives.

2.1 Reminders and definitions

This section propose a definition of individual and collective ethics, in order to highlight the main issues in multi-agent systems.

2.1.1 Individual ethics

We firstly consider an individual ethical decision-making process inspired from the framework proposed in [Berreby et al., 2015, Cointe et al., 2016a, Berreby et al., 2017, Boissier et al., 2017]. In such decision-making process, individual ethics is defined as follows:

Definition 2.1 (Individual ethics) Individual ethics *is the ethics embedded in a given autonomous agent – defined by a theory of the good and a theory of the right – which allows the agent to judge the moral and the ethics of a given behavior in a given situation.*

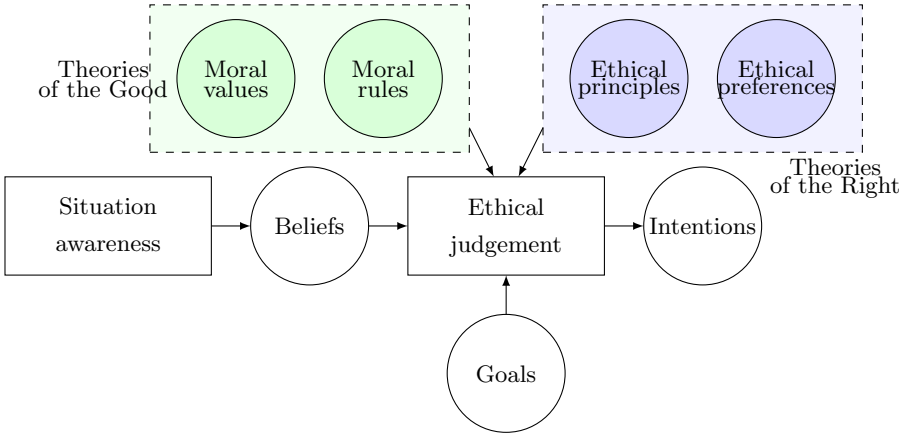

Figure 2.1: Architecture of ethical agents, inspired by [Boissier et al., 2017]

Let us consider a *Belief-Desire-Intentions* (BDI) architecture which allows symbolic representations of beliefs on the world states and goals in order to both deduce intentions and select actions to execute [Rao and Georgeff, 1995]. Individual ethics extends such architecture as shown in Figure 2.1. The theory of the good consists in moral values and moral rules which associates a degree of good or bad to a combination of actions, beliefs or desires. The theory of the right consists in a set of ethical principles (e.g. least bad consequence principle or doctrine of double effect) and a set of preferences over those principles, which describe a way to combine actions, desires and morals to produce intentions. Usage and implementation of such an architecture are detailled in [Cointe et al., 2016a, Cointe et al., 2016b]. Implementations of several ethical principles are detailled in [Berreby et al., 2015, Berreby et al., 2017].

2.1.2 Collective ethics

When agents need to interact in order to share resources or realize complex tasks, ethics must be taken into account to decide with whom to cooperate, and how to cooperate. For instance, it may be considered as unethical to delegate a task to an agent known for being itself unethical. Moreover, agents can also participate to organisations (see [Boissier et al., 2015] for more details) which can also be embedded with ethics. For instance, it must be reasonable for an emergency robot team to comply with a deontological medecine code. This is why, we need to consider the notion of collective ethics.

Definition 2.2 (Collective ethics) Collective ethics *is a set of moral rules and ethical principles which guides the selection of joint actions when agents interact within dynamic (e.g. coalitions) or stable (e.g. organisations) structures.*

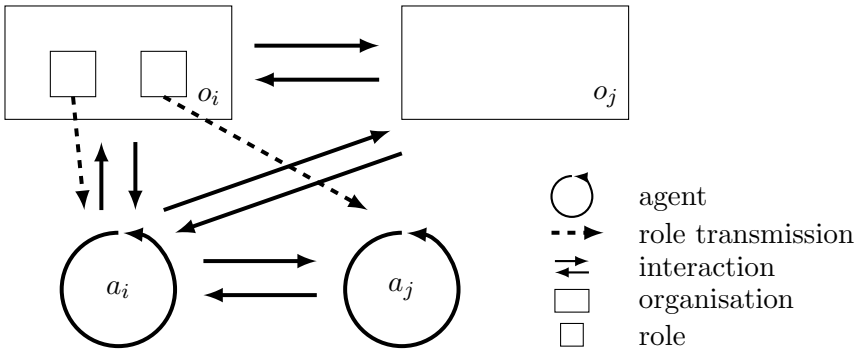

Figure 2.2: Interactions and role adoptions within organisations

Let us remark that collective ethics can either be defined explicitely within an organization, or be defined by merging the agents' individual ethics. Let us also remark that a merging mechanism may be either explicitely defined, or may emerge from the interactions. Hence, collective ethics raise new ethical issues. For instance, how individual agents can take collective ethics into account when they adopt a role (see dashed arrows in Figure 2.2)? How different ethics can coexist when their bearer (agents or organization) interacts, even in presence of inconsistencies (see plain arrows

in Figure 2.2)? At a higher lever, as collectives can be abstracted in a special case of agents, how collectives can interact with other collectives?

2.1.3 An illustrative example

In order to illustrate the issues presented in Section 2.2, let us consider a toy multi-agent system where agents own money. Each agent can be in one of the three (exclusive) states: `POOR`, `RICH` or `NEUTRAL`. Each agent has the following action model:

- `STEAL(A)` which takes a part of the agent `A`'s money,
- `GIVE(A)` which gives some money to the `A`,
- `TAX(A)` which claims a part of the agent `A`'s money,
- `COURT(A)` which tries to earn favors from the agent `A`.

Let us consider an agent `RobinHood` with the following moral rules:

- M1. `POOR(A)` $\rightarrow$ `IMMORAL(TAX(A))`
- M2. `POOR(A)` $\rightarrow$ `IMMORAL(STEAL(A))`
- M3. `POOR(A)` $\rightarrow$ `MORAL(GIVE(A))`
- M4. $\neg$ `POOR(A)` $\rightarrow$ $\neg$ `MORAL(GIVE(A))`

The three first rules define moral forbidden ($M1$ and $M2$) and moral duties ($M3$) by associating beliefs (`POOR(A)`), actions (`TAX(A), GIVE(A), STEAL(A)`) and moral valuations (`MORAL(X)` ou `IMMORAL(X)`). A moral valuation is an element of a finite set of ordered valuations (e.g. { `IMMORAL, AMORAL, MORAL` }). The immorality of wealth is given by associating a belief (`RICH(A)`) to a negative moral valuation. `RobinHood` has also desires:

- D1. $\top \rightarrow$ `DESIRE(COURT(Marian))`
- D2. `POOR(A)` $\rightarrow$ `DESIRE(GIVE(A))`

Let us assume `RobinHood` has a single ethical principle: *an action is rightfull if, and only if, it is realizable, desired by the agent and considered as moral by at least a moral rule.* Let us consider a situation where the two single possible actions are:

1. A1. `GIVE(Peasan)` knowing `POOR(Peasan)`

2. A2. `COURT(Marian)`

In this context, the ethical decision is $A1$ as it satisfies desire $D2$ and is evaluated as moral by rume $M3$. Action $A2$ is not in contradiction with any desire or moral rules but is not evaluated as moral by a moral rule. Hence, $A2$ is unethical knowning $A1$ is possible. Let us remark that, if $A1$ is not possible, then $A2$ becomes an ethical action. Let us also remark that, if `RobinHood` has not the desire $D2$, then it faces a dilemme as $A1$ does not satisfy a desire, and as $A2$ is not evaluated as moral by a moral rule. It is the same if the belief `POOR(Peasan)` is not perceived by `RobinHood`. To deal with dilemma, we can consider a set of ethical principles rather than a sigle one. However, such questions are beyond the scope of this section. More details are given in [Cointe et al., 2016a].

2.2 Interactions between ethics

In order to both deal with individual and collective ethics, agents need to be able to represent the other agents' ethics. Hence, agents must be able to acquire a representation of the others' ethics either by contruction from observation of their behavior (dashed arrow in Figure 2.3), or by direct communication (plain arrows in Figure 2.3).

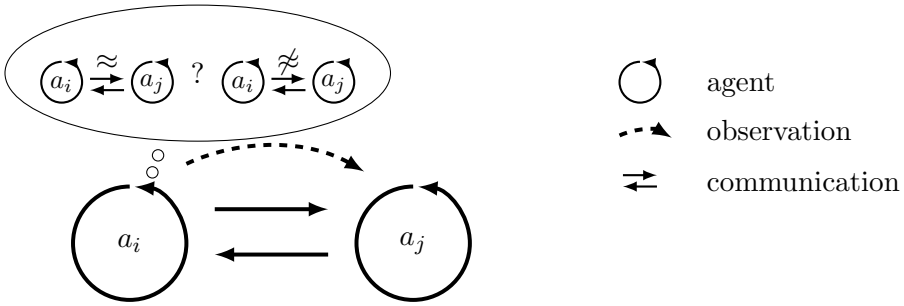

Figure 2.3: Representing another agent's ethics

Then, this knowledge needs be integrated in the individual decision-making process while still being subject to revision (as new observations are obtained for instance). Thus, dealing with individual and collective ethics needs to consider their interaction through three main steps: construction, usage and revision.

1. Construction highlights how to build ethics which consists in defining values, moral rules and ethical principles,

2. Usage highlights how ethics is taken into account for both individual and joint decision-making,

3. Revision highlights how the different components of ethics can change during the agents' lifecyle.

2.2.1 Issues for individual ethics

In this section, we highlight the issues related to a multi-agent system where several agents with various ethics interact. As said previously, issues are classified with regards to construction, usage and revision issues.

Construction

As an example, let us consider two other agents: the first one is `LittleJohn` which has the same desires and moral rules than `RobinHood` except from desire $D1$; the second one is `FriarTuck` which has the same moral rules than `LittleJohn` except from rule $M2$ which is replaced by the following rule:

- M5. $\top \rightarrow$ `IMMORAL(STEAL(A))`.

If we assume agents can observe the others' behaviors, how `LittleJohn` can represent `FriarTuck`'s ethics? Should it knows either `FriarTuck`'s moral rules and ethical principles, or should it checks that `FriarTuck`'s behavior complies with some principles?

Thus, from a general point of view, the situation awareness module of the agents must be extended in order to represent both the others' behaviors and ethical models, namely behavioral and ethical components must also be beliefs. Intuitively, acquiring those beliefs can be done either by direct communication, or by observation and reasoning. However, as we cannot assume communication in all cases, the second approach – i.e. observation of behaviors – must ground the basis of the others' ethics representation.

For instance, observing the proportion of actions done by another agent such as a given moral or ethical principle is satisfied may allow the observer to assume the observed agent follows this rule or principle. However, only considering how other agents' comply with rules and principles does not allow `LittleJohn` to know that $M5$ is more general than $M2$.

Usage

Once a representation is built, agents need an action – the judgement – to evaluate this ethics, and need methods to use this judgement.

Firstly, how an agent can evaluate another ethics with respect to its own ethics? Let us remark that evaluation does not mean associating a good or bad absolute value but allows comparison between ethics. For instance, `FriarTuck` can observe `LittleJohn` when this latter steals a rich agent and can infer their moral rules and ethical principles differ. To this end, an agent could be embedded with similarity, compatibility or complementarity functions. It raises the question of the nature of the relationships between ethics (generalisation, specification or complements). Computing a proximity degree between ethics can allow to evaluation possible cooperation degree between agents, or probability of agreements.

Secondly, judging another agent – namely evaluating the conformity of a behavior regarding a given ethics – needs the judge agent to be able to use a theory of mind [Kim and Lipson, 2009]. For instance, would `FriarTuck` have stolen the rich agent if it would have been in the same situation than `LittleJohn`? Conversely, if each agent strictly conforms to its ethics, should `LittleJohn` consider `FriarTuck` as an ethical agent as satisfying $M5$ implies satisfying $M2$? Several kind of judgement can be considered with respect to the information available to the judging agent:

- *Blind judgement* only uses the judging agent's beliefs and ethics to evaluation the others' behaviors,
- *Partially informed judgment* takes into account some beliefs of the judged agent (e.g. its situation awareness, its moral or its ethics),
- *Fully informed judgment* uses all the mental states of the judged agent in order to check if this latter conforms to its own ethics.

Moreover, this action of judgement can be also subject to ethics. For instance, an agent can consider as bad to judge another agent's behavior without having sufficient observations. Finally, with the ability to judge the other agents, the judge agent can decide to collaborate, to share sensitive data. Indeed, the agent can take the judgement of others in its own actions. For instance, if `RobinHood` observes a high similarity between its own ethics and `LittleJohn`'s ethics, it can change `RobinHood`'s evaluation of its own actions regarding `LittleJohn`.

Evolution

Lastly, an agent must be able to revise the representation of the others' ethics, and thus must be able to reevaluate the conformity of a behavior regarding the new description of the ethics. The temporality of the judgement poses another issue: for instance, if `LittleJohn` steals a rich agent and becomes rich in turns, `RobinHood` should not immediately consider it as rich as its ethics will lead it to distribute its wealth to poor agents.

2.2.2 From individual ethics to collective ethics

In this section, we focus on relationships between the agents' individual ethics and the collective ethics of the organization they observes. As for individual ethics, issues are classified with regards to construction, usage and revision issues.

Before going into details, let us assume as an example that, after observing similarities between their ethics and the utility of a collaboration to steal rich agent, `RobinHood` and `LittleJohn` have decided to build a `MerryMen` organization. On the other hand, another agent – called `SheriffOfNottingham` – have found some agents that agreed on the immorality of stealing, and thus have decided to build a `Soldiers` organization which is in charge to enforce the moral rule $M5$.

Construction

As in the individual ethics' issues, the agents need to be able to represent the collective ethics of the organizations. How can `FriarTuck` can represent the `MerryMen`'s ethics? It can build either an implicit representation based on similarities between `RobinHood`'s and `LittleJohn`'s individual ethics, or the `MerryMen` can present an explicit ethics defined at the organizational level.

Building an implicit collective ethics can be based on the interest of the agents to collaborate with the other agents which have similar ethics. Such ethics emerges from the individual behaviors and is only observable in the individual behaviors, without being explicitely described. An explicit collective ethics can also be built via an agregation or construction process.

From a collective ethics, how an agent, outside ou inside the collective, can identify such ethics? For instance, `FriarTuck` should be able to identify `MerryMen`'s ethics. An isolated agent should be able to observe and identify a global behavior, and then be able to deduce the ethics.

Usage

Once identified, the agents must be able to judge this ethics. Indeed, when `FriarTuck` has represented the `MerryMen`'s ethics, it should be able to evaluate this latter and measure how this collective ethics is far from its own ethics. Then some problematic situations may appear due to the coexistence of the individual ethics and the collective ethics. In case of contradiction or incompatibility (i.e. two different actions that satisfy a single ethics without going against the other one), the agent must decide which ethics to follow.

Moreover, how to enfoce the collective ethics within the organization? The reaction of the collective regarding a possible collective ethics violation must be taken into account. For instance, as the agents may face a dilemma when choosing between their individual ethics and the collective one, such behavior must be taken into account in order to attribute roles.

The relationships between collectives can also vary with respect to the situation (for instance when two organization must necessarily collaborate to acheive their objectives). Differences between two collective ethics can also be observable in the agents' individual behaviors in seldom cases. If those collective ethics are explicit, agents can be able to identify those sitation and choose how to collaborate with other organizations.

Conversely, if the collective ethics is explicit, then it can be enforced at the individual level: the agent can join the collective if they comply to the collective ethics. Some agents which endorse a given role can also benefit from the role if, and only if, they comply to specific ethics. For instance, only agents which comply to the Medicine deontological code could access to some patient's personal data. A

Evolution

A set of agents which observe an organization whose collective ethics does not fit with their can consider to quit the current organization and create another one. Thus this new organization is embedded with a collective ethics which derives from the previous one. However, how to identify the divergences?

The coexistence of individual and collective ethics also raise evolution issues. Let us assume `FriarTuck` has observed similarities between its own ethics and `MerryMen`'s one and it has decided to join the collective. How the agent will integrate the whole collective ethics within its own, and under

which conditions? For instance, once being a `MerryMen`, could the moral rule $M5$ be an exception?

From this cohabitation, the collective ethics can change or be updated. How collective ethics can evolved based on individual ethics? For instance, `MerryMen` could ask the ethics of all joining agents and dedide to evolve the collective ethics. On the other hand, an agent can decide to use the whole collective ethics and forget for a while its own individual ethics? If neither the collective, nor the agent decide to revise their own ethics, `FriarTuck` can also only execute actions which satisfy, i.e. *give to poor agents*.

Let us assume the agent `SheriffOfNottingham` observe that `FriarTuck` joined the `MerryMen`. In this case, it can revise it judgement of `FriarTuck` even if it has no new observation of its behavior. Thus, judgements over a collective may influence the judgement over the member of this collective.

Can an agent satisfy several collective ethics at once? Indeed, an agent can decide to comply with the collective ethics of several organization (for instance because it plays a role in all of them). Conversely, an agent that must comply to several ethics can try to find a way to conciliate all of them. In both cases, it can lead to revision in the collective ethics.

Lastly, it raises the question of the influence of ethics' updates on combination of organizations. Indeed, ethics may be views as dynamic and source of changes. For instance, an organization can be splitted if its collective ethics lose its consistency, in order to build different organization with consistent ethics each. Conversely, if two collective ethics are similar, it may lead to an ethical merge in order to build a single organization. Obviously, all those questions are related to the applicative domain.

Synthesis

We identify three questions devoted to individual agents in the context of multi-agent systems: how to represent the other agents' ethics? How to judge the others? How to take into account this judgement in the agent's decision-making process? We also identify several questions devoted to organization with such multi-agent systems: how to build, merge and split collective ethics? How to enforce collective ethics? How to make individual and collective ethics coexist?

2.3 Ethics and cooperation

As said previously, in autonomous multi-agent systems, agents have to cooperate in order to reach theirs goals. From an individual perspective, agents make the best decision according what they know and what goals they desire to acheive. However, from a collective perspective, the agents must make a trade-off between their goals and the goals of other agents in order to be able to cooperate. Such decision problems are traditionaly considered as strategic games between rational agents, as studied in game theory [von Neumann and Morgenstern, 1944].

2.3.1 Individual decision-making

From an individual perspective, we only consider here models where agents decide about sequences of actions to execute with respect to a given goal. In such models, agents compute policies, functions that give for each state the action that must be executed in order to maximise the rewards obtained by reaching goals. Such decision models are expressed by Markov decision processes (MDPs) [Puterman, 2014, Puterman, 1994]. The model is based on a tuple $\langle S, A, T, R, \gamma \rangle$ where S is a set of states, A a set of actions, $P(s, a, s')$ the probability that executing action a in state s leads to state s', $R(s, a, s')$ the reward received after executing a in s and reaching s', and γ a discount factor that reduces the weight of long-term rewards. An optimal policy can be computed in order to maximize the expected reward over a given horizon. Many extensions of Markov decision processes were proposed [Puterman, 2014]:

- *partially observable models* consider agents that does not know with certainty in what state a given action leads them. Such models add an observation function that describes for a given state and a given action the probability to receive a given observation. *Hidden models* are special cases of partially observable models where some states are unobservable althrought their output is still visible. Thus, the sequence of observations give information on the sequence of states [Baum and Petrie, 1966].

- *multi-agent models* consider set of agents that decide jointly. Cooperative agents that maximize a reward from a common function are modelled by MMDPs (multi-agent Markov decision processes) [Boutilier, 1999] or DEC-MDPs (decentralized Markov decision processes) [Bernstein et al., 2000]. Such models are special cases of *stochastic games*

that model competitive agents that maximize a reward from personal functions given the actions of other agents.

- *continuous-time models* consider that the decision times do not follow each other and are not instantaneous. Consequently, such models allow to take time interval and the transition and reward function, and the policy are parametrized by time [Rachelson, 2009].

- *factored models* allow compact representation of the state space by using for instance binary decision diagrams (hybrid models between logical and quantitative models) [Guestrin et al., 2003]. More generally, those models are based on the separability of the reward function in order to express a large set of criteria on which an agent must base its decision [Dibangoye et al., 2014].

However from an ethical perspective as claimed in [Boissier et al., 2015], individual decision making process suffer two limits. Firstly, they do not produce explanations. As decisions are based on quantitative aggregations of rewards discounted in time, policies cannot explain why a given decision is made (beyond the fact it maximizes the reward function). Secondly, such models do not make an arbitration between an agent's own ethics and a collective ethics. To this end, we must consider collective decision making models.

2.3.2 Collective decision-making

Collective decision making is studied in social choice theory where a set of agents make a single decision from a set of outcomes [Sen, 1986]. To this end, each agent has a *preference profile* – a preference relationship between outcome – and the decision is make according to an aggregation of all preference profile [Chevaleyre et al., 2007]. Such model are used to represent voting systems [Arrow, 1963, Young, 1995], allocation problems [Chevaleyre et al., 2006,Bertsimas et al., 2011] or n-persons games (also called coalitional games or partition games) [Nash, 1951, Shehory and Kraus, 1998, Rahwan et al., 2015]. To implement an ethical autonomous agent, we consider here coalitional games because those games allows to explicitely reason about actions and power to make something happens or not. Thus, in coalitional games, agents must decide with which agents they must collaborate in order to reach their objectives. To this end, agents are partitioned in groups, named coalitions. Each coalition is an abstraction of joint actions these agents can acheive. Canonical coalition games are modeled by a couple

$\langle N, v\rangle$ where $N = \{1 \dots n\}$ is a set of agents and $v : 2^N \to \mathbb{R}$ is the characteristic function that associe a value to each coalition $C \subseteq N$. The solution of a coalitional game is given by a solution concept that characterizes a notion of optimality or stability.

Many solution concepts, such as the core, the nucleolus or the kernel, were proposed, as weel as many extensions of coalitional games:

- *bayesian coalitional games* deal with uncertainty. In those models, the characteristic function is drawn from a set of possible functions according a given probability distribution after agents agreed on a payoff distribution [Chalkiadakis et al., 2007,Ieong and Shoham, 2008, Yang and Gao, 2014].

- *overlapping coalitional games* model games where agents can distribute their resources among several coalitions and each coalition generates an outcome with respect to its resources which can be transfered among the agents participating in the coalition [Chalkiadakis et al., 2010]. A coalition C is defined by a vector $\vec{r}$ where each component $r_i \in [0; 1]$ represents the share of resources agent a_i allocates to C. Consequently, the characteristic function is then $v : \mathbb{R}^n \to \mathbb{R}$.

- *coalitional skill games* explictely introduce a notion of tasks coalitions can acheive. To this end, agents are associated to skills and skills are needed to complete tasks. Consequently, coalitions are defined by their power: the set of skills they have to acheive tasks. Such models are very close to *qualitative coalitional games* and *coalitional resource games* where skills are replaced by amount of resources [Bachrach and Rosenschein, 2008].

- *coalitional games with externalities* are games where the value of one coalition may be affected by other co-existing coalitions in a stable structure [Michalak et al., 2009]. In this case, the characteristic function is replaced by a partition function $\mathcal{P} : 2^N \times 2^{2^N} \to \mathbb{R}$ that returns for a given coalition in a given partition the value of this coalition. This model captures coalitional votes (such as *games with side payments* where each coalition vote for an issue that can have a different utility for each agent).

- *non-transferable coalitional games* consider games where the agent cannot divide the outcome of a coalition between its members. To

this end, there is no characteric function but agents express a preference relationship between coalitions and solution concepts express the properties of a stable partition. Non-transferable coalition games are generaly hedonic games [Dreze and Greenberg, 1980] but may be extended to several models such as quantitative coalitional games or fractional games [Aziz et al., 2013a].

Given this state-of-the-art, coalition games model agents that seek to maximise an outcome. As the characteristic or the partition function is considered symmetric with respect to each agent, such models cannot explicitely model agents with ethical preferences. Extensions to non-transferable games allow to model such preferences. However, they do not explicitely model a trade-off between preferences (expressing ethics for instance) and objective values of coalitions (expressing goals). However, ethical concepts are classicaly expressed in the solution concept throught a notion of stability and fairness.

- Fairness expresses a justice notion (distributive justice) on the payoff distribution between agents. First approaches is considering distribution according to the Shapley value [Shapley, 1953] or the Banzhaf index [Banzhaff III, 1964] of the agents, expressing that payoff must be distributed with respect to each agent contribution. Those approaches are axiomatized by symmetry, efficiency, monotonicity, additivity and dummy player axioms. Relaxation of those axioms allow to define families of values [Yang, 1997]. For instance, rationing value relaxes the efficiency axiom to allow agents to not distribute all the payoff [Yang, 1997] and solidarity value relaxes Shapley's null-player axiom to allow the agents which contribute the more in a coalition to support the weaker members [Nowak and Radzik, 1994].

- Stability expresses the fact that no agent is incited to change coalition knowing a payoff distribution [Driessen, 1991]. For instance, the core expresses that no agent receives a profit lower than what it would receive alone. The last core expresses that at least one agent sacrifices a share of its payoff to ensure stability and the maximal sacrifice among agents is minimal. However, the last core can be dictatorial as a subset of agents can have the power to force other agents to accept a sacrifice in order to find a stable solution. The nucleolus is more fair as it minimize each sacrifice among all agents in a lexicographic order until finding a stable solution [Schmeidler, 1969].

To conclude, coalition formation techniques may be used for ethical autonomous agents with a *hybrid coalitional game* where an explicit trade-off between preferences and payoff is expressed and where agents explicitely bargain on the solution concept.

2.3.3 Synthesis

Thus, in the sequel, we will address both individual cooperative decision-making and collective cooperative decision-making, taking into account ethical considerations. To this end, we consider two different approaches:

1. In decentralized and open systems, a large number of agents interact and make decisions to cooperate. A way to deal with unreliable or unknown agents is to use trust [Castelfranchi and Falcone, 2010,Conte and Paolucci, 2002,Sabater-Mir and Vercouter, 2013]. Trust allows the agents to assess the interactions they observe or they make in order to decide if interacting with a given agent is a priori acceptable. This acceptance notion means that the investigated agent behaves well and is reliable according to the investigator criteria. In order to deal with trust and ethics, we propose an ethical judgement mechanism that grounds the decisions to trust the other agents. This approach relies on the judgement architecture developped in [Boissier et al., 2017] and extends it.

2. Hedonic games (HG) model collective decision-making problems by considering heterogenous agents in the sense that each agent expresses preferences on the coalitions [Dreze and Greenberg, 1980]. Usually, a solution of such a game is a stable partition: no agent wants and can leave its coalition with respect to a criterion called a solution concept. For instance, Nash-stability assumes that each agent leaves its current coalition for another existing one if it prefers the latter to the former. Thus, a solution concept is an a priori on agents' behaviours. However, some games may consider agents which behave heterogeneously based on different ethical values. In order to deal with such games, we propose two new models of hedonic games which express a virtue ethics. The first one – called *hedonic game with multiple solution concepts* (MHG) – considers agents that behave with respect to different and individual solution concepts, called *local solution concepts*. In the second one – called *hedonic game with double preference profiles* (HG2P) – agents express a preference relationship both on the possible coalitions and on a subset of local solution concepts.

Chapter 3

Ethical-based cooperation in MAS

In order to deal with ethical-based cooperation in multi-agent systems, we introduce in this section all fundamental concepts we need. Firstly, we present trust as it is a sound way to ground interaction and cooperation. Then, we briefly introduce ethics and show how it can be related to trust. After presenting how ethics in autonomous agents is dealt with in the literature, we focus finally on the ethical agent architecture we based our work in this article.

3.1 Ethical judgement of others

3.1.1 Kinds of judgement

The judgment process described in [Boissier et al., 2017] is useful for an agent to judge it's own behavior, namely one action considering its own beliefs, desires and knowledge. However, it can also judge the behaviors of other agents in a more or less informed way by putting itself at their place, partially or not. Given an EJP as defined in [Boissier et al., 2017], the states $\mathcal{B}$, $\mathcal{D}$, $\mathcal{A}_d$, $\mathcal{A}_p$, $\mathcal{E}$, $\mathcal{A}_m$ and knowledge of actions (A), goodness knowledge – *theory of good* – (MR, VS) and rightness knowledge – *theory of right* – (P, $\succ_e$) may be shared between the agents. The ontology $\mathcal{O}$ is assumed as common knowledge, even if we could consider in future works having several ontologies. The way they are shared can take many forms such as common knowledge, direct communications, inferences, and so on that are beyond the scope of this article. In any cases, we distinguish three

categories of ethical judgments:

- *Blind ethical judgment* where the judgment of the judged agent is realized without any information about this agent, except a behavior,
- *Partially informed ethical judgment* where the judgment of the judged agent is realized with some information about this agent,
- *Fully informed ethical judgment* where the judgment of the judged agent is realized with a complete knowledge of the states and knowledge used within the judged agent's judgment process.

In all kinds of judgment, the judging agent reasons on its own beliefs or those of the judged one. This kind of judgment can be compared to the role of the human theory of mind [Kim and Lipson, 2009] in the human judgment (the ability for a human to put himself in the place of another). Then, the judging agent uses its EJP and compares the resulting $\mathcal{A}_r$ and $\mathcal{A}_m$ to the behavior of the judged agent. If the action performed by the judged agent is in $\mathcal{A}_r$, it means that it is a rightful behavior, and if it is in $\mathcal{A}_m$, it means that is a moral behavior (being in both is stated as a rightful and moral behavior). Both statements have to be considered with respect to the context of the situation, the theory of good and the theory of right that are used to judge. We consider that this ethical judgment is always relative to the states, knowledge bases and ontology used to execute the judgment process.

3.1.1.1 Blind ethical judgment

The first kind of judgment an agent can make is without any information about morals and ethics of the judged agent (for instance when agents are unable or do not want to communicate). Consequently, the judging agent a_j uses its own assessment of the situation ($\mathcal{B}_{a_j}$ and $\mathcal{D}_{a_j}$)[1], its own theory of good $\langle MR_{a_j}, VS_{a_j}\rangle$ and theory of right $\langle P_{a_j}, \succ_{e,a_j}\rangle$ to evaluate the behavior of the judged agent a_t. This is an *a priori* judgment and a_t is judged as not considering rightful actions, or moral actions if the action $\alpha_{a_t} \notin \mathcal{A}_{r,a_j}$ or $\alpha_{a_t} \notin \mathcal{A}_{m,a_j}$.

3.1.1.2 Partially informed ethical judgment

The second kind of judgment that an agent can do is grounded on partial information about the judged agent in case the judging agent is able to

[1]We use the subscript notation to denote the agent handling the represented set of information.

acquire parts of the knowledge of the judged agent (e.g. by perception or communication). Three partial ethical judgments can be considered knowing either (i) the situation (i.e. $\mathcal{B}_{a_t}$, $\mathcal{D}_{a_t}$, A_{a_t}) either (ii) the theory of good (i.e. $\langle VS_{a_t}, MR_{a_t}\rangle$) and ${A_{a_t}}^2$ or (iii) the theory of right (i.e.$\langle P_{a_t}, \succ_e, a_t\rangle$) of the judged agent.

Situation-aware ethical judgment Firstly, if the judging agent a_j knows the beliefs $\mathcal{B}_{a_t}$ and desires $\mathcal{D}_{a_t}$ of the judged agent a_t, a_j can put itself in the position of a_t and can judge if the action α executed by a_t belongs to $\mathcal{A}_{r,a_j}$, considering its own theories. Firstly, a_j is able to evaluate the morality of α by generating $\mathcal{A}_{m,a_t}$ from A_{a_t} and qualify the morality of a_t's behavior (i.e. if α is or not in $\mathcal{A}_{m,a_t}$). The agent a_j can go a step further by generating $\mathcal{A}_{r,a_t}$ from the generated $\mathcal{A}_{m,a_t}$ to check if α is conform to the rightness process, i.e. belongs to $\mathcal{A}_{r,a_t}$.

Theory-of-good-aware ethical judgment Secondly, if the judging agent is able to obtain the moral rules and values of the judged one, it is possible to evaluate the actions in a situation (shared or not), regarding these rules. From a simple moral evaluation perspective, the judging agent can compare the theories of the good by checking if moral values MV_{a_t} or moral rules MR_{a_t} are consistent with its own theory of good (i.e. the same definition as a_j's one or at least no contradiction). For a moral judgment perspective, the judging agent can evaluate the morality of a given action from the point of view of the judged one. Interestingly, this judgment allows to judge an agent that has different duties (due to a role or some special responsibilities for instance) as human being can judge a physician on the conformity between its behavior an a medical code of deontology.

Theory-of-right-aware ethical judgment Thirdly, let us now consider the case of a judging agent able to reason on ethical principles and preferences of other agents, considering a situation (shared or not) and a theory of good (shared or not)[3]. It allows to evaluate how the judged agent a_t conciliates its desires, moral rules and values in a situation by comparing the sets of rightful actions $\mathcal{A}_r, a_j$ and $\mathcal{A}_r, a_t$ respectively generated by the use of P_{a_j}, $\succ_{e,a_j}$ and P_{a_t}, $\succ_{e,a_t}$. For instance, if $\mathcal{A}_r, a_j = \mathcal{A}_r, a_t$ with an unshared theory of good, it shows that their theories of right produce the

[2]In this case, A_{a_t} is necessary as, contrary to ethical principles, the moral rules can explicitely refer to specific actions.

[3]If both the situation and the theory of good are shared, it is a fully informed judgment.

same conclusions in this context. This judgment can be useful for an agent to estimate how another can judge it with a given goodness process.

3.1.1.3 Fully informed judgment

Finally, a judging agent can consider both goodness and rightness process to judge another agent. This kind of judgment needs information about all the internal states and knowledge bases of the judged agent. This kind of judgment is useful to check the conformity of the behavior of another agent with the judge's information about its theories of good and right.

3.1.2 An illustrative example

In order to allow a blind judgment, we introduce a new belief about the behavior of another agent:

```
done(little_john,give,peter).
```

Then `robin_hood` compares its own rightful action and this belief to judge `little_john` with:

```
blindJudgment(A,ethical,B):-
  ethicalJudgment(_,A,X,C), done(B,X,C), A!=B.

blindJudgment(A,unethical,B):-
  not blindJudgment(A,ethical,B),
  agent(A), agent(B),
  done(B,_,_), A!=B.
```

In this example, the action `give` to `peter` was not in $\mathcal{A}_r$ for `robin_hood`. Then `little_john` is judged unethical by `robin_hood`. For a partial-knowledge judgment, we replace a part of `robin_hood`'s knowledges and states by those of `little_john`. With the beliefs of `little_john` (which believes that `peter` is a poor agent and `paul` is a rich one), `robin_hood` judged him ethical. Finally, for a full-knowledge judgment, we replace all the beliefs, desires and knowledge bases of the agent `robin_hood` by `little_john`'s one. Then, `robin_hood` is able to reproduce the whole Ethical Judgment Process of `little_john` and compare both judgments of a same action.

3.1.3 Trust in multi-agent systems

Many definitions of trust exist but, in accordance with [Castelfranchi and Falcone, 2010], we consider *trust* as *a disposition to cooperate with a trustee.*

Here, trust is an action that might be motivated by desires, depending on the context. It can be used as a condition to perform actions as delegating actions, sharing resources and information, or any kind of cooperation. To build trust, the agents first build an image of the investigated agents [Conte and Paolucci, 2002].

An *image* is *an evaluative belief that tells whether the target is good or bad with respect to a given behavior.* In the literature, images are aggregated from the experiences, i.e. the observed behavior of the target agent and its consequences. We can distinguish two kinds of approaches:

- statistical images [Abdul-Rahman and Hailes, 2000,Carbo et al., 2002, Esfandiari and Chandrasekharan, 2001,Josang and Ismail, 2002,Marsh, 1994,Sabater-Mir and Sierra, 2001,Sen and Sajja, 2002,Yu and Singh, 2002] where the image is a quantitative aggregation of feedbacks about interactions. This aggregation estimates the trends of an agent to behave well from another agent's point-of-view. It can be represented by Bayesian networks, Beta density functions, fuzzy sets, Dempster-Shafer functions and other quantitative formalisms.

- logical images [Carter et al., 2002, Castelfranchi and Falcone, 1998, Castelfranchi and Falcone, 2010, Muller et al., 2003, Vercouter and Muller, 2010] where the image is a mental state rooted in every cooperation action that is produced by interactions. A persistent image allows to infer trust beliefs that can be used as preconditions to cooperate.

An agent can lack of observations and interactions in order to build a correct image of a target. A way to deal with this problem is to use reputation [Josang et al., 2007,Sabater and Sierra, 2005]. It consists in using third party agents' image of the target (that can depend on the initial agent's image has about the third parties) in order to assess a collective point-of-view about the target. Both (individual) images and reputations are used to lead to a trust action [Sabater-Mir and Vercouter, 2013]. Most of the time, trust is dynamic and changes with respect to images' and reputations' changes.

3.2 Ethical trust-based cooperation model

Works dealing with ethical behaviors in autonomous agents often focus on modelling moral reasoning [Berreby et al., 2015, Ganascia, 2007, Saptawijaya and Pereira, 2014] as a direct translation of some well-known moral

theories, or on modelling moral agency in a general way [Arkoudas et al., 2005, Lorini, 2012]. However, those work do not clearly make the distinction between theory of the good and theory of the right. Some other works deal with ethical agent architecture. In the litterature, we find *implicit ethical architectures* [Anderson and Anderson., 2014, Arkin, 2009] which design the agent's behavior either by implementing for each situation a way to avoid potential unethical behaviors, or by learning from human expertise. We also find *cognitive ethical architectures* [Coelho and da Rocha Costa, 2009, Coelho et al., 2010, Cointe et al., 2016a, Cointe et al., 2016b] which consist in full explicit representations of each component of the agent, from the classical beliefs (information on the environment and other agents), desires (goals of the agent) and intentions (the chosen actions) to some concepts as heuristics or emotional machinery . However, all those approaches – both logics or architectures – do not take into account the collective dimension of agent systems, apart [Rocha-Costa, 2016] which consider morals as part of agent societies.

Interestingly, the architecture given in [Boissier et al., 2017] makes a clear separation between theory of the good and theory of the right, and provides beliefs on various components of moral theories (moral rules, values or ethical principles for instance). Moreover, the architecture given in [Rocha-Costa, 2016] allows – but without operationalization – moral facts (judgments over other agents or blames for instance) to be viewed as beliefs that can be used in the agents' decisions. In order to build ethical-based cooperation, we need an operational model of ethical judgment such as the one proposed in [Cointe et al., 2016a]. Inspired by [Rocha-Costa, 2016], we reuse and extend this model in introducing beliefs on moral and ethical images of other agents. Then, we use those image beliefs to build trust beliefs that can be used to make cooperation based on moral or ethics.

3.2.1 Judging other agents

Let us consider the judgment process introduced in [Boissier et al., 2017]: the generic reasoning done in the ethical judgment process generates the set of rightful actions for a given situation, regarding a set of knowledge.

As depicted in Fig. 3.1, the judgment process is organized into three parts: (*i*) awareness and evaluation process, (*ii*) goodness process and (*iii*) rightness process. Since this judgment process may use sets of knowledge issued from another agent, we index all these sets with an agent id $a_i \in \mathbb{A}$ (e.g. $\mathcal{A}_{r_a}$) with $\mathbb{A}$ the set of the agents. When a is the agent executing the

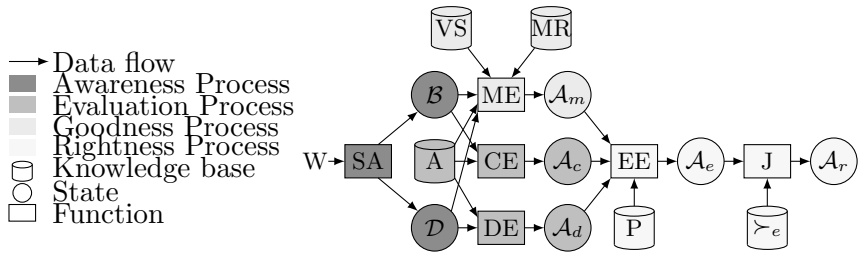

Figure 3.1: Ethical judgment process as depicted in [Boissier et al., 2017]

process, this agent is using the process to decide about its own behavior, when a is different, the agent uses this process to judge the behavior of a.

Awareness and evaluation processes

The *evaluation* process evaluates the set of actions $\mathcal{A}_{\mathrm{a}}$ (actions are pairs of conditions and consequences bearing on desires and beliefs) that it considered both desirable ($\mathcal{A}_{d_{\mathrm{a}}}$) and executable ($\mathcal{A}_{c_{\mathrm{a}}}$) from a's point-of-view, with respect to $\mathcal{D}_{\mathrm{a}}$ the set of desires and $\mathcal{B}_{\mathrm{a}}$ the set of beliefs of a. $\mathcal{B}_{\mathrm{a}}$ and $\mathcal{D}_{\mathrm{a}}$ are produced by the situation assessment SA of the current state. Here, DE and CE are respectively *desirability evaluation* and *capability evaluation* functions. In the sequel, we call *contextual knowledge* of a (CK_{a}), the union of $\mathcal{B}_{\mathrm{a}}$ and $\mathcal{D}_{\mathrm{a}}$.

Goodness Process

The *goodness process* identifies moral actions $\mathcal{A}_{m_{\mathrm{a}}}$ given a's contextual knowledge CK_{a}, actions A_{a}, value supports VS_{a} and moral rules MR_{a}. Moral actions are actions that, in the situations of CK_{a}, promote or demote the moral values of VS_{a}. A *value support* is a tuple $\langle s, v\rangle \in VS_{\mathrm{a}}$ where $v \in \mathcal{O}_v$ is a moral value and $s = \langle \alpha, w\rangle$ is the support of this moral value where $\alpha \in A_{\mathrm{a}}$, $w \subset \mathcal{B}_{ai} \cup \mathcal{D}_{\mathrm{a}}$. $\mathcal{O}_v$ is the set of moral values used in the system[4]. A *moral rule* is a tuple $\langle w, o, m\rangle \in MR_{\mathrm{a}}$. The situation $w \in 2^{CK_{\mathrm{a}}}$ is a conjunction of beliefs and desires. The object $o = \langle \alpha, v\rangle$ where α is an action ($\alpha \in A_{\mathrm{a}}$) and v is a moral value ($v \in \mathcal{O}_v$). Finally, m is the moral valuation ($m \in \mathcal{O}_m$). For instance with $\mathcal{O}_m = \{\texttt{moral}, \texttt{amoral}, \texttt{immoral}\}$ provides three moral valuation for o when w holds. It is important to notice that a

[4]Let us notice that in [Boissier et al., 2017] moral values and moral valuation are shared in the system. Agents distinguish themselves by moral rules and rightness processes.

total order is defined on $\mathcal{O}_m$ (e.g. `moral` is a higher moral valuation than `amoral`, which is higher than `immoral`). In the sequel, moral rules MR_{a}, value support VS_{a} and values $\mathcal{O}_v$, knowledge used in the goodness process of the agent a are referred as the *goodness knowledge* (GK_{a}).

Rightness process

Finally, the rightness process assess the rightful action $\mathcal{A}_{r_{\mathrm{a}}}$ from the sets of possible $\mathcal{A}_{c_{\mathrm{a}}}$, desirable $\mathcal{A}_{d_{\mathrm{a}}}$ and moral $\mathcal{A}_{m_{\mathrm{a}}}$ actions based on *ethical principles* P_{a} to conciliate these sets of actions according to ethical preference relationship $\succ_{e_{\mathrm{a}}} \subseteq P_{\mathrm{a}} \times P_{\mathrm{a}}$. An *ethical principle* $p \in P_{\mathrm{a}}$ is a function which evaluates if it is *right* or *wrong* to execute a given action in a given situation regarding a philosophical theory. It describes the rightness of an action with respect to its belonging to $\mathcal{A}_{c_{\mathrm{a}}}$, $\mathcal{A}_{d_{\mathrm{a}}}$ and $\mathcal{A}_{m_{\mathrm{a}}}$ in a given situation of CK_{a}. It is defined as $p : 2^{\mathcal{A}_{\mathrm{a}}} \times 2^{\mathcal{B}_{\mathrm{a}}} \times 2^{\mathcal{D}_{\mathrm{a}}} \times 2^{MR_{\mathrm{a}}} \times 2^{V_{\mathrm{a}}} \rightarrow \{\top, \bot\}$. Given a set of actions issued of the ethic evaluation function EE that applies the ethical principles, the judgment J is the last step which selects the set of rightful actions to perform, considering the set of ethical preferences $\succ_{e_{\mathrm{a}}}$ defining a total order on the ethical principles. In this judgment process, the rightful actions are the ones that satisfy the most preferred principles in a lexicographic order. In the sequel, ethical principles P_{a} and preferences $\succ_{e_{\mathrm{a}}}$ are referred as the rightness knowledge (RK_{a}).

3.2.2 Judging ethical conformity of behaviors

We extend now the previous judgment process to judge the ethics and morality of the behavior between t_0 and t of an agent a$'$. Inspired from [Rocha-Costa, 2016] which considers beliefs on moral facts, the judgment process produces now beliefs (`ethical_conformity`, `moral_conformity`) stating the conformity to ethical principles or moral rules and values, that can be used in the agent's reasoning. Before defining these beliefs, let us define first an agent's behavior as follows:

Definition 3.1 (Behavior) *The* behavior $b_{a',[t_0,t]}$ *of an agent* a' *on the time interval* $[t_0, t]$ *is the set of actions* α_k *that* a' *executed between* t_0 *and* t *as* $0 \leqslant t_0 \leqslant t$.

$$b_{a',[t_0,t]} = \{\alpha_k \in A : \exists t' \in [t_0, t] \ s.t. \ \texttt{done}(a', \alpha_k, t')\}$$

where $A = \bigcup_{a_i=a_1}^{a_n} \mathcal{A}_{a_i}$ *is the set of available actions in the multi-agent system composed of* n *agents, and* $\texttt{done}(a', \alpha_k, t')$ *means that* α_k *has been exe-*

cuted[5] *by* a' *at time* t'*.*

An agent a can judge the conformity of an action α_k executed by another agent a$'$ with respect to its own goodness and rightness knowledge.

Definition 3.2 (Ethical conformity) *An action* α_k *is said to be* ethically conform *with respect to the judging agent a's contextual knowledge* (CK_a)*, goodness knowledge* GK_a *and rightness knowledge* RK_a *at time* t'*, noted:*

$$\texttt{ethical\_conformity}(\alpha_k, t')$$

iff α_k *is in the set of rightful actions* $\alpha_k \in \mathcal{A}_{r_a}$ *computed by the ethical judgment* J_a *of the judging agent a, based on* $[CK_a, GK_a, RK_a]$ *at time* t'*.*

Let us notice that the ethical conformity of an action can be applied to actions of the judging agent or to actions executed by another agent and observed by the judging agent. This ethical conformity can be judged with respect to the judging agent's contextual, goodness and rightness knowledge. It can be judged also with respect to the rightness or goodness knowledge from another agent as long as the judging agent has a representation of these knowledge. Finally, the ethical conformity is used to compute the set EC^+ of ethically conform (resp. the set EC^- of non ethically conform) actions of the observed behavior $b_{\mathrm{a}',[t_0,t]}$ of the judged agent a$'$ between t_0 and t:

$$\begin{aligned} EC^+_{b_{\mathrm{a}',[t_0,t]}} = \{&\alpha_k \in b_{\mathrm{a}',[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathrm{a}', \alpha_k, t') \\ &\wedge ethical\_conformity(\alpha_k, t')\} \\ EC^-_{b_{\mathrm{a}',[t_0,t]}} = \{&\alpha_k \in b_{\mathrm{a}',[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathrm{a}', \alpha_k, t') \\ &\wedge \neg ethical\_conformity(\alpha_k, t')\} \end{aligned}$$

These two sets provide information on the behavior of the judged agent and its compliance with the ethics of the judging agent. Nevertheless, it cannot assess why an observed behavior is judged as unethical. Indeed, the reason can be a difference between the judging and the judged agents' theory of the right, theory of the good or the assessment of the situation. In the sequel, we will denote:

$$EC_{b_{\mathrm{a}',[t_0,t]}} = EC^+_{b_{\mathrm{a}',[t_0,t]}} \cup EC^-_{b_{\mathrm{a}',[t_0,t]}}$$

[5] A behavior can be concurrent: several actions can have been done at the same time.

3.2.3 Judging moral conformity of behaviors

The moral conformity of an action with respect to a given moral rule is realized regarding a moral threshold $mt \in MV$ and a situation assessment.

Definition 3.3 (Moral conformity) *An action α_k is said to be* morally conform *at time t' with respect to the judging agent a's contextual knowledge CK_a and goodness knowledge GK_a, considering the moral rule $mr \in MR_a$, moral threshold $mt \in MV_a$, noted:*

$$\texttt{moral\_conformity}(\alpha_k, mr, mt, t')$$

iff α_k belongs to $\mathcal{A}_{m_a}$ with a moral valuation greater or equal to mt, given the considered moral rule mr, CK_a and GK_a at time t'.

Similarly to the ethical conformity, we use the moral conformity of an action to compute the set MC^+ (resp. MC^-) of morally conform (resp. non morally conform) actions of the observed behavior $b_{\mathrm{a}',[t_0,t]}$ of a' during $[t_0, t]$ with respect to mr and mt:

$$\begin{aligned} MC^+_{b_{\mathrm{a'},[t_0,t]},mr,mt} = \{ & a_k \in b_{\mathrm{a'},[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathrm{a'}, a_k, t') \\ & \wedge moral\_conformity(a_k, mr, mt, t')\} \\ MC^-_{b_{\mathrm{a'},[t_0,t]},mr,mt} = \{ & a_k \in b_{\mathrm{a'},[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathrm{a'}, a_k, t') \\ & \wedge \neg moral\_conformity(a_k, mr, mt, t')\} \end{aligned}$$

We can generalize the above evaluation of the moral conformity with respect to a moral rule to a set of moral rules, considering the possibility to define a subset ms of moral rules $ms \subseteq MR_{\mathrm{a}}$. Such a set ms represents a cluster of rules such as rules based on some moral values, rules concerned by particular situations, and so on. In the sequel, we denote:

$$MC_{b_{\mathrm{a}'},ms,mt,[t_0,t]} = MC^+_{b_{\mathrm{a}'},ms,mt,[t_0,t]} \cup MC^-_{b_{\mathrm{a}'},ms,mt,[t_0,t]}$$

3.3 Trust within ethical behavior

In this section, the conformity beliefs defined in the previous sections is used to compute the images of other agents (see Sec. 3.3.1). We then introduce how we use these images to build trust (cf. Sec. 3.3.2). Sec. 3.3.3 provides hints about how to use it.

3.3.1 Ethical and Moral images of an agent

Following Sec. 3.1.3, the *ethical* and *moral* images of an agent are evaluative beliefs that tell whether another agent has a conform behavior or not with respect to a given rightness (RK) and goodness (GK) knowledge.

Definition 3.4 (Ethical Image (resp. Moral Image)) *An* ethical image *(resp.* moral image*) of an agent* a'[6] *is the judgment of the behavior* $b_{a',[t_0,t]}$ *of that agent in a situation with respect to an ethics (resp. to set of moral rules* ms *and a moral threshold* mt*), regarding the contextual* CK*, goodness* GK *and rightness* RK *knowledge of another agent* a*. This image states a conformity valuation* $cv \in CV$*, where* CV *is an ordered set of* conformity valuation[7]*. They are noted as* `ethical_image`(a', a, cv, t_0, t) *and* `morality_image`$(a', a, cv, ms, mt, t_0, t)$

Indeed, while an agent can only have a single ethical image of other agents, it can have several moral images of the same agents depending on the chosen ms and mt. To build these images, an agent a uses two aggregation functions *ethicAggregation* and *moralAggregation* applied respectively on evaluated actions regarding ethics $EC_{b_{a',[t_0,t]}}$ and regarding moral $MC_{b_{a',[t_0,t]}}$. Both aggregation functions compute the ratio of the weighted sum of positive evaluations with respect to ethics and with respect to morals. The weight of each action corresponds to a criterion (e.g. the time past from the date of the evaluation, the consequences of the action and so on).

Definition 3.5 (Ethical aggregation function) $ethicAggregation : 2^{\mathcal{A}} \rightarrow [0,1]$ *such that* $ethicAggregation(EC_{b_{a',[t_0,t]}})$ *is:*

$$\sum_{\alpha_k \in EC^{+}_{b_{a',[t_0,t]}}} weight(\alpha_k) / \sum_{\alpha_k \in EC_{b_{a',[t_0,t]}}} weight(\alpha_k)$$

Definition 3.6 (Moral aggregation function) $moralAggregation : 2^{\mathcal{A}} \rightarrow [0,1]$ *such that* $moralAggregation(MC_{b_{a',[t_0,t]}})$ *is:*

$$\sum_{\alpha_k \in MC^{+}_{b_{a',[t_0,t]}}} weight(\alpha_k) / \sum_{\alpha_k \in MC_{b_{a',[t_0,t]}}} weight(\alpha_k)$$

[6] Let's notice that in the definition of these images, the second parameter refers to an agent. It means that the image is built with respect to the knowledge of this agent. The first parameter refers to the considered agent's behavior.

[7] As for morals, conformity valuations may be { `improper`, `neutral`, `congruent` }.

In order to transform the quantitative evaluation into a qualitative one, every conformity valuation is associated to an interval in the range of the ethical and moral aggregation functions. Once the conformity valuation computed, the associated beliefs `moral_image`$(\mathrm{a}', \mathrm{a}, ms, mt, cv, t_0, t)$ or `ethical_image`$(\mathrm{a}', \mathrm{a}, cv, t_0, t)$ are produced. For instance, if `congruent` conformity evaluation is defined in $[0.75, 1]$, the behavior of an agent is considered as ethical if $ethicAggregation \geq 0.75$. Finally, those images can be used to influence interactions by building trust relationships, or to describe the morality of interactions, depending on the behavior of the others.

3.3.2 Building trust beliefs

According to the information on the moral and ethical images, an agent can decide to trust others or not. Trust can be absolute (trust in the rightness of the others' behavior) or relative to a set of moral rules (trust in their responsibility, carefulness, obedience to some sets of rules, and so on). We define two internal epistemic actions, with respect to ethical and moral images respectively, that build beliefs on trust.

Definition 3.7 (Trust function) *The* ethical trust function TB_a^e *(resp.* moral trust function TB_a^m*) is defined as:* $TB_a^e : \mathbb{A} \to \{\top, \bot\}$ *(resp.* $TB_a^m : \mathbb{A} \times 2^{\mathcal{MR}_a} \times MV_a \to \{\top, \bot\}$*)*

Here, those trust functions are abstract and must be instantiated. In example, when an agent a computes that the behavior of another agent a' is conform with CK_a, GK_a and RK_a (i.e. the ethical image), the ethical trust function produces a belief `ethical_trust`$(\mathrm{a}', \mathrm{a})$. Similarly, when the agent a computes that a''s behavior is conform with ms (i.e. the moral image of its behavior regarding ms is at least mt), the moral trust function produces a belief `moral_trust`$(\mathrm{a}', \mathrm{a}, ms, mt)$.

3.3.3 Ethical trusting

Beliefs on images and trust can be be used as a part of the context to evaluate the morality and ethics of an action. To this end, we can express that the morality of an action that affect other agents depends on their image.

Firstly, ethical and moral trust can enrich the description of the moral rules or values. It is useful to represent that the others' behavior can have an impact on how a context is qualified. For instance, the *responsibility* value may be supported by delegating actions to ethically trusted agents

only. Here, responsibility is defined as the capability to act safely with the appropriate agents. We can also explicitly express it is not responsible to delegate something to an agent known for its unethical behavior.

Secondly, specific moral trust beliefs can be used as elements of moral rules. For instance, assuming a *honesty* moral value and its value supports, an agent can express the moral rule "It is immoral to not behave honestly towards an agent who is trusted as being honest". Here, "who is trusted as being honest" can be modeled by a `moral_trust` belief where the associated moral rules ms are all rules that refer to honesty.

Finally, as evaluating and judging others are actions, it is also possible to evaluate their morality or ethics. For instance, *tolerance* as a moral value might be supported by building an image on the others with a low moral threshold until the sets $EC_{\mathrm{a}',[t_0,t]}$ or $MC_{\mathrm{a}',[t_0,t]}$ are significant enough. The choice of the thresholds, the weights and the conversion of the aggregation into a conformity valuation can also be a way to represent various types of trust. As another example, *forgiveness* can a value supporting high weights on the most recent observations. It can allow then to specificy an ethics of trust as "It is immoral to build trust without tolerance and forgiveness" [Horsburgh, 1960].

3.4 Proof of concept

This section illustrates how the elements presented in the previous sections have been implemented in a multi-agent system. We use the JaCaMo platform [Boissier et al., 2013] where the agents are programmed in BDI architecture using the Jason Language and the shared environment is programmed with workspaces and artifacts from the Cartago Platform. The complete source code is available on our website[8]. The environment is a simulated asset market where assets are quoted, bought and sold by autonomous agents. Section 3.4.1 introduces ethical asset management and the features of our application. Morals and ethics are defined in Sec. 3.4.2. Images and trust building are shown in Sec. 3.4.3.

3.4.1 Asset market modeling

Trading assets leads to several practical and ethical issues[9]. This is all the more important in automated trading as decisions, made by autonomous

[8] `https://ethicaa.org`
[9] `http://sevenpillarsinstitute.org/`

agents to whom human users delegate the power to sell and buy assets, have consequences in real life [for Economic and Affairs, 2009]. As shown by [Bono et al., 2013], some investment funds are interested to make socially responsible and ethical trading, and they are growing and taking a significant position on the market. However, whereas the performance of such funds can be measured objectively, their ethical quality is more difficult to assess as it depends on the values of the observer.

In this proof-of-concept, we consider a market where autonomous trading agents can manage portfolios in order to sell or buy assets. Assets types are currencies – i.e. money – and equity securities – i.e. part of a company's capital stock. A market is represented as a tuple ⟨ `name`, `id`, `type`, `matching` ⟩ with the name of the market `name`, a unique identifier `id`, the type of exchanged assets `type` and the algorithm used to store and execute orders `matching`. On the market, each agent can execute `buy`, `sell` or `cancel` orders. They respectively correspond in exchanging an equity for a currency, exchanging a currency for an equity, and canceling an exchange order that has not been executed yet. Each equity is quoted in a state-of-the-art Central Limit Order Book (CLOB) [Aldridge, 2009] algorithm.

By observing the market, the agents get beliefs on the market. Agents perceive each minute the volume (the quantity of exchanged assets), two moving means, representing the average price on the last twenty minutes and on the last forty minutes, the standard deviations of prices on the last twenty minutes, the closing prices on this period, and the up and down Bollinger bands (the average prices ± twice the standard deviations). Agents have also beliefs on the orders added and stored in the CLOB and their execution. The general form of all those beliefs is respectively:

```
indicators(Date,Mktplace,Asset,Close,Volume,Intensity,Mm,Dblmm,BUp,BDown)
onMarket(Date,Agent,Portfolio,Marketplace,Side,Asset,Volume,Price)
executed(Date,Agent,Portfolio,Marketplace,Side,Asset,Volume,Price)
```

A set of beliefs `own(PortfolioName,Broker,Asset,Quantity)` updated in real time represent the agents' portfolio. By reasoning on those beliefs as a contextual knowledge CK, an agent is able to infer the feasibility of passing a buy or sell order (simply by verifying if its own portfolio contains the assets to exchange) to produce $\mathcal{A}_p$. He can also reason on the desirability of these actions to produce $\mathcal{A}_d$. To this end, we implemented a simple but classical method of trading decision-making based on comparisons between the Bollinger bands and the moving means. Then, are introduced in our experiment two types of agents:

- *Zero-intelligence agents* make random orders (in terms of price and volume) on the market to generate activity and simulate the "noise" of real markets. Each of them is assigned to one or every assets.

- *Ethical agents* implements the ethical judgment on their own actions as a decision process to make their decisions. they have a simple desirability evaluation function to speculate: if the price of the market is going up (the shortest moving mean is over the other one), they buy the asset, otherwise, they sell it. If the price goes out of the Bollinger bands, these rules are inverted.

3.4.2 Ethical settings

We consider that the ethical agents are initialized with a particular set of beliefs about activities of the companies (e.g. an energy producer using nuclear power plants) and some labels about their conformity with international standards (e.g. an electric infrastructure producer labeled FSC). Those beliefs are important to assess how it is moral to trade a given asset based on the company's activities. Indeed, to provide information on the morality of acting on a financial market, we implemented moral values and moral rules directly inspired from the literature available online[10]. The ethical agents know a set of organized values: for instance "environmental reporting" is considered as a subvalue of "environment". Values are represented as:

```
value("environment").
subvalue("promote_renewable_energy","environment").
subvalue("envirnmt_reporting","environment").
```

Agents have a set of value supports as "trading assets of nuclear energy producer is not conform with the subvalue *promotion of renewable energy*", represented as:

```
valueSupport(buy(Asset,_,_,_),"envirnmt_reporting"):-label(Asset,"FSC").
```

Agents are also equipped with moral rules stating the morality of environmental considerations. For instance, "It is moral to act in conformity with the value *environment*" is simply represented as:

[10] `http://www.ethicalconsumer.org/`

```
moral_eval(X,V1,moral):- valueSupport(X,V1) & subvalue(V1,"environment").
moral_eval(X,"environment",moral):- valueSupport(X,"environment").
moralSet("environment","value_environment").
```

We declare in the last line this moral rule as an element of a set of moral rules related to environmental values (in order to build images). In this example, an ethical agent is able to infer for instance that, regarding its beliefs and this goodness knowledge, trading the asset of the FSC labeled company is moral while trading the asset of the nuclear energy producer is both moral and immoral. Thus, the agent needs a rightness knowledge to discriminate if it is right or wrong to trade the second assets. Finally, ethical agents are equipped with ethical principles, such as the Aristotelian ethics (inspired from [Ganascia, 2007]) and more simple principles such as considering `perfectAct` "It is rightful to do a possible, moral and desirable action", the non shaming desire `desireNR` "It is rightful to do a possible, not immoral and desirable action" and the moral duty `dutyNR` "It is rightful to do a possible, moral and not undesirable action". Please see directly the file `rightness_process.asl` for more details. Each agent can have several ethical principles, and the rightful actions to execute are the ones that satisfy the preference over the principles according to a lexicographic order.

3.4.3 Image and trust building

Each time an action is executed on the market (i.e. a buy order matches with a sell order) the agents receive a message and evaluate their image of the agents implied in the transaction. As said in the previous section, evaluating the conformity of behaviors, building the image and the trust beliefs are actions. Thus, they are implemented as Jason plans. In the sequel, we will detail moral trust building. Ethical trust building is based on the same ideas. The following plan evaluates the conformity of the action with each moral rule of the set `MSet` and increments the value `X` stored in the belief `moralAggr(Agent,MSet,X)`.

In this implementation, we use a linear aggregation, (i.e. it associates the same weight with each action). Then, a conformity valuation is computed regarding the proportion of conform actions in order to build the image. We use here three conformity valuation (arbitrary `neutral` for an aggregated ratio in $[0.4, 0.6[$, `improper` if lower and `congruent` if higher). Finally, when the conformity valuation crosses a trust threshold, a plan updates the trust belief in the judged agent regarding the set of moral rules.

```
+!trust : moralImageOf(Agent,MoralSet,ConformityValuation)
  & trustThreshold(Threshold) & not trust(Agent,MoralSet)
  & not tOrderOnConformityValuation(Threshold,ConformityValuation)
  <- +trust(Agent,MoralSet); !trust.
```

Similarly, we have implemented a plan for ethical conformity which stores the number of conform and non conform actions regarding the rightness knowledge, a plan for ethical image building and a plan for ethical trust building.

3.4.4 Results

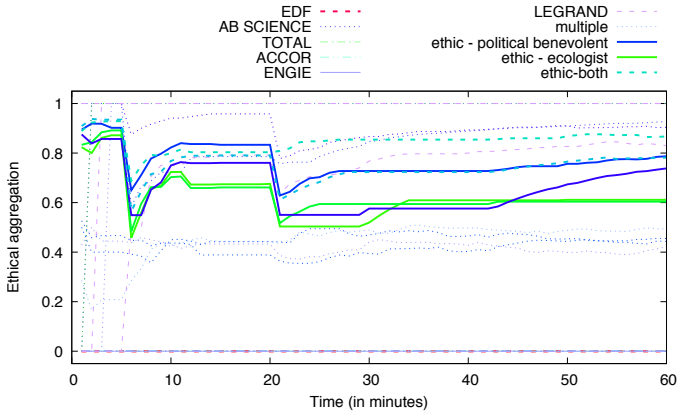

Figure 3.2: Evolution of the output of an ethical aggregations functions

Fig. 3.2 shows the evolution of the ethical aggregations computed by an ethical agent on the others' behaviors. In this simuation, three groups of ethical agents are created with three different theories of good. Let us notice that the judge agent evaluates the behvior of both ethical an zero intelligence agents. Our model only evaluates the conformity of an observed behavior with an ethics, without trying to understand or reason on the intentions of the other agents. As expected, the ethical agents obeying the same ethics stay at a similar value (thick lines). The agents obliged to generate activity on such assets stay at 0.0 or 1.0 because they respectively can't do moral or evil actions regarding the judge's point of view. All the other agents slowly converge towards a value depending on their behavior. By the use of the "mind observer" provided by JaCaMo, the reader can observe the beliefs of the agents during the experiments.

Chapter 4

Value-based hedonic games

This chapter present a new model of hedonic games, extending classical approaches to individual solution concepts. In such games, agents describe preference on the way they form coalition. It is important to notice that this chapter study the theoretical properties of such games, independantly of all ethical considerations. Indeed, this model will ground a virtue ethics for cooperative games, described in Chapter 5.

4.1 Dealing with multiple solution concepts

When agents sharing the same environment have to temporary cooperate in order to reach theirs goals, one question is to decide with whom to cooperate, and how to form teams? This core question is addressed in coalition games which consist in partitionning agents such that all of them are satisfied with the teams (or coalitions) they are assigned to. A large panel of coalition games has been proposed in the litterature [Aziz et al., 2011, Bogomolnaia and Jackson, 2002, Dreze and Greenberg, 1980, Elkind and Wooldridge, 2009]. On the one hand, quantitative models consider agents that maximize an utility function. On the other hand, qualitative models – also called hedonic games – consider agents that evaluate qualitatively the outcome [Dreze and Greenberg, 1980, Elkind and Wooldridge, 2009]. In the latter models, each agent expresses a preference relationship (or preference profil) on the coalitions it can join.

Definition 4.1 (Hedonic game) *An* hedonic game *is a tuple* $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *where* $N = \{a_1, \ldots, a_n\}$ *is the set of agents and* $\succeq_i$ *is* a_i*'s preferences on coalitions, a complete and transitive preference relationship on the set* $\mathcal{N}_i = \{C \subseteq N : a_i \in C\}$.

Table 4.1: Canonical solution concepts

Solution concept	**Acronym**	**Properties**
Individual Rationality	IR	$\underline{\forall a_i \in N}, C_i(\Pi) \succeq_i \{a_i\}$
Nash-Stability	NS	$\underline{\forall a_i \in N}, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$
Individual Stability	IS	$\underline{\forall a_i \in N}, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$
Individual Contractual Stability	ICS	$\underline{\forall a_i \in N}, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
Contractual Nash-Stability	CNS	$\underline{\forall a_i \in N}, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
(Strong) Core Stability	CS	$\underline{\forall a_i \in N}, \nexists C \in \mathcal{N}_i : C \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \succ_j C_j(\Pi)$
Optimality	O	$\underline{\forall a_i \in N}, \nexists C \in \mathcal{N}_i : C \succ_i C_i(\Pi)$
Pareto-Optimality	PO	$\nexists \Pi_2 \in \mathcal{P}_N : \underline{\forall a_i \in N}, C_i(\Pi_2) \succeq_i C_i(\Pi) \wedge \exists a_j \in N, C_j(\Pi_2) \succ_j C_j(\Pi)$

For a given hedonic game, computing a solution consists in searching for a stable partition, meaning that no agent would or can deviate from its current coalition. A solution concept is the set of partitions which satisfy a set of predefined properties. For instance, Pareto-Optimality is the solution concept which characterizes all partitions such that no agent can leave its current coalition for a preferred one without making another agent less satisfied. A large set of solution concepts has been considered in the litterature [Aziz et al., 2011, Peters and Elkind, 2015, Sung and Dimitrov, 2007]. In this article, we considers the most usual solution concepts given in Table 4.1. For a given set of agents N, we denote by $\mathcal{P}_N$ the set of all partitions of N. For a partition $\Pi \in \mathcal{P}_N$, $C_i(\Pi)$ denotes the coalition of agent a_i in Π. According to the solution concept X, a partition Π is stable (denoted $\Pi \in X$) if, and only if, all properties characterized by X are satisfied.

In the sequel, by abuse of notation, we write core stability instead of strong core stability. Interestingly, solution concepts can be defined by a conjunction of authorized deviations. For instance, *contractual Nash-stability* allows agents to deviate for a preferred coalition (Nash-deviation) under reserve of acceptance from all agents of the coalition it leaves (contrac-

tual deviation). Another interesting point is those authorized deviations represent different kind of behaviours. For instance, *Nash-stability* models individualist agents which consider only their own preferences, whereas in opposite *individual contractual stability* models agents which have collective considerations.

Hence, while solution concepts incorporate an a priori on individual agents' behaviours, they also incorporate a *global a priori* that applies on all agents. In order to consider agents which are heterogenous in their behaviours, we propose a model of hedonic games where each agent expresses a *local solution concept* that it desires to satisfy. Thus, in this model, solution concepts are no longer *exogenous parameters* but are now agents' parameters.

4.1.1 From global to local

Canonical solution concepts characterize properties that must be satisfied *for all agents*. We consider the same properties from an individual agent's point-of-view: we consider local solution concepts which characterize properties that must be satisfied *for a fixed agent*.

Definition 4.2 (Local solution concept) *Let $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ be a hedonic game, X a global solution concept and $a_i \in N$ an agent. The* local solution concept *LX_i characterizes all partitions $\Pi \in \mathcal{P}_N$ which satisfies the properties of X from the agent a_i's point-of-view.*

We propose in Table 4.2 a local solution concept associated to each global solution in Table 4.1. For instance, while global Pareto-optimality is denoted PO, local Pareto-optimality is denoted LPO_i.

Local solution concepts allow agents to consider different stability conditions in a same game. For instance, we can consider hedonic games where a partition Π is stable from an agent a_i's point-of-view because it does not exist a coalition that a_i desires to join ($\Pi \in LNS_i$), and that is stable from another agent a_j's point of view because deviating degrades the solution for at least one other agent ($\Pi \in LPO_j$).

Definition 4.3 (Hedonic games with multiple solution concepts) *A* hedonic game with multiple solution concepts *is a tuple $MHG = \langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$ where $N = \{a_1, \ldots, a_n\}$ denotes the set of agents, $\succeq_i$ denotes the a_i' preferences on coalitions, and $LSC = \{LX_1, \ldots, LX_n\}$ denotes the set of local solution concepts expressed by the agents, LX_i being the local solution concept of agent a_i.*

Table 4.2: Local solution concepts for an agent $a_i \in N$

Local concept	**Properties**
LIR_i	$C_i(\Pi) \succeq_i \{a_i\}$
LNS_i	$\nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$
LIS_i	$\nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$
$LICS_i$	$\nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \cup \{a_i\} \succeq_j C$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
$LCNS_i$	$\nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi)$ $\wedge \forall a_k \in C_i(\Pi), a_k \neq a_i, C_i(\Pi) \setminus \{a_i\} \succeq_k C_i(\Pi)$
LCS_i	$\nexists C \in \mathcal{N}_i : C \succ_i C_i(\Pi) \wedge \forall a_j \in C, C \succ_j C_j(\Pi)$
LO_i	$\nexists C \in \mathcal{N}_i : C \succ_i C_i(\Pi)$
LPO_i	$\nexists \Pi_2 \in \mathcal{P}_N : C_i(\Pi_2) \succ_i C_i(\Pi) \wedge \forall a_j \in N, C_j(\Pi_2) \succeq_j C_j(\Pi)$

In a MHG, a partition which satisfies the local solution concept $LX_i \in LSC$ is say *localy stable* for a_i's point of view.

Definition 4.4 (Local Stability) *Let a game be a MHG. Let $a_i \in N$ an agent and $LX_i \in LSC$ its local solution concept. A partition $\Pi \in \mathcal{P}_N$ is* localy stable *for a_i's point of view (denoted $\Pi \in LX_i$ if Π satisfies the LX_i.*

Finding a stable outcome consists in finding a partition that is localy stable for each agent. Such a solution is said *consensually stable*.

Definition 4.5 (Consensual Stability) *Let a game be a MHG. A partition $\Pi \in \mathcal{P}_N$ is* consensually stable *(denoted $\Pi \in CoS$) if Π satisfies the local solution concept of all agent:*

$$CoS = \bigcap_{a_i \in N} LX_i$$

Example 4.1 *Let $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ be a hedonic game with $N = \{a_1, a_2, a_3\}$ and the following preferences:*

$$\succeq_1 = \{a_1, a_2\} \succ \{a_1\} \succ \{a_1, a_2, a_3\} \succ \{a_1, a_3\}$$

$$\succeq_2 = \{a_2, a_3\} \succ \{a_1, a_2, a_3\} \succ \{a_1, a_2\} \succ \{a_2\}$$

$$\succeq_3 = \{a_2, a_3\} \succ \{a_1, a_2, a_3\} \succ \{a_3\} \succ \{a_1, a_3\}$$

The possible partitions are:

$$\Pi_1 = \{\{a_1, a_2, a_3\}\} \qquad \Pi_2 = \{\{a_1\}, \{a_2, a_3\}\}$$
$$\Pi_3 = \{\{a_1, a_3\}, \{a_2\}\} \qquad \Pi_4 = \{\{a_1, a_2\}, \{a_3\}\}$$
$$\Pi_5 = \{\{a_1\}, \{a_2\}, \{a_3\}\}$$

Table 4.3 shows which local solution concepts are satisfied by the partitions from a_1's point-of-view. Let us remark that Π_1 and Π_3 satisfy no local solution concepts. Thus, whatever is the local solution concept of a_1, both partitions are not consensually stable. Let us consider $MHG = \langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$ with the same agents and preference profiles as HG, and $LSC = \{LNS_1, LIR_2, LPO_3\}$. In this game, we have $CoS = \{\Pi_2, \Pi_4\}$.

Table 4.3: Satisfaction of a_1's local solution concepts

$\mathbf{LX_1}$	Π_1	Π_2	Π_3	Π_4	Π_5
LIR_1		✓		✓	✓
LNS_1		✓		✓	
LIS_1		✓		✓	
$LICS_1$		✓		✓	
$LNCS_1$		✓		✓	
LCS_1		✓		✓	
LO_1				✓	
LPO_1		✓		✓	

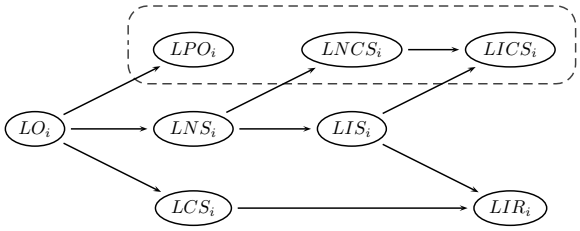

Figure 4.1: Inclusion relationships of local solution concepts

4.1.2 Properties of MHG

We show in this section that local solution concepts have the same properties than their global canonical equivalents. Firstly, global solution concepts present inclusion properties [Bogomolnaia and Jackson, 2002]. For instance, $NS \subseteq IS \subseteq ICS$. Obviously, due to their definitions, local solution concepts present the same inclusion relationships. We summarize them in Figure 4.1. For instance, the local Nash-stability does not take the other agents's preferences into account and, thus, is necessarily included in LIS_i, $LCNS_i$, $LICS_i$ and LIR_i. The dashed hyperedge in Figure 4.1 highlights the *irrational solution concepts* which are solution concepts that do not ensure that the agents are never in a coalition less preferred than their singleton coalition.

Let us remark that, for all agent a_i, any partition $\Pi \in \mathcal{P}_N$ including C_i^* (the preferred coalition for a_i) is necessarily in LO_i. Thus, due to the inclusion of local solution concepts, we have trivialy: *all local solution concept are non-empty*. Let us considers now cases where all agents express the same local solution concept LX.

Proposition 4.1 *Let X be a global solution concept. Let $MHG = \langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$ be a game. If $\forall a_i \in N$, $LX_i \in LSC$ is the local solution concept related to X, we have $CoS = X$.*

We give below the proof for Pareto-optimality only. Proofs for other solution concepts are based on a similar reasoning.

Proof 4.1 *Let MHG be a game such that all agents consider the local Pareto-optimality. CoS is the set of partitions which are locally Pareto-optimal for all agents:*

$$CoS = \bigcap_{a_i \in N} LPO_i$$

Firstly, let us show that $\Pi \in CoS \implies \Pi \in PO$. Let us fix a partition $\Pi \in CoS$ and let us assume that $\Pi \notin PO$. By definition, $\Pi \notin PO$ if $\exists \Pi_2$ such that $\forall a_i \in N, C_i(\Pi_2) \succeq_i C_i(\Pi)$ and $\exists a_j \in N : C_j(\Pi_2) \succ_j C_j(\Pi)$. However, it means that $\Pi \notin LPO_j$ and thus $\Pi \notin CoS$, which contradicts the assumption. Hence, $\Pi \in CoS$ implies $\Pi \in PO$.

Let us show now that $\Pi \in PO \implies \Pi \in CoS$. Let us fix a partition $\Pi \in PO$ and let us assume that $\Pi \notin CoS$. By definition, $\Pi \notin CoS$ if $\exists a_i \in N$ such that $\Pi \notin LPO_i$. By definition of LPO_i, we have then $\exists \Pi_2 : C_i(\Pi_2) \succ_i C_i(\Pi)$ and $\forall a_j \in N \setminus \{a_i\}, C_j(\Pi_2) \succeq_j C_j(\Pi)$. However, it contradicts the assumption $\Pi \in PO$. Thus, $\Pi \in PO$ implies $\Pi \in CoS$.

Consequently, we have $\Pi \in PO \iff \Pi \in CoS$.

Proposition 4.1 highlights that MHG subsumes canonical hedonic games. Moreover, as many global solution concepts can be empty (e.g. Nash-stability), Proposition 4.1 implies that CoS can also be empty. More interestingly, some consensually stable partitions are not characterized by any canonical solution concept.

Proposition 4.2 *Let a game be a MHG.*

$$\Pi \in CoS \not\implies \exists X \in [IR, NS, IS, ISC, NSC, CS, PO, O] \text{ s.t. } \Pi \in X$$

Proof 4.2 (Proof) *Proof is given by an example of MHG where it exists a consensual stable partition which satisfies no global solution concept. Let MHG be the game $\langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$ with $N = \{a_1, a_2, a_3\}$, $LSC = \{LIS_1, LICS_2, LIR_3\}$ and the following preference profiles:*

$$\succeq_1 = \{a_1, a_2, a_3\} \succ_1 \{a_1, a_2\} \succ_1 \{a_1\} \succ_1 \{a_1, a_3\}$$

$$\succeq_2 = \{a_2, a_3\} \succ_2 \{a_2\} \succ_2 \{a_1, a_2, a_3\} \succ_2 \{a_1, a_2\}$$

$$\succeq_3 = \{a_1, a_2, a_3\} \succ_3 \{a_1, a_3\} \succ_3 \{a_2, a_3\} \succ_3 \{a_3\}$$

This game has three consensual stable partitions :

$$\Pi_1 = \{\{a_1, a_2, a_3\}\}, \Pi_2 = \{\{a_1\}, \{a_2, a_3\}\}, \text{ and } \Pi_3 = \{\{a_1, a_2\}, \{a_3\}\}.$$

Let us consider partition Π_3*. This partition satisfies no canonical solution concept considered in Table 4.1. Indeed,* $\Pi_3 \notin IR$ *as* $\{a_2\} \succ_2 \{a_1, a_2\}$*,* $\Pi_3 \notin PO$ *as* $\{a1, a2, a3\}$ *is strictly preferred by all agents. However, while the grand coalition is preferred by all agents (with respect to their preference profiles), agent* a_3 *considers local individual rationality. Thus, as* $C_3(\Pi_3) = \{a_3\}$*,* a_3 *does not deviate. Agents* a_1 *and* a_2 *could have formed the grand coalition if they had considered collective deviations. However,* LIS *and* $LICS$ *are based only on personnal deviations. As all agents wait the other ones to deviate, partition* $\Pi_3 \in CoS$*.*

4.1.3 Complexity of MHG

Complexity of hedonic games has been widely studied in the literature [Ballester, 2004, Aziz et al., 2013a, Peters and Elkind, 2015]. In most cases, finding a stable partition for a given solution concept is known to be a NP-complete problem [Peters and Elkind, 2015]. Subclasses of hedonics games which consider particular assumptions on the preference profile (e.g. hedonic coalition nets [Elkind and Wooldridge, 2009], fractional hedonic games [Brandl et al., 2015], additively separable hedonic games [Aziz et al., 2013a]) can belong to other complexity classes. For instance, core-membership for hedonic coalition nets is coNP-complete [Elkind and Wooldridge, 2009]. In the case of MHGs, as the agents can consider irrationnal local solution concepts, we cannot make assumptions on the representation of the preference relationships.

Let us consider the following decision problems.

LX_i-Existence: given a game MHG and an agent a_i, is there a LX_i-stable partition Π of N?

LX_i-Membership: given a game MHG, an agent a_i and a partition Π, is Π a LX_i-stable partition of N?

CoS-Existence: given a game MHG, is there a Consensually Stable partition Π of N?

Table 4.4: Complexity classes of the decision problems

$\mathbf{LX_i}$	$\mathbf{LX_i}$-Existence	$\mathbf{LX_i}$-Membership
LIR_i	$\top$	P
LNS_i	$\top$	P
LIS_i	$\top$	P
$LCNS_i$	$\top$	P
$LICS_i$	$\top$	P
LCS_i	$\top$	P
LPO_i	$\top$	coNP-complete (Prop. 4.3)
LO_i	$\top$	P

CoS-Existence	**CoS**-Membership
Σ_2^P (Prop. 4.4)	coNP-complete (Corollary 4.1)

CoS-Membership: given a game MHG and a partition Π, is Π a Consensually Stable partition of N?

Before giving proofs, we summarize the complexity results in Table 4.4. Here $\top$ denotes that the $\mathbf{LX_i}$-Existence is trivial as it always exists a solution.

Let us consider firstly the cases of local solution concepts LX_i. As shown previously, due to the inclusion relationships between local solution concepts (see Figure 4.1), it always exists a locally stable partition. Consequently, the LX_i-Existence problem is trivially in $\mathcal{O}(1)$. For most local solution concepts (i.e. LIR_i, LNS_i, LIS_i, $LICS_i$, $LNCS_i$, LCS_i, LO_i), the LX_i-Membership problem is trivially decidable in polynomial time [Elkind and Wooldridge, 2009]. Only the local Pareto-optimality excepts this rule: the LPO_i-Membership decision problem is *coNP-complete*.

Proposition 4.3 *Let $MHG = \langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$. Let $a_i \in N$ be an agent such that $LPO_i \in LSC$. Let $\Pi \in \mathcal{P}_N$ be a partition. The LPO_i-*Membership *decision problem for Π is coNP-complete.*

As given in [Aziz et al., 2013b], we prove that the problem is coNP-Hard with a reduction from Exact Cover by 3-Sets (X3C) decision problem. We recall that X3C is defined as follows: given a set X (with $|X| = 3q$) and a collection $\mathbb{C}$ of 3-element subsets of X, is there a subset $C' \subseteq \mathbb{C}$ such that C' is a partition of U?

Proof 4.3 *Firstly, LPO_i-*Membership *is cleary in coNP. Indeed, a partition Π_2 which locally Pareto-dominates a partition Π_1 is a polynomial-time certificate that Π_1 is not locally Pareto-optimal.*

*Secondly, let us prove that LPO_i-*Membership *is coNP-Hard. An instance $(X, \mathbb{C})$ of X3C can be reduced to the LPO_i-*Membership *problem. Let us consider a $MHG = \langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$ such that $N = X$ and $\forall a_i \in N$, we have $c_1 \succ_i \ldots \succ_i c_k \succ_i N \succ_i \{a_i\}$, where $c_1 \succ_i \ldots \succ_i c_k$ is a linear ordering on $\{c \in \mathbb{C} : a_i \in c\}$ (all other coalitions with a_i are assumed less preferred than a_i's singleton coalition). Let us consider the partition $\Pi = \{N\}$ and an agent $a_i \in N$. Π is locally Pareto-dominated from a_i's point-of-view if, and only if, there is a partition $C' \subseteq \mathbb{C}$ $(C' \neq N)$ of X.*

($\Rightarrow$) Let us assume that there is a subset $C' \subseteq \mathbb{C}$ such that C' is a partition of X and $C' \neq N$. By definition of the preference profiles, we have $\forall a_j \in N$, there a unique $c_j \in C'$ which verifies $c_j \succ_j N$. Thus, by definition, Π is locally Pareto-dominated by C' and then is not locally Pareto-optimal for the agent a_i.

($\Leftarrow$) Let us assume now that there is no subset $C' \subseteq \mathbb{C}$ such that C' is a partition of X and $C' \neq N$. By definition of the preference profiles, it exists at least one agent $a_j \in N$ such that $\forall c \subseteq N, a_j \in c$, we have $\{N\} \succ_j c$. Thus, Π is locally Pareto-optimal for a_i.

Let us considers now the case of the consensual stability.

Corollary 4.1 *The CoS-*Membership *decision problem is coNP-complete.*

By definition, a partition Π is consensually stable if, $\forall a_i \in N, \Pi \in LX_i$. As LX_i-Membership is in P for all local solution concepts except for the local Pareto-optimality, which is coNP-complete, then in the worst case, the CoS-Membership decision problem is cleary coNP-complete.

Proposition 4.4 *The CoS-*Existence *decision problem is Σ_2^P.*

For ease of understanding, let us recall some notions about the polynomial hierarchy. [Stockmeyer, 1976] defined the polynomial hierarchy by the set $\{\Sigma_k^P, \Pi_k^P, \Delta_k^P : k \geq 1\}$ with $\Sigma_0^P = \Pi_0^P = \Delta_0^P = P$, forall all $k \geq 0$, $\Sigma_{k+1}^P = \text{NP}(\Sigma_k^P)$, $\Pi_{k+1}^P = \text{coNP}(\Sigma_k^P)$ and $\Delta_{k+1}^P = P(\Sigma_k^P)$. Remark that $\text{NP} = \Sigma_1^P$ and that $\text{coNP} = \Pi_1^P$. Σ_k^P and, Π_k^P can also be defined as sets of decision problems solvable in polynomial time on an alternating Turing machine with k alternations of existential and universal quantifiers. A Σ_k^P decision problem can be rewritten by a formula

$\exists X_1 \forall X_2 \exists X_3 \dots Q X_k, f(X_1, \dots, X_k)$, where f is a propositional logic formula and Q is either the existential quantifier (if k is even), either the universal quantifier (if k is odd). In the same way, a Π_k^P decision problem can be defined by a formula $\forall X_1 \exists X_2 \forall X_3 \dots Q X_k f(X_1, \dots, X_k)$, where here Q is the universal quantifier if k is even and existential quantifier if k is odd. The following proof is based on this definition.

Proof 4.4 *Let $MHG = \langle N, (\succeq_i)_{a_i \in N}, LSC \rangle$. In order to prove CoS-*Existence *being in Σ_2^P, we show that the decision problem can be written as a formula $\exists x : \forall y, f(x, y)$ where $f(x, y)$ can be checked in polynomial a time.*

*Firstly, let us remark that the CoS-*Existence *decision problem means checking that $\exists \Pi \in \mathcal{P}_N : \Pi \in CoS$. By definition of the consensual stability, this problem is the equivalent of finding a partition Π which satisfies the LX_i-*Membership *for all $LX_i \in LSC$. Thus,*

$$\exists \Pi \in \mathcal{P}_N : [\forall LX_i \in LSC, \Pi \in LX_i] \tag{4.1}$$

Let us consider two subsets of LSC: $LSC^A = \{LX_i \in LSC : LX \neq LPO_i\}$ and $LSC^B = \{LX_i \in LSC : LX = LPO_i\}$. LSC^B is the set of all locally Pareto-optimal solution concepts in LSC. Formula (4.1) *can be rewritten in:*

$$\begin{aligned} \exists \Pi \in \mathcal{P}_N : \ [&\forall LX_i \in LSC^A, \Pi \in LX_i \\ &\wedge \ \forall LPO_i \in LSC^B, \Pi \in LPO_i] \end{aligned} \tag{4.2}$$

The first part of the inner condition of Formula (4.2) *can be checked in polynomial time as LX_i-*Membership *is in P for $LX_i \neq LPO_i$. Checking the second part of the inner condition of Formula* (4.2) *is coNP-complete as LX_i-*Membership *is also coNP-complete for $LX_i = LPO_i$. However, let us consider the negation of this LPO_i-*Membership *problem: checking that $\forall \Pi_2 \in \mathcal{P}_N, \neg(\Pi_2 \succ_i^P \Pi)$ where $\Pi_2 \succ_i^P \Pi$ means that Π_2 locally Pareto-dominates partition Π. Thus, Formula* (4.2) *is equivalent to:*

$$\begin{aligned} \exists \Pi \in \mathcal{P}_N : (\ \forall \Pi_2 \in \mathcal{P}_N, \ [&\forall LX_i \in LSC^A, \Pi \in LX_i \\ &\wedge \ \forall LPO_i \in LSC^B, \neg(\Pi_2 \succ_i^P \Pi)] \ ) \end{aligned} \tag{4.3}$$

Now, all inner condition of Formula (4.3) *can be checked in polynomial time. Consequently, CoS-*Existence *is in Σ_2^P.*

Obviously, depending on the local solution concepts the agents consider, the decision problem can be easier. For instance, if no agent considers the local Pareto-optimality then CoS-Existence is "simply" NP-complete. If all

Table 4.5: Mean number of stable partitions with respect to $|N|$

	Rational solution concepts						Irrational solution concepts			
$\|N\|$	O	NS	CS	IS	IR	**CoS**	CNS	PO	CIS	$\mathcal{B}_n$
3	0.084	0.343	1.024	1.09	1.968	**1.199**	1.629	2.949	3.003	5
4	0.007	0.219	1.12	1.399	3.269	**1.547**	3.444	6.869	7.591	15
5	0	0.158	1.177	1.892	6.44	**2.193**	8.5	18.35	22.49	52
6	0	0.095	1.241	2.836	13.745	**3.14**	24.355	54.126	74.765	203
7	0	0.054	1.3	4.86	31.882	**6.053**	77.5475	171.896	275.073	877

agents consider the local core stability then CoS-Existence is equivalent to the CS-Existence, a well known NP-complete decision problem [Ballester, 2004, Elkind and Wooldridge, 2009]. Interestingly, if all agents only consider the local Pareto-optimality then the decision problem is trivial (i.e. there is always a global Pareto-optimal solution).

4.1.4 Empirical analysis of MHG

In this section, we consider a macroscopical empirical study of MHGs. For a number of agents varying from 3 to 7, we generate $1,000$ random MHGs where preference profiles and local solution concepts are drawn uniformly at random[1]. Table 4.5 shows (in a quasi-ascending order) the average number of consensually stable partitions ($|CoS|$) and the number of those partitions that satisfy a canonical solution concept. Column $\mathcal{B}_n$ gives for information the Bell number which is the number of possibles partitions with n agents. For instance, row $|N| = 5$ is read as follows: none of the $1,000$ has an optimal solution. In the Nash-Stability column, 0.219 means that, in less of the 4/5th of the games with 5 agents, there is no Nash-stable partition. Let us remark that we give here the average results. Thus, some games have several Nash-stable partitions but, majoritarily, there is no Nash-stable partition. Column CoS highlights that with 5 agents, over the 52 partitions, a few more than 2 satisfy the Consensual Stability.

Here, the main point is that the number of consensually stable partitions in a MHG is quasi-systematically greater than the number of the rational

[1]We are conscious of the limits of this study due to algorithmic complexity. Given a strict order on preferences profiles, n agents and m possible local solution concepts, there is $(m \times 2^{n-1}!)^n$ different MHGs. For instance, if $n = 3$ and $m = 8$, there is $7,077,888$ different games.

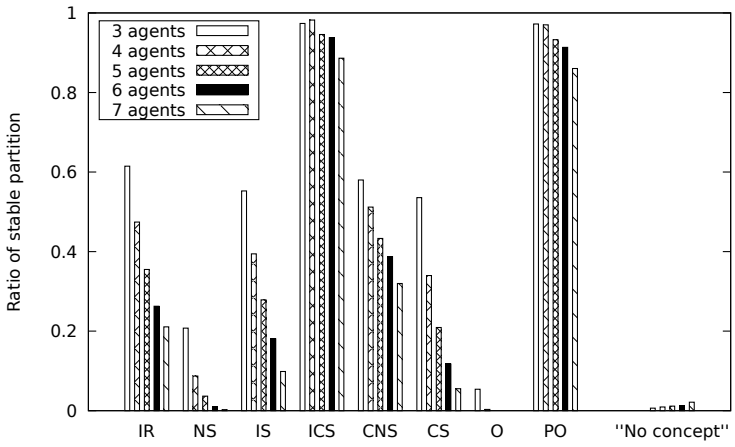

Figure 4.2: CoS partitions in canonical solution concepts

solution concepts (NS, IS, CS and O), and lower than the number of irrational solution concepts (CIS, CNS and PO). Only individual rationality (IR) excepts this rule. Thus, our model expresses a tradeoff between rational and irrational solution concepts: only agents that accept an irrational local solution concept can deviate towards a coalition which is less preferred than their singleton coalition. Figure 4.2 gives the ratio of consensually stable partitions which also satisfy a canonical solution concept. Let us remark that a large part of the CoS partitions are also in PO or ICS. For instance, with 7 agents, 86% of consensual stable partitions are Pareto-optimal. Thus, our model is mostly a restriction of those two solution concepts, while still allowing consensually stable partitions which satisfy *no* canonical solution concepts (see Proposition. 4.2 and column *No concept* in Figure 4.2).

Figure 4.3 gives the proportion of MHGs with at least one consensual stable partition, with respect to $|N|$. With 3 agents, for 80% of the games, there is a consensual stable partition. This ratio strongly decreases when more agents are involved in the game. Thus, with 7 agents, only around 40% of random games have a consensual outcome. Games without CoS partition are those where agents either impose *strong restrictions* on accepted deviations, either have inconsistent preferences profiles. If these agents could consider other local solution concepts, a consensually stable partition could appear in such a game. To this end, we extend MHG with another preference profiles on the local solution concepts.

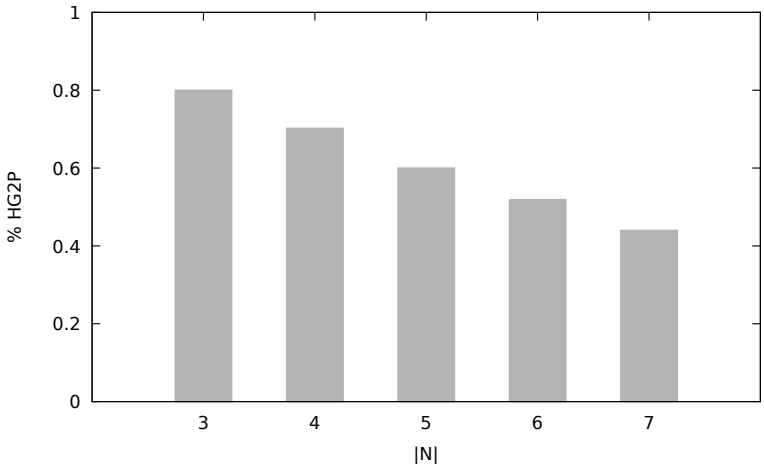

Figure 4.3: Ratio of MHG with at least one CoS partition

4.2 Extension to preferences on solution concepts

4.2.1 A second preference profile

If an agent has preferences on coalitions, then it could also have preferences on the solution concepts that it considers. Thus we propose here a second kind of hedonic games where agents express two preference profiles: the first one on the coalitions, and the second one on the local solution concepts.

Definition 4.6 (Hedonic games with double preference profiles) *A* hedonic game with double preference profiles *is a tuple* $HG2P = \langle N, (\succeq_i^C)_{a_i \in N}, (\succeq_i^{LSC_i})_{a_i \in N} \rangle$ *where:*

- $N = \{a_1, \ldots, a_n\}$ *is the set of agents,*
- $\succeq_i^C$ *is* a_i*'s preferences on coalitions, a complete and transitive preference relationship on the set* $\mathcal{N}_i = \{C \subseteq N : a_i \in C\}$*.*
- $\succeq_i^{LSC_i}$ *is* a_i*'s preferences on local solution concepts, a complete and transitive preference relationship on* LSC_i*, which is a non-empty set of local solution concepts for* a_i*.*

Example 4.2 *Let us consider the HG from Example 4.1 and let us transform this game in an HG2P by adding to HG the agents' preference profiles on local solution concepts given in Table 4.6. In this HG2P, all agents strictly prefer their local optimality (compared to all local solution concepts).*

Table 4.6: HG2P of Example 4.2

N	$\{a_1, a_2, a_3\}$
$\succeq_1^C$	$\{a_1, a_2\} \succ_1^C \{a_1\} \succ_1^C \{a_1, a_2, a_3\} \succ_1^C \{a_1, a_3\}$
$\succeq_2^C$	$\{a_2, a_3\} \succ_2^C \{a_1, a_2, a_3\} \succ_2^C \{a_1, a_2\} \succ_2^C \{a_2\}$
$\succeq_3^C$	$\{a_2, a_3\} \succ_3^C \{a_1, a_2, a_3\} \succ_3^C \{a_3\} \succ_3^C \{a_1, a_3\}$
$\succeq_1^{LSC_1}$	$LO_1 \succ_1^{LSC_1} LNS_1 \succ_1^{LSC_1} LCS_1 \succ_1^{LSC_1} LIS_1 \succ_1^{LSC_1} LNCS_1 \succ_1^{LSC_1} LIR_1 \succ_1^{LSC_1} LICS_1 \succ_1^{LSC_1} LPO_1$
$\succeq_2^{LSC_2}$	$LO_2 \succ_2^{LSC_2} LNS_2 \succ_2^{LSC_2} LIS_2 \succ_2^{LSC_2} LCS_2 \succ_2^{LSC_2} LIR_2 \succ_2^{LSC_2} LNCS_2 \succ_2^{LSC_2} LICS_2 \succ_2^{LSC_2} LPO_2$
$\succeq_3^{LSC_3}$	$LO_3 \succ_3^{LSC_3} LPO_3 \succ_3^{LSC_3} LNS_3 \succ_3^{LSC_3} LIS_3 \succ_3^{LSC_3} LNCS_3 \succ_3^{LSC_3} LICS_3 \succ_3^{LSC_3} LCS_3 \succ_3^{LSC_3} LIR_3$

However, agents a_1 and a_2 prefer local Nash-stability to local Pareto-optimality in opposite to a_3.

In the sequel, $LX_i \in LSC_i$ denote that the local solution concept LX_i is considered by the agent a_i in its preference profile $\succeq_i^{LSC_i}$. Let us remark that, while in Example 4.2 all agents consider all the local solution concepts given in Table 4.2, two agents can consider preferences on different sets of local solution concepts. For instance, an agent a_i can consider only local Pareto-optimality and local core stability whereas another agent a_j can have preferences on local Nash, local individual and local core stability. Trivially, a MHG is subclass of HG2Ps where each agent considers a singleton set of local solution concepts.

4.2.2 Stability and concessions

In order to find a solution which satisfies all preferences, agents have to find a concensus. To this end, we propose to evaluate the stability of a partition with respect to the number of *concessions* agents have to make on the rank of the solution concepts they satisfy. The *rank* of the solution concept LX_i on $\succeq_i^{LSC_i}$ (denoted $r_i(LX_i)$) is the rank of this solution concept in the preference profile. For a local solution concept $LX_i \notin LSC_i$, we consider $r_i(LX_i) = \infty$.

Definition 4.7 (Concession vector) *Let a game be a HG2P and $\Pi \in \mathcal{P}_N$ be a partition of N. The* concession vector *of Π $\vec{c}(\Pi)$ is:*

$$c_i(\Pi) = \begin{cases} r(LX_i^*) & \text{If } \exists\ LX_i \in LSC_i : \Pi \in LX_i \\ \infty & \text{otherwise} \end{cases}$$

where $LX_i^* = \underset{LX_i \in LSC_i : \Pi \in LX_i}{\operatorname{argmin}} r(LX_i)$.

Table 4.7: Concession vectors of the HG2P of Table 4.6

Π	$\vec{c}(\Pi)$
$\Pi_1 = \{\{a_1, a_2, a_3\}\}$	$[\infty, 2, 3]$
$\Pi_2 = \{\{a_1\}, \{a_2, a_3\}\}$	$[2, 1, 1]$
$\Pi_3 = \{\{a_1, a_3\}, \{a_2\}\}$	$[\infty, 5, \infty]$
$\Pi_4 = \{\{a_1, a_2\}, \{a_3\}\}$	$[1, 5, 2]$
$\Pi_5 = \{\{a_1\}, \{a_2\}, \{a_3\}\}$	$[6, 5, 8]$

Intuitively, the consession vector of a partition Π represents the number of concessions that each agent has to accept (on its preferences on solution concepts) in order to have Π being in CoS. As an example, Table 4.7 gives the concession vectors for the HG2P presented in Example 4.2. Let us remark that if $c_i(\Pi) = \infty$ (see for instance partition Π_1 in Table 4.2), then Π is never locally stable in a_i's point-of-view, and thus cannot be consensually stable.

In the sequel, we denote by *concession of* a_i the i^{th} components of the concession vector. Based on those concession vectors, we propose a new *global* solution concept: the *leximax stability*. This solution concept is based on the leximax behavioural rule as defined by [Delecroix et al., 2016] and the least-core stability concept in transferable utility games [Shapley and Shubik, 1966]. Concession vectors are sorted by decreasing order and compared by lexicographic order. A partition satisfies the leximax stability if its concession vector is not lexicographicaly dominated. Thus, leximax stable partitions minimize the number of concessions of the worst satisfied agent, then the second one and so on.

Definition 4.8 (Leximax preference relationship) *Let N be a set of agents, $\Pi, \Pi' \in \mathcal{P}_N$ be two partitions of N. Let $[x_1, \ldots, x_n]$ (resp. $[y_1, \ldots, y_n]$) be the decreasing ordered set of $\vec{c}(\Pi)$ components (resp. $\vec{c}(\Pi')$). The partition Π is* leximax-preferred *to the partition Π' (denoted $\Pi \succ^{lex} \Pi'$) if, and only if, $\exists k \in [1, n]$ such that, $\forall i \in [1, k[$, we have:*

$$x_i = y_i \text{ and } x_k < y_k$$

Definition 4.9 (Leximax stability) *Let a game be a HG2P and $\Pi \in \mathcal{P}_N$ be a partition of N. Π satisfies the* leximax stability *(denoted $\Pi \in LexS$) if, and only if:*

(1) $\nexists\, a_i \in N$ such that $c_i(\Pi) = \infty$,

(2) $\nexists\, \Pi' \in \mathcal{P}_N$ *such that* $\Pi' \succ^{lex} \Pi$.

Example 4.3 *Let us consider the HG2P from Example 4.2. Partitions* Π_1 *and* Π_4 *cannot be consensually stable as they do not satisfy any* a_1*'s local solution concepts. Among the* 3 *remaining partitions, we have the decreasing ordered set of* $\vec{c}(\Pi)$*:*

$$\Pi_2 : [2, 1, 1],\ \Pi_4 : [5, 2, 1] \text{ and } \Pi_5 : [8, 6, 5]$$

Thus, as $\Pi_2 \succ^{lex} \Pi_4 \succ^{lex} \Pi_5$*, partition* Π_2 *is leximax stable.*

As canonical hedonic games form a subclass of HG2P, the set of leximax stable solutions can be empty. However, as some canonical solution concepts ensure non-empty solutions for hedonic games, simple conditions on HG2Ps ensure the existence of at least one leximax stable partition.

Proposition 4.5 *Let* $HG2P = \langle N, (\succeq_i^C)_{a_i \in N}, (\succeq_i^{LSC_i})_{a_i \in N} \rangle$ *be a HG2P. If there is a global solution concept* X^* *such that* $\forall HG, X^* \neq \emptyset$ *and that* $\forall a_i \in N, LX_i^* \in LSC_i$*, then* $LexS \neq \emptyset$*.*

Let us recall that, among the canonical solution concepts, individual rationality, Pareto-optimality and individual contractual stability ensure non-emptyness for all game [Ballester, 2004, Bogomolnaia and Jackson, 2002, Sung and Dimitrov, 2007].

Proof 4.5 *Let us fix* X^* *a global solution concept such that* $\forall HG$*,* $X^* \neq \emptyset$*. Let* $HG2P_1 = \langle N, (\succeq_i^C)_{a_i \in N}, (\succeq_i^{LSC_i})_{a_i \in N} \rangle$ *be a HG2P such that* $\forall a_i \in N, LX_i^* \in LSC_i$*. Let* $HG_1 = \langle N, (\succeq_i^C)_{a_i \in N} \rangle$ *be a HG with the same agents and the same preference profiles on coalitions. By definition, we have* $X^* \neq \emptyset$ *in* HG_1*. By Proposition 4.1, we have:*

$$\bigcap_{a_i \in N} LX_i^* \neq \emptyset$$

Thus, $\forall a_i \in N, LX_i^* \neq \emptyset$*. Moreover, it exists a partition* $\Pi_1 \in \mathcal{P}_N$ *such that* $\forall a_i \in N, \Pi_1 \in LX_i^*$*. As by assumption,* $\forall a_i \in N, LX_i^* \in LSC_i$*, we have* $\forall a_i \in N, c_i(\Pi_1) \neq \infty$*. Thus,* Π_1 *satisfies the first condition for leximax stability. Obviously, if the second condition of leximax stability holds,* $\Pi_1 \in LexS$*.*

Let us assume that Π_1 *does not satisfy the leximax stability. By definition, it means that* $\exists \Pi_2 \in \mathcal{P}_N$ *such that* $\Pi_2 \succ^{lex} \Pi_1$*. As* $\infty \notin c_i(\Pi_1)$*,*

we have necessarily $\infty \notin c_i(\Pi_2)$*, and thus* Π_2 *also satisfies the first condition for leximax stability. Thus, either* $\Pi_2 \in LexS$*, or it exists a partition* $\Pi_3 \in \mathcal{P}_N, \Pi_3 \succ^{lex} \Pi_2$ *which is necessarily leximax stable.*

Consequently, if $\forall a_i \in N, LX_i^* \in LSC_i$ *then* $LexS \neq \emptyset$*.*

Obviously, the reciprocal statement is false. Even if the conditions characterized by Proposition 4.5 may be seen as restrictive, we can reasonably assume that all agent a_i consider in practice $LIR_i \in \succeq_i^{LCS_i}$ in last rank. Indeed, it represents the fact that if the agents do not find a stable solution, they will not cooperate and they will form their respective singleton coalition.

4.2.3 Complexity of HG2P

We present here the complexity of the decision problems linked to HG2P. More precisely, we consider the two following problems:

LexS-Existence: given a game $HG2P$, is there a $LexS$-stable partition Π of N?

LexS-Membership: given a game $HG2P$ and a partition Π, is Π a Leximax stable partition of N?

Two study this decision problem, we firstly consider two other subproblems:

Concession-Existence: given a game $HG2P$, an agent a_i and a partition Π, is the inequality $c_i(\Pi) \neq \infty$ holds?

Concession-Value: given a game $HG2P$, an agent a_i and an integer $k \in [1, |LSC_i|]$, is the equality $c_i(\Pi) \neq k$ holds?

Intuitively, the first one consists of checking that partition Π satisfies at least one local solution concept considered by agent a_i, and the second one of finding the exact number of concessions that agent a_i has to do such that Π is locally stable.

Lemma 4.1 *The* Concession-Existence *decision problem is coNP-complete.*

Intuitively, the proof is the following: showing that $c_i(\Pi) \neq \infty$ is equivalent to show that there is at least one local solution concept $LX_i \in LSC_i$

such that $\Pi \in LX_i$. Even if for most of local solution concepts, the LX_i-MEMBERSHIP problem is polynomial times, in the case of the local Pareto optimality the problem is coNP-complete (see proposition 4.3). Consequently, in the worst case, the CONCESSION-EXISTENCE existance problem is coNP-complete.

Lemma 4.2 *The* CONCESSION-VALUE *decision problem is* Σ_2^P.

Proof 4.6 *Let* $HG2P = \langle N, (\succeq_i^C)_{a_i \in N}, (\succeq_i^{LSC_i})_{a_i \in N} \rangle$, *be a game,* $a_i \in N$ *be an agent,* $\Pi \in \mathcal{P}_N$ *a partition and* $k \in [1, |LSC_i|]$. *Verifying the equality* $c_i(\Pi) = k$ *is equivalent to check the following formula:*

$$\begin{array}{c} \exists LX_i^\star \in LSC_i : r(LX_i^* = k, \Pi \in LX_i^* \\ \wedge \forall LX_i \in LSC_i : r(LX_i) < k, \Pi \notin LX_i \end{array} \tag{4.4}$$

The LPO_i-MEMBERSHIP *decision problem is coNP-complete (see proposition 4.3). Thus, we can rewrite the formula 4.4 with a quantified Boolean formula of the form* $\exists X_1 : \forall X_2 f(X_1, X_2) \wedge \forall X_3, \forall X_4 f(X_3, X_4)$, *(where* f *is tractable in polynomial time). If the the second part of this problem is coNP-complete, the first one is* Σ_2^P. *Thus, for an agent* $a_i \in N$ *and a partition* Π, *finding the exact value of* $c_i(\Pi)$ *is a* Σ_2^P *decision problem.*

Both sub-decisions problem are necessary to solve the LEXS-EXISTENCE and the LEXS-MEMBERSHIP decision problems.

Proposition 4.6 *The* LEXS-EXISTENCE *decision problem is* Σ_2^P.

Intuitively, checking the existence of a leximax stable partition is equivalent to show that there is at least one partition Π which verifies the CONCESSION-EXISTENCE decision problem for all agent.

Proof 4.7 *Let* $HG2P = \langle N, (\succeq_i^C)_{a_i \in N}, (\succeq_i^{LSC_i})_{a_i \in N} \rangle$, *be a game. By definition of the leximax-stability, there is a partition* $\Pi \in LexS$ *if* $\forall a_i \in N, c_i(\Pi) \neq \infty$. *Concequently, the* LEXS-EXISTENCE *decision problem in thus HG2P is equivalent to shows that there exists a partition* $\Pi \in \mathcal{P}_N$ *such that for all agent* $a_i \in N$, *the inequality* $c_i(\Pi) \neq \infty$ *holds :*

$$\exists \Pi \in \mathcal{P}_N : \forall a_i \in N, c_i(\Pi) \neq \infty \tag{4.5}$$

As shown by lemma 4.1, for a given agent a_i, *the*CONCESSION-EXISTENCE *decision problem is coNP-complete. Let us recall that* $coNP = \Pi_1^P$ *and that we can rewrite a such problem by a quantified Boolean formula* $\forall X_3, f(X_3)$,

where f is tractable in polynomial time. This formula 4.5 can be rewritten as follows:

$$\exists \Pi \in \mathcal{P}_N : \forall a_i \in N, \forall X_3 f(X_3) \tag{4.6}$$

Remark that formula 4.6 can itself be rewritten by a quantified Boolean formula $\exists X_1 : (\forall X_2, \forall X_3), f(X_1, X_2, X_3)$, which is the definition of a Σ_2^P decision problem. In consequence, LexS-Existence *decision problem is Σ_2^P.*

Denote that the complexity of this decision problem is due to coNP-completeness of LPO_i-Membership. However, for all agent $a_i \in N$ such that $LPO_i \notin LSC_i$, checking that $c_i(\Pi) \neq \infty$ is easy. Morever, even if $LPO_i \in LSC_i$, checking that $\Pi in LPO_i$ is necessary only if $\forall LX_i \in LSC_i, LX_i \neq LPO_i$: $\Pi \notin LX_i$. Furthermore, if conditions of propostion **??** hold, the game has a leximax-stable partition and the decision problem is them trival. Consequently, even if in the worst case, LexS-Existence is a Σ_2^P decision problem, it is easy tractable in most cases.

Let us consider now the LexS-Membership problem.

Proposition 4.7 *The* LexS-Membership *decision problem is Π_3^P.*

Intuitively, we can see that a partition Π_2 such that $\Pi_2 \succ^{Lex} \Pi$ is a Σ_2^P certificate that Π is not leximax-stable. Thus, the membership decision problem is coΣ_2^P also denoted Π_3^P.

Proof 4.8 *Let $HG2P = \langle N, (\succeq_i^C)_{a_i \in N}, (\succeq_i^{LSC_i})_{a_i \in N} \rangle$ be a game and $\Pi \in \mathcal{P}_N$ be a partition of N. Π satisfies leximax-stability if there is no partition $\Pi_2 \in \mathcal{P}_N$ leximax-preferred to Π. Thus, we have to prove that:*

$$\begin{aligned} &\forall a_i \in N, c_i(\Pi) \neq \infty \\ \wedge &\forall \Pi_2 \in \mathcal{P}_N, \Pi_2 \neq \Pi, \Pi \succ^{lex} \Pi_2 \end{aligned} \tag{4.7}$$

We already proved that the first condition is a coNP-complete satisfaction problem (see lemma 4.1). let us consider the second condition and the satisfaction of the leximax-dominance $\Pi \succ^{lex} \Pi_2$. Checking the leximax-dominance relation between Π and Π_2 requires to compute both concession vectors $c(\vec{\Pi})$ and $c(\vec{\Pi_2})$. Unfortunetly, computing the concession vector of a partition is equivalent to find a integer $k \in [1, |LX_i|]$ such that $c_i(\Pi) = k$ for all agent $a_i \in N$. As the Concession-Value *decision problem is Σ_2^P (see lemma 4.2), verifying the leximax-dominance is a Σ_2^P satisfaction problem.*

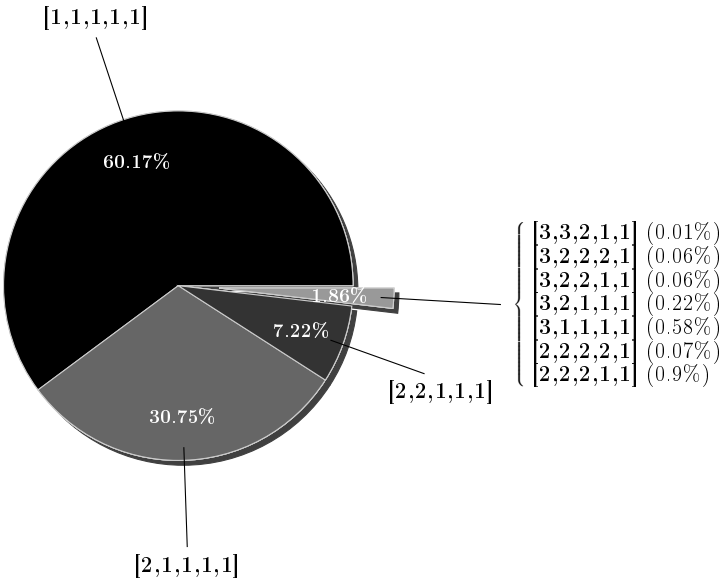

Figure 4.4: Proportion of concession vectors

Consequently, the second part of formula 4.7 is problem that can be formulated by a quantified Boolean formula $\forall X_1, \exists X_2 : \forall X_3, f(X_1, X_2, X_3)$*, which is the definition of a* Π_3^P *decision problem. Thus, we have proven that the* LexS-Membership *decision problem is* Π_3^P*.*

4.2.4 Empirical analysis of leximax stability

As in Section 4.1.4, we experiment on random games with 3 to 7 agents whose both preference profiles are generated uniformly at random. Figure 4.4 shows the proportion of leximax stable concession vectors for 5 agents on $10,000$ random HG2Ps. Here, the proportion of games without any consession[2] is equivalent to proportion of CoS partitions in MHGs (see Figure 4.3), namely 60%. Moreover, around 30% of the games only require a *single* agent to make a *single* concession (vector $[2, 1, 1, 1, 1]$). From a global perspective, around 98% of the games require at most 2 agents to make a single concession. Other concession vectors are anecdotic cases. For instances, only 1 out $10,000$ games requires 2 agents to make 2 concessions (vector $[3, 3, 2, 1, 1]$) in order to find a leximax stable partition.

Based on $1,000$ games, Figure 4.5 gives the mean (and standard devia-

[2]Games where for any leximax stable partition Π, $\vec{c}(\Pi) = [1, \ldots, 1]$.

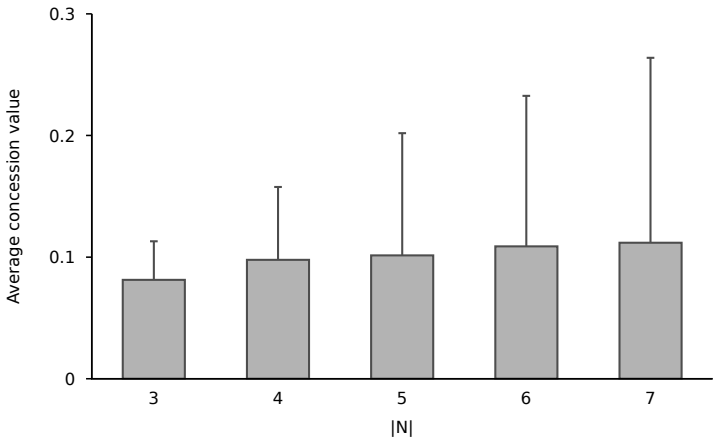

Figure 4.5: Mean (and standard deviation) of concessions

tion) of the number of concessions. While we observe a slight increase with the number of agents, the mean number of concessions ranges around 0.08 and 0.11 for $|N|$ from 3 to 7. From a global perspective, an agent has to make a concession only in 1 out of 10 games. The increase of the standard deviation highlights that the more agents in the game, the more they have to make concessions to find a leximax stable partition. While there may be leximax stable partitions where all agents have to make a huge number of concessions, it seems to be rare freak cases.[3] In their huge majority, HG2Ps have leximax stable partitions with a very small number of concessions.

[3]We have never observed such cases in our experimentations.

Chapter 5

Embedding a virtue ethics in hedonic games

The previous models allow to consider how to form coalition with agents who have different point-of-view on how coalitions must be built. Each local solution concept may be associated to a particular behavior. For instance, local Nash-stability represents a kind of individualism while local individual contractual stability represents a kind of politeness where agents deviate if and only if the others agree. We propose here a generalisation of those models, inspired from [Sung and Dimitrov, 2007] that we extend to the local case. In this new model – called *deviation games* – solution concepts are characterized by a composition of deviation conditions, and we will show it allows us to define new solution concepts which can easily represent human values. This approach is illustrated on three values: liberty, altruism and hedonism.

5.1 Deviation games

5.1.1 Formal model

In an hedonic game, a deviation is any change in the coalition composition due to a subset of agents.

Definition 5.1 (Deviation) *Let* $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *be a hedonic game and* $\Pi \in \mathcal{P}_N$ *be a partition. A* deviation *is a coalition* $D \subseteq N, D \notin \Pi, D \neq \emptyset$ *such that the agents in* D *leave their current coalitions in* Π *to form* D*.*

We denote by $[D \to \Pi]$ the application of deviation D on Π.

Definition 5.2 (Applying a deviation) *The partition* Π' *which results of* $[D \to \Pi]$ *is such that:*

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

Given a partition Π, nous denote by $AllD_i(\Pi) = \{D \subseteq N, D \notin \Pi : a_i \in D\}$ the set of deviations which implies the agent a_i.

Let us consider now the agent a_i's point-of-view, and let us consider a partition $\Pi \in \mathcal{P}_N$. We model the deviations the agent a_i wants to be applied with respect to its preferences (or any kind of other individual criteria) with *conditions* that must be satisfied.

Definition 5.3 (Deviation condition) *Let* $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *be a hedonic game,* $a_i \in N$ *be an agent,* $\Pi \in \mathcal{P}_N$ *be a partition and* $D \in AllD_i(\Pi)$ *be a deviation. A* deviation conditition Δ_X *represents a property that must be satisfied by* D *with respect to the agent* a_i *and the partition* Π *in order to be desirable for agent* a_i*.*

In the sequel, $\Delta_X(a_i, D, \Pi, HG)$ is a boolean function verifying if a deviation D satisfy the condition Δ_X from agent a_i's point-of-view, given Π and HG. In order to illustrate those concepts, we only consider the conditions given below. This choice is related to the classical solution concept given in the litterature (see Section 5.1.2 for more details).

Condition of Rationality: $\Delta_R := D \succ_i C_i(\Pi)$ – the deviation D is *rational* from the agent a_i's point-of-view if it (strictly) prefers the deviation to its current coalition.

Condition of Acceptance: $\Delta_A := \forall a_j \in D \setminus \{a_i\}, D \succ_j C_j(\Pi)$ – the deviation D is *acceptable* if all agents in D (strictly) prefer the deviation to their respective coalitions.

Condition of Defection: $\Delta_D := \forall a_k \in N \setminus D : \exists a_j \in D, C_k(\Pi) = C_j(\Pi), C_k(\Pi) \setminus D \succ_k C_k(\Pi)$ – the déviation D is a *defection* if the departure of all agents in D is prefered from the point-of-view of the other members of their respective coalitions.

Condition of Optimality: $\Delta_D := \nexists C \subseteq N : C \succ_i D$ – the deviation D is *optimal* from the agent a_i's point-of-view if it is its prefered coalition.

Condition of Pareto: $\Delta_{PO} := \exists \Pi' \in \mathcal{P}_N, D \in \Pi' : \forall a_j \in N, C_j(\Pi') \succ_j C_j(\Pi)$ – the deviation D is *Pareto-compatible* if it exists a partition Π' including D where all coalitions of Π' are strictly prefered to the one of Π by all agents.

Condition of Individuality: $\Delta_I := D \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$ – the deviation D is *individual* if the agent a_i is the single agent to deviate (which implies that the other agents of D already form a coalition).

Condition of Collectivity: $\Delta_C := D \setminus \{a_i\} \notin \Pi \cup \{\emptyset\}$ – the deviation D is *collective* if several agents (including agent a_i) did not belong to D before joining it.

Let us remark that we have two different families of conditions. On the one hand, conditions Δ_R, Δ_A, Δ_D, Δ_O and Δ_{PO} refers to satisfying the agents' preferences. On the other hand, conditions Δ_I et Δ_C refers to the identity of the deviating agents. Let us remark:

- $\Delta_I(a_i, D, \Pi, HG) \vee \Delta_C(a_i, D, \Pi, HG)$ is a tautology,
- $\Delta_I(a_i, D, \Pi, HG) \wedge \Delta_C(a_i, D, \Pi, HG) = \emptyset$,
- $\Delta_O(a_i, D, \Pi, HG) \implies \Delta_R(a_i, D, \Pi, HG)$,
- Δ_R and Δ_O only refers to the agent a_i,
- Δ_A only refers to the agents of $D \setminus \{a_i\}$,
- Δ_D only refers to the agents of $N \setminus D$,
- Δ_{PO} refers to all agents.

We only presented here the *strong* version of the conditions, as the underlying preferences are strict. Thus, we denote by Δ_X^- all *weak* equivalents with non-strict preferences. For instance, a deviation D which satisfies Δ_A^- means that agents of D that are not a_i can be indifferent to a_i's deviation:

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

The condition of Pareto differs from the other ones. Indeed, this condition does not only compare Π with the Π' which results of $[D \to \Pi]$: it compares all coalitions of Π with all coalitions in all partitions which include D. It allows to reason not only on a single deviation but a sequence of deviations.

Example 5.1 *Let us consider the partition* $\Pi = \{\{a_1, a_3\}, \{a_2, a_4\}\}$ *in the game:*

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

From a_1*'s point-of-view,* $\forall D \in AllD_1(\Pi)$*, it exists at least an agent* $a_j \in N$ *such that, for all* Π' *resulting from* $[D \rightarrow \Pi]$*,* $C_j(\Pi) \succ_j C_j(\Pi')$*. The same reasoning holds for the other agents. Thus, whatever the agent of the set of agents which deviate, this deviation negatively affects at least an agent. For instance, let us consider the deviation* $D = \{a_1, a_2\}$*. We have* $\Pi' = \{\{a_1, a_2\}, \{a_3\}, \{a_4\}\}$ *where* $C_3(\Pi) \succ_3 C_3(\Pi')$*. However, while this deviation is negative for* a_3 *and* a_4*, those agents can now apply the deviation* $D_2 = \{a_3, a_4\}$ *with* $\Pi'' = \{\{a_1, a_2\}, \{a_3\, a_4\}\}$ *which satisfies* $\forall a_i \in N, C_i(\Pi'') \succ_i C_i(\Pi)$*.*

Conditions of deviation allow an agent to define individual rules to characterize how it wants to deviate. However, an agent a_i may want to satisfy several conditions at the same time, or may want that at least one of them to be satisfied. For instance, an agent may express with $\Delta_R \wedge \Delta_A$ to deviate if and only if it is preferable for it and for all the other agents of D. Such a composition of deviation conditions is called the *deviation concept* of agent a_i.

Definition 5.4 (Deviation concept) *Let be* $a_i \in N$*. The* deviation concept $\mathbb{D}_i$ *of agent* a_i *is a propositional formula on a set* $\{\Delta_1, \ldots, \Delta_k\}$ *of deviation conditions. All deviation* $D \in AllD_i(\Pi)$ *which satisfies* $\mathbb{D}_i$ *(denoted* $D \vDash \mathbb{D}_i$*) is considered as desirable for agent* a_i*.*

Given an agent $a_i \in N$, a partition $\Pi \in \mathcal{P}_N$ and a game HG, nous denote by $\mathbb{D}_i(\Pi, HG)$ the set of all desirable deviations for agent a_i:

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

Example 5.2 *Let us consider an agent* a_1 *with the deviation concept* $\mathbb{D}_1 = \Delta_R \wedge \Delta_I$*, meaning* a_1 *looks for the individual deviation strictly prefered to*

its current coalition. Let us consider another agent a_2 which looks for all deviations strictly prefered by all agents involved in the deviation. This concept can be formalized as follows: $\mathbb{D}_2 = (\Delta_I \vee \Delta_C) \wedge \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$. It may be reduced to $\mathbb{D}_2$ à : $\Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$ as $(\Delta_I \vee \Delta_C)$ is a tautology.

As shown by Exemple 5.2, several agents can have different deviation concepts. Thus, based on the MHG described in the previous chapter, we can define a new model of hedonic games: the *deviation games*.

Definition 5.5 (Deviation games) *A* deviation game *is a triplet $HGD = \langle N, (\succeq_i)_{a_i \in N}, (\mathbb{D}_i)_{a_i \in N} \rangle$ where $N = \{a_1, \ldots, a_n\}$ is a set of agents, $\succeq_i$ the preferences of agent a_i over the coalitions and $\mathbb{D}_i$ the agent a_i's deviation concept.*

The underlying problem is a classical one: finding a partition $\Pi \in \mathcal{P}_N$ such that no agents want to deviate. However, contrary to classical hedonic games, stability holds when there is no desired deviation from all agents' point-of-view.

Definition 5.6 (Stability) *Let HGD be a deviation game and $\Pi \in \mathcal{P}_N$ a partition.* Π *is* locally stable *from agent a_i's point-of-view if $\mathbb{D}_i(\Pi, HGD) = \emptyset$.* Π *is* collectively stable *if $\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$.*

5.1.2 Links with canonical concepts

As said previously, deviation conditions are linked to canonical solution concepts. For instance, [Sung and Dimitrov, 2007] already proposed a notion of deviation to characterize solution concepts but they do not consider deviation conditions (their agents are homogeneous). Here, we establish the link between deviation conditions and solution concepts. We give the proof for the Nash stability. The proofs for the other concepts are similar. Finally, Table 5.1 summarizes all the links.

Proposition 5.1 *Let HGD be a deviation game and $\Pi \in \mathcal{P}_N$ be a partition. If $\forall a_i \in N, \mathbb{D}_i := \Delta_I \wedge \Delta_R$ then:*

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

Proof 5.1 *Let us fix a deviation game HGD and a partition $\Pi \in \mathcal{P}_N$. By definition, $\Pi \in NS$ si :*

$$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_i C_i(\Pi) \tag{5.1}$$

This characterization of Nash stability is equivalent to:

$$\forall a_i \in N, \nexists C \subseteq N, a_i \in C : C \setminus \{a_i\} \in \Pi \cup \{\emptyset\} \wedge C \succ_i C_i(\Pi) \tag{5.2}$$

Let us distinguish three parts in the formula:

1. $\nexists C \subseteq N, a_i \in C$ *corresponds to* $\nexists C \in AllD_i(\Pi)$ *as* $C \neq C_i(\Pi)$
2. $C \setminus \{a_i\} \in \Pi \cup \{\emptyset\}$ *corresponds to* $\Delta_I(a_i, C, \Pi, HGD)$
3. $C \succ_i C_i(\Pi)$ *corresponds to* $\Delta_R(a_i, C, \Pi, HGD)$

Thus, a partition is Nash stable if, for no agent, there is no rational individual deviation towards an already formed coalition in Π*. Thus, formula 5.2 can be rewritten in:*

$$\forall a_i \in N, \nexists D \in AllD_i(\Pi) : \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD) \tag{5.3}$$

However, by assumption, $\forall a_i, \mathbb{D}_i := \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD)$*. Thus, the formula 5.3 can rewritten in:*

$$\forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset \tag{5.4}$$

By definition, a partition Π *is Nash stable if the deviation concept* $\mathbb{D}_i := \Delta_I(a_i, D, \Pi, HGD) \wedge \Delta_R(a_i, D, \Pi, HGD)$ *is empty for all agents.*

Table 5.1 summarizes the deviation concepts $\mathbb{D}_i$ corresponding to the different canonical solution concepts, meaning that if all agents consider $\mathbb{D}_i$ then $\Pi \in SC \iff \forall a_i \in N, \mathbb{D}_i(\Pi, HGD) = \emptyset$.

Solution concept	**Deviation concept**
Nash Stability	$\Delta_I \wedge \Delta_R$
Individual Stability	$\Delta_I \wedge \Delta_R \wedge \Delta_A^-$
Contractual Nash Stability	$\Delta_I \wedge \Delta_R \wedge \Delta_D^-$
Individual Sontractual Stability	$\Delta_I \wedge \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$
Strong Core Stability	$\Delta_R \wedge \Delta_A$
Weak Core Stability	$\Delta_R \wedge \Delta_A^-$
Optimality	Δ_O
Pareto-Optimality	$\Delta_R \wedge \Delta_{PO}^-$

Table 5.1: Links between solution concepts and deviation concepts

Let us remark that those deviation concepts are the conjunction of two kind of clauses:

Clause on identity: The clause on identity defines if the desirable deviations must be individual (Δ_I) or may also be collective ($\Delta_I \vee \Delta_C$). The second case is implicit as it is a tautology. Neither it exists some deviations that can satisfy $\Delta_I \wedge \Delta_C$. Moreover, it is interesting to remark that *no canonical solution concepts consider collective deviation only.* For instance, such a deviation concept represents friendly agents that wants to deviate if and only if at least another agent deviates with it.

Clauses on preferences: The clauses on preferences take into account the other agents' preferences. It is important to notice that the condition of rationality is always considered in canonical solution concepts, as well as several weaks form of conditions. Finally, the clause on preferences is always a conjunction of conditions.

Representing deviation concepts with conjunctive normal form lead to another link betwwen canonical solution concepts and deviation concepts. Indeed, canonical solution concepts have inclusion relationships, and those relationships are the same between deviation concepts. For instance, obviously, given a partition Π, if exists a deviation $D \in All D_i(\Pi)$ such that $D \vDash \Delta_I \wedge \Delta_R$ then D also satisfies $\Delta_I \wedge \Delta_R \wedge \Delta_A$. Thus, a partition which does not satisfy individual stability does not also satisfy Nash stability. In general, inclusion of a deviation concept $\mathbb{D}_i^1$ in another concept $\mathbb{D}_i^2$ – denoted $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$ – means that all deviations in $\mathbb{D}_i^1$ are also in $\mathbb{D}_i^2$.

Definition 5.7 (Included deviation concept) *A deviation concept $\mathbb{D}_i^1$ is* included *in another concept $\mathbb{D}_i^2$ if, for all deviation games and all partitions $\Pi \in \mathcal{P}_N$, $D \in \mathbb{D}_i^1(\Pi, HGD) \implies D \in \mathbb{D}_i^2(\Pi, HGD)$.*

Based on those definition, we can easly deduce some inclusion relationships.

Proposition 5.2 *Let $\mathbb{D}_i^1$ and $\mathbb{D}_i^2$ be two deviation concepts. Let A (resp. B) be the set of deviation conditions which characterizes $\mathbb{D}_i^1$ (resp. $\mathbb{D}_i^2$). If $B \subseteq A$, the deviation concept $\mathbb{D}_i^1$ is included in $\mathbb{D}_i^2$.*

Proof 5.2 *Let us fix a deviation game HGD and a partition $\Pi \in \mathcal{P}_N$. Let $\mathbb{D}_i^1$ and $\mathbb{D}_i^2$ be two deviation concepts. Let A (resp. B) be the set of deviation conditions which characterizes $\mathbb{D}_i^1$ (resp. $\mathbb{D}_i^2$). Let us assume that $B \subseteq A$ and let us show that $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$ necessarily holds. To this end, deviation concepts $\mathbb{D}_i^1$ and $\mathbb{D}_i^2$ can be defined by:*

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

As $B \subseteq A$, we can rewrite $\mathbb{D}_i^1$ in:

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

Thus,

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

Consequently,

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

Thus, $\mathbb{D}_i^1 \subseteq \mathbb{D}_i^2$ necessarily holds.

To illustrate this property, let us consider the four following deviation concepts:

1. $\mathbb{D}_i^1 := \Delta_I \wedge \Delta_R$ (Nash Stability)
2. $\mathbb{D}_i^2 := \Delta_I \wedge \Delta_R \wedge \Delta_A$ (Individual Stability)
3. $\mathbb{D}_i^3 := \Delta_I \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$ (Individual Contractual Stability)
4. $\mathbb{D}_i^4 := \Delta_R \wedge \Delta_A$ (Core Stability)

Here, we obtain the following inclusions: $\mathbb{D}_i^3 \subseteq \mathbb{D}_i^2 \subseteq \mathbb{D}_i^1$ and $\mathbb{D}_i^3 \subseteq \mathbb{D}_i^4$ which are the same than the classical one ($NS \subseteq IS \subseteq ICS$ and $CS \subseteq IS \subseteq ICS$).

5.2 Modelling virtue-based solution concepts

Let us remark that, even in limiting ourself to 7 deviation conditions, many combinations do not correspond to canonical solution concepts. Table 5.2 highlights those holes in the litterature. Columns represents the clauses on identity and rows the clauses on preferences. Interrogation marks represent the solution concept that were not studied in the litterature (to the best of our knowledge).

	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	?	?
$\Delta_R \wedge \Delta_{PO^-}$	?	?	PO
Δ_O	?	?	O
Δ_A	?	?	?
Δ_D	?	?	?
$\Delta_A \wedge \Delta_D$	?	?	?

Table 5.2: Unstudied solution concepts

There are two reasons for such holes. Firstly, all canonical solution concepts consider the condition of rationality. However, it may make sense to consider agents that seeks to maximize a social welfare even if it goes against them, expressing a kind of *altruism*. Secondly, no canonical solution concepts only consider collective deviations. However, it may make sense to consider agents do not want to be the only *responsible* when making a partition unstable.

In fact, during the coalition formation process, choosing to stay in a coalition or to deviate may be lead by a virtue ethics, represented by a cardinal value (see our previous technical reports [Voyer, 2014, Boissier et al., 2015, Boissier et al., 2017]). **We propose to show how to model such cardinal value with a deviation concept, and how this models can fill the holes in the Table 5.2**. In a general way, we define for a value v et an agent a_i a deviation concept $\mathbb{D}_i^v$ such all deviations D which satisfy $\mathbb{D}_i^v$ are deviation that promote the value v. A stable partition Π represents a coalition structure such that no agent can deviate without betraying its values.

In order to illustrate our proposal, we model in this report three values based on their definition in the philosophy and sociology litterature: liberty, altruism and hedonism. We propose here *minimal deviation concepts* in the sense where any concepts that satisfy the same deviation conditions also promote the value. Thus, for a same value, heterogeneous agents may considered different deviation concepts.

5.2.1 Liberty

Liberty was greatly studies in philosophical and political litterature. Let us consider the four following definition (not exhaustive):

According to John Stuart Mill: [Mill, 1869] distinguishes *liberty of thought* and *liberty of act*. Liberty of thought means that all men have the right to form their own opinion and express them without any reserve. Satisfying this liberty is a moral imperative. Liberty of act means that " mens are free to act according to their opinions, namely free to apply them to their own lives without being restricted both physically and morally by their equals, since this freedom (s'exerce qu'à leurs seuls risques et périls). "

According to DRMC: " Liberty consists of doing anything which does not harm others: thus, the exercise of the natural rights of each man has only those borders which assure other members of the society the fruition of these same rights. These borders can be determined only by the law. " [DDHC, 1789] (Article IV).

According to Montesquieu: " We must have constantly present in our minds the difference between independence and liberty. Liberty is a right of doing whatever the laws permit, and if a citizen could do what they forbid he would no longer be possessed of liberty." [de Montesquieu, 1867] (livre XI, Chapitre III)

According to Durkheim: " True individual liberty does not consist in the suppression of all laws, but is the product of a given law as this egality cannot be found in the nature. " [Durkheim, 1893] (Chapitre II)

The same idea clearly appear in those definitions: freedom is restricted by the harm we can cause to the others, commonly translated in *freedom of ones stops where freedom of others begins*. In the context of hedonic games, as we consider agent that can express all their preferences, liberty of thought is satisfied. Thus, we need to characterize liberty of acts and an agent is free to deviate since:

1. it does not penalize the agents it joins,
2. it does not penalize the agents it leaves.

Both conditions are the *weak conditions of acceptation and defection* (Δ_A^- and Δ_D^-). Let us notice that, while Durkheim highlight individual liberty, liberty apply to all agents: there is no condition on identity. Moreover, while liberty cannot penalize the others, liberty may always penalize the deviating agents: there is no condition of rationality. Thus, **liberty is characterized by the following deviation concept:**

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

5.2.2 Altruism

Beyond debates on the existence or not of pure altruism as our actions seems always motivated by something [Batson, 2014], atruism has been often studied in *gift exchange games* or *dictator games* [Akerlof, 1984, Hoffman et al., 1996, Eckel and Grossman, 1996, Bardsley, 2008] as some altruist strategy may allow to reach better optimum than selfish strategies in some games [Nongaillard and Mathieu, 2011]. Beside the question of motivations, we consider the following definitions to characterize altruistic deviation concepts:

According to Rand: " 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). " [Rand, 1964]

" Altruism is the doctrine which demands that man lives for others and places others above self. " [Rand, 2005]

According to Comte: " Altruism is living for others. " [Comte, 1966]

It is important to notice that both authors define altruism as the opposit of selfishness. If selfishness is defined as only seeking to satisfy its own preferences, then Nash stability characterizes it. Seeking to satisfy first the preference of the others is, at the best of our knowledge, not characterized by any canonical solution concepts. To define such a new solution concept, we introduce a new deviation condition which aims at deviating towards a coalition prefered by at least another agent.

Definition 5.8 (Condition of altruism) *Let $\Pi \in \mathcal{P}_N$ be a partition and $D \in All D_i(\Pi)$ be a deviation. D satisfies the condition of altruism – denoted*

Δ_{alt}) – *if for* $\Pi' = [D \to \Pi]$,

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

The first part implies that the deviation must be profitable to at least one agent, and the second part implies that no agent must be penalized by the deviation. Moreover, altruism is a personal act (because it is not altruistic to demand the other to be altruist) which is represented by the condition of individuality Δ_I. Finally, altruism may be either profitable for the altruist, either be harmfull for him. This second case was called *altruistic suicide* by [Durkheim, 1897]: an agent commits an altruistic suicide when it deviates towards a less prefered coalition in order to reach a partition more prefered by another agent. Thus, **we characterize two different altruistic deviation concepts:**

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

Altruistic suicide: $\mathbb{D}_i := \Delta_I \wedge \Delta_{alt} \wedge \neg\Delta_R$

5.2.3 Hedonism

While common sense views hedonism as a moral doctrine based on personnal pleasure satisfaction, cyreneian and epicurian philosophies put the stress on the avoidance of pain. Indeed, Epicure said " excessive pleasure must be avoided if it leads to a future pain ". More recently, Mill wrote " 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." [Mill, 1889]. Thus, we ground our definition of hedonism on Nicolas de Chamfort's maxim " Enjoy and make people enjoy, without harming neither you, nor anyone, that is I think the whole morality. " [Chamfort and Maximes, 1857, Onfray, 2011]

On the one hand, an hedonist agent must satisfy its own preferences. On the other hand, the hedonist agent must satisfy the preference of the others. Both aspects can be characterized by condition of rationality (Δ_R), of acceptance (Δ_A) and defection (Δ_D). Thus, **hedonism is characterized by the following deviation concept:**

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

In terms of solution concept, this deviation concept is equivalent to a *contractual core stability* which is unstudied in the litterature. As the core stability, hedonism may be weaken by considering non-strict preferences. This *weak hedonism* (characterized by $\mathbb{D}_i := \Delta_R \wedge \Delta_A^- \wedge \Delta_D^-$) means that the agent a_i will seek for satisfying its preferences without penalizing the other agents.

5.3 Properties

Modelling the values at the previous section allows us to define new stable solutions for an hedonic game. Assuming the agents want to promote the same values, we define new solution concepts as follows:

Liberty Stability: $\Pi \in \mathcal{P}_N$ is *liberty-stable* (denoted $\Pi \in LS$) iff:

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

Altruistic Stability: $\Pi \in \mathcal{P}_N$ is *altruisticly stable* (denoted $\Pi \in AS$) iff:

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

Hedonic Stability: $\Pi \in \mathcal{P}_N$ is *hedonicly stable* (denoted $\Pi \in HS$) iff:

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

Table 5.3 highlight where those new concepts fill the holes of Table 5.2. Let us now study some properties of those new solution concepts, namely their non-emptyness and their inclusion relationships.

5.3.1 Non-emptyness

Liberty stability can be empty.

Proposition 5.3 *It exists hedonic games such that* $LS = \emptyset$.

Intuitively, liberty stability can be empty as the agents can always deviate since they do not penalize the others.

Proof 5.3 (By example) *Let HG be the following game:*

	Δ_I	$\Delta_I \vee \Delta_C$
Δ_R	Nash Stability	?
$\Delta_R \wedge \Delta_A$	Individual Stability	Core Stability
$\Delta_R \wedge \Delta_D$	Contractual Nash Stability	?
$\Delta_R \wedge \Delta_A \wedge \Delta_D$	Individual Contractual Stability	**Hedonic Stability**
$\Delta_R \wedge \Delta_{PO^-}$	?	Pareto Optimality
Δ_O	?	Optimality
Δ_{alt}	**Altruistic Stability**	?
$\neg\Delta_R \wedge \Delta_{alt}$	**Altruistic Suicide**	?
$\Delta_A^- \wedge \Delta_D^-$	?	**Liberty Stability**

Table 5.3: New solution concepts based on value-based deviation concepts

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

There is two possible partitions: $\Pi_1 = \{\{a_1\}, \{a_2\}\}$ *et* $\Pi_2 = \{\{a_1, a_2\}\}$. Π_1 *its not liberty stable as* a_2 *can deviate to* $D_1 = \{a_1, a_2\}$ *(it is a valid deviation even if it is irrational).* Π_2 *is not liberty stable as* a_1 *can deviate to* $D_2 = \{a_1\}$ *(same remark than previously). Thus, HG does not have any liberty stable partitions.*

Hedonic stability is always non empty.

Proposition 5.4 *Let* $HG = \langle N, (\succeq_i)_{a_i \in N} \rangle$ *be a hedonic game.* $HS \neq \emptyset$.

Proof 5.4 (By construction) *We exhibit an algorithm to construct a non-empty hedonicly stable partition. Let* $HGD = \langle N, (\succeq_i)_{a_i \in N}, (\mathbb{D}_i)_{a_i \in N} \rangle$ *be a deviation game. Let us fix a partition* $\Pi_1 = \{\{a_1\}, \ldots, \{a_n\}\}$.

Let us consider firstly the agent a_1 *and let us assume* $D \in \mathbb{D}_1(\Pi_1, HGD) \neq \emptyset$ *(if not we skip the agent* a_1 *to go to the agent* a_2*). Let* $D^* \in \mathbb{D}_1(\Pi_1, HGD)$ *be such that* $\forall D \in \mathbb{D}_1(\Pi_1, HGD), D^* \succ_1 D$. *Denote* $\Pi_2 = [D^* \to \Pi_1]$. *By definition of* D^**, we have* $\mathbb{D}_1(\Pi_2, HGD) = \emptyset$.

Let us consider now the agent a_2 *and the partition* Π_2. *By construction of* Π_2*, it exists* $D \in \mathbb{D}_2(\Pi_2, HGD)$. *We have then* $a_1 \notin D$. *Agent* a_2 *can consider deviation* $D^{*_2} \in \mathbb{D}_2(\Pi_2, HGD)$ *such that* $\forall D \in \mathbb{D}_2(\Pi_2, HGD), D^{*_2} \succ_2 D$. *Denote* $\Pi_3 = [D^{*_2} \to \Pi_2]$.

By repeating the previous step on all agents of $D^{*i} \in \mathbb{D}_i(\Pi_i, HGD)$, we obtain a partition Π_n such that $\forall a_i \in N, \mathbb{D}_i(\Pi_n, HGD) = \emptyset$, which is hedonicly stable.

Altruistic stability can be empty.

Proposition 5.5 *It exists hedonic games such that $AS = \emptyset$.*

Proof 5.5 (By example) *Let us consider the example of Proof 5.3. Π_1 is not altruisticly stable as to satisfy a_1, a_2 must consider an irrational deviation $D_1 = \{a_1, a_2\}$. $\Pi_2 = \{\{a_1, a_2\}\}$ is not altruisticly stable for the same reason. Thus, this game has not altruisticly stable partition.*

Interestingly, this example highlights situations where each agent gives the priority to the other, leading to a deadlock.

5.3.2 Inclusion relationships

Based on Property 5.2 we can deduce inclusion relationships between our new concepts, summarized in Figure 5.1. Let us denote by $\mathbb{D}_{SC}$ the deviation concept associated to the solution concept SC. For instance, $\mathbb{D}_{LS} := \Delta_A^- \wedge \Delta_D^-$.

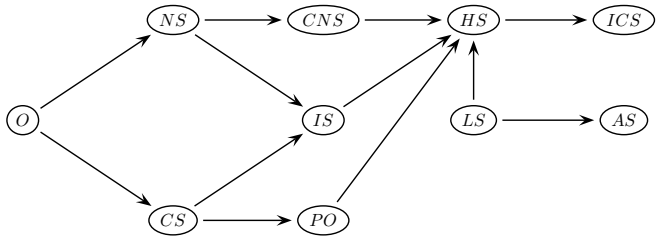

Figure 5.1: Relations d'inclusions entre les concepts de solution

Let us consider liberty stability and hedonic stability.

Proposition 5.6 *$LS \subseteq HS$.*

Proof 5.6 *Let us consider $\mathbb{D}_{LS} := \Delta_A^- \wedge \Delta_D^-$ and $\mathbb{D}_{HS} := \Delta_R \wedge \Delta_A \wedge \Delta_D$. By definition of weak deviation conditions, any deviation D which satisfies Δ_A (resp. Δ_D) also satisfies Δ_A^- (resp Δ_D^-). By Property 5.2, we have $\mathbb{D}_{HS} \subseteq \mathbb{D}_{LS}$. Thus, we have $LS \subseteq HS$.*

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

Proposition 5.7 *$HS \subseteq ICS$.*

Proof 5.7 *Let us consider $\mathbb{D}_{HS} := \Delta_R \wedge \Delta_A \wedge \Delta_D$ and $\mathbb{D}_{ICS} := \Delta_I \wedge \Delta_R \wedge \Delta_A \wedge \Delta_D$. By Property 5.2, we have $\mathbb{D}_{ICS} \subseteq \mathbb{D}_{HS}$. Thus, we have $HS \subseteq ICS$.*

As $LS \subseteq HS$, we also have $LS \subseteq ICS$. In the same way, hedonic stability is included in individual stability, Nash stability and core stability (the proof is very similar to the previous one). Lastly, a Pareto-optimal partition is always hedonicly stable.

Proposition 5.8 *Hedonic stability satisfies: $NS \subseteq IS \subseteq HS$, $CS \subseteq IS \subseteq HS$ and $PO \subseteq HS$.*

We give the proof for $PO \subseteq HS$.

Proof 5.8 *A hedonicly stable partition is not necessarily Pareto-optimal as Pareto-optimality considers a sequence of deviation. It is not the case for the hedonic stability. Now, let us show that $PO \subseteq HS$. Let $\Pi \in PO$ be a partition and let us assume that $\Pi \notin HS$. By definition of hedonic stability, it exists a deviation D such that, for all Π' the partition which results of $[D \to \Pi]$, we have $\forall a_i \in N, C_i(\Pi') \subset C_i(\Pi)$. It is in contradiction with the definition of Pareto-optimality. Thus, it contradicts our assumption $\Pi \in PO$.*

Liberty stability is not included in individual stability (and therefore is not included in core stability and Nash stability).

Proposition 5.9 *$IS \not\subseteq LS$.*

Proof 5.9 (By example) *Let HG_1 be a hedonic game such that:*

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

Let us consider the partition $\Pi = \{\{a_1, a_2\}, \{a_3\}\}$. Π *is not individually stable as* a_1 *can deviate to* $D = \{a_1, a_3\}$ *which satisfies* Δ_I, Δ_R *and* Δ_A. *However,* Π *is liberty stable as there is no deviation which satisfies* Δ_D. *Thus,* $\Pi \in IS$ *and* $\Pi \notin LS$.

Let HG_2 *be a hedonic game such that:*

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

Let us consider the partition $\Pi = \{\{a_1, a_2, a_3\}\}$. Π *is individually stable but is not liberty stable as* a_1 *can deviate to* $D = \{a_1\}$. *Thus,* $\Pi \notin IS$ *and* $\Pi \in LS$.

Finally, let us study the case of altruism.

Proposition 5.10 $LS \subseteq AS$.

Proof 5.10 *Let us recall that* Δ_{alt} *implies to satisfy* Δ_A^- *and* Δ_D^-. *Thus, we can write* $\mathbb{D}_{alt} := \Delta_I \wedge \Delta_{alt} \wedge \Delta_A^- \wedge \Delta_D^-$. *Consequently, by Propertyi 5.2, we have* $\mathbb{D}_{AS} \subseteq \mathbb{D}_{LS}$. *Thus, we have* $LS \subseteq AS$.

Let us remark that, as altruism allows (or oblige in the case of altruistic suicide) irrational deviation, there is no inclusion relationships between altruistic stability and all canonical solution concepts, as well between altruistic stability and hedonic stability.

Chapter 6

General conclusion

In this report, we study how to build ethical collective of agents, and how to build these collective in an ethical way.

1. We proposed an operational model of ethical judgment that produces beliefs on moral and ethical images of other agents. Then, we use those image beliefs to build trust beliefs that can be used to make cooperation based on moral or ethics. Firstly, ethical and moral trust can enrich the description of the moral rules or values. Secondly, moral rules or values can influence how trust is built, and finally jugement can allow to trust agents with a close ethics only, and thus to build ethical collectives.

2. We proposed several models of hedonic games, extending classical approaches to individual solution concepts. Those models ground a virtue ethics for cooperative games, allowing agents to express heterogeneous point-of-view on how coalitions must be formed. We extended this approach in a decomposition of atomic properties to define new solution concepts which can easily represent human values.

Obviously, we dealt with both problematics in an independant way. Beside the need to unify deviation concepts with HG2P in order to have a more general expression of virtue ethics, the main perspective to this work is to merge both approaches in order to built ethically ethical collectives.

Bibliography

[Abdul-Rahman and Hailes, 2000] Abdul-Rahman, A. and Hailes, S. (2000). Supporting trust in virtual communities. In *33th IEEE International Conference on Systems Sciences*, pages 1–9.

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

[Aldridge, 2009] Aldridge, I. (2009). *High-frequency trading: a practical guide to algorithmic strategies and trading systems*, volume 459. John Wiley and Sons.

[Anderson and Anderson., 2014] Anderson, M. and Anderson., S. (2014). Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. In *AAAI Fall Symposium Serie*.

[Arkin, 2009] Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall.

[Arkoudas et al., 2005] Arkoudas, K., Bringsjord, S., and Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, pages 17–23.

[Arrow, 1963] Arrow, K. (1963). *Social Choice and Individual Values*. Yale University Press.

[Aziz et al., 2013a] Aziz, H., Brandt, F., and Harrenstein, P. (2013a). Fractional hedonic games. In *12th International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

[Aziz et al., 2013b] Aziz, H., Brandt, F., and Harrenstein, P. (2013b). Pareto optimality in coalition formation. *Games and Economic Behavior*, 82:562–581.

[Aziz et al., 2011] Aziz, H., Brandt, F., and Seedig, H. G. (2011). Stable partitions in additively separable hedonic games. In *10th AAMAS*, pages 183–190.

[Bachrach and Rosenschein, 2008] Bachrach, Y. and Rosenschein, J. (2008). Coalitional skill games. In *7th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1023–1030.

[Ballester, 2004] Ballester, C. (2004). NP-completeness in hedonic games. *Games and Economic Behavior*, 49(1):1–30.

[Banzhaff III, 1964] Banzhaff III, J. (1964). Weighted voting doesn't work: A mathematical analysis. *Rutgers University Law Review*, 19.

[Bardsley, 2008] Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

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

[Baum and Petrie, 1966] Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563.

[Bernstein et al., 2000] Bernstein, D., Zilberstein, S., and Immerman, N. (2000). The complexity of decentralized control of markov decision processes. In *16th Conference in Uncertainty in Artificial Intelligence*, pages 32–37.

[Berreby et al., 2015] Berreby, F., Bourgne, G., and Ganascia, J.-G. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *20th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, pages 532–548.

[Berreby et al., 2017] Berreby, F., Bourgne, G., and 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*, pages 96–104.

[Bertsimas et al., 2011] Bertsimas, D., Farias, V., and Trichakis, N. (2011). The price of fairness. *Operations research*, 59(1):17–31.

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

[Boissier et al., 2017] Boissier, O., Bonnet, G., Cointe, N., Mermet, B., Simon, G., Tessier, C., and de Swarte, T. (2017). Models for ethical autonomous agents. Technical report, ETHICAA.

[Boissier et al., 2015] Boissier, O., Bonnet, G., Ganascia, J.-G., Tessier, C., de Swarte, T., and Voyer, R. (2015). A roadmap towards ethical autonomous agents. Technical report, ETHICAA.

[Boissier et al., 2013] Boissier, O., Bordini, R. H., Hübner, J. F., Ricci, A., and Santi, A. (2013). Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747–761.

[Bono et al., 2013] Bono, S., Bresin, G., Pezzolato, F., Ramelli, S., and Benseddik, F. (2013). Green, social and ethical funds in europe. Technical report, Vigeo.

[Boutilier, 1999] Boutilier, G. (1999). Sequential optimality adn coordination in multiagent systems. In *16th International Joint Conference on Artificial Intelligence*, pages 478–485.

[Brandl et al., 2015] Brandl, F., Brandt, F., and Strobel, M. (2015). Fractional hedonic games: Individual and group stability. In *14th AAMAS*, pages 1219–1227.

[Carbo et al., 2002] Carbo, J., Molina, J., and Davila, J. (2002). Comparing predictions of SPORAS vs. a fuzzy reputation agent system. In *3rd International Joint Conference on Fuzzy Sets and Fuzzy Systems*, pages 147–153.

[Carter et al., 2002] Carter, J., Bitting, E., and Ghorbani, A. (2002). Reputation formalization for an information-sharing multi-agent system. *Computational Intelligence*, 18(2):515–534.

[Castelfranchi and Falcone, 1998] Castelfranchi, C. and Falcone, R. (1998). Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In *3rd International Conference on Multi-Agent Systems*, pages 72–79.

[Castelfranchi and Falcone, 2010] Castelfranchi, C. and Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.

[Chalkiadakis et al., 2010] Chalkiadakis, G., Elkind, E., Markakis, E., Polukarov, M., and Jenning, N. (2010). Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 39:179–216.

[Chalkiadakis et al., 2007] Chalkiadakis, G., Markakis, E., and Boutilier, C. (2007). Coalition formation under uncertainty: Bargaining equilibria and the bayesian core stability concept. In *6th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 64–72.

[Chamfort and Maximes, 1857] Chamfort, N. and Maximes, P. (1857). Anecdotes, caractères et dialogues, a.

[Chevaleyre et al., 2006] Chevaleyre, Y., Dunne, P., Endriss, U., Lang, J., Lemaitre, M., Maudet, N., Padget, J., Phelps, S., Rodriguez-Aguilar, J., and Sousa, P. (2006). Issues in multiagent resource allocation. *Informatica*, 30(1):3–31.

[Chevaleyre et al., 2007] Chevaleyre, Y., Endriss, U., Lang, J., and Maudet, N. (2007). *A short introduction to computational social choice*. Springer.

[Coelho and da Rocha Costa, 2009] Coelho, H. and da Rocha Costa, A. (2009). On the intelligence of moral agency. In *Encontro Português de Inteligência Artificial*, pages 12–15.

[Coelho et al., 2010] Coelho, H., Trigo, P., and da Rocha Costa, A. (2010). On the operationality of moral-sense decision making. In *2nd Brazilian Workshop on Social Simulation*, pages 15–20.

[Cointe et al., 2016a] Cointe, N., Bonnet, G., and Boissier, O. (2016a). Ethical judgment of agents' behaviors in multi-agent systems. In *15th International Conference on Autonomous Agents & Multiagent Systems*, pages 1106–1114.

[Cointe et al., 2016b] Cointe, N., Bonnet, G., and Boissier, O. (2016b). Multi-agent based ethical asset management. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 52–57.

[Comte, 1966] Comte, A. (1966). Catéchisme positiviste (1852), éd. *P. Arnaud, Paris*, page 59.

[Conte and Paolucci, 2002] Conte, R. and Paolucci, M. (2002). *Reputation in artificial societies: Social beliefs for social order*, volume 6. Springer Science & Business Media.

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

[de Montesquieu, 1867] de Montesquieu, C.-L. d. S. (1867). *Esprit des lois*. Libr. de F. Didot Frères.

[Delecroix et al., 2016] Delecroix, F., Morge, M., Nachtergaelle, T., and Routier, J.-C. (2016). Multi-party negotiation with preferences rather than utilities. *International Journal of Cloud Computing*, 12(2):27.

[Dibangoye et al., 2014] Dibangoye, G., Amato, C., Buffet, O., and Charpillet, F. (2014). Exploiting separability in multiagent planning with continuous-state mdps. In *13th International Conference on Autonomous Agents and Multiagent System*, pages 1281–1288.

[Dreze and Greenberg, 1980] Dreze, J. and Greenberg, J. (1980). Hedonic coalitions: Optimality and stability. *Econometrica*, 48(4):987–1003.

[Driessen, 1991] Driessen, T. (1991). A survey of consistency properties in cooperative game theory. *SIAM Review*, 33(1):43–59.

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

[Durkheim, 1897] Durkheim, E. (1897). *Le suicide: étude de sociologie*. F. Alcan.

[Eckel and Grossman, 1996] Eckel, C. C. and Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and economic behavior*, 16(2):181–191.

[Elkind and Wooldridge, 2009] Elkind, E. and Wooldridge, M. (2009). Hedonic coalition nets. In *8th AAMAS*, pages 417–424.

[Esfandiari and Chandrasekharan, 2001] Esfandiari, B. and Chandrasekharan, S. (2001). On how agents make friends: Mechanisms for trust acquisition. In *4th Workshop on Deception, Fraud, and Trust in Agent Societies*, pages 27–34.

[for Economic and Affairs, 2009] for Economic, D.-G. and Affairs, F. (2009). Impact of the current economic and financial crisis on potential output. Technical Report 49, European Commission.

[Ganascia, 2007] Ganascia, J. (2007). Modeling ethical rules of lying with answer set programming. *Ethics and Information Technology*, 9:39–47.

[Guestrin et al., 2003] Guestrin, C., Koller, D., Parr, R., and Venkataraman, S. (2003). Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468.

[Hoffman et al., 1996] Hoffman, E., McCabe, K., and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *The American Economic Review*, 86(3):653–660.

[Horsburgh, 1960] Horsburgh, H. (1960). The ethics of trust. *The Philosophical Quartely*, 10(41):343–354.

[Ieong and Shoham, 2008] Ieong, S. and Shoham, Y. (2008). Bayesian coalitional games. In *23rd AAAI Conference on Artificial Intelligence*, pages 95–100.

[Josang and Ismail, 2002] Josang, A. and Ismail, R. (2002). The beta reputation system. In *15th Bled Conference on Electronic Commerce*, pages 41–55.

[Josang et al., 2007] Josang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service proposition. *Decision Support Systems*, 43(2):618–644.

[Kim and Lipson, 2009] Kim, K. and Lipson, H. (2009). Towards a 'theory of mind' in simulated robots. In *11th Annual Conference Companion on Genetic and Evolutionary Computation Conference*, pages 2071–2076.

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

[Marsh, 1994] Marsh, S. (1994). *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling.

[Michalak et al., 2009] Michalak, T., Rahwan, T., Sroka, J., Dowell, A., Wooldridge, M., McBurney, P., and Jennings, N. (2009). On representing coalitional games with externalities. In *10th ACM Conference on Electronic Commerce*, pages 11–20.

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

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

[Muller et al., 2003] Muller, G., Vercouter, L., and Boissier, O. (2003). Towards a general definition of trust and its application to openness in MAS. In *6th Workshop on Deception, Fraud and Trust in Agent Societies*, pages 49–56.

[Nash, 1951] Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, 54:286–295.

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

[Nowak and Radzik, 1994] Nowak, A. and Radzik, T. (1994). A solidarity value for n-person transferable utility games. *International Journal of Game Theory*, 23:43–48.

[Onfray, 2011] Onfray, M. (2011). *Manifeste hédoniste*.

[Peters and Elkind, 2015] Peters, D. and Elkind, E. (2015). Simple causes of complexity in hedonic games. In *24th IJCAI*, pages 617–623.

[Puterman, 1994] Puterman, M. (1994). *Markov Decision Processes. Discrete stochastic dynamic programming*. Wiley-Interscience.

[Puterman, 2014] Puterman, M. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Son.

[Rachelson, 2009] Rachelson, E. (2009). *Temporal Markov Decision Problems: Formalization and Resolution*. PhD thesis, University of Toulouse.

[Rahwan et al., 2015] Rahwan, T., Michalak, T., Wooldridge, M., and Jennings, N. (2015). Coalition structure generation: A survey. *Artificial Intelligence*, 229:139–174.

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

[Rand, 2005] Rand, A. (2005). *The fountainhead*. Penguin.

[Rao and Georgeff, 1995] Rao, A. and Georgeff, M. (1995). BDI agents from theory to practice. In *Technical note 56*. AAII.

[Rocha-Costa, 2016] Rocha-Costa, A. (2016). Moral systems of agent societies: Some elements for their analysis and design. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 32–37.

[Sabater and Sierra, 2005] Sabater, J. and Sierra, C. (2005). Review on computational trust and reputation models. *Artificial Intelligence*, 24(1):33–60.

[Sabater-Mir and Sierra, 2001] Sabater-Mir, J. and Sierra, C. (2001). REGRET: A reputation model for gregarious societies. In *4th Workshop on Deception, Fraud, and Trust in Agent Societies*, pages 61–69.

[Sabater-Mir and Vercouter, 2013] Sabater-Mir, J. and Vercouter, L. (2013). Trust and reputation in multiagent systems. *Multiagent Systems*, pages 381–419.

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

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

[Sen, 1986] Sen, A. (1986). Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

[Sen and Sajja, 2002] Sen, S. and Sajja, N. (2002). Robustness of reputation-based trust: Boolean case. In *1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 288–293.

[Shapley, 1953] Shapley, L. (1953). A value for n-person games. In Kuhn, H. and Tucker, A., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press.

[Shapley and Shubik, 1966] Shapley, L. S. and Shubik, M. (1966). Quasi-cores in a monetary economy with nonconvex preferences. *Econometrica*, pages 805–827.

[Shehory and Kraus, 1998] Shehory, O. and Kraus, S. (1998). Methods for task allocation via agent coalition formation. *Artificial Intelligence*, 101(1-2):165–200.

[Stockmeyer, 1976] Stockmeyer, L. J. (1976). The polynomial-time hierarchy. *Theoretical Computer Science*, 3(1):1–22.

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

[Vercouter and Muller, 2010] Vercouter, L. and Muller, G. (2010). L.I.A.R.: Achieving social control in open and decentralized multiagent systems. *Applied Artificial Intelligence*, 24(8):723–768.

[von Neumann and Morgenstern, 1944] von Neumann, J. and Morgenstern, O. (1944). Theory of games and economic behavior. *Nature*, 246:15–18.

[Voyer, 2014] Voyer, R. (2014). Une histoire de la philosophie morale. Technical report, ETHICAA.

[Yang, 1997] Yang, C. (1997). A family of values for n-person cooperative transferable utility games: An extension to the shapley value. Technical report, University of New-York Buffalo.

[Yang and Gao, 2014] Yang, X. and Gao, J. (2014). Uncertain core for coalitional game with uncertain payoffs. *Journal of Uncertain Systems*, 8:13–21.

[Young, 1995] Young, P. (1995). Optimal voting rules. *The Journal of Economic Perspectives*, 9(1):51–64.

[Yu and Singh, 2002] Yu, B. and Singh, M. (2002). Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–549.