Préferences psychologiques et nouvelle économie politique

Antoine Billot, Chantal Marlats

PARIS SCHOOL OF ECONOMICS
ÉCOLE D'ÉCONOMIE DE PARIS

WORKING PAPER N° 2009 - 04

Préférences psychologiques et Nouvelle Economie Politique

Antoine Billot

Chantal Marlats

Codes JEL : D01, D03, C7
Mots-clés : Préférence, altruisme, comportement

PARIS-JOURDAN SCIENCES ECONOMIQUES
LABORATOIRE D'ECONOMIE APPLIQUÉE - INRA

48, BD JOURDAN – E.N.S. – 75014 PARIS
TÉL. : 33(0) 1 43 13 63 00 – FAX : 33 (0) 1 43 13 63 10
www.pse.ens.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE – ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES
ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES – ÉCOLE NORMALE SUPÉRIEURE

Preferences Psychologiques et Nouvelle Economie Politique*

Antoine Billot† et **Chantal Marlats**‡

February 3, 2009

1 Introduction

L'intuition première de la **T**héorie des **P**références **P**sychologiques (**TPP**) est qu'il existe des déterminants du comportement individuel qui puisent leur source dans la perception et l'appréhension psychologique d'autrui - "autrui" désignant ici l'ensemble des partenaires sociaux, qu'ils soient des adversaires (en cas de conflit) aussi bien que des candidats à une éventuelle coordination (en cas d'entente)... Cette dépendance, agressive ou bienveillante, outre qu'elle révèle la présence d'une disposition empathique qui rappelle celle à l'oeuvre dans les modèles épistémiques à la Aumann et que fonde le raisonnement spéculaire[1], traduit de manière plus générale l'influence du contexte social sur les choix individuels, l'empreinte de l'environnement sur chacune des décisions que prennent les individus. En d'autres termes, non seulement Robinson ne se comporte pas de la même façon selon qu'il se croit seul sur l'île Désespoir ou qu'il noue commerce avec Vendredi (cela, la théorie standard des jeux le disait déjà) mais, plus encore, il adapte son propre comportement au gré des sentiments qu'il nourrit à l'égard de Vendredi - lesquels dépendent largement de sa conception personnelle de la solidarité, de la justice voire de ce qu'il comprend (ou veut comprendre) des intentions de Vendredi... Libre de toute considération anthropologique et/ou psychanalytique, la **TPP** propose en réalité une re-présentation de la plupart des concepts de la théorie des préférences à la lumière de l'hypothèse selon laquelle l'agent économique est un "animal" social doué de sentiments, l'héritier d'une culture, intime et historique, qui a façonné son rapport au monde et qui, en conséquence, conçoit sa relation à l'autre de manière essentiellement psychologique, qu'il soit empathique ou agressif, reconnaissant ou défiant...

*... laquelle n'a rien à voir avec la célèbre NEP que Lénine recommanda d'adopter à partir de 1921 et qui eut pour conséquence un (très relatif et très provisoire) assouplissement de la conception bolchévique des mécanismes économiques...

†Université Panthéon-Assas (Paris 2) et PSE-Jourdan (Ecole d'économie de Paris)

‡Université Panthéon-Sorbonne (Paris 1) et CES

[1] Dupuy (1989) appelle *spécularité* l'acte mental par lequel un agent se met à la place d'un autre afin d'expérimenter virtuellement son raisonnement. Il s'agit d'une sorte d'empathie stratégique propre à des situations d'entente ou à l'inverse de conflit. Chaque joueur doit être capable de reproduire (d'imiter ?) la logique de la pensée de son adversaire de façon à se coordonner avec lui au mieux de l'intérêt collectif (entente) ou de façon à neutraliser ses calculs et ses ruses en les anticipant le plus efficacement possible (conflit) : en somme, contrefaire pour faire ou défaire.

1.1 Motivations Théoriques

1.1.1 Les déterminants nouveaux du comportement mis en avant par la **TPP** renvoient à la psychologie profonde des individus, à leur mode d'évaluation de l'environnement, c'est-à-dire finalement à leur perception du réseau des interactions sociales au sein duquel ils sont amenés à prendre des décisions. Au delà donc de l'habituelle exigence de rationalité micro-économique conçue en tant que dispositif méthodologique (*methodological device*) - laquelle correspond fondamentalement à une véritable avidité individuelle (*greediness* et *selfishness*[2]) - nombre d'expériences et d'observations empiriques semblent indiquer qu'une telle exigence normative, pour commode qu'elle soit dans le langage de la théorie (celui du moins de la théorie standard), n'en reste pas moins incapable *per se* de capturer d'autres dimensions du choix individuel, sans doute plus sociales qu'économiques, plus émotionnelles que rationnelles, d'autres mécanismes à l'oeuvre dans le processus d'optimisation que cette seule avidité. Bien que motivée par la volonté de rationaliser certains de ces résultats expérimentaux et de ces faits stylisés qui contredisent apparemment les prédictions de la théorie standard, la **TPP** n'est pas à proprement parler une remise en cause de la théorie des préférences standard - *i.e.* celle qui postule que les préférences sont *selfish* et le comportement maximisateur - (au sens où, par exemple, elle proposerait une nouvelle norme de rationalité) mais bien plutôt une extension psychologique, c'est-à-dire une approche capable d'intégrer de nouvelles motivations humaines (en particulier, les différentes formes que peuvent prendre respectivement l'altruisme, le sens de l'équité ou l'aversion à l'inégalité...) dans la liste (pour l'heure, assez étroite) des facteurs déterminants du comportement économique, et partant, d'enrichir la définition habituelle d'une fonction d'utilité.

1.1.2 La tentation est grande d'opposer les deux approches, standard *vs* psychologique, et de les cantonner à l'intérieur d'une sorte de conflit plus idéologique que réellement scientifique entre *selfishness* et altruisme, choix mécanique et choix psychologique - l'évidente non neutralité des termes ne manquant pas alors de donner l'avantage à la seconde sur la première au nom de cette interdépendance (généreuse et solidaire plutôt qu'agressive et prédatrice) dont d'aucuns regrettent qu'elle soit souvent ignorée par la réflexion économique et toujours absente des modèles censés l'alimenter. Pourtant, loin de n'être que concurrentes et antagonistes, ces deux approches étayent plutôt deux facettes d'une même réalité, deux aspects d'une même conception, celle du contrat social[3] par lequel se lient les agents, implicitement ou explicitement, dès lors qu'ils décident et agissent au sein d'un même univers (économique, politique, stratégique...). L'ambition de notre contribution pourrait être ici entendue en tant que tentative de délimitation des champs respectifs de pertinence de ces deux approches - l'intuition étant que l'altruisme d'origine psychologique n'a de réelle conséquence qu'à la condition que le poids social de chaque individu soit suffisant pour contrarier l'atomicité traditionnelle qui le condamne à n'exercer aucune sorte d'influence au niveau global[4].

[2] On conservera ici certains mots anglais, *selfish*, *fairness equilibrium*..., lorsque la définition conceptuelle est fondamentalement plus restrictive (plus précise) que la traduction ne risquerait de le faire croire.

[3] ... entendu comme un agrément de volontés faisant naître des obligations entre elles...

[4] Ce que résument Fehr et Schmidt (2003) en écrivant: *These models explain why in some*

1.2 Motivations Empiriques

Les deux concepts les plus féconds que l'économie comportementale a mis en évidence après de nombreuses expérimentations et qui nourrissent la **TPP** (voir Fehr et Schmidt 2003, Sobel 2005) sont incontestablement ceux d'"équité" (Kahneman, Knetsch et Thaler 1986, Bewley 1999) et de "réciprocité" (Segal et Sobel 2007a,b).

1.2.1 (Equité) Le désir d'équité traduit une sorte de sensibilité individuelle à ce que l'on pourrait appeler "la justice (ou l'injustice) spontanée"[5]. Il est de multiples circonstances où le désir d'équité se manifeste par delà l'expérience intime de l'injustice: ainsi de l'épouse qu'accable la gestion quotidienne des responsabilités familiales auxquelles son conjoint se dérobe, ainsi d'enfants d'une même fratrie qui constatent une différence notable de traitement entre eux, ainsi de collègues de travail dont les trajectoires de carrière demeurent irréversiblement disjointes quoique leurs mérites objectifs soient identiques, ainsi encore d'étudiants n'affrontant pas la même difficulté à se faire embaucher au sortir de l'université bien que leurs cursus soient semblables... Tous ont en commun de ressentir intuitivement (pour s'en féliciter éventuellement mais surtout pour le regretter) que leur situation n'est pas "éthiquement" juste. Le ressort de cette sensibilité est de même nature que celui qui affecte les agents en matière d'incertitude: la dimension psychologique est ici fondamentale (*i.e.* exogène au modèle de comportement en ce que prédéterminée par des facteurs non nécessairement économiques mais sociaux, religieux ou culturels...) et les préférences se font alors l'instrument naturel de la capture théorique de cette sensibilité - comme elles le font au sein de la théorie de la décision pour le risque ou l'incertitude.

1.2.1.1 (Le jeu de l'ultimatum) Considérons l'expérience suivante qui met en scène deux individus i et j. On se propose de remettre $\mathbf{10}$ euros à l'individu i à condition que celui-ci se désaisisse d'une partie de la somme pour la donner à j. Cette indemnité $\mathbf{x}$ versée par i à j doit être entière et comprise entre $\mathbf{0}$ et $\mathbf{10}$ euros. L'agent j peut alors accepter l'indemnité $\mathbf{x}$ que lui propose i et le marché est alors conclu sur la base d'un partage du type $(\mathbf{10}-\mathbf{x},\mathbf{x})$ ou la refuser et les deux individus i et j repartent alors les mains vides, *i.e.* l'issue est $(\mathbf{0},\mathbf{0})$. Le choix rationnel est logiquement le suivant: comme i et j sont censés être de parfaits agents néo-classiques standard (*selfish*), i cherche à maximiser son seul gain tandis que j est fondamentalement indifférent entre $(\mathbf{10},\mathbf{0})$ et $(\mathbf{0},\mathbf{0})$ (puisque dans les deux cas, il ne gagne rien) et l'issue théorique est alors $(\mathbf{10},\mathbf{0})$. Mais l'expérience infirme largement la théorie. En pratique, c'est-à-dire lors des expériences effectuées en "laboratoire", i donne en général $\mathbf{4}$ euros à j. La raison en est intuitivement assez simple: i sait que s'il ne donne rien à j, j considèrera que i n'a pas tenu compte du pouvoir de nuisance qu'il peut exercer en refusant tout partage qu'il estimera injuste. Aussi, i choisit-il un partage objectivement plus juste, comme $(\mathbf{6},\mathbf{4})$ voire $(\mathbf{5},\mathbf{5})$. Toutefois, à première vue, *i.e.* à la lumière crue de la théorie *selfish*, la décision de j est aussi irrationnelle lorsqu'il refuse $(\mathbf{10},\mathbf{0})$ que celle de i quand il propose finalement $(\mathbf{6},\mathbf{4})$.

strategic settings almost all people behave as if they are completely selfish, while in others the same people will behave as if they are driven by fairness.

[5] "Spontané" signifiant ici "qui ne procède pas d'une vision théorique" *a priori* ni d'un raisonnement mais bien plutôt d'une conviction culturelle ou d'une croyance.

1.2.2 (Réciprocité) La réciprocité - ce que, du moins, les économistes entendent introduire dans leur discours sous ce terme[6] - est un phénomène social qui procède de l'existence d'une relation d'un premier terme à un second terme telle que cette relation a la capacité de s'inverser du second terme au premier. La réciprocité est ici une relation symétrique entre deux agents reposant sur le postulat (illustré par de nombreuses expériences) selon lequel l'*homo-oeconomicus* n'agit pas pour satisfaire son seul intérêt matériel (comme le suppose la théorie standard) mais parvient le plus souvent à le transcender pour considérer aussi bien l'intérêt des autres membres du groupe auquel il appartient dès lors qu'il sait (qu'il croit) que les autres membres de ce groupe considèrent symétriquement son intérêt lorsqu'ils déterminent leurs propres choix. Plus spécifiquement, l'économie comportementale a récemment mis en évidence des comportements traduisant un "désir de réciprocité" qui se distinguent de ceux rencontrés dans les modèles économiques standard (Fudenberg et Maskin 1986). Les termes de "réciprocité forte" et de "réciprocité faible" sont parfois utilisés pour distinguer la réciprocité envisagée par la **TPP** de la réciprocité admise par la théorie standard. Il s'agit toutefois, dans les deux cas, d'une prédisposition à coopérer mais aussi à punir (ou à récompenser) les agents qui se détournent du (ou qui adhèrent au) principe même de la coopération. Toutefois, lorsque les comportements sont empreints de réciprocité forte, ils induisent des coûts dont il est raisonnable de penser qu'ils ne seront pas "couverts" dans le futur, des coûts tels qu'un agent purement *selfish* ne choisit donc jamais de s'y conformer. En outre, à l'inverse de ce qui advient dans les modèles standard, la réciprocité forte peut demeurer effective quoique affaiblie dans des situations à la fois anonymes et non-répétées (voir, par exemple, §**1.2.2.2** *infra*).

1.2.2.1 (Le jeu de l'échange de don) Supposons un employeur i (*i.e.* le principal) qui propose une rémunération w à son employé j (*i.e.* l'agent) en échange d'un certain niveau d'effort e. L'agent j peut accepter ou refuser l'offre du principal i et, en conséquence, décider de retourner sur le marché du travail. Le niveau d'effort demandé $\hat{e}$ et le niveau d'effort choisi e appartiennent à l'ensemble $\{\mathbf{0}, \mathbf{0.1}, \mathbf{0.2}, \ldots, \mathbf{1}\}$ et le salaire w qui est alors versé à j appartient à l'ensemble $\{\mathbf{0}, \mathbf{1}, \ldots, \mathbf{100}\}$. A chaque niveau d'effort e est associé un coût $c(e)$ qui est croissant avec e et supposé convexe. Les gains sont donc respectivement $\mathbf{100}.e - w$ pour le principal i et $w - c(e)$ pour l'agent j. L'issue en cas d'échange est formée du couple $[\mathbf{100}.e - w, w - c(e)]$. A l'inverse, si l'échange n'a pas lieu, alors i comme j gagnent $\mathbf{0}$. Si l'on analyse cette situation dans le cadre d'un jeu standard, il apparaît que le principal a toujours intérêt à offrir le salaire le plus bas, *i.e.* $\mathbf{10}$, et l'agent à fournir le niveau d'effort le plus faible, *i.e.* $\mathbf{0.1}$, sachant que l'agent ne rejette jamais l'offre du principal. En effet, i sait que, quelle que soit la rémunération w, le niveau d'effort qui maximise la fonction d'utilité de j est le minimum des w possibles. Afin de maximiser sa propre utilité, i propose donc le minimum des salaires possibles. Toutefois, les résultats expérimentaux ont tendance à montrer que le niveau d'effort demandé $\hat{e}$ croît avec la rémunération versée et qu'en outre, la rente du principal est d'autant plus importante que le niveau d'effort demandé $\hat{e}$ est plus élevé (en d'autre termes, le niveau d'effort fourni par l'agent augmente significativement avec le salaire).

[6] L'ethnologie utilisait déjà - longtemps avant que l'économie s'en saisisse - le terme de réciprocité afin de désigner l'ensemble des prestations économiques mais aussi symboliques qui caractérisent l'univers praxéologique des sociétés traditionnelles.

Les employés tendraient donc, en moyenne, à répondre de manière réciproque à l'offre des employeurs. (Cependant, il convient d'ajouter que l'effort effectif est inférieur à l'effort désiré et qu'un certain nombre d'individus se comportent conformément aux modèles standard.)

1.2.2.2 (Le jeu de la contribution à un bien public) Ce jeu est particulièrement instructif quant à la mise en évidence (et, partant, la résolution) du conflit entre intérêt individuel et intérêt collectif, compétition et coopération (voir Camerer 2003). Il existe dans la littérature deux versions emblématiques de ce jeu. La comparaison de ces deux variantes permet d'appréhender le rôle particulier des motivations non directement pécuniaires (*i.e.* psychologiques) dans le comportement des joueurs. Considérons n individus i qui choisissent simultanément leur niveau de contribution g_i à la fourniture d'un bien public quelconque et qui reçoivent en retour un payoff $x_i = y_i - g_i + m \sum_{j=1}^{n} g_j$ (où y_i est la dotation initiale du joueur i et m le paiement monétaire par unité de bien public tel que $m < \mathbf{1} < n \times m$). Dans un tel contexte, les joueurs *selfish* ont immédiatement une stratégie dominante à leur disposition - laquelle consiste à ne contribuer en rien à la fourniture du bien public. La prédiction théorique (standard) de ce jeu est alors qu'à l'équilibre aucun joueur n'accepte de contribuer de manière positive (*i.e.* $g_i = \mathbf{0}$, pour tout i et $x_i = y_i$). Expérimentalement, en revanche, les choses ne sont pas aussi simples. Pratiquement, on répète ce jeu une dizaine de fois et, à chaque période, les joueurs se trouvent affectés de manière aléatoire à une coalition donnée. Durant les premières phases du jeu, on remarque qu'ils consacrent de **40** à **60**% de leurs dotations à la fourniture du bien public tandis qu'à la dernière période, environ **75**% d'entre eux refusent de contribuer à quelque hauteur que ce soit et les **25**% restants choisissent de ne consacrer qu'une très faible partie de leurs dotations. Ce phénomène peut certes s'expliquer *via* un argument standard de simple répétition du jeu (et conséquemment d'apprentissage de la rationalité *selfish*). Cependant, on peut aussi bien avancer une autre explication, à savoir que la plupart des joueurs sont enclins à contribuer significativement dès qu'ils anticipent que les autres joueurs contribueront eux-aussi de manière significative. Mais, en contrepartie, dès lors qu'ils observent des comportements déviants ("anti-sociaux", pour faire bref) de type passager clandestin - *i.e.* certains agents profitent de la coopération sans pour autant contribuer à la fourniture du bien public -, ils révisent aussitôt leurs croyances à l'endroit du niveau d'implication sociale des autres joueurs et choisissent donc, à leur tour, de ne pas coopérer. Le raisonnement fondé sur la réciprocité conditionnelle des comportements individuels peut ainsi être assimilé à celui qui est à l'oeuvre dans un jeu de bien public "avec punition". Lors d'une première étape, les agents jouent ici dans le cadre standard d'un modèle de contribution à la fourniture d'un bien public - tel que présenté ci-dessus. Puis, ils observent le niveau de contribution des membres de leur groupe. Ensuite, dans une seconde étape, ils décident (ou non) d'infliger une punition à certains de leurs partenaires, moyennant un coût - lequel est défini de façon telle que le joueur *selfish* est dissuadé d'agir de façon "antisociale". D'un point de vue théorique, il s'agit là, toutefois, d'une menace non crédible et, à l'équilibre, aucun joueur ne participe positivement à la fourniture du bien public... Pourtant, les résultats expérimentaux montrent qu'à la deuxième période les agents contribuent généralement à la hauteur de **75**% de leurs dotations (comme dans la variante précédente du jeu, *i.e.* sans punition) et même que

plusieurs parmi eux décident effectivement d'infliger une punition à certains de leurs partenaires. Une interprétation de ce résultat pour le moins déroutant est la suivante: les joueurs sensibles à la réciprocité (*ex-ante* et *ex-post*) se manifestent en tant qu'ils sont sensibles à la réciprocité grâce à la possibilité qui leur est offerte de punir les agents *selfish* et, ainsi, de les amener à coopérer...

1.3 Plan

Dans une première partie, nous présentons la **TPP** à travers, d'une part, l'axiomatique proposée par Sandbu (2008) pour les décisions individuelles pures et, d'autre part, celle de Segal et Sobel (2007a) pour les décisions stratégiques. Dans une seconde partie, nous essayons de caractériser l'apport potentiel de cette littérature (*i.e.* celle consacrée aux préférences psychologiques) à la définition d'une économie politique nouvelle et nous cherchons à délimiter le champ pertinent d'investigation d'une telle approche qui combinerait à la fois les exigences "micro" que véhiculent les préférences psychologiques avec l'objet plus "macro" de l'économie politique (comprise comme la branche de la science économique qui étudie les conséquences de l'intervention d'un décideur public et les conditions optimales de l'action collective). En conclusion, nous défendons la thèse selon laquelle l'étude de l'action collective au niveau particulier des "communautés" ou des systèmes dits "polycentriques" peut *a priori* profiter des résultats abondants produits par la **TPP** - tout autre niveau d'investigation semblant *a contrario* inadapté en l'état actuel des développements de cette théorie.

2 La Théorie des Préférences Psychologiques

A la lumière de ces nombreuses observations expérimentales éminemment paradoxales, plusieurs approches ont émergé qui proposent une alternative théorique au modèle standard de maximisation *selfish*. La plupart d'entre elles ont pour vocation première de décrire et d'organiser les résultats produits par ces expériences. Leur dénominateur commun n'est donc pas de remettre en cause la rationalité supposée des agents mais bien plutôt d'admettre une définition moins contraignante de leurs préférences et donc des fonctions d'utilité qui les représentent, c'est-à-dire une description plus riche de leurs motivations, des mobiles de leur comportement, *i.e.* de leur psychologie. Certes, la **TPP** est récente et souffre incontestablement d'un certain manque d'unité. C'est en effet la grande variété des hypothèses introduites en matière de motivation individuelle qui semble la caractériser de prime abord. Aussi, afin de distinguer les différentes approches qui en constituent le coeur et de cerner les implications précises de chacune des hypothèses spécifiques qui les fondent, nous présentons d'abord le cadre axiomatique proposé par Sandbu (2008) avant de commenter celui de Segal et Sobel (2007a). Celui de Sandbu (2008), de par son homogénéité, est une première réponse à la suspiscion d'incohérence qui flotte autour de la **TPP**. Il nous permet d'identifier en amont les propriétés générales de la **TPP** et facilite en aval la comparaison interne de ses différentes composantes. Toutefois, les modèles d'"identité" à la Akerlof et Kranton (2000), comme ceux de "motivation intrinsèque et extrinsèque" à la Bénabou et Tirole (2003), postulent l'existence de préférences qui ne requièrent pas de s'écarter significativement, ni de manière conceptuelle ni de manière formelle, de la théorie standard (à

l'inverse de la **TPP**). Quand bien même les relations de préférence décrites par Akerlof et Kranton (2000) ou Bénabou et Tirole (2003) ont vocation à intégrer une importante dimension psychologique (et, partant, à prendre en compte des effets jusque là négligés tels que la confiance en soi - *self-esteem* -, la participation à des activités charitatives, le rôle désincitatif des récompenses monétaires...), elles sont toutefois capables d'assimiler ces phénomènes sans rien abandonner du cadre micro-économique traditionnel. Ici, l'innovation théorique consiste donc plutôt à considérer les croyances à propos des différents types possibles d'agent - du plus *selfish* au plus altruiste - comme un bien de consommation indirect et à introduire le niveau d'investissement dans ce bien au coeur même des préférences individuelles *via* un paramètre affectant la fonction d'utilité qui les représente.

2.1 Le Cadre Général et les Axiomes de la TPP

La présentation des axiomes suit ici la division usuelle d'entre les préférences psychologiques définies sur les distributions, et qui concernent les comportements altruistes ainsi que ceux manifestant une aversion à l'inégalité (Sandbu 2008), et celles qui traitent de l'altruisme conditionnel et de la réciprocité intrinsèque (Segal et Sobel 2007a) - quand bien même les dernières peuvent s'interpréter en tant que généralisation des premières. Dans un cas comme dans l'autre, l'agent représenté ne choisit pas une distribution sur la seule base des conséquences matérielles (monétaires) que cette distribution implique à son endroit mais en prenant aussi bien en compte, d'une part, les conséquences matérielles qu'un tel choix ne manquera pas d'avoir sur le bien-être des autres agents et, d'autre part, la charge éthique induite voire l'évaluation de la situation dans lequel s'inscrira son choix - toutes motivations qui participent des ingrédients de la maximisation de sa propre satisfaction.

2.1.1 Le Modèle de la TPP

2.1.1.1 Formellement, la **TPP** se construit de la manière suivante: soit une société $\mathbf{N}$ formé de $n+1$ membres - un décideur i et n autres agents j. Une distribution quelconque $\mathbf{x}$ est un vecteur $(x_i, x_1, ..., x_j, ..., x_n)$ de conséquences matérielles, *i.e.* de payoffs associés aux $n+1$ membres de la société $\mathbf{N}$. Comme les payoffs sont toujours positifs, on considère que $\mathbf{x} \in \mathbb{R}_+^{n+1}$. L'agent i est doté d'une relation de préférence, notée $\succsim_i$, définie sur $\mathbb{R}_+^{n+1} \times \mathbb{R}_+^{n+1}$. Ainsi $\mathbf{x} \succsim_i \mathbf{y}$ doit être lu: l'agent i préfère la distribution $\mathbf{x}$ à la distribution $\mathbf{y}$.

2.1.1.2 La **TPP** s'éloigne de la vision standard des préférences en ce qu'elle offre la possibilité de définir une relation de préférence qui ne soit pas invariante aux payoffs des autres agents. Pour autant, les préférences psychologiques ne sont pas identiques à celles représentées par les fonctions de bien-être social et d'évaluation sociale. En effet, elles s'en distinguent par leur caractère éminemment subjectif: l'évaluation de la distribution des payoffs ne procède pas d'un individu hors la société mais d'un individu à l'intérieur de la société. La question, centrale en théorie du choix social, de savoir s'il est plus cohérent de fonder l'évaluation d'une distribution de payoffs sur les payoffs eux-mêmes ou sur l'utilité qu'ils confèrent aux individus qui les percoivent ne peut plus être examinée sous le même angle. Ce qui importe désormais, c'est bien plutôt la

personnalité psychologique de l'individu évaluateur: que signifie "juste" pour lui? Quel est est son degré d'empathie à l'endroit des autres agents?...

2.1.2 Les Axiomes Généraux

Les axiomes qui suivent sont traditionnels.

Complétude: La relation de préférence $\succsim_i$ est complète si: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$,

$$\mathbf{x} \succsim_i \mathbf{y} \text{ ou } \mathbf{y} \succsim_i \mathbf{x}.$$

Transitivité: La relation de préférence $\succsim_i$ est transitive si: $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}_+^{n+1}$,

$$[\mathbf{x} \succsim_i \mathbf{y} \text{ et } \mathbf{y} \succsim_i \mathbf{z}] \Rightarrow [\mathbf{x} \succsim_i \mathbf{z}].$$

Continuité: La relation de préférence $\succsim_i$ est continue si: $\forall \mathbf{x} \in \mathbb{R}_+^{n+1}$,

$$\left\{\mathbf{y} \in \mathbb{R}_+^{n+1} | \mathbf{y} \succsim_i \mathbf{x}\right\} \text{ et } \left\{\mathbf{y} \in \mathbb{R}_+^{n+1} | \mathbf{y} \succsim_i \mathbf{x}\right\}$$

sont des sous-ensembles fermés.

En résumé, la relation de préférence $\succsim_i$ de l'agent i est un préodre complet et continu.

2.2 Les Préférences D-b

Au sein de la **TPP**, on distingue les approches dites "**D**istribution-**b**ased" (**D-b**) - qui relèvent de la théorie de la décision individuelle - des approches "**C**ontext-**b**ased" (**C-b**) - dont l'essentiel des contributions relève plutôt de la théorie des jeux (voir §**2.3** *infra*). La première conception permet de modéliser des comportements altruistes ou envieux mais aussi bien de rendre compte de motivations plus complexes telles que l'aversion à l'inégalité, l'arbitrage entre efficacité et justice... (Voir sur ce sujet Fehr et Schmidt 1999, Bolton et Ockenfels 2000, Charness et Rabin 2002.) La seconde, quant à elle, peut être vue comme une généralisation de la première en ce qu'elle fait dépendre les propriétés de la relation de préférence psychologique du contexte général (y compris, donc, du comportement des agents qui entourent le décideur).

2.2.1 Les Axiomes D-b

2.2.1.1 La plupart des axiomes spécifiques des préférences **D-b** sont voisins de ceux utilisés traditionnellement dans les axiomatiques du choix en univers risqué (ou incertain) et du choix social. Le premier d'entre eux - la séparabilité forte - est ainsi trés proche dans sa formulation et son esprit du fameux **I.I.A.**, *i.e.* l'axiome d'indépendance des alternatives non-pertinentes d'Arrow (1951). Il requiert que l'agent comparant deux distributions particulières néglige la partie commune de celles-ci. Il implique par ailleurs que si certains payoffs évoluent, alors l'évaluation de cette évolution ne dépend que des payoffs qui ont évolués.

(SF) Séparabilité Forte: Pour tout sous-ensemble d'agents $\mathbf{I} \subseteq \mathbf{N}$ et pour toutes distributions $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ telles que $\mathbf{x_I} = \mathbf{y_I}$,

$$\forall \mathbf{z_I} \in \mathbb{R}_+^{|\mathbf{I}|}: \ [\mathbf{x} \succsim_i \mathbf{y} \Leftrightarrow (\mathbf{z_I}, \mathbf{x_{-I}}) \succsim_i (\mathbf{z_I}, \mathbf{y_{-I}})]$$

où $\mathbf{x_I} = (x_j)_{j\in\mathbf{I}}$, $|\mathbf{I}|$ désigne le cardinal de $\mathbf{I}$ et $-\mathbf{I}$, l'ensemble $\mathbf{N}$ privé des agents j appartenant à $\mathbf{I}$.

2.2.1.2 A moins qu'il ne soit accompagné d'un axiome supplémentaire caractérisant son attitude face aux différences de payoffs, un agent respectant (**SF**) demeure insensible aux inégalités. En réalité, (**SF**) ne dévie guère de l'hypothèse standard. C'est pourquoi il est pertinent de lui adjoindre un nouvel axiome (introduisant par exemple une préférence pour les distributions procédant de transferts à la Pigou) ou bien d'en autoriser une version plus faible. Dans ce dernier cas, l'affaiblissement de l'axiome de séparabilité passe par l'introduction de la notion (classique en théorie des choix en univers risqué comme en théorie de la mesure des inégalités) de "dépendance par rapport au rang": il s'agit de procéder à des comparaisons de distributions qui ne demeurent insensibles aux payoffs non-pertinents qu'à la seule condition que la position relative des agents, en termes de richesse, soit identique dans les deux distributions que l'on compare.

2.2.1.4 Pour toute partition $\mathbf{H} \subseteq \mathbf{N}$ de l'ensemble $\mathbf{N}$, l'ensemble des distributions qui "maintiennent le rang relativement à l'agent i" est défini par:

$$\mathbb{R}_{\mathbf{H}} \equiv \left\{\mathbf{x} \in \mathbb{R}_{+}^{n+1} | x_j \leq x_i \Leftrightarrow j \in \mathbf{H} \subseteq \mathbf{N}\right\}. \tag{1}$$

L'affaiblissement de (**SF**) implique l'invariance relative (l'identité en rang, donc) du payoff des agents non concernés:

(Sf) Séparabilité faible: Pour toutes partitions $\mathbf{H} \subseteq \mathbf{N}$, $\mathbf{I} \subseteq \mathbf{N}$, et toutes distributions $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}_{\mathbf{H}}$ telles que $\mathbf{x_I} = \mathbf{y_I}$, $(\mathbf{z_I}, \mathbf{x_{-I}}) \in \mathbb{R}_{\mathbf{H}}$ et $(\mathbf{z_I}, \mathbf{y_{-I}}) \in \mathbb{R}_{\mathbf{H}}$:

$$\left[\mathbf{x} \succsim_i \mathbf{y} \Leftrightarrow (\mathbf{z_I}, \mathbf{x_{-I}}) \succsim_i (\mathbf{z_I}, \mathbf{y_{-I}})\right].$$

2.2.1.6 L'axiome de neutralité concerne la façon dont l'agent i conçoit les autres membres de la population $\mathbf{N}$. Bien que cet axiome semble raisonnable en situation théorique, il n'en va pas de même en situation expérimentale. En effet, la neutralité implique que l'agent i considère de manière impartiale les agents j qui l'entourent. Autrement dit, l'anonymat est ici la règle. Toute dimension affective, toute forme de favoritisme ou de solidarité reposant sur l'identité précise d'un agent ou d'un groupe d'agents, est *a priori* exclue des motivations qui fondent les choix individuels. Cet axiome, déjà hautement critiquable dans un contexte de choix social, l'est plus encore dans un contexte purement subjectif.

(N) Neutralité: Soit $\sigma : \mathbf{N}\backslash\{i\} \longrightarrow \mathbf{N}\backslash\{i\}$ une permutation, où $\mathbf{N}\backslash\{i\}$ désigne l'ensemble $\mathbf{N}$ privé de l'agent i. Pour toute permutation σ et toutes distributions $\mathbf{x}, \mathbf{y} \in \mathbb{R}_{+}^{n+1}$:

$$\left[\mathbf{x} \succsim_i \mathbf{y} \Leftrightarrow (x_i, x_{\sigma(1)}, ..., x_{\sigma(j)} ..., x_{\sigma(n)}) \succsim_i (y_i, y_{\sigma(1)}, ..., y_{\sigma(j)}, ..., y_{\sigma(n)})\right].$$

2.2.1.7 L'axiome d'homothétie est une hypothèse d'invariance: la multiplication de tous les payoffs par un même réel positif ne modifie pas l'ordre de préférence entre deux distributions. On suppose donc que les agents ne sont sensibles qu'aux seuls payoffs relatifs.

(H) Homothétie: Pour toutes distributions $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ et tout $\lambda > 0$:

$$[\mathbf{x} \succsim_i \mathbf{y} \Leftrightarrow \lambda\mathbf{x} \succsim_i \lambda\mathbf{y}]\,.$$

2.2.1.8 L'axiome d'accroissement minimal stipule que si deux distributions sont uniformes au sens où elles affectent le même payoff à tous les membres de la société y compris l'agent i lui-même, alors l'agent i préfère celle des deux distributions qui est la plus généreuse. Il ressemble clairement à l'axiome de monotonie en théorie du choix social (ou encore à la dominance stochastique du premier ordre en théorie du choix en univers risqué). Toutefois, son interprétation est différente en ceci qu'il préconise fondamentalement de préférer une distribution $\mathbf{x}$ qui domine au sens de Pareto une distribution $\mathbf{y}$ quand bien même $\mathbf{x}$ serait beaucoup plus inégalitaire que $\mathbf{y}$: il signifie donc que comparer deux distributions à l'aide du critère de Pareto n'a de sens que si elles sont parfaitement égalitaires.

(AM) Accroissement Minimal: Soient $\overline{\mathbf{x}}$ et $\overline{\mathbf{y}}$ deux distributions telles que, pour tout agent $i \in \mathbf{N}$, $x_i = \overline{x}$ et $y_i = \overline{y}$:

$$[x_i > y_i \Leftrightarrow \overline{\mathbf{x}} \succ_i \overline{\mathbf{y}}]\,.$$

2.2.2 Deux Théorèmes de Représentation des Préférences D-b

La satisfaction de ces axiomes permet de représenter des préferences psychologiques de type **D-b** par une fonction d'utilité de type **CES**. Les fonctions de type **CES** sont d'usage fréquent en économie. Elles se caractérisent, d'une part, par l'influence de deux paramètres internes - lesquels garantissent une grande flexibilité d'utilisation - et, d'autre part, par la constance de l'élasticité de substitution entre ses arguments - ici, les payoffs. Ceci implique alors qu'une des caractéristiques notables des préférences psychologiques (telles que représentées ci-dessous, **Th. 1**) est l'insensibilité du classement des distributions au niveau de la richesse de l'agent.

Théorème de représentation 1: *Si $n \geq 2$, alors une relation de préférence psychologique $\succsim_i$ de type **D-b** satisfait (**SF**), (**N**), (**H**) et (**AM**) si et seulement si elle peut être représentée par la fonction d'utilité $U_i(.)$ telle que:*

$$U_i(\mathbf{x}) = \begin{cases} sign(\rho)[(1-n\alpha)x_i^\rho + \alpha\sum_{j=1}^n x_j^\rho] & \text{si } \rho \neq 0, \\ (1-n\alpha)\ln x_i + \alpha\sum_{j=1}^n \ln x_j & \text{sinon,} \end{cases}$$

avec $\mathbf{x} \succsim_i \mathbf{y} \Leftrightarrow U_i(\mathbf{x}) \geq U_i(\mathbf{y})$ pour toutes distributions $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$.

2.2.2.1 Le paramètre α évalue le poids de la tendance intrinséquement altruiste de l'agent i et $\alpha/(1-n\alpha)$ mesure le taux marginal de substitution entre le payoff de l'agent i et celui de n'importe quel autre agent j lorsque le même payoff est alloué aux deux individus. On remarque que, si $\alpha = 0$, $U_i(.)$ décrit alors des préférences purement *selfish*. Le second paramètre caractéristique de ces fonctions d'utilité de type **CES**, en l'occurrence ρ, mesure, quant à lui, l'aversion à l'inégalité. Si $\rho < 1$, l'agent i valorise très significativement le transfert marginal de son payoff vers un autre agent j quand ce dernier est moins bien loti

que lui mais, en revanche, très peu significativement quand il est mieux loti que lui. Le cas où $\rho \to -\infty$ correspond à la parfaite complémentarité, c'est-à-dire à des préférences de type "rawlsien" (voir Rawls 1971). Un agent "rawlsien" considère toujours l'agent le plus pauvre de la société comme prioritaire. Ainsi, selon la valeur que l'on donne à ce paramètre ρ, il est possible de modéliser une aversion aux inégalités substituables ou complémentaires (voir §**2.2.2.3** *infra*).

2.2.2.2 L'intuition selon laquelle les individus ont en général un comportement altruiste envers les membres de la société qui sont plus pauvres qu'eux et un comportement envieux envers ceux qui sont plus riches qu'eux est vérifiée par un grand nombre d'études expérimentales. Cette attitude face aux inégalités se modélise par un *kink*, *i.e.* une fonction d'utilité coudée. Cette caractérisque des fonctions d'utilité est notablement utilisée en théorie des choix dans l'incertain et découle de (**Sf**).

Théorème de représentation 2: *Si $n \geq 2$, alors une relation de préférence psychologique $\succsim_i$ de type **D-b** satisfait les axiomes (**Sf**), (**N**), (**H**) et (**AM**) si et seulement si elle peut être représentée par la fonction d'utilité $U_i(.)$ telle que:*

$$U_i(\mathbf{x}) = \begin{cases} sign(\rho)[(1-\gamma\alpha-(n-\gamma)\beta)x_i^\rho + \alpha\sum_{\{j\in\mathbf{N}:x_j\leq x_i\}} x_j^\rho \\ +\beta\sum_{\{j\in\mathbf{N}:x_j>x_i\}} x_j^\rho] \text{ si } \rho \neq 0, \\ (1-\gamma\alpha-(n-\gamma)\beta)\ln x_i + \alpha\sum_{\{j\in\mathbf{N}:x_j\leq x_i\}} \ln x_j \\ +\beta\sum_{\{j\in\mathbf{N}:x_j>x_i\}} \ln x_j \text{ sinon,} \end{cases}$$

avec $\gamma \equiv |j \in \mathbf{N} : x_j \leq x_i|$ et $\mathbf{x} \succsim_i \mathbf{y} \Leftrightarrow U_i(\mathbf{x}) \geq U_i(\mathbf{y})$ pour toutes distributions $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$.

2.2.2.3 La plupart des préférences psychologiques admettent pour représentation une fonction d'utilité correspondant à la formulation ci-dessus. Autrement dit, elles ne se distinguent les unes des autres que par les valeurs prises par les paramètres α, β et ρ. Toutefois, l'axiome de neutralité (**N**) impose naturellement l'égalité des coefficients venant pondérer les payoffs $(x_i, x_1, ..., x_n)$. Et cela ne correspond pas exactement à l'ambition affichée par les différentes théories de la préférence psychologique. En effet, celles-ci désirent plutôt que ces coefficients varient avec l'identité de l'agent. C'est le cas, notamment, du modèle d'aversion à l'inégalité de Fehr et Schmidt (1999).

L'Aversion aux Inégalités Substituables (Fehr et Schmidt 1999)

$$U_i(\mathbf{x}) = x_i^\rho + \alpha\sum_{j=1}^{n}\min[(x_j - x_i, 0)] + \beta\sum_{j=1}^{n}\min[(x_i - x_j, 0)]. \qquad (2)$$

Telle que formulée par Fehr et Schmidt (1999), cette formulation de l'aversion à l'inégalité, lorsque les agents sont indifférents à l'identité des autres agents, est un cas particulier du théorème **2** (*i.e.* $\rho = 1$). Le comportement face à l'inégalité se traduit ici par une manifestation d'altruisme à l'égard des plus pauvres (plus pauvres que l'agent i) et d'envie à l'égard des plus riches. Toutefois, Fehr et Schmidt (1999) imposent quelques hypothèses supplémentaires quant aux valeurs prises par les coefficients α et β, à savoir: $\alpha \geq \beta$ et $\beta \leq 1$. Ainsi,

les agents ont une plus grande propension à être envieux que *selfish*: la désutilité due à une inégalité favorable est, à la marge, plus grande que celle due à une inégalité défavorable. Fehr et Schmidt (1999) soutiennent que ce modèle permet d'expliquer un nombre important de résultats expérimentaux. (On peut notamment penser à ceux que révèlent les jeux de fourniture de bien public avec punition - *cf.* §**1.2.2.2** *supra* -, quand la contribution est supérieure à la prédiction théorique standard. La présence d'une petite fraction de joueurs hostiles aux inégalités est alors suffisante pour rendre crédible une menace que des joueurs purement *selfish* n'auraient jamais mise à exécution.)

Les Préférences Quasi-Maximin (Charness et Rabin, 2002)

$$U_i(\mathbf{x}) = (1-\lambda)x_i + \lambda \left[\delta \min_j x_j + (1-\lambda) \sum_{j=1}^{n} x_j \right]. \tag{3}$$

Charness et Rabin (2002) proposent une fonction d'utilité dont la forme (c'est une combinaison convexe) traduit la présence simultanée d'une motivation *selfish* et d'un désir à la fois d'équité (où l'on retrouve le critère maximin rawlsien) et d'efficacité (de type utilitariste). La fonction d'utilité qui apparaît dans le Théorème 1 peut elle aussi rendre compte d'un arbitrage entre ces différentes motivations: en effet, α y joue un rôle identique à celui de λ, $\rho << 0$ correspond à un δ proche de 1 et $\rho \to 1$ est équivalent à $\delta \to 0$. Toutefois, tandis que l'aversion à l'inégalité, telle que formulée par Fehr et Schmidt (1999), permet de décrire certains comportements de type "sensible à la réciprocité" (ceux qui visent à réduire les écarts entre les payoffs des agents), la fonction $U_i(.)$ proposée par Charness et Rabin (2002) est incompatible avec les observations expérimentales selon lesquelles les joueurs essaient de punir systématiquement leurs opposants - même si cela leur est coûteux.

L'Aversion aux Inégalités Complémentaires (Heidhues et Riedel 2007)

$$U_i(\mathbf{x}) = \min_j [x_i, \alpha x_j]. \tag{4}$$

Cette fonction d'utilité est un cas particulier de celle mise en évidence dans le Théorème 1 (*i.e.* $\rho \to -\infty$). Elle s'interprète de la manière suivante: si les inégalités ne sont pas excessives, l'agent i adopte des préférences standard. En revanche, lorsque les inégalités sont déraisonnables, il adopte un comportement rawlsien, *i.e.* α, le paramètre d'altruisme, est supposé supérieur à 1. Et, quand $\alpha \to \infty$, l'agent se comporte alors de manière quasi-*selfish*.

2.2.2.4 L'une des théories de la préférence psychologique qui ne peut pas être produite par les axiomes exposés ci-dessus est celle de Bolton et Ockenfels (2000). La raison en est que l'axiome de séparabilité (**SF**) est violé. Bolton et Ockenfels (2000) introduisent plutôt une fonction de motivation, notée $v = v_i(x_i, \sigma_i)$, dont les deux arguments sont, d'une part, le payoff x_i et, d'autre part, la proportion que ce dernier représente au sein de la distribution $(x_j)_{j=1}^n$. Ainsi, tout se passe comme si l'agent i comparait sa situation à celle d'un agent représentatif dont la dotation serait constituée d'un agrégat des payoffs dis-

tribués à l'ensemble de tous les agents[7]:

$$\sigma_i = \begin{cases} x_i / \sum_{j=1}^{n} x_j & \text{si } \sum_{j=1}^{n} x_j > 0, \\ 1/n & \text{si } \sum_{j=1}^{n} x_j = 0. \end{cases} \quad (5)$$

La fonction v est croissante en σ_i lorsque $\sigma_i < 1/n$ et décroissante en σ_i lorsque $\sigma_i > 1/n$. Ces propriétés traduisent alors sans ambiguïté une aversion à l'inégalité.

2.3 Les Préférences C-b et les Jeux Psychologiques

L'approche **C-b** permet aux préférences d'un agent de dépendre du contexte de ses décisions, c'est-à-dire du jeu même au sein duquel il pense évoluer. Les croyances qu'il forme à l'endroit de la survenance des différentes distributions de $\mathbb{R}_+^{n+1}$ sont ainsi en mesure d'influencer directement sa satisfaction. En cas de réciprocité intrinsèque, par exemple, on conçoit aisément que s'il interprète une distribution particulière en tant qu'elle manifeste l'intention bienveillante de son adversaire, un agent soit tenté de récompenser celui-ci. A l'inverse, il peut vouloir le punir dès lors qu'il estime que cette distribution n'est pas équitable. Ce qui importe fondamentalement aux joueurs, ce sont donc non seulement les conséquences matérielles de leurs décisions conjuguées avec celles de leurs adversaires, mais aussi la signification qu'ils leur prêtent en termes d'équité. Nous exposons ci-dessous l'axiomatique de Segal et Sobel (2007a) au sein duquel les agents sont dotés de préférences définies à la fois sur les payoffs et les stratégies. Cette approche peut, en outre, servir d'introduction à la théorie des jeux psychologiques.

2.3.1 Segal et Sobel (2007a) proposent un cadre axiomatique à la modélisation de la réciprocité intrinsèque qui nécessite que les joueurs aient - et c'est cela qui est crucial - des préférences définies simultanément sur les loteries et sur les profils de stratégies. Les préférences concernent alors naturellement les stratégies mixtes et non plus directement les payoffs (qui ne sont plus que les conséquences du profil choisi). L'intérêt de cette approche réside non seulement dans sa capacité à produire un cadre axiomatique adéquat pour décrire les relations de préférence psychologique qui dépendent des croyances mais également à autoriser un rapprochement avec la théorie des jeux dits (eux-aussi) psychologiques (**G**eanakoplos, **P**earce et **S**tacchetti (**GPS**) 1989, Rabin 1993, Dufwenberg et Kirchsteiger 2004, Battigalli et Dufwenberg 2007). En effet, l'introduction de préférences psychologiques dans l'étude des interactions stratégiques ne pose pas - dans le cas des préférences de type **D-b** du moins - de problème théorique insurmontable. Cependant, dès lors que ces préférences dépendent des croyances croisées que forment les agents à l'endroit du comportement d'autrui, les concepts traditionnels d'équilibre (au sein de la théorie des jeux) ne sont plus guère adaptés. Il est ainsi nécessaire de caractériser un nouveau concept - plus spécifique. Tel est l'objectif poursuivi par **GPS** (1989) lorsqu'ils définissent l'"équilibre de Nash psychologique" (voir *infra*).

2.3.2 Dans Segal et Sobel (2007a), les joueurs sont dotés d'une relation de préférence sur les statégies mixtes et les payoffs espérés qui découlent du

[7] Ceci constitue une différence majeure d'avec la version de Fehr et Schmidt (1999) où l'agent i compare sa situation à celle de chacun des autres agents.

choix de ces stratégies mixtes. Ces stratégies peuvent s'interpréter en tant qu'elles représentent les croyances de l'agent sur les choix d'autrui. Formellement, pour un profil de stratégie mixte $\boldsymbol{\sigma}^* = (\sigma_i^*, \boldsymbol{\sigma}_{-i}^*) \in \Sigma = \Sigma_i \times \Sigma_{-i}$, où Σ_i (respectivement $\Sigma_{-i} = \times_{j \neq i} \Sigma_j$) est l'ensemble des stratégies mixtes de l'agent i (respectivement $-i$, *i.e.* tous les agents $j \neq i$), Segal et Sobel (2007a) émettent l'hypothèse que le joueur i manifeste des préférences, notées $\succsim_{i,\boldsymbol{\sigma}^*}$, définies sur son ensemble de stratégies mixtes en plus de préférences sur les loteries, notées $\succsim^{out}$ et représentées par la fonction $u(.)$ - ces dernières correspondant aux préférences *selfish* et vérifiant les axiomes traditionnels du modèle von-Neumann-Morgenstern (**vNM**). De cette manière, $\boldsymbol{\sigma} \succsim_{i,\boldsymbol{\sigma}^*} \boldsymbol{\sigma'}$ s'interprète ainsi: dans le contexte $\boldsymbol{\sigma}^*$, le joueur i préfère jouer la stratégie $\boldsymbol{\sigma}$ à la stratégie $\boldsymbol{\sigma'}$. Segal et Sobel (2007a) imposent alors à ces préférences de satisfaire des axiomes de Continuité et d'Indépendance[8] qui se différencient néanmoins de l'acception standard de la continuité et de l'indépendance (*cf.* §**2.1.2** et §**2.2.1**) en ceci que l'objet du classement effectué par les préférences n'est plus une distribution de payoffs mais un profil de stratégies mixtes. Segal et Sobel (2007a) ajoutent un dernier axiome qui permet de faire dépendre l'utilité du joueur i des gains espérés par les autres joueurs; ils le nomment ***S**elf **I**nterest* (**SI**). (**SI**) requiert que si deux profils de stratégies mixtes conduisent à des loteries qui confèrent le même payoff à tous les agents $j \neq i$, alors i préfère le profil qui lui apporte le payoff espéré le plus important[9].

Théorème de représentation 3: *Une relation de préférence psychologique de type* ***C-b*** $\succsim_{i,\boldsymbol{\sigma}^*}$ *satisfait les axiomes de Continuité, d'Indépendance, (**vNM**) et (**SI**) si et seulement si elle peut être représentée par la fonction d'utilité* $U_{i,\boldsymbol{\sigma}^*}(.)$ *telle que:*

$$U_{i,\boldsymbol{\sigma}^*}(\boldsymbol{\sigma}_i) = u_i(\sigma_i, \boldsymbol{\sigma}_{-i}^*) + \sum_{j=1}^{n} \alpha_{i,\boldsymbol{\sigma}^*}^j u_j(\sigma_i, \boldsymbol{\sigma}_{-i}^*)$$

avec $\boldsymbol{\sigma} \succsim_{i,\boldsymbol{\sigma}^*} \boldsymbol{\sigma'} \Leftrightarrow U_{i,\boldsymbol{\sigma}^*}(\boldsymbol{\sigma}) \geq U_{i,\boldsymbol{\sigma}^*}(\boldsymbol{\sigma'})$ *pour toutes stratégies* $\boldsymbol{\sigma}, \boldsymbol{\sigma'} \in \Sigma$.

De plus si le sous-ensemble $A_i(\boldsymbol{\sigma}_{-i}) = \left\{u_i(\sigma_i, \boldsymbol{\sigma}_{-i}^*) : \sigma_i \in \Sigma_i\right\}$ est un ouvert non vide alors les poids $\left\{\alpha_{i,\boldsymbol{\sigma}^*}^j\right\}_{j=1}^n$ sont uniques.

L'Equilibre de Nash avec des Préférences dépendant des Croyances

Un équilibre de Nash dans le contexte des préférences **C-b** (dépendant des croyances) correspond à un profil de stratégies $(\sigma_i^*, \boldsymbol{\sigma}_{-i}^*)$ pour lequel, quelque soit un joueur $i \in \mathbf{N}$, σ_i^* est maximal pour la relation $\succsim_{i,\boldsymbol{\sigma}^*}$. Ici, le joueur i interprète en termes d'équité l'action du joueur j en se fondant sur ce que j pense que lui, i, va décider. Il compare ensuite les déviations possibles de σ_i qui laissent $\boldsymbol{\sigma}^*$ inchangé - c'est-à-dire celles qui ne modifient pas le contexte dans lequel il évalue le comportement de son adversaire. A l'équilibre, à contexte donné (*i.e.* à $\boldsymbol{\sigma}^*$ fixé), le joueur i choisit la stratégie σ_i, telle que cette stratégie soit la meilleure réponse qu'il puisse faire - étant données ses croyances sur les stratégies choisies par ses adversaires; en outre, les croyances de j sur les

[8] La relation $\succsim_{i,\boldsymbol{\sigma}*}$ satisfait l'axiome d'Indépendance si et seulement si, $\forall \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2, \boldsymbol{\sigma}$ et $0 < \alpha \leq 1$, $\boldsymbol{\sigma}^1 \succsim_{i,\boldsymbol{\sigma}*} \boldsymbol{\sigma}^2 \Leftrightarrow \alpha\boldsymbol{\sigma}^1 + (1-\alpha\boldsymbol{\sigma}) \succsim_{i,\boldsymbol{\sigma}*} \alpha\boldsymbol{\sigma}^2 + (1-\alpha)\boldsymbol{\sigma}$.

[9] (**SI**) rappelle ici l'axiome de Pareto-indifférence - standard en théorie du choix social.

choix de i concordent avec les actions effectivement choisies par i et le profil de stratégies mixtes définissant le contexte correspond à celles des stratégies qui sont effectivement jouées. Notons que le cadre de Segal et Sobel (2007a) est assez proche de celui des jeux psychologiques proposé par **GPS** (1989) dans lequel les utilités dépendent des hiérachies infinies de croyances[10]. L'apport de cette modélisation tient essentiellement au fait qu'il est possible d'analyser les interactions stratégiques en prenant en compte l'influence des intentions fondées sur les croyances (culpabilité, revanche...). A l'équilibre de Nash psychologique, tous les joueurs ont ainsi des croyances "collectivement cohérentes"[11]. Segal et Sobel (2007a) ont une approche similaire: les préférences qu'ils introduisent dépendent du contexte *via* les profils de stratégies mixtes adoptés par les joueurs (avec des croyances de niveau $k = 1$). Qu'il soit de connaissance commune que ceux-ci utilisent un profil de stratégies particulier permet alors d'associer une hiérarchie de croyances à chaque contexte donné. Selon Segal et Sobel (2007a), limiter le niveau spéculaire des hiérarchies de croyances au seul cas $k = 1$ n'implique pas de différence significative d'entre le modèle défendu par **GPS** (1989) et le leur. Les résultats en termes de représentation obtenus par Segal et Sobel (2007a) ont ainsi des propriétés telles qu'ils peuvent aussi bien être appliqués à la classe des jeux psychologiques à la **GPS** (1989).

Les Préférences pour la Réciprocité (Rabin, 1993)

L'application la plus célèbre de la théorie des jeux psychologiques est certainement le modèle de réciprocité développé par Rabin (1993). Ici, les décisions résultent d'un compromis rationnel entre une motivation standard, *i.e.* purement *selfish* (la maximisation des payoffs individuels) et une motivation plus psychologique, *i.e.* celle de réciprocité - laquelle repose sur les intentions individuelles. Afin d'analyser les modalités sous lesquelles s'effectue cette sorte d'arbitrage, Rabin (1993) défend une notion nouvelle: celle de *fairness equilibrium* qui est en réalité un équilibre de Nash psychologique où la condition de cohérence collective (telle que définie note **10** *supra*) n'est satisfaite qu'à la seule hauteur des deux premiers niveaux de croyance. Les résultats principaux de ce modèle permettent de mettre en évidence l'existence d'équilibres de Nash menant à des gains dits *mutual*-max ou *mutual*-min[12]. En outre, si les payoffs sont faibles, on peut montrer que l'ensemble des *fairness equilibria* est très proche de l'ensemble des résultats *mutual*-max et *mutual*-min. L'intégration de la notion d'équité a donc des conséquences en termes de bien-être social. Et, si les payoffs sont importants, la motivation purement *selfish* l'emporte sur la motivation d'équité. L'ensemble des *fairness equilibria* correspond alors (approximativement) à l'ensemble des équilibres de Nash. De nombreux travaux ont depuis proposé différentes modélisations alternatives de la préférence pour la réciprocité (voir Charness et Rabin 2002, ou Falk et Fischbacher 2006) ou en-

[10] Une hiérarchie d'ordre 2 correspond à la situation suivante: i pense que j pense que i joue une stratégie donnée... Lorsque l'ordre de la hiérarchie tend vers l'infini, il s'agit alors naturellement de ce que l'on appelle une "hiérarchie infinie" (voir aussi note **1**).

[11] Un joueur possède une hiérarchie de croyances cohérente si la distribution marginale d'une croyance de niveau $k+1$ est égale à la croyance correspondante de niveau k. Une hiérarchie infinie ($k \to \infty$) est collectivement cohérente s'il est de "connaissance commune" que toutes les croyances sont cohérentes.

[12] Un gain *mutual*-max (resp. min) est un résulat du jeu qui maximise (resp. minimise) les payoffs de chacun des joueurs.

core des adaptations du modèle de Rabin à des jeu extensifs (voir Dufwenberg et Kierchsteiger 2004). Ainsi Segal et Sobel (2007a) développent un modèle de la réciprocité intrinsèque que l'on peut interpréter en tant qu'il serait une version simplifiée du modèle de Rabin. En considérant la représentation des préférences telle qu'elle apparaît dans le Théorème 3, il convient de spécifier ici les poids $\alpha^j_{i,\boldsymbol{\sigma}^*}$ de la manière suivante:

$$\alpha^j_{i,\boldsymbol{\sigma}^*} = \begin{cases} \lambda \frac{U^h_j(\sigma^*) - F^G}{\overline{U}_j - \underline{U}_j} & \text{si } \overline{U}_j - \underline{U}_j > 0, \\ 0 & \text{si } \overline{U}_j - \underline{U}_j = 0, \end{cases} \tag{6}$$

où: $U^h_j(\boldsymbol{\sigma}^*) = \max_{s_j \in S_j} U_j(s_j, \boldsymbol{\sigma}^*_{-j})$ avec S_j l'ensemble des stratégie pures de j, $\overline{U}_j = \max_{\mathbf{s}_{-i} \in \times^n_{j=1} S_j} U_j(\mathbf{s})$ et $\underline{U}_j = \min_{\mathbf{s}_{-i} \in \times^n_{j=1} S_j} U_j(\mathbf{s})$. F^G s'interprète comme l'issue équitable du jeu. Cette modélisation demeure fondamentalement flexible quant à la définition de la préférence pour la réciprocité. En effet, Segal et Sobel (2007) ne cherchent pas à construire une théorie complète de l'équité: selon la spécification de F^G et la valeur prise par le paramère λ, il est possible d'engendrer différentes théories, *i.e.* d'intégrer les principales intuitions du modèle de Rabin (1993) mais aussi celles de Charness et Rabin (2002), Dufwenberg et Kierchsteiger (2004) ou Falk et Fischbacher (2006).

3 Vers une Nouvelle Economie Politique

La **TPP** présentée dans la partie précédente constitue une sorte de boîte à outils conceptuels - laquelle permet de produire des modèles dont le pouvoir à la fois explicatif, prédictif et normatif, est incontestablement plus grand que celui des modèles standard - eu égard aux faits expérimentaux que les premiers parviennent à intégrer après avoir constaté que les seconds s'y essayaient sans succès. Toutefois, l'économie politique (au sens traditionnel du terme[13]) ne semble pas avoir pris la pleine mesure des avancées faites dans le domaine de l'économie comportementale (dont la **TPP** est l'un des fleurons) non plus que des profondes mutations théoriques que ces avancées rendent possibles - dès lors qu'elles paraissent souhaitables. Nous proposons donc ici une réflexion (très préliminaire) à propos des conditions sous lesquelles l'intégration et l'utilisation des principaux résultats de la **TPP** pourraient enrichir l'économie politique - sans néanmoins l'alourdir exagérément de définitions abstraites et de raisonnements obscurs... Il s'agit en réalité d'estimer l'apport potentiel de l'économie comportementale à la mise en place des politiques économiques et de suggérer, le cas échant, quelques pistes d'investigation...

3.1 Incitations et Préférences Psychologiques

L'ignorance manifeste des comportements empreints d'aversion à l'inégalité, de réciprocité intrinsèque ou de toute autre forme de motivation déviant de celles traditionnellement postulées par la théorie standard - en bref, la négligence des intuitions dont la **TPP** est porteuse -, révèle dans certaines situations concrètes (décisions publiques, investissements...) une incompréhension

[13] *I.e.* la branche de la science économique qui étudie de façon théorique les moyens d'action de la politique économique des gouvernements, des États et des collectivités territoriales.

majeure de la tension qui oppose coopération et compétition et, partant, qu'un système d'incitations insensible à la nature fondamentale des préférences individuelles prend le risque de produire des effets fâcheux (voire nuisibles en termes d'efficacité sociale) que la théorie standard ne peut guère anticiper.

3.1.1 L'Effet *Crowd-Out*

L'effet "*Crowd-Out*" est un parfait exemple de ces conséquences à la fois imprévisibles par la théorie standard et socialement nuisibles. Ce phénomène fait directement référence aux effets néfastes des interventions extérieures (incitations monétaires, punitions...) sur les motivations intrinsèques. Cependant que, pour la théorie standard, le décideur public doit pallier les inefficacités dues aux comportements de type "passager clandestin" grâce à des incitations monétaires, il n'est pas évident *a priori* que ces mêmes incitations monétaires soient encore opportunes afin de résoudre les problèmes d'action collective lorsque les agents que l'on envisage ont cessé d'être de simples maximisateurs d'intérêt individuel (voir, par exemple, Andreoni 1993).

3.1.1.1 Frey et Oberholzer-Gee (1997) ont ainsi mené en Suisse une étude empirique à propos de la disposition des citoyens à accepter "l'entassement" des déchets nucléaires dans leur voisinage immédiat. Puisqu'il s'agit là d'une situation classique de type "*Not In My Backward*", on sait que la théorie standard prédit une plus grande fréquence d'acceptation dès l'instant qu'une compensation monétaire suffisante est proposée. Pourtant, l'étude de Frey et Oberholzer-Gee (1997) montre exactement l'inverse! Et cette réaction, typiquement irrationnelle dans le cadre standard, s'avère en revanche tout à fait rationnelle pour des individus dont les préférences expriment une sensibilité à la réciprocité intrinsèque. En effet, le seul fait de proposer une compensation monétaire change la nature des relations économiques en jeu et les transforme aussitôt en des relations purement transactionnelles, *i.e.* compétitives plutôt que coopératives. Toute incitation monétaire affecte la perception individuelle du contexte économique et ne constitue donc pas seulement un flux monétaire, socialement neutre, mais également un message quant à la nature de l'enjeu qui les justifie en amont. Elles peuvent, en outre, heurter certaines des valeurs morales et des normes sociales qui structurent l'univers psychologique des agents.

3.1.1.2 De même, les incitations de nature monétaire agissent parfois en tant que catalyseur d'informations dont l'impact est purement ignoré par la théorie standard**.** Dans le cas d'un jeu de contribution à la fourniture d'un bien public, la simple occurrence d'une incitation matérielle prouve que les agents ne sont pas naturellement enclins à coopérer. Or, un agent ayant des préférences psychologiques (exprimant une sensibilité à la réciprocité) coopère d'autant plus facilement qu'il estime que les autres agents vont eux-aussi coopérer. Aussi, l'existence d'une incitation monétaire, interprétée en tant que signal de la présence potentielle d'un ou plusieurs passagers clandestins, engendre des inefficacités sociales en ce qu'elle diminue le niveau de contribution de tous les individus dont le comportement exprime une préférence conditionnelle pour la réciprocité. De plus, l'existence d'une telle incitation masque les contributions volontaires (à la fourniture du bien public) et annule donc les comportements spontanément coopératifs qui visent à récompenser les actions jugées équitables ou "*nice*".

3.1.1.2 (Les préférences psychologiques comme réponse au contexte) De manière plus générale, le comportement décrit en **3.1.1.1** correspond au type **C-b** des préférences (présenté §**2.3** *supra*). Considérons, par exemple un jeu standard, type dilemme du prisonnier, dans lequel "ne pas coopérer" est une stratégie strictement dominante. Un individu i dont les préférences sont telles qu'il peut classer les distributions de manière différente selon que les autres agents sont enclins à coopérer avec lui ou non peut alors être amené à choisir de coopérer en ceci que la stratégie de coopération n'est plus nécessairement dominée. En effet, cet agent i n'évalue pas l'équité d'une distribution de manière absolue. Son jugement se fonde plutôt sur une sorte de norme dépendant d'une situation donnée - laquelle prend en compte le contexte singulier dans lequel les payoffs sont distribués. Sa relation de préférence est théoriquement fonction d'"un point de référence"[14]. Imaginons à présent que ce point de référence soit défini par la proportion, inconnue de l'agent i, d'individus prêts à coopérer volontairement avec lui et supposons qu'il choisisse d'autant plus facilement une distribution favorisant le bien-être des autres agents que cette proportion est forte. Dès lors que l'agent i interprète la présence d'une incitation matérielle comme le signal de la présence d'un assez grand nombre de passagers clandestins, il révise aussitôt ses croyances et se comporte de manière nettement plus *selfish*. Formellement, il pondère plus fortement son propre payoff dans la distribution des poids relatifs de la fonction d'utilité qui représente ses préférences de sorte que, finalement, son comportement ressemble à celui d'un passager clandestin - et ceci alors même qu'intrinsèquement il ne l'est pas. On voit ici toute la capacité théorique des préférences **C-b**: les motivations qui soutendent le comportement ne sont pas invariantes mais évoluent au gré des situations. Ainsi un même agent peut être *selfish* dans une situation et altruiste dans une autre.

3.1.1.3 La préférence pour la réciprocité et l'apprentissage progressif des normes sociales dépendent des croyances et des anticipations que forme un agent à propos de ce que vont décider les autres agents qui définissent son environnement social. Un comportement intrinséquement réciproque repose sur la perception psychologique du contexte (*i.e.* des agents qui le définissent) et donc sur l'information ayant trait à la fiabilité comportementale de ce contexte. On peut alors soutenir que le rôle du décideur public est de concevoir et de promouvoir des politiques telles qu'elles aient, ces politiques, la capacité à produire un contexte où motivations matérielles, désir d'équité et informations interagissent harmonieusement. Ceci suggère aussitôt qu'il serait sans doute fécond de repenser les modèles de l'économie politique à la lumière de la **TPP**. Une telle rénovation de l'économie politique permettrait alors d'anticiper efficacement les

[14] La modélisation d'une telle structure de préférences trouve son origine dans Tversky et Kahneman (1991) et Munro et Sugden (2003). Il s'agit en fait de lier les éléments d'un ensemble de points de référence à ceux d'un ensemble de relations de préférence. Le point de référence de l'agent i, noté r_i, est déterminé par le payoff jugé équitable par i. Pour chaque distribution de points de référence $\mathbf{r} = (r_i, r_1 ..., r_n)$, une structure de préférences $\succsim_i^{\mathbf{r}}$ est spécifiée. Les fonctions d'utilité qui représentent ces $\succsim_i^{\mathbf{r}}$ sont alors similaires à celles apparaissant dans les différents théorèmes de représentation (voir *supra*) à ceci près que les poids ne sont plus fixes mais varient avec $\mathbf{r}$. Ce modèle permet en particulier d'aborder plusieurs conceptions théoriques de la justice telle, par exemple, celle fondée sur la notion introduite par Amartya Sen d'"*act of choice*". Sen (1997) suppose ainsi que la manière dont un payoff est évalué dépend de celui qui est à l'origine du choix qui a conduit à la détermination de ce payoff (*chooser dependance*) et des choix alternatifs (*alternative dependance*).

effets des décisions publiques et de concevoir les institutions idoines à leur mise en oeuvre.

3.1.2 La Formation Endogène des Institutions

Puisqu'il est avéré qu'une politique économique peut avoir des effets non anticipés par la théorie standard - lesquels s'expliquent par la présence d'un certain nombre d'agents manifestant un désir d'équité, une aversion à l'inégalité ou des préférences sensibles à la réciprocité -, la question se pose à présent de savoir sous quelles conditions et de quelle manière la présence de tels agents aux préférences psychologiques peut affecter la formation d'institution.

3.1.2.1 (Les institutions centralisées: le jeu) Pour répondre à cette question, Kosfeld *et al.* (2006) étudient la formation endogène des institutions dans un jeu de bien public. Dans ce cadre, les agents ont certes intérêt à élaborer une institution ayant un pouvoir de sanction mais chaque agent peut augmenter ses payoffs s'il demeure en dehors de l'institution (*outsider*) cependant que les autres agents y participent (*insider*). Le jeu se décompose en trois étapes. Dans un premier temps, les agents choisissent ou non de participer à une institution centralisée ayant un pouvoir de sanction sur les joueurs dont la contribution à la fourniture du bien public est jugée trop faible. Dans un second temps, ils apprennent le nombre des *insiders* potentiels et l'institution est concrètement instaurée (*implemented*) si et seulement si aucun de ces *insiders* potentiels ne change d'avis et préfère finalement demeurer un *outsider* (sachant qu'instaurer cette institution induit un coût monétaire). Enfin, les agents choisissent leur niveau de contribution à la fourniture du bien public.

3.1.2.2 (Les institutions centralisées: les expériences) Les résultats expérimentaux montrent que les institutions n'apparaissent qu'à condition qu'elles aient un impact positif sur le taux de coopération et le bien-être du groupe. Les agents économiques semblent en effet reticents à instaurer des institutions où certains individus conservent l'opportunité d'adopter un comportement de passager clandestin; ils favorisent plutôt la création d'institutions ayant la possibilité de sanctionner ces derniers. Bien que les principaux résultats expérimentaux soient simultanément compatibles avec des préférences standard et des préférences psychologiques, le processus même de formation d'une institution semble répondre à des principes fondamentaux que la théorie standard ignore, faute de pouvoir en tenir compte. En contrepartie de quoi, certaines institutions - qui induisent pourtant une amélioration matérielle de la situation des agents - sont rejetées pour des raisons sociales (ou psychologiques au sens de la **TPP**).

3.1.2.3 Qu'en est-il de la formation d'institutions lorsque certains agents manifestent des préférences psychologiques qui traduisent une aversion aux inégalités substituables (comme dans Fehr et Schmidt 1999) et dont les fonctions d'utilité sont telles que le paramètre β y est assez fort (*i.e.* le paramètre d'altruisme envers les moins bien lotis, *cf.* §**2.2.2** *supra*)? Un agent *outsider* s'expose ainsi à ressentir une importante désutilité s'il adopte un comportement de passager clandestin. Cependant, l'aversion à l'inégalité agit dans les deux sens: si tous les *outsiders* contribuent moins que lui à la fourniture du bien public, l'agent subit une perte d'utilité due aux inégalités désavantageuses. Devenir *insider* peut donc améliorer son bien-être. Ainsi, le nombre nécessaire d'*insiders*

pour qu'une institution soit instaurée dépend non seulement du niveau de contribution des *outsiders* mais aussi bien des paramètres d'aversion à l'inégalité dans les fonctions d'utilité individuelles.

3.1.2.4 (Les institutions décentralisées) Peut-on comparer l'efficacité relative d'une institution centralisée avec celle d'une institution décentralisée lorsque les agents ont des préférences psychologiques? Poteete et Ostrom (2004) proposent quelques éléments de réponse à cette question en comparant les diverses institutions en charge de la gestion des ressources forestières dans dix pays différents. Ils montrent que plus les agents concernés ont un réel pouvoir discrétionnaire, quant à l'élaboration et l'instauration des règles qui encadrent leurs comportements, plus ils coopèrent. De même, de nombreux travaux empiriques - ayant trait par exemple à la gestion des systèmes agricoles au Népal - montrent que, si les règles et les sanctions sont conçues par les agriculteurs eux-mêmes (ou leurs représentants) plutôt que par une autorité extérieure - bien souvent perçue comme contraignante -, elles sont plus fréquemment appliquées et respectées. Donner un rôle actif dans la gestion des ressources aux utilisateurs eux-mêmes de ces ressources permet en effet à ceux-ci d'acquérir de l'information sur leurs semblables, d'améliorer la communication entre eux, d'appliquer plus facilement les normes et les règles, et finalement de construire des relations proches de celles stipulées par un contrat incomplet - tout cela à un coût relativement faible. En d'autres termes, les institutions décentralisées permettent de créer un contexte favorable à ceux des agents dont les préférences sont sensibles à la réciprocité conditionnelle.

3.1.2.5 Toutes ces études vont donc l'encontre de la vision standard de ce que Hardin (1968) a appelé *the Tragedy of Commons*, une tragédie métaphorique illustrant le conflit traditionnel entre intérêt individuel et intérêt collectif lors de la compétition pour l'accès à une ressource naturelle. Le postulat néo-classique selon lequel tous les agents adoptent un comportement *selfish* conduit ici à des niveaux de contribution individuelle tout à fait inefficaces. Pour pallier ce problème, la théorie standard suggère en retour qu'une agence centralisée soit créée dont la fonction est de concevoir des incitations susceptibles de ramener la contribution de chaque agent à la fourniture d'un bien public à un niveau significativement plus élevé. Cependant, puisque les études expérimentales et empiriques montrent sans ambiguïté que les préférences *selfish* ne sont pas les seules à se manifester, qu'il existe des agents dont les préférences sont proches de celles décrites par la **TPP**, il convient, en conséquence, de repenser la nature des incitations ainsi que le mode de leur instauration.

3.2 Globalement *Selfish vs.* Localement Altruiste

3.2.1 Les exemples précédents montrent que la **TPP** est *a priori* en mesure de fournir des fondements théoriques aux (*i.e.* en amont des) différents modèles grâce auxquels l'économie comportementale peut espérer enrichir l'économie politique. Il est toutefois assez sage de s'interroger à l'endroit de la robustesse des observations empiriques et des faits expérimentaux mis ainsi en lumière. Deux arguments principaux s'imposent qui invitent à la prudence: en premier lieu, la nature des protocoles expérimentaux et des études empiriques et, en second lieu, la cohérence interne et l'homogénéité de la **TPP**.

3.2.1.1 La plupart des études expérimentales ainsi qu'empiriques présentées dans ce papier n'impliquent qu'un petit nombre d'agents. Or, rien ne permet de supposer (sans état d'âme ni inquiétude) que l'on puisse en extrapoler les résultats à des groupes de plus large dimension sans coût ni perte de signification. Cet élément est d'importance car si ces études n'étaient pas robustes à l'extension de la dimension du groupe considéré, il serait alors préférable, sous certaines conditions, de restreindre leur champ d'analyse au seul niveau local.

3.2.1.2 La **TPP** ne constitue pas une théorie cohérente et homogène. Bien au contraire, elle se caractérise par une relative immaturité et un manque de logique interne. Aucune de ses composantes ne parvient à capturer l'ensemble des déviations par rapport à la théorie standard que l'on observe lors des expériences et chacune demeure *ad hoc* dans sa formalisation. Les différentes tentatives d'axiomatisation ne dessinent pas encore les limites claires d'un cadre unique. L'utilisation de la **TPP** est donc coûteuse en ceci qu'elle induit automatiquement une perte en généralité. Pourquoi en effet choisir telle formulation plutôt qu'une autre et comment décider qu'une certaine déviation est plus fondamentale que sa voisine empirique? En outre, l'accroissement du nombre des paramètres dont il conviendrait de tenir compte si l'on désirait absorber la totalité des déviations révélées en amont de la **TPP** accroîtrait simultanément la complexité du modèle global, ce que l'on nomme parfois sa "tractabilité", dans des proportions considérables - dès lors que l'on parviendrait à l'écrire... Force est de constater donc que la "vieille" hypothèse *selfish*, si elle caricature la substance des comportements et en érase la diversité au point de n'en proposer qu'une représentation fruste et même rudimentaire, a pour elle de simplifier la modélisation.

3.2.2. (Poids de la décision, anonymat: effet taille et nature des connections entre les individus) Les jeux expérimentaux de marché et les analyses théoriques de Sobel (2005, 2008) ou Heidhues et Riedel (2007) suggèrent que la théorie standard parvient à fournir, dans certaines situations, de bonnes prédictions. Davis et Holt (1994) ont montré que des sujets artificiellement placés en situation concurrentielle se comportaient exactement comme des agents *selfish* maximisateurs. Sobel (2008) montre toutefois que ces résultats ne sont pas incompatibles avec une classe particulière de préférences psychologiques - classe qu'il identifie et caractérise. Il décrit ainsi une famille de relations de préférence qui peuvent à la fois dépendre de la distribution des payoffs et subir éventuellement l'influence des intentions conçues par les autres agents. La fonction d'utilité mise en évidence par Fehr et Schmidt, celles proposées par Charness et Rabin (dans les deux cas, *cf.* §**2.2.2.3** *supra*) et Segal et Sobel (2007) (*cf.* eq.(6), §**2.3.2** *supra*) appartiennent à cette famille. Il démontre que, quand bien même il y aurait des agents manifestant de telles préférences (*i.e.* psychologiques), l'équilibre de marché est le même que celui prédit par la théorie standard. La principale raison de ce résultat est intuitive: puisque les agents n'ont aucun pouvoir de marché en situation concurrentielle, les préférences non-*selfish* sont "atomisées" et ne peuvent donc influer sur l'équilibre. On peut noter, de même, que l'axiome d'Indépendance (voir note **8**) entraînant une séparabilité des différents arguments de la fonction d'utilité de ces agents aux préférences psychologiques, les décisions socialement agrégées ne peuvent que conduire, dans un jeu de marché, à un équilibre similaire à celui prédit par la théorie standard, c'est-à-dire celui procédant du fait que tous les agents ont des

préférences *selfish*. En bref, si l'on s'intéresse à des situations économiques où les décisions individuelles n'ont que peu d'impact sur le bien-être des autres (ou dans lesquelles les agents opèrent des échanges scrupuleusement anonymes) et si l'on postule l'axiome d'Indépendance, alors on peut s'attendre à ce que les résultats en termes d'équilibre soient identiques à ceux produits par la théorie standard.

3.2.3 En ce qui concerne la délimitation (par exclusion) du champ d'analyse d'une économie politique comportementale - laquelle définirait alors l'embryon théorique d'une nouvelle économie politique - il semblerait donc que si l'on cherche à caractériser des politiques économiques visant à influencer le comportement d'agents (**i**) indépendants et socialement déconnectés les uns des autres (**ii**) dont les préférences vérifient l'hypothèse de séparabilité, l'approche standard soit suffisante. Ces circonstances singulières renvoient clairement (à nos yeux, du moins) à des politiques économiques menées au niveau global en cela qu'alors les agents-atomes n'ont pas la possibilité d'impressionner directement le bien-être de leurs partenaires à travers ces politiques.

3.2.4 La taille de la population, la répétition des intéractions sont des facteurs déterminants quant à la capacité des agents à signaler leurs intentions ou les principes auxquels ils adhèrent. Et ceci concerne aussi bien les agents *selfish* que non-*selfish*. Ainsi, les premiers ont parfois intérêt à se faire passer pour les seconds afin d'éviter des "représailles" ou pour attirer la sympathie (économique) de ceux des agents qui ont des préférences sensibles à la réciprocité. Pour qu'un agent aux préférences conditionnellement sensibles à la réciprocité adopte un comportement coopératif, il faut que s'y prêtent les croyances qu'il forme à propos de l'environnement au sein duquel il prend ses décisions. La possibilité d'observation et d'interprétation des décisions est ici déterminante. Une décision économique a non seulement des conséquences matérielles, objectives, mais aussi des conséquences éminement informationnelles: lorsqu'un agent est incertain de son environnement (et aussi bien de son évolution), il sait que les actions des autres agents autour de lui sont éventuellement porteuses d'information. La compilation de cette information lui permet alors de réviser ses croyances et, partant, d'adapter son comportement. Toutefois, afin que cela se produise de façon efficace, il est nécessaire que les actions des autres agents soient univoques, *i.e.* clairement informatives, c'est-à-dire que chacun soit à même de les interpréter sans commettre d'erreur. Or, le type de politique économique, et notamment l'échelle à laquelle on souhaite la mener, influence la capacité informative des messages que reçoivent les agents. Ainsi, une politique de niveau global (ou s'adressant à l'ensemble des agents au sein d'un réseau d'interactions anonymes) n'augmentera en rien la coopération car une telle politique n'est pas capable *a priori* de produire des signaux informatifs perceptibles au niveau individuel. Un agent purement *selfish* ne construisant pas sa propre réputation n'aura plus, en conséquence, à craindre de "représailles" - *i.e.* d'incitations négatives. Aussi, dans un contexte où l'information est limitée, les comportements procédant de préférences sensibles à la réciprocité sont plus timorés voire, dans certaines situations, identiques à ceux d'un agent purement *selfish*: l'altruisme même rationnel ne s'apprend pas. L'effet *Crowd-Out* est réduit et l'impact des incitations monétaires est alors semblable à celui que prédit la théorie standard.

3.2.5 Les croyances des agents et l'information relative au contexte qu'ils

détiennent sont des éléments déterminants pour l'émergence de comportements issus de préférences sensibles à la réciprocité (de comportements réciproques, pour faire vite...). Ces comportements dépendent, d'une part, des croyances, d'autre part, de l'information acquise *via* l'observation des décisions des autres agents et, enfin, de l'interprétation qui peut en découler. En résumé, ils dépendent donc du contexte - ce qui nous amène à considérer avec prudence les effets *Crowd-Out* ou la substitution de la confiance aux contrats complets.

3.2.5.1 (**La sensibilité des effets *Crowd-Out* au contexte**) Il est faux de croire que n'importe quel système standard d'incitations monétaires est caduc dès lors qu'il existe un nombre suffisant d'agents adoptant un comportement intrinsèquement réciproque. En effet, l'impact des incitations économiques sur la décision finale des agents peut être tel que les motivations standard demeurent prépondérantes. En outre, ces incitations sont en mesure de modifier la perception que les agents ont de la disposition de leurs partenaires à coopérer avec eux.

3.2.5.2 (**La sensibilité de la substitution de la confiance aux contrats complets**) Il est tout aussi faux de croire que des comportements réciproques constituent des substituts parfaits aux contrats complets lorsque ces derniers ne sont pas parfaitement applicables ou encore très coûteux. L'expérience d'Andreoni (2005) révèle qu'une intervention extérieure est nécessaire afin qu'apparaissent des comportements coopératifs. Un contrat de type "satisfaction garantie" améliore certes l'efficacité et aide à construire une relation de confiance mais à la condition qu'il soit instauré par une instance neutre - par exemple, la loi. Il est toutefois suffisant qu'un seul des deux côtés de la transaction soit soumis à cette contrainte. Le contrat "satisfaction garantie", même si son application est volontaire, augmente toujours la fiabilité des vendeurs mais n'assure en revanche la confiance des acheteurs qu'à condition qu'il soit protégé de façon exogène: quand bien même certains agents manifestent des préférences sensibles à la réciprocité, ils n'acceptent donc de coopérer qu'en s'estimant protégés par une autorité étrangère à la transaction elle-même.

3.2.6 Que des agents aient un désir d'équité n'annule pas le rôle privilégié du décideur public centralisé mais le redéfinit plutôt: sa fonction est à présent de créer et de garantir un environnement, un contexte, favorable aux agents dont les préférences traduisent qu'ils sont intrinséquement enclins à coopérer - en garantissant en particulier qu'ils ne seront pas souterrainement lésés par des passagers clandestins opportunistes. Autrement dit, le décideur public doit ici se faire l'avocat d'une intrication optimale des intentions privées et des institutions publiques afin de faciliter et de clarifier les interactions économiques entre des agents divers et diversement sensibles au bien-être de leur environnement social.

4 Conclusion

La **TPP** produit des prédictions différentes de celles que propose l'approche standard sous deux conditions principales. D'abord, si l'on suppose que les préférences psychologiques vérifient l'axiome d'Indépendance, il est crucial, pour qu'un agent se distingue par son comportement d'un individu *selfish*, qu'il ait une influence non négligeable sur le bien-être des autres agents (condition locale). Ensuite, en ce que la capacité à échanger de l'information sur les car-

actérisques ou les intentions des autres agents est un facteur déterminant du comportement, la taille du groupe considéré, les interactions existant entre les différents agents, la répétition et la fréquence de ces interactions, tout cela conditionne fondamentalement l'opportunité théorique de la **TPP**.

4.1 (Communautés) Les groupes d'agents présentant un niveau d'interaction suffisant (intensité et fréquence) pour ne pas être aussitôt dilué dans l'anonymat ou la multitude de leur environnement social sont typiquement de petites sociétés - ou ce que l'on appelle des "communautés". Il est alors naturel de considérer l'économie politique de ces communautés à la lumière de la **TPP**.

4.1.1 (Systèmes polycentriques) Les systèmes polycentriques (*i.e.* ceux dont le principe de gouvernance est décentralisé auprès des différentes communautés qui partitionnent la société) sont ici particulièrement intéressants en ce qu'ils apportent une solution originale au problème de la coordination lorsque ni le marché ni un decideur centralisé ne sont en mesure d'y répondre de manière adéquate. C'est notamment le cas dès que l'écriture et l'instauration d'un contrat se révèlent coûteuses (*cf.* §**3.2.5.2** *supra*). Or, la complexité des interactions et la part grandissante des informations invérifiables ou privées nécessaires à la prise de décision dans les économies modernes conduisent à reconsidérer ces structures d'un oeil nouveau. La compréhension de leur mode de fonctionnement semble en effet requérir assez naturellement une approche en termes de préférences psychologiques.

4.1.2 (Gouvernance et TPP) Les systèmes polycentriques génèrent un environnement propice à la manifestation des désirs individuels d'équité. Le type d'interaction qu'ils supposent crée, en effet, un contexte favorable à l'adoption de comportements réciproques motivés par un souci d'équité; certains agents sont ici davantage enclins à coopérer volontairement et à sacrifier une partie de leurs payoffs pour récompenser (ou pour punir) leurs partenaires - les caractéristiques sociales d'une communauté rendant cet environnement naturellement propice aux comportements réciproques car la probabilité que les membres d'une même communauté intéragissant aujourd'hui ensemble soient amenés à se rencontrer dans le futur est élevée. En outre, la fréquence des interactions parmi un petit nombre d'individus réduit en amont le coût d'acquisition de l'information et accroît en aval les bénéfices associés à une meilleure connaissance des caractéristiques et des motivations qui animent les membres de la communauté: les agents sensibles à la reciprocité voient plus clairement les opportunités de coopération mais aussi de punition des comportements "anti-sociaux" (de type passager clandestin). La présence d'individus sensibles à la réciprocité augmente alors la valeur de l'information dispersée dans la communauté. Par ailleurs, on sait que permettre aux agents de jouer un rôle actif dans la création et l'instauration des règles sociales a des effets positifs sur l'efficacité des actions collectives (*cf.* §**3.1.2.4** *supra*). Or, la taille réduite des entités concernées et la nature des interactions entre les agents au sein des systèmes polycentriques rendent sans aucun doute crédible que le pouvoir discrétionnaire de ces structures soit délégué aux agents eux-mêmes. La visibilité des décisions et la possibilité de participer à l'élaboration des normes régulant l'action collective constituent alors un moyen tout à fait pertinent d'atténuer les problèmes liés à la mise en place d'incitations.

4.1.3 Toutefois, on sait aussi que les agents préfèrent être membres d'une institution centralisée ayant un pouvoir de punition et qu'ils rejettent celles des

institutions où les passagers clandestins demeurent impunis parce qu'invisibles: quand même ils manifesteraient des préférences conditionnellement sensibles à la réciprocité, ils exigent néanmoins d'être mis à l'abri des comportements "anti-sociaux". C'est la raison pour laquelle une intervention extérieure a ici sa place. La gouvernance par les communautés n'est donc pas un substitut mais un complément opérationnel au décideur public centralisé.

4.2 Le mariage entre la **TPP** et l'économie politique paraît donc *a priori* fertile comme semble le montrer l'étude des modes de gouvernance par les communautés. Cependant, la combinaison de ces deux "langages" de la science économique que sont l'économie comportementale et l'économie politique (pour résumer) est à manipuler avec beaucoup de prudence: l'économie comportementale - au moins sa partie qui concerne les préférences psychologiques - manque encore de robustesse et de cohérence interne (*cf.* §**3.2.1.2** *supra*) si l'économie politique souffre, quant à elle, de sénescence. Pourtant, ce qui est au coeur de la **TPP**, à savoir le relâchement de l'hypothèse d'avidité rationnelle dont la logique économique se nourrit au point d'oublier parfois qu'il ne s'agit là que d'une posture théorique (et non d'un constat empirique) devrait mener rapidement à l'élaboration de nouveaux modèles pour peu que les applications attendues (et apparemment souhaitées par une vaste partie du monde académique) parviennent à se libérer de leur défiance à l'égard de cette abstraction sans laquelle il n'est d'avancée que transitoire - défiance qui, telle la loi de la gravitation, ramène toujours le nouveau vers l'ancien, l'audacieux vers le frileux, l'inédit vers l'éprouvé...

5 Bibliographie

Akerlof, G. et R. Kranton (2000). "Economics and Identity," *Quarterly Journal of Economics* **115**: 715-753.

Andreoni, J. (1993). "An Experimental Test of the Public-goods Crowding-out Hypothesis," *American Economic Review* **83**: 1317-1327.

Andreoni, J. (2005). "Trust, Reciprocity, and Contract Enforcement: Experiments on Satisfaction Guaranteed," *miméo.*

Arrow., K. (1951). *Social Choice and Individual Values.* John Wiley: New York (2^{nde} édition, 1963).

Battigalli, P. et M. Dufwenberg (2007): "Dynamic Psychological Games," *Journal of Economic Theory*, à paraître.

Bénabou, R. et J. Tirole (2003). "Intrinsic and Extrinsic Motivation," *Review of Economic Studies* **70**: 489-520.

Bolton, G. et A. Ockenfels (2000). "ERC: A Theory of Equity, Reciprocity and Competition," *American Economic Review* **90**: 66-193.

Bewley, T. (1999). *Why Wages Don't Fall During a Recession.* Harvard University Press: Cambridge, MA.

Camerer, C. (2003). *Behavioral Game Theory.* Princeton University Press: Princeton.

Charness, G. et M. Rabin (2002). "Understanding Social Preferences With Simple Tests," *Quarterly Journal of Economics* **117**: 817-869.

Davis, D. et C. Holt (1994). "Market Power and Mergers in Laboratory Experiments with Posted Prices," *RAND Journal of Economics* **25**: 467-487.

Dufwenberg, M. et G. Kirchsteiger (2004). "A Theory of Sequential Reciprocity," *Games and Economic Behavior* **47**: 268-298.

Dupuy, J.P. (1989): "Convention et common knowledge," *Revue Economique* **2**: 361-400.

Falk, A. et U. Fischbacher (2006). "A Theory of Reciprocity," *Games and Economic Behavior* **54**: 293-315.

Falk, A. et M. Kosfeld (2006). "The Hidden Costs of Control," *American Economic Review* **96**: 1611-1630.

Fehr, E., S. Gachter et G. Kirchsteiger (1997). "Reciprocity as a Contract Enforcement Device: Experimental Evidence," *Econometrica* **65**: 833-860.

Fehr, E. et K. Schmidt (1999). "A Theory of Fairness, Competition and Cooperation," *Quarterly Journal of Economics* **114**: 817-868.

Fehr, E. et K. Schmidt (2003). "Theories of Fairness and Reciprocity - Evidence and Economic Applications," Invited Lecture at the 8th World Congress of the Econometric Society 2000, dans M. Dewatripont, L. Hansen et St. Turnovsky Eds., *Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs*, Cambridge University Press, Cambridge, MA.

Fehr, E. et G. Simon (2000). "Cooperation and Punishment in Public Goods Experiments," *American Economic Review* **90**: 980-994.

Fischbacher, U., S. Gachter et E. Fehr (2001). "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment," *Economics Letters* **71**: 397-404.

Frey, B. et F. Oppenheimer-Gee (1997). "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding Out," *American Economic Review* **87**: 746-755.

Fudenberg, D. (2006). "Advancing Beyond Advances in Behavioral Economics," *Journal of Economic Literature* **44**: 694-711.

Fudenberg, D. et E. Maskin (1986). "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* **54**: 533-554.

Geanakoplos J., D. Pearce et E. Stacchetti (1989). "Psychological Games and Sequential Rationality," *Games and Econonomic Behavior* **1**: 60-79.

Gibson, C., M. McKean et E. Ostrom (2000). *People and Forests: Communities, Institutions and Governance*. MIT Press: Cambridge, MA.

Gintis, H., S. Bowles, R. Boyd et E. Fehr Eds. (2005). *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economics*. MIT press: Cambridge, MA.

Groves, T. et J. Ledyard (1977). "Optimal Allocation of Public Goods: A Solution to the 'Free-rider' Problem," *Econometrica* **45**: 783-810.

Hardin, G. (1968). "The Tragedy of the Commons," *Science* **162**: 1243-1248.

Heidhues, P. et F. Riedel (2007). "Do Social Preferences Matter in Competitive Markets?," *miméo*.

Kahneman, D., J. Knetsch et R. Thaler (1986). "Fairness and the Assumptions of Economics," *Journal of Business* **59**: 285-300.

Kosfeld, M., A. Okada et A, Riedl (2006). "Institution Formation in Public Goods Games," *miméo*.

Kranton, R.(1996). "Reciprocal Exchange: A Self-Sustaining System.," *American Economic Review* **86**: 830-51.

Munro, A. et R. Sugden (2003). "On the Theory of Reference-dependent Preferences," *Journal of Economic Behavior and Organization* **50**: 407-428.

Poteete, A. et E. Ostrom (2004). "In Pursuit of Comparable Concepts and Data about Collective Action," *Agricultural Systems* **82**: 215-232.

Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics," *American Economic Review* **83**: 1281-1302.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press: Cambridge, MA.

Sandbu, M.E. (2008). " Axiomatic Foundations for Fairness-motivated Preferences," *Social Choice and Welfare*, à paraître.

Segal U. et J. Sobel (2007a). "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," *Journal of Economic Theory* **136**: 197-216.

Segal, U. et J. Sobel (2007b). "A Characterization of Intrinsic Reciprocity," *International Journal of Game Theory*, à paraître.

Sen, A. K. (1997). "Maximization and the Act of Choice," *Econometrica* **65**: 745-779.

Sobel J. (2005). "Interdependent Preferences and Reciprocity," *Journal of Economic Literature* **XLIII**: 392-436.

Sobel J. (2008). "Do Markets Make People Selfish?," *miméo*.

Tversky A. et D. Kahneman (1991). "Loss Aversion in Riskless Choice: a Reference-dependent Model," *Quarterly Journal of Economics* **106**: 1039-1061.

Vernon, L. S. (1962). "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy* **72**: 923-955.