GREThA
Groupe de Recherche en
Économie Théorique et Appliquée

Dynamique des préférences et valeurs morales : une contribution de la théorie des émotions à l'analyse économique

Emmanuel PETIT

GREThA UMR CNRS 5113

Cahiers du GREThA
n° 2008-11

GRETHA UMR CNRS 5113
Université Montesquieu Bordeaux IV
Avenue Léon Duguit - 33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75 - Fax : +33 (0)5.56.84.86.47 - www.gretha.fr

Dynamique des préférences et valeurs morales : une contribution de la théorie des émotions à l'analyse économique

Résumé

L'objectif de cet article est d'étudier comment la prise en compte des processus émotionnels permet de représenter la dynamique des préférences individuelles et collectives dans l'analyse économique. Nous illustrons l'apport de la théorie des émotions de Livet (2002) dans le cas du modèle théorique de l'obéissance à l'autorité d'Akerlof (1991) et dans celui du jeu expérimental du bien public. Nous montrons que les émotions nous sont utiles pour nous révéler consciemment ou non nos vraies préférences ou les valeurs morales que nous sommes prêts à défendre. Cela implique que les émotions soient intégrées dans l'analyse économique comme un vecteur d'information des changements de préférences et non pas seulement comme un simple argument de la fonction d'utilité des individus.

Mots-clés : Théorie des émotions, obéissance à l'autorité, coopération, jeu du bien public, valeurs morales.

Dynamic preferences, moral values and emotions in economical analysis

Abstract

Our analysis is about the role of emotions and moral values in economical analysis. We use the theory of emotions of Livet (2002) in order to understand the behaviour of individuals alternatively in the Akerlof (1991)'s model of undue obedience and in a public good experiment. We argue that the emotional process permits us to reveal to ourselves our true preferences or values. We thus claim that recent economic analysis should take into account the role of emotion not only as a psychic cost or even a rational tool but also as a useful warning process.

Keywords: Theory of emotions, obedience to authority, cooperation, public goods game, moral values.

JEL : B49 ; A13 ; C90 ; H41

GRETHA UMR CNRS 5113
Université Montesquieu Bordeaux IV
Avenue Léon Duguit - 33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75 - Fax : +33 (0)5.56.84.86.47 - www.gretha.fr

Introduction

Dans la théorie du choix rationnel, les préférences des agents économiques sont supposées déterminées une fois pour toutes et, le plus souvent, stables dans le temps [Stigler et Becker (1977)]. On suppose également que les préférences « révélées » peuvent être déduites *a posteriori* des observations sur les marchés. Les préférences déterminent les choix des individus qui motivent leurs comportements. Dans cette analyse, comme le souligne Lewin (1996), les motivations des individus sont évacuées de telle façon que seule la logique des comportements perdure : la rationalité des individus maintient en effet la cohérence de la séquence si bien, qu'en définitive, ce sont bien les comportements qui révèlent la structure des préférences. La théorie économique standard admet cependant que les préférences individuelles ne sont pas nécessairement constantes. En revanche, lorsque les goûts évoluent, ils doivent respecter à tout moment une règle de cohérence dans le temps. En particulier, ces changements de goûts doivent être prévus. On suppose ainsi, qu'au lieu de véritablement changer de préférences à une date ultérieure, nous sommes capables d'anticiper à l'instant présent l'évolution de nos priorités. La théorie ne nous indique pas, par conséquent, comment nous ajustons nos préférences lorsque nos choix ont eu des conséquences néfastes et qu'ils sont en contradiction avec d'autres préférences. En ce sens, un modèle économique plus élaboré des préférences individuelles devrait rendre compte de la façon dont nous sommes amenés à modifier l'ordre de nos priorités.

L'introduction des émotions dans la théorie économique peut contribuer à améliorer la prise en compte des changements de goûts. Par nature, les « passions » sont en effet susceptibles de s'interposer entre le processus de décision et les comportements qu'il implique. Davantage, les émotions ont été plus récemment reconnues comme un facteur essentiel du processus de choix [Damasio (1995)]. Dans l'analyse économique, les émotions ont ainsi été introduites sous la forme d'un coût psychique, d'une préférence temporaire ou même d'un outil stratégique (voir Elster (1998) pour une revue). Le plus souvent, les tentatives de modélisation consistent à intégrer les émotions directement dans la fonction d'utilité des individus. Dans ce cadre, les émotions se limitent en fait à un simple argument de cette fonction. Elles n'ont donc pas vocation à nous signaler ou à nous rappeler nos priorités comme le suggérait Simon (1967). Elles ne décrivent pas non plus la complexité du processus émotionnel.

Dans cet article, notre objectif est d'étudier comment la prise en compte d'une théorie des émotions plus élaborée permet de représenter la dynamique des préférences individuelles et collectives et des comportements dans l'analyse économique. Notre étude repose sur la théorie des émotions de Pierre Livet (*Emotions et rationalité morale*, Sociologies, Puf, 2002). D'une part, celle-ci propose une conception dynamique et rationnelle du processus émotionnel qui a vocation, à terme, à réviser les préférences des individus. Elle introduit également le rôle des valeurs, qui correspondent à nos préférences les plus enracinées et qui sont une source de motivation pour l'action. D'autre part, la théorie des émotions (morales) est proche de la façon dont nos actions révèlent nos préférences dans la théorie économique académique. Nos valeurs y conditionnent en effet les stimuli émotionnels mais, *in fine*, c'est l'émotion (sa récurrence) qui a la faculté de nous révéler à nous-mêmes nos valeurs. L'émotion est nécessaire et joue le rôle d'un signal. Elle n'en est pas pour autant suffisante puisque la révélation des valeurs nécessite le concours de la raison.

Nous illustrons l'apport de la théorie des émotions à l'analyse économique dans deux cas polaires. Le premier concerne le rôle des émotions individuelles dans le modèle théorique

d'obéissance irrationnelle à l'autorité proposé par Akerlof (1991). Le second traite de l'effet des émotions collectives dans le jeu expérimental du bien public. Nous décrivons tout d'abord le cadre théorique de la théorie des émotions avant d'aborder ensuite ces illustrations.

I. Emotions, préférences et valeurs : le cadre théorique

1.1 L'émotion, un processus de révision de nos préférences

Dans son analyse sur les valeurs morales, Livet (2002) oppose les croyances et les préférences de l'individu à la réalité imposée par le monde. Nos croyances s'ajustent généralement à la réalité du monde, elles n'ont pas vocation à transformer le monde. A contrario, la direction d'ajustement des désirs consiste à ajuster le monde à nos préférences, c'est-à-dire à changer le monde. Dans ce schéma, les émotions jouent le rôle d'une interface entre le monde et nos préférences (figure 1).

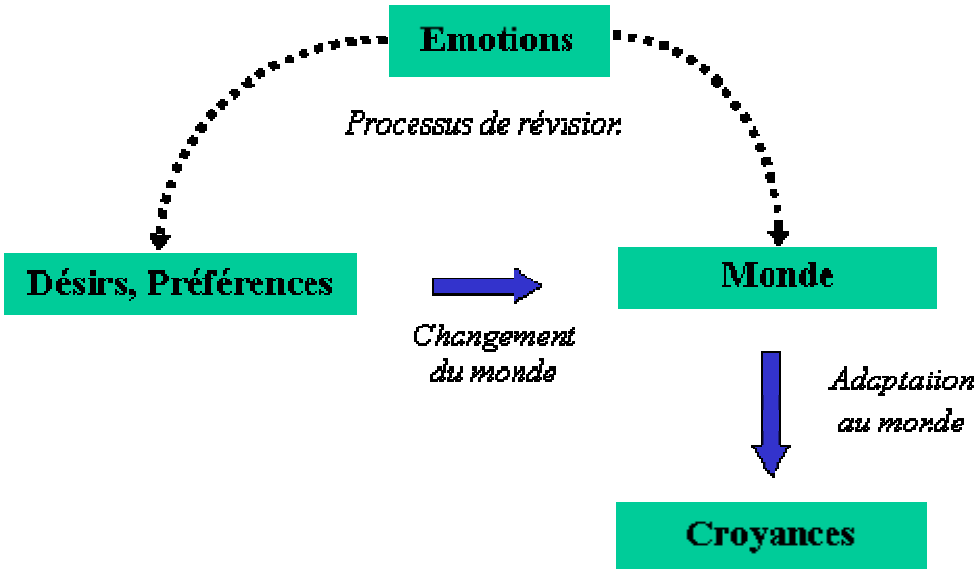

Figure 1 : L'émotion et la révision des préférences

Les émotions nous ajustent au monde, comme les croyances, mais en fonction des ajustements que nous souhaiterions de la part du monde. Elles indiquent dans quelle mesure le monde peut satisfaire ou non nos préférences, du point de vue de nos préférences. Les émotions fonctionnent comme un signal d'alarme, une mémoire d'origine sensitive : elles sont sensibles, réceptives, à ce que propose le monde. Elles font écho à la réalité du monde et nous incitent à adapter nos croyances et nos préférences. En ce sens, Livet (2002) définit le processus émotionnel comme un processus de révision des croyances et des préférences. L'émotion se déclenche lorsque la réalité que nous percevons, imaginons ou reconnaissons, ne correspond pas à nos attentes en cours ou à nos préférences. Le processus de révision comporte trois caractéristiques essentielles.

Il dispose tout d'abord, dans une perspective de long terme, d'une certaine forme de rationalité1. Adossée à un objectif simple de révision, l'émotion nous donne la motivation nécessaire pour changer l'ordre de nos préférences. Ni les croyances, ni les choix, ni même les désirs n'ont cette capacité là. Livet (2002) suppose que l'intensité du signal envoyé par l'émotion est d'autant plus forte que le différentiel entre la réalité et nos préférences est élevé ; il admet également comme hypothèse que le signal persistera tant que la révision n'a pas été accomplie. Le processus se poursuit donc en dynamique via la récurrence des émotions. Inversement, lorsque l'émotion se répète et que la révision s'effectue graduellement, l'émotion tend à disparaître : on dira qu'il y a accoutumance.

L'émotion nous avertit que quelque chose est à changer dans l'ordre de nos priorités sans toutefois nous indiquer la manière de changer ces priorités. Le processus vise donc à déstabiliser l'ordre de nos préférences. Une condition de réussite de la révision est qu'elle soit par conséquent non consciente. Nous pouvons avoir conscience de son résultat, lorsque le processus est terminé, mais pas du processus qui y a conduit. La seconde caractéristique du processus est donc qu'il est, par nature, le plus généralement inconscient. Un cas particulier important correspond cependant à la situation dans laquelle nous imaginons consciemment les conséquences émotionnelles de nos actions futures pour guider nos décisions présentes et nous révéler la véritable hiérarchie de nos préférences. Le ressenti émotionnel d'une situation imaginée peut en particulier nous aider à résoudre le problème de l'incohérence temporelle de nos préférences2. Les émotions nous servent alors en tant que force de rappel (consciente) de nos désirs et nous fournissent une solution plus souple et moins risquée que celle que la raison nous donne via la stratégie d'engagement préalable proposée par Elster (1982). La solution des émotions implique, cependant, que nous ayons déjà fait l'expérience d'une émotion similaire dans un cas semblable et que cette émotion ait été, et soit toujours, une incitation à la révision de nos actions. En se sens, « les émotions ne soutiennent la raison que si leur articulation avec un processus de révision s'est déjà enclenchée » Livet (2002).

Enfin, nous retenons que le processus est limité. La finalité de la révision est celle de notre adaptation au monde. Aussi nécessaire soit-elle, il n'est pas concevable que nous ayons à modifier nos croyances ou nos préférences chaque fois que le monde nous est hostile ou contraire. Il est ainsi logique, selon Livet (2002), que « l'évolution nous (ait) dotés d'une résistance à la révision plus forte que celle d'un calculateur probabiliste ». Pour l'auteur, ceci implique toutefois deux choses. Tout d'abord, qu'il existe des émotions qui n'ont pas vocation à réviser nos préférences, mais au contraire à résister à la pression du monde pour les préserver. C'est le cas notamment de la colère, de l'indignation, voire du dégoût. Ces émotions peuvent renforcer nos actions, et donc les réviser, et nous indiquent l'existence de préférences bien enracinées auxquelles le monde est particulièrement hostile. La limitation du processus implique également que son blocage puisse être effectué par le mécanisme même qui est à l'origine de la révision, à savoir l'émotion. Livet (2002) décrit ainsi un certain nombre de cas dans lesquels l'angoisse, qui est le symptôme qu'une révision difficile est en jeu, nous détourne vers une révision de dérivation pour éviter de remettre en cause une préférence profondément enracinée. Le blocage se traduit donc par une dérivation, un détournement non conscient de l'objectif implicite de la révision. Concrètement, la révision de dérivation implique un processus cognitif inconscient dont les phénomènes de « duperie de

[1] Au sens où les émotions seraient « raisonnables » [De Sousa (1987) ; Damasio (1995)].

[2] Lorsque, par exemple, notre désir à l'instant présent nous incite à prolonger une soirée entre amis mais que nous imaginons et ressentons les conséquences de ce choix sur notre activité du lendemain (rage impuissante, contrariété, regret). L'émotion imaginée nous révèle (et nous permet d'accéder à) notre préférence de long terme.

soi » [Fingarette (1998)], de névrose obsessionnelle ou de « faiblesse de la volonté » [Davidson (1991)] sont les principales manifestations.

Rationalité, inconscience et autolimitation sont les caractéristiques principales du processus de révision. L'émotion, nous l'avons vu, ne se limite cependant pas à la révision. Elle se définit également par opposition à ce processus. Dans ce cas, elle représente une interface entre le monde et nos désirs d'une autre nature : elle prétend résister au monde et/ou changer le monde. Livet (2002) fait naître de cette opposition au monde la révélation de nos valeurs.

1.2 L'émotion, un révélateur de nos valeurs

En suivant Livet (2002), nous posons qu'une valeur correspond à « une orientation de nos préférences, orientation que nous prétendons pouvoir justifier face à des positions contraires ». Implicitement, une valeur doit pouvoir résister à des tentatives de révision pour justifier d'une certaine stabilité. Or, l'émotion nous suggère des révisions lorsque nous percevons un différentiel entre un aspect d'une situation et nos orientations en cours. Le blocage de cette révision nous révèle à nous-mêmes que nos orientations sont profondément enracinées et qu'il est par conséquent difficile, voire impossible, de les rétrograder dans notre hiérarchie des préférences. Il est ainsi logique de postuler que nos valeurs sont à rechercher au sein de ces préférences enracinées. L'émotion serait donc effectivement un « révélateur de nos valeurs ».

Le principe de révélation proposé par Livet (2002) s'apparente à la manière dont nos choix révèlent nos préférences dans la théorie de la décision (*i*). Le principe repose sur un processus dynamique dans lequel l'individu prend conscience de ses valeurs en imaginant les implications émotionnelles (partageables) de ses décisions ou de ses préférences (*ii*). Il introduit de surcroît un critère de démarcation qui permet de séparer les valeurs des préférences temporaires, mais aussi des habitudes (*iii*).

(*i*) Dans la théorie du choix rationnel, on postule qu'il ne nous est pas indispensable de faire des choix pour avoir des préférences. Nous sommes dotés d'une structure des préférences que l'on suppose connue *a priori*. Cependant, c'est en effectuant nos choix que nous nous révélons à nous-mêmes nos préférences. De même, nous pouvons effectuer des jugements de valeurs sans être ému. C'est, cependant, en éprouvant des émotions que nous pouvons déceler quelles sont nos valeurs. L'émotion est nécessaire, elle n'en est pas pour autant suffisante. Car, la perception des valeurs a une dimension plus active que celle que présuppose l'émotion, qui reçoit du monde et y fait simplement écho. Porter des jugements de valeurs, *a contrario*, c'est « étalonner le monde à l'aune de nos préférences les plus enracinées » (op. cit.), c'est l'approuver ou le désapprouver. L'expérience de valeurs implique que nous nous opposions au monde et qu'inversement le monde s'oppose à nos valeurs : c'est en vivant consciemment cette double opposition que nous pouvons percevoir la robustesse de nos valeurs. Livet (2002) pose ainsi que « toute préférence qui présente la dynamique croisée en question aura donc le statut de valeur ».

(*ii*) Comme dans le cas de la révision des préférences, le processus de révélation des valeurs est dynamique : il s'effectue par « des coups de sonde que sont nos essais de jugement de valeurs, coups de sonde ou ondes de chocs qui déclenchent ces échos radars que sont les émotions » (op. cit.). C'est donc bien la résistance répétée des émotions à l'accoutumance qui peut nous révéler à nous-mêmes nos valeurs. Cependant, dans l'expérience de valeurs morales, la résistance n'est plus inconsciente mais consciente : résister, c'est « avoir la

capacité, une fois la révision menée jusqu'où elle peut aller, de maintenir une position » (op. cit.), et donc de la justifier. Résister, c'est imposer une figure de test à nos valeurs, en imaginant les conséquences émotionnelles induites par nos préférences les plus enracinées. Enfin, résister, c'est également partager ses émotions avec autrui. C'est le partage de nos émotions, du moins sa potentialité, qui nous permet d'ériger notre préférence en valeur. L'émergence et la reconnaissance d'une valeur impliquent en particulier, argument ultime, une « révision intersubjective » dans laquelle « nos conséquences émotionnelles imaginées sont soumises à la révision émotionnelle d'autrui, et inversement » (op. cit.). Une expérience de valeurs morales est *complète* lorsqu'elle a donné lieu à un jugement qui s'expose à la critique et au débat. Partager ses émotions, c'est donc « les transformer en valeurs socialement reconnues et résistantes à un destin contraire » (op. cit.).

(*iii*) Le principe de révélation fournit également un critère permettant de distinguer les valeurs des préférences temporaires mais aussi des habitudes. Nos valeurs sont l'objet d'un débat, d'une confrontation, d'une opposition au monde mêlant nos émotions, nos croyances et nos désirs. Les émotions ne sont pas en ce sens la source des valeurs, elles y donnent simplement accès. La résistance des émotions à la révision dévoile en effet nos préférences les plus ancrées et nous choisissons nos valeurs parmi ces préférences. Les valeurs sont ainsi le fruit de toute une histoire de révision passée inachevée et d'un débat conscient, qui témoignent que la préférence que l'on érige en valeur résistera à tout processus émotionnel futur. La valeur s'impose au monde contrairement à une préférence temporaire, qui résiste à la révision et qui demeure cependant inadaptée à notre environnement (les « faux positifs », les « valeurs illusoires » ou plus généralement les résistances cognitives inconscientes). Le critère de révélation distingue également les valeurs des habitudes. Les valeurs ne sont pas en effet simplement des lieux de résistance au monde, comme peuvent l'être de façon passive les habitudes. Elles présentent également un caractère prescriptif d'universalisation : elles s'opposent aux changements du monde et cherchent à s'imposer à lui, elles portent en elle l'attente que l'exigence de valeur soit satisfaite. Les valeurs rempliraient en ce sens la fonction inverse des émotions, à savoir réaffirmer nos ajustements souhaités face aux changements du monde ; « elles seraient alors symétriques des émotions au lieu d'être simplement ce à quoi elles donnent accès » (op. cit.).

II. Obéissance, valeurs et rationalité des émotions individuelles

2.1 Obéissance et dissonance cognitive

Dans son analyse de l'obéissance immorale, Akerlof (1991) propose un modèle de procrastination qu'il utilise pour décrire le comportement des sujets au cours de la célèbre expérience de psychologie sociale de Milgram (1974)[3]. Akerlof (1991) insiste en particulier sur l'irrationalité du comportement des sujets qui accordent trop d'importance à leur décision

[3] Rappelons, qu'au cours de l'expérience, les sujets (les « professeurs ») sont progressivement amenés à infliger des chocs électriques (qu'ils croient douloureux) à des élèves (en fait, les complices de l'expérimentateur) de façon à tester les « effets de la punition sur l'apprentissage ». Les sujets sont dupés en ce qui concerne l'objectif réel de l'expérience qui vise l'étude de la soumission à l'autorité. Les résultats surprenants de Milgram (1974) montrent que 65% des sujets vont jusqu'au bout de l'expérience (30ème curseur, 450 volts). En revanche, les résultats sur les prédictions des individus (qu'ils soient psychiatres, étudiants, salariés, etc.) montrent qu'ils estiment unanimement qu'eux-mêmes ou autrui n'auraient pas obéi. Pour une revue récente des nombreux travaux sur l'expérience de Milgram, voir Blass (1999).

initiale. A l'origine de cette forme d'irrationalité se trouve l'incapacité des agents à tenir compte consciemment de leurs structures cognitives. Les sujets choisissent d'obéir au début de l'expérience, planifient de désobéir si celle-ci continue et ne se rendent pas compte que cette stratégie les conduit à obéir jusqu'à la fin.

Akerlof (1991) modélise le choix du sujet au cours de l'expérience comme le résultat d'un processus de maximisation de son utilité intertemporelle, Vt. La fonction d'utilité prend en compte le coût de la désobéissance, qui dérive essentiellement du contrat explicite que le sujet a passé initialement avec l'expérimentateur, et le coût de l'obéissance, qui traduit la désutilité associée à la mise en œuvre de la punition de la victime. Si, à la période t, le sujet désobéit, il obtient un niveau d'utilité équivalent à :

$$(1) \qquad V_t = -bD(1+\delta),$$

où D est le coût de la désobéissance et où le paramètre δ représente le coût supplémentaire induit par une désobéissance à la période présente[4]. Si, en revanche, il repousse sa décision à la période suivante et ne désobéit qu'à la période $T \geq t+1$, son utilité espérée est égale à :

$$(2) \qquad V_t = -bD - c\sum_{k=t}^{T-1}(W_k - W_{t-1}) \qquad T \geq t+1,$$

où W_k est le voltage administré par le sujet à l'élève à la période k, W_{t-1} le voltage correspondant à la période précédente.

Pour une valeur suffisamment élevée du paramètre δ, le sujet qui maximise son utilité intertemporelle a tendance, à chaque période, à remettre à la période suivante le choix de désobéir à l'expérimentateur. Le sujet obéit tout en projetant de désobéir à la période suivante (procrastination). Ses anticipations sont en fait naïves au sens où le sujet croit que son comportement initial d'obéissance n'a aucun impact sur son comportement ultérieur. Le sujet naïf anticipe donc à tort qu'il pourra concrétiser sa préférence future (différente de la préférence initiale) à une étape ultérieure de l'expérience. Akerlof (1991) montre ainsi qu'il est victime d'une incohérence temporelle. L'incohérence se traduit par une obéissance « immorale » et « irrationnelle ». Elle est immorale au sens où, comme l'attestent les prédictions des individus, les sujets obéissants agissent à l'encontre de leurs valeurs morales (ne pas faire souffrir une personne innocente). Elle est irrationnelle dans la mesure où le comportement des sujets contredit le principe des préférences révélées : ces derniers ne maximisent pas leur vrai niveau d'utilité (« *true* utility »), leurs choix effectifs ne révèlent pas leurs préférences réelles.

Dans l'équation (2), le coût de l'obéissance dépend de la variation du voltage par rapport à la période précédente (15 volts) et non pas de son niveau absolu (du 1er au 30ème curseur, c'est-à-dire de 15 à 450 volts). Conformément à la théorie de la dissonance cognitive [Festinger (1957)], le sujet rationalise l'action qu'il a entreprise à la période précédente. Dans le modèle, cette hypothèse est cruciale : la dissonance révèle le conflit auquel le sujet est confronté entre le désir d'arrêter la souffrance de la victime et le besoin de se conformer aux ordres de l'autorité. La rationalisation non consciente de ce conflit permet au sujet de continuer l'expérience. Inversement, l'absence de rationalisation ne ferait qu'accroître la

[4] Le paramètre δ représente le biais qui s'exerce en faveur du présent (aujourd'hui plutôt que demain) caractéristique des modèles sur l'incohérence temporelle des préférences [O'Donoghue et Rabin (1999)].

tension émotionnelle ressentie par le sujet qui est une incitation à la désobéissance. En ce sens, Akerlof (1991) propose une formulation de la « vraie » fonction d'utilité de long terme du sujet (désobéissant) qui élimine la dissonance cognitive :

$$(3) \qquad V_0 = \sum_k \{-bD_k - cW_k\}$$

Dans l'équation (3), la désutilité associée à l'obéissance est une fonction croissante du voltage administré à l'élève. En effet, la hausse de l'intensité des chocs amplifie la souffrance (simulée) de la victime et accroît en conséquence la tension émotionnelle du sujet. Une forte tension émotionnelle est donc symptomatique d'une désutilité élevée. Dans cette formulation, les émotions du sujet sont intégrées directement dans la fonction d'utilité de l'individu ; elles en sont un argument. En maximisant son utilité intertemporelle, le sujet peut ainsi anticiper que la poursuite de l'expérience augmentera suffisamment les coûts liés à l'obéissance pour qu'il renonce à s'engager dans le schéma de soumission proposé par l'expérimentateur. La formalisation de la vraie fonction d'utilité représentative du comportement du sujet dans l'expérience de Milgram n'est cependant pas adéquate pour trois raisons essentielles : (*i*) au cours de l'expérience, la tension n'augmente pas linéairement avec le voltage notamment en raison de mécanismes de résolution de la tension utilisés par le sujet[5], (*ii*) Milgram (1974) n'a pas observé de lien entre le niveau de tension, indiqué par le sujet au cours de l'interview post-expérimentale, et l'acte de désobéissance et (*iii*) la tension ressentie par le sujet dépend de la pression exercée par l'expérimentateur, et donc, formellement du coût de la désobéissance (D)[6].

L'interprétation d'Akerlof met l'accent sur l'incapacité des sujets de l'expérimentation à tenir compte consciemment de leurs structures cognitives ; elle ne rend cependant pas compte de la complexité du processus émotionnel. En utilisant la théorie des émotions de Livet (2002), nous montrons que la désobéissance peut être perçue comme une expérience victorieuse de valeurs que les émotions permettent de révéler.

2.2 La désobéissance, une expérience de valeurs victorieuse

Alors que les sujets entrent volontairement, sans conflit apparent, dans le système d'autorité proposé par l'expérimentateur, la poursuite de l'expérience les confronte de façon répétée à la souffrance morale et physique de leur victime. La tension émotionnelle récurrente fait écho à l'hostilité du monde et, fonctionnant comme un signal d'alerte, avertit le sujet qu'une de ses préférences, le « besoin d'obéir », est à rétrograder dans l'ordre de ses priorités. Plus encore, elle montre que le processus de révision est engagé : les manifestations émotionnelles observées en laboratoire représentent ainsi autant de preuves que le sujet envisage de désobéir. La tension émotionnelle est également le signe que les sujets s'opposent à la volonté de l'expérimentateur et que, dans le même temps, ce dernier exerce une pression sur eux. Le sujet s'oppose consciemment au monde, il le désapprouve, et le monde s'oppose au sujet, via l'attitude inflexible et insistante de l'expérimentateur. La récurrence de l'émotion

[5] La dérobade, le refus de l'évidence, la désapprobation feinte, la recherche de la confirmation (auprès de l'autorité) de sa non-responsabilité ou encore les manifestations psychosomatiques sont des mécanismes cognitifs ou sensitifs qui permettent au sujet de continuer l'expérience. Dans le modèle d'Akerlof (1991), ces mécanismes sont retranscrits au travers du principe de la dissonance cognitive.

[6] Rochat, Maggioni et Modigliani (1999) modélisent ce lien en rappelant que, conformément au protocole expérimental, l'expérimentateur renforce la pression sociale de l'autorité (« vous devez continuer ! ») en cas d'hésitation, de résistance ou de désapprobation du sujet.

permet la révélation d'une valeur réelle, qui tient face à l'hostilité du monde. Pour les sujets désobéissants, l'expérimentation est ainsi une expérience de valeurs dont ils sortent vainqueurs. Le sujet s'indigne face à une violence envers une personne innocente et résiste à la pression de l'autorité en désapprouvant son comportement. La désapprobation réelle (et non feinte) constitue ainsi le premier stade d'un conflit progressif entre le sujet et l'expérimentateur, une façon de sonder sa conviction ou de l'amener à réviser sa position. Rochat et Modigliani (1995) montrent ainsi qu'une résistance verbale précoce est prédictive du refus d'obéissance du sujet : tous les sujets qui s'arrêtent à 150 volts (10ème curseur) ou avant le font alors qu'ils ont manifesté leur désapprobation avant 150 volts. Or, les auteurs établissent un lien entre la tension émotionnelle et la désapprobation puisque celle-ci appelle une réponse de l'autorité prévue dans le protocole. On en déduit que plus les processus de tension émotionnelle et d'opposition consciente à l'autorité démarrent tôt, plus le sujet a des chances de ne pas obéir à l'autorité.

Les nombreuses variantes de l'expérience originelle montrent également que la désobéissance s'accroît lorsque l'expression ou la transmission des émotions est favorisée par le protocole[7]. Les psychologues sociaux y voient notamment un élément de preuve que les réactions d'empathie émotionnelle chez le sujet lui donnent une compréhension plus aiguë de ce qu'endure la victime. Nous dirons, de notre côté, que le partage de l'émotion avec la victime donne au sujet un accès à la valeur de non violence ou d'absence de cruauté.

Le rapprochement de l'élève et du professeur est également l'occasion pour le sujet de tester sa capacité d'obéissance via le processus de révision. Au contact de l'élève, en effet, le sujet se rend compte qu'il est devenu un élément plus saillant dans le champ cognitif de la victime et il lui est plus difficile, en conséquence, de lui infliger des chocs électriques. Puisque la victime est désormais le témoin de son action, le sujet peut ressentir de la gêne, de la honte voire de la culpabilité, en fonction de l'état d'avancement du processus de révision. La gêne, en effet, « c'est l'impression que les autres voient notre comportement comme à réviser » Livet (2002). Contrairement à la honte ou à la culpabilité, cependant, le sujet qui ressent de la gêne ne sait pas forcément sur quelle propriété de son action porte le jugement négatif de l'élève. La gêne est donc le signe de l'amorce du processus de révision. Après la gêne peut venir la honte qui est associée à « une prise de conscience de ce que les autres considèrent un de nos actes comme mauvais, même si de notre point de vue cet acte est bon » (op. cit.). Dans le cas de la honte, cependant, « nous adoptons une attitude incohérente à l'égard de la responsabilité. D'un côté, nous la reconnaissons, sans quoi nous n'aurions pas honte. De l'autre, nous la fuyons : notre premier souci, dans la honte, est de nous cacher, de disparaître provisoirement ou définitivement » Ogien (2002). Ce n'est pas le cas de la culpabilité qui implique que nous reconnaissions ouvertement nos actions. Le sens de la responsabilité associée à l'émotion de culpabilité indiquerait donc que le processus de révision est davantage engagé que lorsque le sujet ressent de la honte ou de la gêne. Au cours de l'expérience, cependant, l'occurrence de la culpabilité demeure plus improbable que celle de la honte, puisque dans la honte, ce qui nous réveille, « c'est une secousse extérieure, le regard de l'autre, alors que dans la culpabilité, c'est une sorte de sonnerie intérieure (…) que nous *entendons* » Ogien (2002).

Accès à la valeur de non violence, d'un côté, test de la valeur d'obéissance, de l'autre, les émotions suscitent effectivement un débat sur les valeurs. Une des conditions pour que

[7] Dans la variante n°3 dite de « proximité de la victime », le taux de désobéissance s'élève à 60% et, dans celle dite de « contact » (n°4), il monte à 70%.

l'émotion-valeur de refus d'obéissance l'emporte sur la réalité imposée par le monde (la volonté de l'expérimentateur), c'est que ce débat ait lieu consciemment. L'expérience de valeurs vécue par les sujets qui s'opposent à l'autorité ne s'impose qu'à une petite minorité des sujets. Nous montrons ci-dessous que l'obéissance à l'autorité provient du blocage émotionnel du processus de révélation des valeurs.

2.3 L'obéissance, un blocage du processus de révélation des valeurs

Au cours de l'expérience, l'angoisse est le symptôme que la révision d'une préférence (le besoin d'obéir) est difficile à réaliser. Pour des raisons très largement décrites par Milgram (1974), cette préférence est effectivement profondément enracinée[8]. Puisqu'il s'agit d'une révision délicate, l'angoisse peut par conséquent également être l'un des facteurs bloquants de la révision. L'obéissance est en effet un comportement usuel et adéquat en présence d'une autorité dans un contexte social. Par conséquent, lorsqu'il envisage de désobéir, le sujet est en proie à une inquiétude vague provenant de la crainte que lui inspire l'inconnu. Que se passe-t-il si je transgresse la règle sociale implicite d'obéissance ? Les situations qui nous angoissent sont effectivement celles « dont l'indécidabilité elle-même n'est pas assurée, et donc où nous ne disposons d'aucune procédure de repérage en laquelle nous puissions avoir confiance » Livet (2002). L'angoisse témoigne donc du fait que le sujet ne sait pas s'il sera capable de réviser cette préférence fondamentale. En conséquence, l'angoisse bloque la révision en orientant le sujet vers une révision de dérivation (qui dissipe provisoirement l'angoisse). Tous les mécanismes de résolution de la tension en sont des manifestations diverses : par exemple, l'utilisation de subterfuges (minimisation de la durée du choc électrique ou de son intensité en l'absence de l'expérimentateur) donne au sujet l'impression, l'illusion, qu'il brave l'autorité et qu'il fait preuve d'un sentiment de compassion vis-à-vis de l'élève ; de même, la focalisation obsessionnelle du sujet vers les tâches requises par l'expérience lui permettent de détourner son attention de la victime et, en conséquence, de la révision qui reste à faire ; plus généralement, le sujet s'engloutit peu ou prou dans un processus de « duperie de soi » dont la dérobade, le refus de l'évidence, ou la désapprobation feinte en sont les manifestations. Le blocage émotionnel du processus, via une révision de dérivation, correspond à la conversion du sujet à l'état « d'agent » Milgram (1974).

Dans notre analyse, l'obéissance est donc le résultat du blocage du processus de révélation des valeurs. Ce blocage intervient lorsque l'angoisse détourne le sujet de la révision nécessaire de l'ordre de ses priorités. Hors contexte expérimental, cependant, les prédictions des sujets concernant leur comportement ou celui d'autrui dans une telle situation contredisent les résultats obtenus en laboratoire [Mixon (1972) ; Freedman (1969) ; Maughan et Higbee (1981)]. Milgram (1974) analyse l'écart entre ses résultats en laboratoire et les prédictions des sujets comme le signe pour ces derniers « d'une singulière méconnaissance du réseau complexe des forces qui interviennent dans une situation sociale réelle » et comme une incapacité de l'individu à « transformer convictions et valeurs en actes ». Nous y voyons également, de façon différenciée, le rôle de l'émotion imaginée dans le processus de révélation des valeurs.

[8] L'obéissance à l'autorité découle d'un mode d'organisation social (familial et hiérarchique) dont le résultat est « *l'intériorisation de l'ordre social*. Autrement dit, l'individu adopte pour son compte personnel l'ensemble des axiomes qui régissent la vie collective, le principal étant : 'Faites ce que votre supérieur vous dit' » Milgram (1974).

2.4 L'émotion imaginée peut-elle nous aider à défier l'autorité ?

Constater l'écart manifeste entre l'auto prédiction d'un sujet et son comportement *attendu* au cours de l'expérience peut nous conduire à nous demander pour quelle raison il a tant de mal à admettre qu'il aurait probablement obéi. Une autre façon de s'interroger sur la pertinence de cet écart consiste, à notre sens, à nous demander pour quelle raison il lui est si facile *d'imaginer* qu'il désobéirait dans la situation proposée en laboratoire. Quel est le processus cognitif et émotionnel qui permet à l'individu d'affirmer sans aucun doute sa position morale ? Notre hypothèse ici est que la méthode d'auto prédiction fournit à l'individu le cadre le plus propice au test de ses valeurs contradictoires : elle concilie la possibilité d'un débat moral intra subjectif avec le ressenti émotionnel de situations que nous imaginons. Cependant, contrairement à une situation vécue, l'imagination sert de modérateur dans ce débat, comme le soulignait Adam Smith (1790), et nous permet idéalement de nous révéler à nous-mêmes les valeurs auxquelles nous adhérons. Dans un tel débat, en effet, « nous ne nous laissons par submerger par une première émotion (…), nous recherchons les autres émotions qui pourraient être les conséquences de la situation envisagée. Sortir d'une émotion pour passer à une autre nous est ici possible, parce que nous sommes seulement en train d'imaginer des situations au lieu de les vivre et que l'imagination nous donne une capacité de variation qu'une émotion imposée par une situation nous laisserait plus difficilement » Livet (2002). L'imagination d'émotions partageables nous suffit donc généralement pour éviter le blocage par les émotions du processus de comparaison des valeurs. En l'absence de blocage et de confusion autour des valeurs, l'individu se positionne sans ambiguïté sur une valeur morale ce qui signifie, en un sens, qu'il se sent prêt à la défendre et à la prescrire. Pourtant, nous savons que, dans le contexte réel, il ne suivra pas cette prescription et ne transformera pas cette valeur en action. A notre sens, l'écart entre sa prédiction et son comportement réel s'explique par le rôle joué, d'une part, par l'émotion, imaginée ou non, et d'autre part, par la valeur morale, dans le processus qui conduit à l'action. Comme nous l'avons déjà souligné, les émotions nous incitent à réviser nos préférences ainsi que nos comportements et nous aident également à « transformer la croyance dans une valeur en une action qui réalise cette valeur » Livet (2002). Face au monde, l'émotion est vécue comme une résistance mais aussi comme un moteur de l'action pour changer le monde. En revanche, notre imagination émotionnelle ne produit pas des émotions qui ont toute la richesse d'une émotion produite par une situation réelle. Elle autorise une certaine modération dans le débat moral mais ne nous donne pas la motivation pour agir. En revanche, « la motivation des valeurs tend à nous faire franchir une (…) distance, celle entre la simple orientation de nos préférences et la prescription, qui impose l'action – *ce qui, cependant, ne suffit pas toujours à nous faire agir* » Livet (2002, c'est nous qui soulignons). Au cours de l'auto prédiction, la désobéissance est donc le fruit d'un débat conscient (la révélation) qui ne suffit cependant pas à garantir la mise en œuvre de cette valeur en situation réelle. Au cours de l'expérience, l'émotion qui nous pousserait à agir n'est pas associée à un processus de délibération suffisamment avancé. C'est la raison pour laquelle, probablement seuls les sujets déjà instruits par un débat moral préalablement à l'expérience (le cas 'Gretchen Brandt') ou qui disposent d'un niveau de jugement moral [Kohlberg (1969)] et/ou d'intelligence sociale [Burley et McGuinness (1977)] élevé(s), sont en mesure de concrétiser la valeur en actes. Ceci est conforme à la réalité d'un processus d'évaluation de nos valeurs long et toujours inachevé.

2.5 Implications

Dans son modèle, Akerlof (1991) propose une interprétation du comportement d'obéissance qui repose sur le principe de l'incohérence temporelle des préférences. Les comportements observés ne révèlent pas les vraies préférences du sujet. A l'origine de cette irrationalité se trouve l'incapacité du sujet à anticiper les effets de la pression sociale sur sa structure cognitive. La tension émotionnelle crée une dissonance cognitive qui incite le sujet à rationaliser l'action (d'obéissance) qu'il a entreprise précédemment[9]. Akerlof (1991) en conclut, qu'en l'absence de dissonance, l'accroissement progressif de la tension devrait conduire à la désobéissance et révéler ainsi les vraies préférences de long terme du sujet. Implicitement, dans le modèle, l'émotion est ainsi intégrée directement dans la fonction d'utilité, le niveau de voltage étant un indicateur de la tension émotionnelle et du désagrément qu'elle procure au sujet. L'émotion module les préférences du sujet et motive en conséquence l'action : le comportement d'obéissance est le fruit de la dissonance induite par la tension (équation 2), celui de défiance provient des effets anticipés des coûts émotionnels croissants de la soumission (équation 3).

Comme l'indique notre interprétation de l'expérience de Stanley Milgram, les effets des émotions sur le comportement sont en réalité plus ambigus et plus variés : elles peuvent éventuellement perturber le processus de révision des préférences dans un contexte d'incertitude ou d'ambiguïté mais elles sont essentiellement à l'origine de nos changements de priorité et participent au mécanisme de révélation de nos valeurs. Au cours des processus de révision ou de révélation, l'émotion représente un signal, récurrent et le plus souvent inconscient. La rationalité des émotions consiste, pour l'individu, soit à prendre acte *a posteriori* de la modification de l'ordre de ses préférences, soit à constater l'enracinement d'une de ses préférences résistantes qu'il érige consciemment en valeur. Dans le premier cas, nous nous rendons compte, par exemple, qu'une fois le changement opéré, nous préférons désormais les œuvres littéraires contemporaines aux œuvres classiques. Dans le second, nous restons ferme quant à celles de nos préférences qui témoignent de nos valeurs (sens de l'équité, altruisme, individualisme, etc.). L'information transmise par nos émotions peut cependant être consciente lorsque nous les ressentons en imagination : elles sont utiles pour nous révéler consciemment nos vraies préférences ou nos valeurs morales ; elles sont cependant insuffisantes pour motiver une action future, puisque l'anticipation n'implique pas nécessairement l'action. Le cas échéant, la distance entre l'orientation de nos préférences et l'action justifierait donc le principe d'engagement préalable [Elster (1982)] : en l'absence notamment d'un solide engagement moral, la rationalité de l'émotion anticipée impliquerait que l'on restreigne les choix de l'individu à sa place.

Notre analyse s'est limitée jusqu'ici au cas des émotions individuelles. Nous la prolongeons en envisageant le rôle des émotions collectives dans une expérimentation.

[9] Myers (2006) rappelle ainsi qu'au moment de la première protestation de l'élève (à 75 volts), le sujet a déjà obéi cinq fois.

III. Coopération et émotions collectives dans le jeu du bien public

3.1 Les principaux enseignements des études expérimentales

Le jeu du bien public expose une situation de dilemme social dans laquelle l'intérêt individuel entre en conflit avec l'intérêt collectif[10]. Sur le plan théorique, la solution à ce dilemme privilégie les comportements opportunistes au détriment des comportements coopératifs. La théorie standard prédit en effet que les individus cherchent à profiter du bien public tout en évitant, dans la mesure du possible, de participer à son financement : ils se comportent ainsi comme des « passagers clandestins »[11]. Sur le plan expérimental, ces prédictions théoriques ont cependant été systématiquement invalidées[12] : les expériences en laboratoire montrent, en particulier, 1°) que les sujets contribuent volontairement à la cagnotte commune entre 40 et 60% de leur dotation initiale (lors du premier tour de jeu), 2°) que lorsque le jeu est répété, le niveau de contribution est initialement proche des 50% mais décroît en revanche rapidement et 3°) que trois types de comportement peuvent être identifiés, soit (*i*) les contributeurs inconditionnels, (*ii*) les opportunistes qui ne contribuent jamais, et, enfin, plus majoritairement, (*iii*) les contributeurs conditionnels qui adaptent leur contribution en fonction du niveau moyen observé dans la période précédente [Keser et van Winden (2000)]. L'instauration initiale d'une coopération entre joueurs ainsi que leur incapacité à la maintenir de façon stable au cours du temps posent deux questions différenciées : comment expliquer la contribution volontaire des participants ? Quels sont les facteurs qui jouent sur leur taux de contribution ?

Dans la littérature, la coopération provient essentiellement de comportements altruistes[13] ou moraux[14], de la satisfaction individuelle issue de l'*acte* de contribution en lui-même[15], mais aussi d'erreurs de décision induites par le protocole expérimental[16]. L'influence de ces facteurs sur le taux de contribution dépend naturellement de certains paramètres clefs

---

[10] Le jeu se joue avec *n* joueurs, chacun d'eux recevant une dotation initiale de Y jetons à chaque tour. Chaque joueur, i = 1, …, n, décide à chaque tour du montant de sa contribution à la cagnotte commune, g_i. Les paiements du joueur sont : $(i) \quad \pi_i = Y - g_i + a\sum_{j=1}^{n} g_j \quad avec \quad 0 < a < 1 < na$ où a représente le rendement individuel marginal du bien public.

[11] $g_i = 0$ est la stratégie dominante du jeu puisque $\partial\pi_i / \partial g_i = -1 + a < 0.$ Les paiements agrégés sont maximisés lorsque chaque joueur contribue l'intégralité de sa dotation $(g_i = Y)$ puisque $\partial\sum_{i=1}^{n}\pi_i / \partial g_i = -1 + na > 0.$

[12] Pour une revue, voir Ledyard (1995).

[13] Les préférences altruistes impliquent que l'utilité du sujet augmente en fonction du gain monétaire réalisé par le groupe [Rabin (1993) ; Palfrey et Prisbey (1997) ; Anderson, Goeree et Holt (1998)].

[14] Qui incluent notamment l'aversion à l'inégalité [Fehr et Schmidt (1999) ; Bolton et Ockenfels (2000)], les vertus de la réciprocité [Axelrod (1992) ; Keser et van Winden (2000)] mais aussi l'éthique de groupe [Dawes, van de Kragt et Orbell (1988), dans le dilemme des prisonniers].

[15] Ce qu'Andreoni (1995) nomme le « *warm-glow effect* » dans lequel l'acte coopératif augmente l'utilité du contributeur d'un montant fixe (qui ne dépend donc pas des gains réalisés par le groupe).

[16] Andreoni (1995) suggère ainsi, avec d'autres [Palfrey et Prisbey (1997)], que les individus peuvent se méprendre sur le sens des instructions fournies par l'expérimentateur ou de la structure d'incitation définie dans le protocole. Cet impact, qui tend à s'estomper au fur et à mesure que se déroule l'expérience, est appelé « effet de confusion ».

définis lors du protocole : en particulier, la hausse du rendement marginal du bien public, du nombre de périodes ou de la taille du groupe [Isaac, Walker et Williams (1994)], la possibilité de communication entre les membres du groupe [Palfrey et Rosenthal (1991)], leur interaction sociale avant (ou après) le début du jeu [Gächter et Fehr (1999)] ainsi que certains effets de présentation [Park (2000)] augmentent les taux de contribution. D'autre part, des études récentes montrent que la mise en place d'une procédure de punition non stratégique[17] [Fehr et Gächter (2000a)], la menace d'une expulsion du groupe [Cinyabugama, Page et Putterman (2005)], la simple désapprobation punitive [Masclet, Noussair, Tucker et Villeval (2003)], ou encore l'approbation [Gatcher et Fehr (1999)] ont une influence notable sur les comportements de coopération.

Dans ces expériences, le rôle des émotions est souvent souligné : les émotions jouent ainsi un rôle stratégique (par exemple, en crédibilisant des menaces non monétaires dans Fehr et Gächter (2000a)) mais sont surtout à l'origine de motivations morales ou d'incitations sociales[18]. En économie comportementale, les émotions ont ainsi été intégrées formellement dans les préférences individuelles pour prendre en compte ces motivations morales ou ces incitations sociales. L'introduction des « sentiments moraux » (Serra, 2007) a donné naissance aux nouveaux modèles de préférences hétérogènes et « non auto-centrées ». La culpabilité, la honte, mais aussi la sympathie (au sens de David Hume) ou l'envie, y jouent un rôle central. La culpabilité, qui est une source de désutilité pour le joueur qui a conscience qu'il pénalise ses partenaires [Fehr et Schmidt (1999) ; Bolton et Ockenfels (2000)], ou la honte, lorsque le comportement opportuniste est dénoncé par les autres participants, incitent les passagers clandestins à rectifier leur attitude non-coopérative [Bowles et Gintis (2001)]. Dans le modèle de Rabin (1993), la prise en compte d'un coefficient endogène de sympathie (envie) dans la fonction d'utilité sociale des joueurs, lorsque ceux-ci anticipent de bonnes (mauvaises) intentions chez les autres joueurs, permet la réalisation d'un équilibre équitable coopératif (non-coopératif)[19]. D'autres types de modèles introduisent les émotions pour tenir compte de la pression sociale impulsée par les pairs [Kandel et Lazear (1992)] ou du rôle de l'approbation sociale [Holländer (1990)]. Cependant, comme le soulignent Masclet et al. (2003), il n'existe pas actuellement de consensus sur la façon de modéliser les processus émotionnels. Dans la plupart des approches, les émotions sont intégrées dans la fonction d'utilité des agents dans la lignée des travaux de Frank (1988), de Hirshleifer (1987) ou de Becker (1996). Dans son analyse du dilemme du prisonnier, Frank (1988) postule, par exemple, que la défection crée un sentiment de honte qui empêche l'agent de suivre le raisonnement rationnel de l'opportuniste. Il s'agit là, cependant, d'une conception statique, qui intègre les émotions dans les préférences en modifiant la matrice des gains du jeu. Cette conception rigide des émotions ne permet pas en particulier d'identifier tous les liens existant entre les émotions, les valeurs morales et les comportements de coopération dans le jeu du bien public. En revanche, la théorie des émotions morales met en avant la dynamique des

[17] La punition est qualifiée de non stratégique dans la mesure où chaque joueur sait qu'il est confronté à chaque tour de jeu à des partenaires différents déterminés de façon aléatoire.

[18] Par exemple, dans Gächter et Fehr (1999), l'intensité des émotions du joueur est un indicateur de son jugement moral (de désapprobation) ; dans Cinyabugama et al. (2005), la crainte de l'expulsion est la source de motivation de la coopération.

[19] Comme le souligne Meidinger (2000, p. 47), le jeu du dilemme du prisonnier se transforme dans ce cas (via les coefficients) en un jeu de coordination analogue au jeu de la chasse au cerf : dès lors, « il semble que, pour Hume, l'institution de promesse ait un rôle important de coordination ». La solution humienne impliquerait ainsi qu'une communication réciproque s'instaure entre les (deux) joueurs et qu'une telle promesse soit considérée comme crédible. Dans la littérature, la solution philosophique inspirée de l'impératif catégorique de Kant fait aussi figure de référence (à ce sujet, voir l'analyse éclairante de Wolfelsperger (1999)).

comportements et des valeurs ainsi que l'hétérogénéité des préférences : dans cette perspective, le jeu du bien public opposerait des individus qui affirment leurs valeurs (les contributeurs inconditionnels, qui plébiscitent la norme de coopération, ou les opportunistes qui défendent une valeur individualiste), à d'autres individus qui sont soumis à un conflit de valeurs (les contributeurs conditionnels). Les enjeux de la fourniture d'un bien public consisteraient ainsi à convaincre les contributeurs potentiels du bien fondé de la norme coopérative pour ensuite les engager à affirmer cette valeur.

3.2 Les vertus des émotions collectives[20]

La structure dynamique du jeu du bien public oppose les valeurs d'une majorité d'individus, les contributeurs potentiels, à la réalité d'un monde (peuplé d'opportunistes) qui leur est défavorable[21]. Le contributeur potentiel est ainsi soumis, au cours du jeu, à une « double révision » qui déstabilise ses repères mais qui peut le conduire en revanche à affirmer ses valeurs. D'une part, chaque coopérateur en puissance perçoit qu'il est de l'intérêt de chacun de ne pas contribuer tout en espérant que les autres le feront pour bénéficier du bien public. Il anticipe par conséquent que l'issue la plus favorable à la collectivité, la réalisation du bien public ou la contribution massive de tous (dans le jeu), est inespérée et très aléatoire. En ce sens, la concrétisation d'un bien public revêt bien le statut d'une valeur qu'il est nécessaire de défendre pour qu'elle puisse voir le jour. Dans le cadre expérimental, celui qui contribue prescrit simultanément une valeur et espère que son engagement moral s'imposera face au monde incarné par les opportunistes. La valeur a vocation ici à changer, c'est-à-dire à réviser, le réel. D'autre part, lorsque le contributeur est confronté à un nombre important de profiteurs, la valeur de coopération qu'il défend se confronte inéluctablement au principe de réalité : moins les contributions du groupe au bien public sont élevées, plus il est dans l'intérêt du joueur de renoncer à la coopération. Les émotions du contributeur (dépit, contrariété, déception, indignation) l'incitent à réviser son ordre de priorités. En ce sens, la baisse progressive et prononcée de la coopération dans un jeu répété ne proviendrait pas d'un effet d'apprentissage (du comportement de passager clandestin) mais davantage de tentatives avortées de coopération (Andreoni, 1995). La faiblesse des contributions serait le résultat de la pusillanimité des joueurs. Le réel constitue donc ici un test (de révision) pour la valeur de coopération.

Selon Livet (2002), l'existence d'une hétérogénéité entre les valeurs et la réalité est rendue nécessaire « si l'on veut conserver la puissance de transformation pratique propre aux valeurs ». Les valeurs sont, rappelons-le, symétriques des émotions : dans le jeu, l'opposition du monde fournit au contributeur, via les mécanismes de (double) révision, la possibilité de réaffirmer ses valeurs. Ce mécanisme assure ainsi l'homogénéité du groupe de ceux qui soutiennent le bien public, qui savent que la présence de comportements égoïstes est inéluctable mais qui puisent dans cette adversité la conviction nécessaire à la concrétisation de leurs valeurs. Comme en témoigne les résultats expérimentaux, la (valeur de) coopération n'est cependant pas stable. Son renforcement est en fait rendu possible par le partage d'une émotion collective positive. Ce partage permet en effet d'entretenir la coopération en faisant vivre un autre monde que celui prescrit par l'opportunisme : « retrouver nos émotions chez les autres nous assure que notre monde de valeurs reste bien une réalité psychologique collective » Livet (2002), le partage assure ainsi la stabilité des valeurs et la conviction que

[20] Cette section est inspirée de Livet (2002, chapitre 3).
[21] Dans leur expérience, Gächter et Fehr (1999) évaluent à 30% la proportion des « égoïstes purs » au sein des joueurs.

ces valeurs ont la capacité de changer le monde. Le joueur tire profit de l'acte de contribution volontaire (« *warm-glow effect* ») en affirmant une valeur ; il affiche une préférence résistante face à un destin contraire en partageant avec les autres contributeurs ses émotions.

Au cours des expériences, le partage social des émotions doit cependant réunir certaines conditions. Il implique en particulier que les contributeurs puissent échanger des signes de reconnaissance d'appartenance à un groupe [Gächter et Fehr (1999) ; Harbaugh et Krause (2000), dans le cas d'enfants ou d'adolescents]. Dans les études citées précédemment, l'approbation (ou la désapprobation), la communication, l'interaction sociale, etc., fédèrent d'autant mieux le groupe qu'elles se font sous couvert d'anonymat et qu'elles concernent un groupe d'une taille suffisante [Isaac et al. (1994)]. L'anonymat est en effet un élément clef d'une émotion collective positive puisque les participants à l'expérience ne peuvent savoir si les autres membres ont bien les mêmes objectifs. Lorsque des signes de coopération et de convergence vers un même objectif sont tangibles, la dissipation de cette incertitude est elle-même un facteur d'émotion et de partage d'émotion qui assure la cohésion du groupe[22]. Dans le scénario de désapprobation proposé par Masclet et al. (2003), dans lequel chaque joueur a la possibilité de communiquer un niveau de désapprobation sur le niveau de contribution des autres joueurs, on remarque ainsi que la coopération au sein des groupes mélangés à chaque période du jeu (« *stranger condition* ») est significativement inférieure à celle des groupes qui ne le sont pas (« *partner condition* »). De même, dans le scénario de punition monétaire non stratégique de Fehr et Gächter (2000a), les joueurs d'un même groupe sanctionnent ceux qui s'écartent de la norme de contribution qui semble s'être mise en place. Dans les questionnaires post-expérimentaux [Fehr et Gächter (2000b)], les joueurs motivent la sanction par un ressentiment à l'égard de ceux qui ne respectent pas la norme. En revanche, les joueurs des groupes mélangés à chaque session ne parviennent pas à établir la formation d'un groupe autour d'une norme de coopération. Certains contributeurs, en particulier, punissent les autres joueurs de façon à augmenter le niveau de contribution moyen qu'ils jugent insuffisant. L'émotion a en fait deux effets distincts : elle renforce la conviction des coopérants inconditionnels (via le partage) et incite également les opportunistes à rejoindre le groupe des contributeurs. L'émotion collective fédère donc le groupe des coopérants mais s'adresse également aux opportunistes.

Dans le jeu, les « égoïstes purs » n'ont pas vocation à vivre une émotion collective associée au partage de la valeur qu'ils incarnent, l'individualisme. Ils sont par définition concurrents, ce qui ne laisse aucune place *a priori* au partage. En revanche, ils peuvent être l'objet d'une émotion collective comme c'est le cas dans le scénario de désapprobation : les joueurs qui ne contribuent pas ou peu perçoivent en effet que leur attitude est l'objet d'un jugement moral négatif de la part des autres membres du groupe. Ils peuvent stratégiquement interpréter cette désapprobation punitive comme un indice de représailles futures (qui se traduirait par une baisse des contributions des autres joueurs) et donc augmenter leur contribution pour l'éviter. La désapprobation a cependant surtout la faculté de créer un sentiment de honte qui incite les opportunistes à accepter la norme de coopération suivie par le groupe. Les effets de la honte sont d'autant plus efficaces que le profiteur décèle des indices fiables sur l'opinion des autres en ce qui le concerne et qu'il est lui-même facilement identifiable. Contrairement à ce que nous avons vu précédemment, c'est l'absence d'anonymat [Rege et Telle (2001)] qui favorise les effets de la honte et donc la coopération. Andreoni et Petrie (2005) suggèrent ainsi que l'identification photographique des joueurs

[22] D'un point stratégique, des signes immédiats de coopération permettent également aux joueurs d'anticiper la coopération future des autres membres du groupe.

dissuade les comportements de passager clandestin. Lorsqu'elle est associée à la divulgation d'une information sur les taux de contribution respectifs de chaque joueur du groupe, l'identification a dès lors une influence notable positive sur la contribution moyenne.

La majorité des expériences décrites jusqu'à présent insistent sur les mécanismes incitatifs (punition, désapprobation, interaction sociale) qui permettent d'obtenir une coopération socialement souhaitable entre les joueurs. Ces mécanismes sont efficaces pour modifier les comportements des opportunistes ou des contributeurs intermittents. Mais, sont-ils pour autant suffisants pour inscrire durablement la norme de coopération ? Autrement dit, les incitations suscitent-elles une valeur (coopérative) que les joueurs seraient prêts à défendre en l'absence de ces mécanismes ? A notre connaissance, les protocoles existants montrent l'absence d'un tel apprentissage des valeurs. Une fois levée, la sanction morale n'a plus ainsi la capacité à mobiliser les contributions volontaires dans l'expérience de Masclet et al. (2003). De même, dans Fehr et Gächter (2000a), lorsque la punition monétaire non stratégique est éliminée, les joueurs qui avaient coopéré précédemment, tendent à limiter sévèrement leur contribution. Plus symptomatique encore, Cinyabugama et al. (2005) montrent l'exceptionnelle capacité de la peur de l'exclusion à augmenter les contributions (proches des 100%). Comme le suggèrent les auteurs, ce mécanisme n'instaure pas pour autant une norme établie de coopération : jusqu'à l'avant-dernière période du jeu, la menace d'éviction du groupe se substitue avec efficacité à la confiance réciproque entre les joueurs coopératifs mais tend cependant à l'amoindrir lorsque se présente la période finale (la contribution moyenne chutant à environ 25%). Ces résultats corroborent l'existence d'un comportement de « fin de jeu » que l'on retrouve très majoritairement chez les contributeurs conditionnels [Keser et van Winden (2000)]. Il semble donc que ni la menace ni la sanction ne procurent véritablement l'engagement (au sens de Beauvois et Joule (2006)) nécessaire à l'établissement d'une norme durablement inscrite dans les comportements. Les protocoles existants n'ont pas, semble-t-il, de vertu éducative permettant l'affirmation des valeurs. Van Dijk, Sonnemans et van Winden (2002) montrent pourtant que la participation au jeu du bien public favorise la formation des liens sociaux entre les joueurs (définis comme l'écart entre l'orientation sociale générale du joueur (Ring-test) et celle établie avec leur partenaire à la suite du jeu). Lorsque l'expérience est profitable (c'est-à-dire, lorsque les gains sont importants), les joueurs aident davantage leur partenaire de jeu lors du Ring-test qu'ils ne le feraient pour un joueur anonyme. En revanche, une expérience décevante renforce leur orientation sociale individualiste et pénalise davantage leur partenaire. En un sens, le déroulement du jeu façonne les valeurs des individus dans un sens coopératif lorsque la coopération s'est installée durablement et dans un sens individualiste lorsque le comportement de passager clandestin s'est imposé. L'apprentissage de la norme coopérative semble donc être possible, mais aussi fragile et peut-être aléatoire[23].

IV. Conclusion

Dans notre analyse, nous avons cherché à comprendre comment la prise en compte des émotions pouvait rendre compte de la dynamique des préférences individuelles et collectives dans l'analyse économique. En nous appuyant sur la théorie des émotions de Livet (2002), nous montrons, à partir de deux illustrations distinctes (l'une théorique, l'autre expérimentale) que les émotions possèdent une véritable rationalité. Les émotions sont en particulier utiles

[23] Keser et van Winden (2000) indiquent en effet que la contribution moyenne au cours d'un jeu répété est fortement liée à la contribution lors de la première période. Tout se jouerait donc au début du jeu, notamment parce que plus de 50% des comportements demeurent inchangés.

pour nous révéler (consciemment ou non) nos vraies préférences ou les valeurs morales que nous sommes prêts à défendre. Contrairement à ce que propose, à notre connaissance, la théorie académique, l'émotion aurait ainsi vocation à intégrer l'analyse non pas uniquement sous la forme d'un argument de la fonction d'utilité mais aussi en tant qu'information (ou signal) permettant d'ajuster nos préférences. La modélisation de l'émotion sous cette forme dépasse naturellement le cadre limité de cet article. Elle est également, de notre point, sans doute prématurée compte tenu de la complexité des mécanismes émotionnels mis en jeu. En revanche, il nous semble envisageable de tester expérimentalement certaines implications de la théorie des émotions morales. L'une des hypothèses de la théorie est en effet que les valeurs révélées par nos émotions sont *a priori* (plus) stables dans le temps, résistantes à une réalité contraire imposée par le monde et donc propices à l'action. Or, dans les travaux expérimentaux, nous avons vu justement que les mécanismes incitatifs, qui sont suffisants pour modifier les comportements, n'inculquent pas pour autant de vraies valeurs. Dans la continuité de ce travail de recherche, la définition d'un protocole expérimental, par exemple dans un jeu de l'ultimatum ou du bien public, pourrait ainsi mettre en évidence les liens entre les réactions émotionnelles des sujets et leur développement moral.

Bibliographie

Akerlof G.A., 1991, Procrastination and Obedience, *American Economic Review* (Papers and Proceedings), vol. 81, n°2.

Anderson S.P., Goeree J.K. et Holt C.A., 1998, A theoretical analysis of altruism and decision error in public goods games, *Journal of Public Economics*, vol. 70, n°2.

Andreoni J., 1995, Cooperation in Public-Goods Experiments: Kindness or Confusion?, *American Economic Review*, vol. 85, n°4.

Andreoni J. et Petrie R., 2004, Public goods experiment without confidentiality: a glimpse into fund-raising, *Journal of Public Economics*, vol. 88, n°7-8.

Axelrod R., 1992, *Donnant Donnant – La théorie du comportement coopératif*, O. Jacob.

Becker G.S., 1996, *Accounting for tastes*, Cambridge, MA: Harvard University Press.

Beauvois J-L. et Joule R-V., 2006, *La soumission librement consentie*, P.U.F.

Blass T., 1999, *Obedience to Authority: Current Perspectives on the Milgram Paradigm*, Mahwah, Lawrence Erlbaum.

Burley P.M. et McGuinness J., 1977, Effects of social intelligence on the Milgram paradigm, *Psychological Reports*, vol. 40.

Bolton G.E. et Ockenfels A., 2000, ERC: A theory of equity, reciprocity and competition, *American Economic Review*, vol. 90, n°1.

Bowles S. et Gintis H., 2001, The Economics of Shame and Punishment, *Working Paper*, University of Massachussetts.

Cinyabugama M., Page T. et Putterman L., 2005, Cooperation under the threat of expulsion in a public goods experiment, *Journal of Public Economics*, vol. 89, n°8.

Damasio A.R., 1995, *L'erreur de Descartes : La raison des émotions*, Sciences, O. Jacob.

Davidson D., 1991 [1982], *Paradoxes de l'irrationalité*, Combas : Editions de l'éclat.

Dawes D.D., van de Kragt A. et Orbell J., 1988, Not Me or Thee but We: The importance of Group Identity in Eliciting Cooperation in Dilemma Situations, *Acta Psychologica*, vol. 68.

De Souza R., 1987, *The rationality of emotions*, Cambridge (Mass.), The MIT Press.

Dijk van F., Sonnemans J. et van Winden F., 2002, Social Ties in a public good experiment, *Journal of Public Economics*, vol. 85, n°2.

Elster J., 1982, *Ulysses and the Sirens*, New York: Cambridge University Press.

Elster J., 1998, Emotions and Economic Theory, *Journal of Economic Literature*, vol. 36, n°1.

Fehr E. et Gächter S., 2000a, Cooperation and Punishment in Public Good Experiments, *American Economic Review*, vol. 90, n°4.

Fehr E. et Gächter S., 2000b, Fairness and Retaliation: The Economics of Reciprocity, *The Journal of Economic Perspectives*, vol. 14, n°3.

Fehr E. et Schmidt K.M., 1999, A theory of Fairness, Competition, and Cooperation, *The Quarterly Journal of Economics*, vol. 114, n°3.

Festinger L., 1957, *A theory of cognitive dissonance*, Standford, CA: Standford U. Press.

Fingarette H., 1998, Self-Deception Needs No Explaining, *The Philosophical Quarterly*, vol. 48, n° 192.

Frank R.H., 1988, *Passions within reason: The Strategic role of emotions*, N.Y.: Norton.

Freedman J.L., 1969, Role playing: Psychology by consensus, *Journal of Personality and Social Psychology*, vol. 13.

Gächter S. et Fehr E., 1999, Collective action as a social exchange, *Journal of Economic Behavior and Organization*, vol. 39, n°4.

Harbaugh W.T. et Krause K., 2000, Children's altruism in public good and dictator experiments, *Economic Inquiry*, vol. 38, n°1.

Hirshleifer J., 1987, On the emotions as guarantors of threats and promises, dans Dupré J., (eds.), *The Latest on the Best: Essays on Evolution and Optimality*, MIT Press, Cambridge.

Holländer H., 1990, A Social Exchange Approach to Voluntary Cooperation, *American Economic Review*, vol. 80.

Isaac M.R., Walker J. et Williams A.W., 1994, Group size and voluntary provision of public goods – Experimental evidence utilizing large groups, *Journal of Public Economics*, vol. 54, n°1.

Kandel E. et Lazear E.P., 1992, Peer Pressure and Partnership, *Journal of Political Economy*, vol. 100, n°4.

Keser C. et van Winden F., 2000, Conditional cooperation and voluntary contributions to public goods, *Scandinavian Journal of Economics*, vol. 102, n°.

Kohlberg L., 1969, Stage and sequence: The cognitive-developmental approach to socialization, dans Goslin D.A. (Ed.), *Handbook of socialization theory and research*, p. 347-480, Chicago: Rand-McNally.

Ledyard J.O., 1995, Publics Goods: A Survey of Experimental Research, dans Kagel J. et Roth A., eds., *Handbook of experimental economics*, Princeton, NJ: Princeton U. Press.

Lewin S. B., 1996, Economics and Psychology: Lessons For Our Own Day From the Early Twentieth Century, *Journal of Economic Literature*, vol. 34, n°3.

Livet P., 2002, *Emotions et rationalité morale*, Sociologies, PUF.

Maughan M.R.C. et Higbee K.L., 1981, Effects of subjects' incentives for participation on estimated compliance for self and others, *Psychological Reports*, vol. 49.

Masclet D., Noussair C., Tucker S. et Villeval M-C., 2003, Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism, *American Economic Review*, vol. 93, n°1.

Meidinger C., 2000, Vertus artificielles et règles de justice chez Hume : une solution au dilemme du prisonnier en termes de sentiments moraux, *Revue de Philosophie Economique*, vol. 1, n°1.

Milgram S., 1974, *Soumission à l'autorité*, Calmann-Levy.

Mixon D., 1972, Instead of Deception, *Journal for the Theory of Social Behaviour*, vol. 2.

Myers D.G. (adapté par Guéguen N.), 2006, *Psychologie sociale pour managers*, Dunod.

O'Donoghue T. et Rabin M., 1999, Doing it now or later, *American Economic Review*, vol. 89, n°1.

Ogien R., 2002, *La honte est-elle immorale ?*, Le temps d'une question, Bayard.

Palfrey T.R. et Prisbey J.E., 1997, Anomalous Behavior in Public Goods Experiments: How Much and Why?, *American Economic Review*, vol. 87, n°5.

Palfrey T.R. et Rosenthal H., 1991, Testing for effects of a cheap talk in a public good game with private information, *Games and Economic Behavior*, vol. 3, n°2.

Park S-O., 2000, Warm-glow versus cold-prickle: a further experimental study of framing effects on free-riding, *Journal of Economic Behavior & Organization*, vol. 43, n°4.

Rabin M., 1993, Incorporating Fairness into Game Theory and Econometrics, *American Economic Review*, vol. 83, n°5.

Rege M. et Telle K., 2001, An Experimental Investigation of Social Norms, *Working Paper*, Case Western Reserve University.

Rochat F., et Modigliani A., 1995, The role of interaction sequences and the timing of resistance in shaping obedience and defiance to authority, *Journal of Social Issues*, vol. 51, n°3.

Rochat F., Maggioni O. et Modigliani A., 1999, The Dynamics of Obeying and Opposing Authority: A Mathematical Model, dans *Obedience to Authority: Current Perspectives on the Milgram Paradigm*, Blass T., chapitre 10.

Serra. D., 2007, Sentiments moraux et économie expérimentale, à paraître dans Livet P. et Leroux A. (eds.), *Leçons de Philosophie Economique*, tome III, De Boeck Université.

Simon H.A., 1967, Motivational and emotional controls of cognition, *Psychological Review*, vol. 74, n°1.

Stigler G.J. et Becker G.S., 1977, De Gustibus Non Est Disputandum, *American Economic Review*, vol. 67, n°2.

Smith A., 1999 [1790], *Théorie des Sentiments Moraux*, Paris : P.U.F.

Wolfesperger A., 1999, Sur l'existence d'une solution kantienne du problème des biens publics, *Revue Economique*, vol. 50, n° 4.

Cahiers du GREThA
Working papers of GREThA

GREThA UMR CNRS 5113

Université Montesquieu Bordeaux IV
Avenue Léon Duguit
33608 PESSAC - FRANCE
Tel : +33 (0)5.56.84.25.75
Fax : +33 (0)5.56.84.86.47

www.gretha.fr

Cahiers du GREThA (derniers numéros)

2007-18 : DOUAI Ali, *Wealth, Well-being and Value(s): A Proposition of Structuring Concepts for a (real) Transdisciplinary Dialogue within Ecological Economics*
2007-19 : AYADI Mohamed, RAHMOUNI Mohieddine, YILDIZOGLU Murat, *Sectoral patterns of innovation in a developing country: The Tunisian case*
2007-20 : BONIN Hubert, *French investment banking at Belle Epoque: the legacy of the 19th century Haute Banque*
2007-21 : GONDARD-DELCROIX Claire, *Une étude régionalisée des dynamiques de pauvreté Régularités et spécificités au sein du milieu rural malgache*
2007-22 : BONIN Hubert, *Jacques Laffitte banquier d'affaires sans créer de modèle de banque d'affaires (des années 1810 aux années 1840)*
2008-01 : BERR Eric, *Keynes and the Post Keynesians on Sustainable Development*
2008-02 : NICET-CHENAF Dalila, *Les accords de Barcelone permettent- ils une convergence de l'économie marocaine ?*
2008-03 : CORIS Marie, *The Coordination Issues of Relocations? How Proximity Still Matters in Location of Software Development Activities*
2008-04 : BERR Eric, *Quel développement pour le 21ème siècle ? Réflexions autour du concept de soutenabilité du développement*
2008-05 : DUPUY Claude, LAVIGNE Stéphanie, *Investment behaviors of the key actors in capitalism : when geography matters*
2008-06 : MOYES Patrick, *La mesure de la pauvreté en économie*
2008-07 : POUYANNE Guillaume, *Théorie économique de l'urbanisation discontinue*
2008-08 : LACOUR Claude, PUISSANT Sylvette, *Medium-Sized Cities and the Dynamics of Creative Services*
2008-09 : BERTIN Alexandre, *L'approche par les capabilités d'Amartya Sen, Une voie nouvelle pour le socialisme libéral*
2008-10 : CHAOUCH Mohamed, GANNOUN Ali, SARACCO Jérôme, *Conditional Spatial Quantile: Characterization and Nonparametric Estimation*
2008-11 : PETIT Emmanuel, *Dynamique des préférences et valeurs morales : une contribution de la théorie des émotions à l'analyse économique*

La coordination scientifique des Cahiers du GREThA est assurée par Sylvie FERRARI et Vincent FRIGANT. La mise en page est assurée par Dominique REBOLLO.