LEARRY GAGNÉ

JON ELSTER ET LES LIMITES DE LA RATIONALITÉ: CHOIX RATIONNEL ET NORMES DANS L'EXPLICATION DES COMPORTEMENTS INDIVIDUELS.

Mémoire
présenté
à la Faculté des études supérieures
de l'Université Laval
pour l'obtention
du grade de maître ès arts (M.A.)

Département de science politique
FACULTÉ DES SCIENCES SOCIALES
UNIVERSITÉ LAVAL

MAI 1997

© Learry Gagné, 1997.

of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-25580-8

Canada

RÉSUMÉ

La théorie du choix rationnel est l'un des outils de recherche les plus utilisés en scienc politique contemporaine. Elle propose d'analyser les phénomènes sociaux par le comportements des individus; ceux-ci étant supposés agir rationnellement, en maximisant leu utilité sous diverses contraintes. Ce mémoire est consacré à une étude critique de la théorie du choix rationnel. Suivant les recherches du politologue Jon Elster, il sera question d l'introduction d'une théorie des normes sociales servant à l'explication de phénomènes pou lesquels le choix rationnel apparaît peu convaincant. Ces critiques seront appliquées au jeu du Dilemme du Prisonnier, une mise en situation sociale qui continue toujours de bafouer l'analys rationnelle. Deux hypothèses seront mises à l'épreuve, d'abord il s'agit de préciser jusqu'à que point le choix rationnel peut analyser correctement le Dilemme et ensuite de voir si une théori des normes sociales permettrait de combler les lacunes.

TABLE DES MATIÈRES

RÉSUMÉ i

TABLE DES MATIÈRES ii

CHAPITRE I: INTRODUCTION 1

CHAPITRE 2: STRUCTURES ET LIMITES DE LA THÉORIE DU CHOIX RATIONNEL 10

I - Épistémologie du choix rationnel 11

I.1 - L'individualisme méthodologique 11

I.2 - L'action individuelle 12

I.3 - La rationalité 15

II - Éléments de la théorie elstérienne 18

II.1 - La théorie restreinte 18

II.2 - La théorie étendue 25

III - Limites de la théorie du choix rationnel 31

III.1 - Indétermination et inadéquation 31

III.2 - Alternatives à la théorie du chox rationnel 36

IV - Conclusion 40

CHAPITRE 3: SOLUTIONS RATIONNELLES AU DILEMME DU PRISONNIER 43

I - La théorie de l'utilité 43

I.1 - Utilité et rationalité 43

I.2 - Utilité ordinale et cardinale 45

I.3 - Intensité des préférences 46

II - La théorie des jeux 48

II.1 - Notions de base 48

II.2 - Le Dilemme du Prisonnier 52

III - Solutions rationnelles au Dilemme du Prisonnier 54

III.1 - Deux joueurs, un coup 55

III.2 - Plusieurs joueurs, un coup 60

III.3 - Deux joueurs, plusieurs coups 66

III.4 - Plusieurs joueurs, plusieurs coups 71

IV - Conclusion 79

CHAPITRE 4: MOTIVATIONS EXTRA-RATIONNELLES ET SOLUTIONS ALTERNATIVES AU DILEMME DU PRISONNIER 81

I - Qu'est-ce qu'une norme sociale ? 82

I.1 - Normes sociales et individu 82

I.2 - Normes sociales et collectivité 87

II - Les normes comme alternative au choix rationnel 91

II.1 - Les "motivations mixtes" de Elster 91
II.2 - Alternatives au modèle elstérien 96
III - Une typologie des motivations 101
III.1 - Normes sociales, personnelles et émotions 101
III.2 - Solutions normatives au Dilemme du Prisonnier 104

CHAPITRE 5: CONCLUSION 109

BIBLIOGRAPHIE 112

CHAPITRE 1

INTRODUCTION

La théorie du choix rationnel représente un courant de pensée dominant des sciences sociales anglo-saxonnes depuis le milieu du XXe siècle. Son application la plus fructueuse fut à l'origine l'économie, mais d'autres disciplines comme la science politique et la sociologie l'ont adoptée comme instrument de recherche empirique, autant au niveau descriptif que prescriptif. Ses prémisses clairement en faveur de l'individualisme et de la notion de responsabilité ont fortement influencées la philosophie politique et les théories de la justice. Ses ramifications sont telles aujourd'hui que l'on peut la percevoir plus comme une espèce d'idéologie, une proto-théorie à partir de laquelle d'autres théories plus spécifiques prennent naissance. Etant donné que le premier chapitre de cet ouvrage comporte une définition très détaillée de cette théorie, nous nous contenterons ici d'une présentation succincte.

Tout d'abord, qu'entend-on par « choix rationnel » ? Précisions avant tout que les applications de la théorie ne sont supposées s'effectuer qu'au niveau *individuel* tout en remarquant qu'il demeure possible de parler d'agrégats d'individus en supposant qu'ils se comportent *comme* des individus. Donc, les notions de « choix » et de « rationnel » suggèrent une prise de décision individuelle et l'existence de décisions rationnelles et irrationnelles. Ultimement, il s'agit pour la théorie d'évaluer, pour une fin donnée, les différents moyens à la portée de l'individu et de sélectionner le meilleur, attribuant une valeur d'utilité à chacune des alternatives. Monroe (1991 : 4) pose quatre postulats fondamentaux au choix rationnel : 1) les acteurs poursuivent un but, 2) ces buts émanent de l'intérêt particulier des acteurs, 3) le comportement des acteurs équivaut à un choix conscient et 4) les phénomènes sociaux peuvent toujours être expliqués par les comportements des acteurs.

Les origines de la théorie du choix rationnel dans l'histoire de la philosophie remonte très loin, jusqu'à Aristote même (Cudd 1993 : 103). L'action rationnelle aristotélicienne se nomme *délibération*, la recherche du meilleur moyen pouvant conduire l'individu à une fin

désirée. Comme son influence fut énorme et ce dans de multiples branches de la philosophie nous nous concentrerons sur les origines plus récentes, dont les formulations s'apparenten d'assez près avec la forme actuelle de la théorie. L'un des tout premiers penseurs à se servir de notions de choix et de maximisation d'utilité fut Adam Smith. Monroe résume ainsi la méthod de Smith : « In its purest form, it refers to behavior by an individual, be it a person, a firm, or political entity, designed to further the actor's perceived self-interest, subject to informatio and opportunity costs » (Monroe 1991 : 1). Cette définition respecte les trois premier postulats, mais pas le quatrième qui ne s'imposera que plus tard.

L'intérêt particulier demeure toutefois une notion trop vague pour qu'une quelconqu théorie s'en inspirant puisse être opérationnelle. L'évolution de la théorie devra passer par un définition de l'utilité. La première fut proposée au XIXe siècle par les utilitaristes « classiques » représentés par J. Bentham et J.S. Mill. Pour eux, l'utilité est une mesure du bonheur et d l'absence de souffrances ; la maximisation de l'utilité correspond donc à une recherche d bonheur absolue (Mill 1978 : 251-53 ; Kymlicka 1990 : 12-13). Ici l'utilité est *cardinale*, c'es à dire qu'elle permet de quantifier le niveau de bonheur, et *interpersonnellement comparable* donc permettant de déterminer si l'individu A est objectivement plus heureux que B. Vers la fi du siècle, l'école marginaliste en économie va expurger la théorie de cette conception hédoniste en précisant que l'agent cherche à maximiser ses choix en vue d'une fin subjective qui peut être autre que le bonheur. L'utilité de chaque option devient alors relative et se mesure par son coût d'opportunité par rapport aux alternatives. La révolution marginaliste consistera alors à réuni les notions de rationalité et d'utilité (Cudd 1993 : 106). Alors que la micro-économie se développe rapidement, les théoriciens recherchent des formulation de plus en plus simples. En 1900, V. Pareto remet en question la cardinalité des utilités et propose que l'on retienne l'*ordre* des préférences et non sa *quantité* de bonheur ou de valeur marginale (Cudd 1993 : 108). Plus tard, J. Von Neumann et O. Morgenstern vont élaborer les axiomes de la théorie « moderne » de l'utilité, encore largement valides aujourd'hui. Nous verrons ces axiomes avec la définition de la théorie des jeux. Précisions seulement que dans la théorie moderne, la cardinalité refait son apparition comme préférence sur des loteries d'options mais que l'utilité n'est plus

interpersonnellement comparable[1]. Jumelée à l'idéal-type wébérien de la rationalité instrumentale[2], la théorie individuelle du choix rationnel est désormais complète.

Une fois les éléments de la théorie bien en place, les chercheurs s'intéresseront à l'interdépendance des acteurs en société et créeront pour l'occasion la théorie des jeux. L'étude des relations entre actions individuelles et société ne date pas d'hier. La célèbre « lutte de tous contre tous » de T. Hobbes s'explique par le constat que la maximisation de l'utilité personnelle (représentée par la liberté) engendre l'anarchie un résultat désastreux pour tous. La solution de Hobbes à ce problème sera l'institution d'un souverain ou Léviathan et l'abandon de la liberté individuelle en retour de la paix. Cette idée de « bon pour chacun, mauvais pour tous » ainsi que la justification hobbesienne d'une puissante autorité hantera longtemps les philosophes. Elle sera formalisée en théorie des jeux par le Dilemme du Prisonnier. La première élaboration articulée de la théorie des jeux en économie provient de l'ouvrage de Von Neumann et Morgenstern (1953). Quelques années plus tard, R.D. Luce et H. Raiffa (1957) introduiront la théorie en sciences sociales, déclenchant ainsi une avalanche de recherches empiriques et conceptuelles modélisant des situations politiques sous forme de jeux. Une étude statistique de contenu de la revue *American Political Science Review* révèle que les articles inspirés du choix rationnel passèrent de quasi-inexistant en 1957 à 5% de la revue dans les années '60, 20% en '70 et plus de 35% en '80 (Green, Shapiro 1995 : 97).

La popularité de ce nouveau courant de pensée va chambarder la conception des sciences sociales. Dans les décennies '60 et '70, les chercheurs vont s'adapter au modèle des sciences économiques tout en délaissant peu à peu les autres disciplines (Almond 1991 : 37). Parmi les plus célèbres notons A. Downs (1957) qui analysa le processus électoral comme un

---

[1] Le sujet de la comparaison d'utilité entre agents refait régulièrement surface. L'argument principal contre l'idée est que chaque individu possède sa propre échelle de valeurs incompatible avec celles des autres. Beaucoup encore de nos jours, dont A. Gibbard (1986) et D. Davidson (1986), ne sont pas convaincus de l'impossibilité de modéliser des échelles comparables d'utilité.

[2] Weber fut l'un des premiers à reconnaître le rôle des désirs et des croyances dans la détermination de l'action. Aussi, son élaboration des idéaux-types a contribué à la formalisation du concept de rationalité. Mais son défaut fut de concevoir les idéaux-types comme des lois purement causales, l'équivalent social de la causalité en sciences pures (Rosenberg 1988: 27).

échange de votes contre des politiques publiques dans un « marché » politique, ou G. Becker (1976) qui qualifia de nombreuses relations humaines - dont le mariage et la criminalité - en termes de coûts/bénéfices. Le Dilemme du Prisonnier fit son entrée en science politique lorsque M. Olson (1971) démontra que dans bien des cas, un bien public ne pouvait être produit par un groupe composé d'individus rationnels car chacun souhaite que les autres le produise pour ainsi en jouir gratuitement.

A partir des années '70 toutefois, les chercheurs rationalistes commencèrent à douter des prétentions à l'universalisme de la théorie du choix rationnel et à dénoncer la domination des méthodes de l'économie sur les sciences sociales. Les critiques prirent toutes sortes de formes. D'abord, A.O. Hirschman (1970) distingua l'action économique de l'action politique lorsque les participants veulent exprimer un désaccord : dans la sphère économique les acteurs se retirent du marché (*exit*) alors qu'en politique ils s'organisent en groupes de pression (*voice*). Le critère d'égoïsme fut remis en question par plusieurs, notamment par A. Sen (1977). J. Harsanyi (1990) proposa une théorie générale tripartite du comportement rationnel comprenant la théorie Von Neumann-Morgenstern de l'utilité, la théorie des jeux et l'éthique qu'il définit par la maximisation de l'utilité moyenne de tous (Harsanyi 1990 : 278). Enfin, E. Ullman-Margalit (1977) ouvrit la porte à une interprétation de certains jeux à l'aide de normes. Ce courant « révisionniste » (Monroe 1991 : 3) semble suggérer que les difficultés inhérentes à la méthode économique ressortent peu à peu de l'ombre maintenant que l'euphorie s'est estompée (Griffith, Goldfarb 1991 : 46). C'est dans ce courant que Jon Elster va amorcer sa contribution à la théorie du choix rationnel.

L'intérêt de Jon Elster pour la théorie moderne du choix rationnel fut initié curieusement par deux philosophes classiques que l'on n'associerait guère au mouvement rationaliste. Après avoir déposé en 1971 une thèse de doctorat sur la production et la reproduction économique chez Marx, Elster publie son premier ouvrage, *Leibniz et la formation de l'esprit capitaliste* en 1975. Là il tentera de coller l'analyse rationaliste, en particulier la théorie des jeux, sur les théorèmes exprimés par Leibniz et aussi Descartes au passage. Il affirmera que Leibniz est le précurseur de l'analyse mathématique des sciences

sociales (LFEC : 122), qui demeurera lettre morte jusqu'à la parution de l'ouvrage capital de Von Neumann et Morgenstern, *The Theory of Games and Economic Behavior*. Elster exprime sa foi sans bornes en ce courant de pensée lorsqu'il voit en ces deux théoriciens des jeux les Newton des sciences sociales (LFEC : 121) !

La première critique sérieuse du choix rationnel arrive un an plus tard en 1976, avec *Boudon, Education and the Theory of Games* et *Some Conceptual Problems in Political Theory*. Une constante entre ces deux articles : l'identification de la sous-optimalité et des conséquences inattendues comme faillites possibles de la rationalité. Cette dernière concerne l'information imparfaite pouvant provoquer des erreurs de jugement. La sous-optimalité va se révéler plus déterminante ; elle apparaît dans un jeu lorsque les participants ne gagnent pas le maximum possible (résultat Pareto-inférieur à l'optimum) tout en respectant à la lettre les canons de la rationalité, dont l'exemple parfait est le Dilemme du Prisonnier. L'enjeu chez Elster est de taille car selon lui, « Politics is the study of ways of transcending the Prisoners' Dilemma » (CPPT : 249).

Avec *Logic and Society* (1978), Elster pose les premières balises de son épistémologie ; il y fera figurer les notions d'intentionnalité et de causalité. En choix rationnel l'individu est doté d'intentions, c'est en effet ce qui confère la signification du terme « choix ». Elster va insérer dans ce modèle deux contraintes causales : la causalité « sub-intentionnelle » affectant les désirs et les croyances de l'individu et la causalité « supra-intentionnelle » où l'action de l'un entre en conflit avec celles des autres (LS : 158). Cette dernière est déjà reconnue par le choix rationnel, il s'agit de la théorie des jeux. La première toutefois est originale, et Elster y consacrera énormément d'efforts par la suite. Il est aussi question dans cet ouvrage du comportement normatif, qu'il décrit aux antipodes du choix rationnel et affichant une double caractéristique causale/intentionnelle : les normes - provenant de la société - causent l'action mais l'individu conserve un certain pouvoir intentionnel sur celles-ci (LS : 161). Pour cet ouvrage, Elster n'ira pas plus loin sur le sujet.

Par la suite Elster se lance dans trois ouvrages complémentaires sur la causalité sub-intentionnelle qui le rendra célèbre : *Ulysses and the Sirens* (1979), *Sour Grapes* (1983) et *Solomonic Judgments* (1989). *Ulysses* traite du *character planning*, les moyens de lutte rationnelles à la disposition de l'individu contre ses propres comportements irrationnels. Mais plus important pour nous reste le chapitre consacré aux limites du choix rationnel qui est en fait une version augmentée de la critique de CPPT. Elster s'en prend particulièrement aux tendances universalistes qui stipulent que le choix rationnel est en mesure d'expliquer n'importe quel comportement humain ou de fournir des conseils justes dans toutes les situations. Ces limites formeront la première ébauche d'une théorie « étendue » du choix rationnel par opposition à la théorie orthodoxe ou « restreinte ». Cette dichotomie revient de façon beaucoup plus élaborée dans *Sour Grapes*. Ici Elster élabore son modèle de l'individu, largement inspiré de Donald Davidson[3]. Il situe la théorie restreinte dans la cohérence logique du comportement alors que la théorie étendue devra s'intéresser au « bien-fondé » du comportement. La distinction tourne autour de l'analyse à l'interne des motivations de l'action, interdite en théorie restreinte. Elster reprend la même discussion sous d'autres termes dans l'article *Rationality* (1982). On retrouve aussi dans SG des études exhaustives sur plusieurs formes d'irrationalité. *Solomonic Judgments* prendra la forme d'un manifeste pour l'utilisation de mécanismes aléatoires en choix social lorsque nous ne pouvons compter sur la rationalité pour des raisons théoriques. Cette même période verra aussi la parution de *Explaining Technical Change* sur la philosophie de l'explication avec comme exemple les théories du progrès technique. Bien que ce sujet soit assez éloigné du choix rationnel, c'est là que l'on retrouve la description la plus complète de son épistémologie des sciences sociales.

Avec l'introduction au collectif *Rational Choice* (1986) dont il est l'éditeur, tout aura été dit sur la théorie du choix rationnel. Il répétera la même chose en langage simplifié dans son manuel de classe, *Nuts and Bolts for the Social Sciences* (1989). L'intérêt pour la pensée de Elster prend une toute autre tournure avec la publication quatorze ans plus tard de sa thèse de doctorat, complètement révisée on s'en doute, qui porte maintenant le titre *Making Sense of Marx* (1985). La réaction chez les philosophes marxistes sera forte et les critiques virulentes.

---

[3] Lire en particulier les cinq premiers essais de Davidson (1980).

Elster aurait semble-t-il péché gravement en insinuant que la pensée de Marx s'apprête à l'analyse rationaliste et qu'il est possible de réduire les intérêts de classe à de simples jeux ! Ce livre, pourtant le plus étoffé de sa bibliographie, sombrera dans l'oubli par la suite, en partie parce que les marxistes ne le retiendront pas et aussi car Elster ne lui donnera pas suite, n'écrivant plus aucun article sur le marxisme sauf pour répondre occasionnellement à ses détracteurs.

Elster va s'engager dans la tâche de trouver un moyen d'arriver à la coopération dans un Dilemme du Prisonnier. Il élabore une stratégie en trois volets : se servir du choix rationnel d'abord et en cas de faillite, rechercher dans deux autres domaines, la sociologie et la psychologie. Les outils conceptuels pertinents pour ce type de tâche sont respectivement la théorie des normes sociales et le comportement émotif. De ces deux alternatives, celle des normes sociales recevra la plus grande attention. Elle fera l'objet de l'ouvrage *The Cement of Society* (1989) et de l'article complémentaire *Norms of Revenge* (1990). On y retrouvera une analyse élaborée des liens entre normes et rationalité et une croyance manifeste en la puissance explicative de ces deux théories prises conjointement dans l'étude de l'action collective. Les émotions jouiront d'un traitement moindre ; Elster ne proposera pas de solutions « émotives » au Dilemme mais tentera plutôt de faire la part des choses entre « agir rationnellement » et « agir sous le coup de l'émotion ». En d'autres termes, l'analyse demeurera individuelle sans déborder vers le social. On retrouve cette étude dans l'article *Sadder But Wiser* (1985). En 1994 paraîtra une tentative d'unification des trois théories, *Rationality, Emotions and Social Norms*. En ce qui concerne le problème fondamental de l'action collective, ce texte n'est pas convaincant mais il replace les choses en perspective en précisant davantage le rôle des émotions. En fait, ce texte concerne surtout les émotions car Elster semblait insatisfait de *Sadder But Wiser* comme il l'indique dans *Solomonic Judgments*.

Le paradigme classique de la théorie des jeux en science politique est le Dilemme du Prisonnier. C'est la modélisation d'une situation où l'utilité la plus élevée pour tous est inaccessible si les acteurs cherchent à maximiser leur gains personnels chacun de leur côté. Pour atteindre l'utilité maximale, les acteurs devront s'unir pour un effort collectif.

On a vu précédemment des exemples de ce problème : l'état de nature chez Hobbes, les bien publics de Olson. D'autres phénomènes sociaux se prêtent à merveille à l'analogie du Dilemme, comme le désarmement nucléaire[4], l'investissement dans la recherche scientifique[5], etc. La clé du Dilemme est qu'il est toujours *irrationnel* pour un individu de coopérer. Et pourtant, lorsque nous rencontrons une situation réelle s'apparentant à ce jeu, on remarque que dans bien des cas, les participants ont atteint la solution coopérative avec succès. En réalité, nous ne vivons pas l'état de nature hobbesien (ni sa solution extrême du Léviathan), les biens publics sont produits en grande quantité, la science se développe rapidement, etc. Il en est tout autrement du désarmement, mais voyons là une preuve que les prédictions du choix rationnel continuent d'être pertinentes !

Ces particularités du Dilemme du Prisonnier soulève deux importantes interrogations qui serviront de toile de fond à notre travail : 1) Comment en arrive-t-on à une action collective rationnelle ? et 2) La théorie du choix rationnelle est-elle adéquate pour expliquer les instances concrètes d'action collective ? Ces questions seront le sujet des deuxième et troisième chapitres respectivement. Le second chapitre sera entièrement consacré à une élaboration de la théorie du choix rationnel actuellement en vigueur, y incluant les critiques de Elster et d'autres. Dans le troisième chapitre, nous explorerons les diverses propositions des théoriciens des jeux, car le développement de la théorie formelle des jeux a permis d'élaborer quelques situations particulières où l'action collective devient rationnellement atteignable. Nous aurons par la suite à juger si ces propositions se révèlent suffisantes et satisfaisantes pour expliquer l'action collective. Jon Elster prétend que non, mais nous allons chercher une réponse par nous-même. Le chapitre suivant exposera les alternatives de Elster au choix rationnel pouvant servir à une analyse plus juste de l'action collective. Comme nous en avons déjà fait mention, Elster place sa

---

[4] Les nations ont un intérêt collectif dans le désarmement mais personne ne veut amorcer le processus. Un accord unanime pourrait résoudre le problème mais encore là, il sera toujours rationnel de conserver quelques bombes en secret.

[5] Ce cas s'applique surtout pour les entreprises privées faisant partie d'un même secteur d'activité. Le progrès scientifique est bien sûr profitable à tous, mais chacun cherche à profiter des innovations des autres à moindre coût plutôt que d'investir des sommes considérables dans la recherche.

foi dans les normes sociales comme principale alternative. Nous verrons, avec plusieurs autres théoriciens de même acabit, comment une théorie structurée des normes pourrait surpasser la théorie du choix rationnel dans le domaine précis du Dilemme du Prisonnier.

CHAPITRE 2

STRUCTURE ET LIMITES DE LA THÉORIE DU CHOIX RATIONNEL

Les sciences sociales constituent un domaine plutôt étrange des sciences, car on y retrouve des théories parfois radicalement opposées sans qu'aucune d'elles ne jouisse de la main haute sur les autres, du moins à long terme. Elles se distinguent par leurs axiomes fondamentaux qui posent chacun le statut de la personne et de l'environnement social en des termes différents. La théorie du choix rationnel repose sur les principes de l'individualisme méthodologique et du caractère intentionnel de l'action individuelle. A l'aide du concept de rationalité, un principe découlant des deux précédents, elle prétend expliquer, évaluer et même prédire le comportement humain. Elle embrasse un vaste domaine, et des divergences théoriques apparaîtront au sein même de la théorie. Ses applications multiples demandent aussi des théorèmes ajustés aux champs particuliers d'étude.

Dans son esquisse schématique du choix rationnel, Jon Elster va partager la théorie en deux domaines : la théorie restreinte aux axiomes simples et directs s'inscrivant dans le mouvement objectiviste en sciences sociales et la théorie étendue, plus ouverte et plus prétentieuse au prix de théorèmes parfois nébuleux. Elster se situe lui-même dans la mouvance étendue, critiquant vertement les tenants d'une discipline trop près des sciences exactes pour refléter toute la complexité humaine. Il croit la version étendue de la théorie capable de mener à bien de nombreuses analyses sociales, mais il la considère quand même d'application limitée. Il manifestera le désir de compléter ou même de remplacer le choix rationnel par d'autres outils des sciences sociales lorsque cela s'avérera nécessaire. La première section de ce texte portera sur l'épistémologie de la théorie du choix rationnel. La seconde mettra aux prises les théories étroites et larges, suivant Elster avec un penchant favorable pour cette dernière. Cet exposé se terminera sur les limites internes et externes de la théorie et les suggestions de Elster et d'autres théoriciens pour y remédier.

I - Epistémologie du choix rationnel

La théorie du choix rationnel en sciences sociales se propose avant tout d'expliquer le phénomènes sociaux par les actions individuelles rationnelles. Une première approximation du contenu de la théorie et des attentes des chercheurs l'employant consiste naturellement à définir les concepts de base tout juste mentionnés. D'abord, le recours à l'*individu* comme unité explicative confirme une adhésion claire à l'individualisme méthodologique. Ensuite l'explication par l'*action* individuelle demande une base théorique de ce que constitue justement une action. Cette notion porte en elle les critères d'intentionnalité et, chez certaines écoles de pensée, de subjectivité. Si ces deux concepts, individu et action, constituent le squelette du choix rationnel, la notion de *rationalité* apparaît comme son coeur. Tentons maintenant de définir avec plus de précision ces termes, tout en tentant de cerner les positions particulières de Elster.

I.1 - L'individualisme méthodologique -

L'individualisme méthodologique constitue l'acte de foi des théoriciens rationalistes. Il ne peut être dérivée d'aucune autre conception de la société[1], tandis qu'à partir d'elle il est possible d'extrapoler vers l'étude de l'action, de la responsabilité, etc. Pour Elster l'individualisme méthodologique est une « vérité triviale » (NB : 13). Il la définit ainsi : « (...) the doctrine that all social phenomena - their structure and their change - are in principle explicable in ways that only involve individuals - their properties, their goals, their beliefs and their actions » (MSM : 5). L'individualisme méthodologique (« anti-collectivisme ») ne doit pas être confondu avec l'atomisme social ou la souveraineté complète de l'individu (« anti-structuralisme », Hindess 1988 : 36). La seconde conception réfère à l'homme totalement libre des diverses forces sociales. Lorsque Elster mentionne que l'explication ne met seulement en cause que les « propriétés » des individus, il semble adhérer aux deux variantes à la fois. Comme nous le verrons plus loin, c'est cependant loin d'être le cas. Soulignons

[1] Sauf peut-être la tendance générale de la science contemporaine à toujours chercher les "micro-fondations" d'un événement. Cela correspond à toujours pousser du plus général vers le plus précis et à chercher à éliminer les délais de temps entre *explanans* et *explanandum* (ETC: 23).

simplement que pour Elster, les propriétés internes au sujet peuvent être influencées de l'extérieur. Il sera toujours question d'un individualisme descriptif ou explicatif en choix rationnel, jamais (autant que possible) d'un individualisme moral (Kolm 1986 : 82-83). L'individualisme méthodologique ne présuppose donc pas une philosophie libérale de l'homme et de la société.

I.2 - L'action individuelle -

La question fondamentale en philosophie de l'action est l'explication de l'action individuelle à partir des dispositions de l'individu et des contraintes externes. La théorie du choix rationnel accorde la primauté aux dispositions. Pour bon nombre de théoriciens - Elster inclus - le modèle de Donald Davidson (1980) de l'action constitue la meilleure approche. Celui-ci est tellement commun dans la littérature que son emploi devient presqu'un automatisme ; il remonte en fait au syllogisme pratique d'Aristote qui considérait la délibération comme une méthode rationnelle de détermination des moyens et des fins (Cudd 1993 : 103). Il existe chez Davidson une multitude d'états mentaux permettant le déclenchement d'une action. Il les regroupe en deux catégories principales : les désirs (ou *pro attitudes*) et les croyances. Les deux « lois » générales reliant ces deux attitudes à l'action se lisent comme suit :

> « C1. *R* is a primary reason why an agent performed the action *A* under the description *d* only if *R* consists of a pro attitude of the agent towards actions with a certain property, and a belief of the agent that *A*, under the description *d*, has that property » (Davidson 1980 : 5).
>
> « C2. A primary reason for an action is its cause » (Davidson 1980 : 12).

Donc, il existe un lien causal entre d'une part les désirs (D) et les croyances (C), et l'action d'autre part. Ce lien s'avère difficile à conceptualiser car il diffère beaucoup de la causalité physique ordinaire. D'abord, les désirs et les croyances ne sont pas en eux-mêmes des événements ; ce sont à proprement parler des états mentaux reliés (causalement) à certains événements sans y être coextensifs. A ceci Davidson répond que ces états orientent, coordonnent notre façon d'agir alors que l'arrivée d'information sur le monde extérieur

déclenche la chaîne causale de l'action en opérant un changement dans les désirs et les croyances (Davidson 1980 : 13). De plus, nous ne pouvons directement évaluer objectivement ou « mesurer » les raisons de l'action, car les mesures ne s'appliquent qu'aux phénomènes de nature extensionnelle. Les désirs et les croyances ont un contenu propositionnel intensionnel, c'est à dire qu'on ne peut substituer une proposition par son équivalent exact sans changer la signification de l'action[2]. Davidson qualifie le modèle de « quasi-intensionnel » car l'action étant observable, nous pouvons nous en servir pour en déduire sa cause (Davidson 1980 : 5). Cette reconstruction à rebours de la chaîne causale se nomme *rationalisation* (fig. 1)[3].

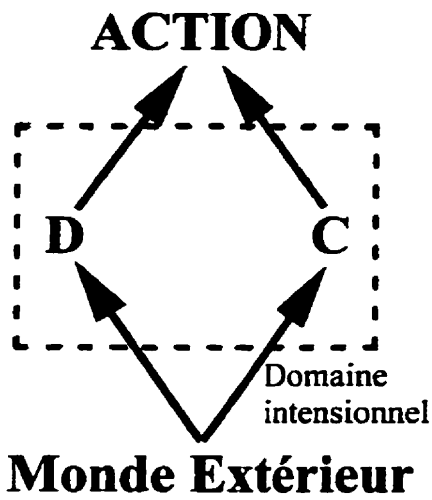


Fig. 1

Certaines lois causales couvrent les actions et leurs raisons, mais nous n'avons pas espoir d'en révéler un jour la nature exacte. En rationalisant l'action, nous ne pouvons découvrir qu'une raison ayant causé celle-ci tandis qu'en réalité, une multitude de raisons furent à son origine. Il faudrait, comme première condition de l'élaboration de (vraies) lois causales,

---

[2] Un exemple impliquant les croyances: Paul croit que Jackie Robinson a débuté sa carrière professionnelle de baseball à Montréal. Robinson fut aussi le premier Noir à jouer dans les ligues majeures. Si Paul ne sait pas cette information, la substitution d'équivalent devient impossible car la phrase "Paul croit que le premier Noir à jouer dans les ligues majeures a débuté sa carrière professionnelle de baseball à Montréal" est fausse.
Les croyances sont donc de nature intensionnelle.

[3] Pour reprendre l'exemple précédent, la rationalisation nous permet d'attribuer à Paul la croyance que Robinson fut le premier Noir à jouer dans les ligues majeures. Comme nous le verrons plus loin, la théorie restreinte du choix rationnel ne permet pas l'étude des motivations précises de l'individu. Celle-ci se satisfait de n'importe quelle motivation correspondant aux conséquences observables de l'action. Dans un effort d'objectivité maximale, la théorie restreinte "extensionnalise" les désirs et les croyances.

pouvoir cerner et identifier toutes ces raisons, ce qui nous est impossible (Davidson 1980 : 15-16). Pour qu'une action puisse être qualifiée d'intentionnelle, le but de l'action doit être voulu par l'individu, c'est à dire que l'on retrouve dans ses désirs et ses croyances une représentation des conséquences de l'action que ces états mentaux causent.

L'adoption d'un modèle intentionnel tel que celui-ci va irrémédiablement cantonner la théorie du choix rationnel dans la méthode herméneutique des sciences, avec des conséquences déterminantes sur son rôle en sciences sociales. En termes simples, l'herméneutique est la méthode hypothético-déductive appliquée aux comportements intentionnels (Follesdal 1979 : 320, ETC : 15). Donc, en choix rationnel tout comme en sciences « naturelles », nous allons nous servir d'hypothèses de base et de déductions logiques avec en plus, l'hypothèse d'intentionnalité. C'est ici que le choix rationnel diffère fondamentalement des sciences naturelles, car le critère essentiel de l'intentionnalité est son intensionnalité, ayant comme conséquence l'impossibilité de mesurer ses composantes, les désirs et les croyances. Toutes les variantes du choix rationnel, en particulier la science économique, se retrouve dans un étau, à la fois visant l'élaboration de théories naturalistes hautement prédictives et se butant immanquablement à l'incommensurabilité des désirs et des croyances. Et pourtant, la science économique comme les autres sciences sociales ne peut exister sans postuler l'intentionnalité de l'agent. C'est la raison pour laquelle les tenants du choix rationnel doivent en général abandonner leur visées prédictives pour se concentrer sur la description et l'explication[4] (Rosenberg 1992 : 150-151).

Elster pose ainsi le rôle fondamental des sciences sociales « The goal of the social sciences is the liberation of man » (LS : 158). Cette formule grandiloquente peut porter à confusion : par « libération » Elster signifie l'abandon de théories purement causales de l'action

[4] "En général" car la place de la notion de prédiction en sciences sociales est un interminable sujet de débats. Rosenberg semble convaincu qu'aucune théorie fondée sur l'intentionnalité ne peut prédire quoi que ce soit de significatif, mais il admet que l'économie ait réussi à quelques reprises dans ce domaine. Elster, quant à lui, prétend que la prédiction constitue une "exception" en sciences sociales sans préciser plus à fond sa pensée (NB: 6-7). Ou encore cette citation de Follesdal: "(...) one can *normally* not predict the exact outcome in the humanities and social sciences" (1979: 336, italiques rajoutées).

humaine pour des théories intentionnelles ; autrement dit un passage de l'homme-robot à l'individu responsable. La méthode herméneutique n'exclut pas cependant l'explication causale[5] ; celle-ci demeurera l'explication privilégiée des actions non intentionnelles, sujet de prédilection chez Elster comme nous le verrons plus loin.

I.3 - La rationalité

Comme nous l'avons vu précédemment, le modèle d'action de Davidson devient intentionnel lorsque l'individu se représente les conséquences de son action avant d'agir. On dira que ce même individu agit *rationnellement* si l'action qu'il entreprend correspond à la meilleure façon, selon lui, d'atteindre le but représenté. A partir de cette définition sommaire, il ne s'agit que d'y rajouter quelques prémisses concernant la motivation des individus et nous obtenons la théorie du choix rationnel en bonne et due forme. Le reste de ce chapitre sera entièrement consacré à l'analyse des prémisses de la rationalité. Mais d'abord, il serait utile de clarifier le rôle de la théorie. Elster lui reconnaît trois types d'usages : normatif, descriptif et explicatif.

D'abord, le rôle *normatif* revient à se servir de la théorie comme prescription. Ici, la théorie stipule comment le sujet *devrait* agir pour recevoir le sceau de la rationalité. Ceci implique que nous soyons en possession d'un modèle non tautologique de l'action rationnelle, un modèle qui définit donc le comportement irrationnel. Pour ce faire nous n'avons d'autre choix que de définir les caractéristiques intensionnelles du modèle. La répugnance innée (et justifiée jusqu'à un certain point) des chercheurs naturalistes (ou positifs) à s'engager dans les sables mouvants des conditions de rationalité aux désirs et aux croyances les porte à mettre plutôt l'emphase sur le rôle *descriptif* de la théorie. Elle sert dans ce cas à élaborer des régularités statistiques reliant des situations sociales à des comportements individuels (ou vice versa) ; pour cette raison la théorie descriptive porte souvent dans la littérature le nom de « choix rationnel béhavioriste ». Est rationnel ce qui est régulier, sans recours aux états

---

[5] "Herméneutique" doit alors être compris au sens large, à ne pas confondre avec la méthode interprétative taylorienne qui semble exclure toute causalité dans l'explication sociale (Little 1991: 68-74).

intentionnels de l'agent. Par exemple, l'observation générale en micro-économie que dans un marché libre plus le prix d'un bien monte, moins les gens en consomment permet de qualifier cette réaction particulière de rationnelle, et son inverse d'irrationnel. Ce qui cause cette réaction n'est d'aucun intérêt pour le chercheur descriptif[6]. Le descriptif analyse l'action *ex post* alors que le normatif détermine *ex ante* la rationalité d'une action. A mi-chemin entre le normatif et le descriptif se situe le rôle *explicatif* de la théorie du choix rationnel. Tout comme le descriptif, l'explicatif analyse les régularités des comportements individuels avec en plus la prétention d'en connaître la *cause* à l'interne, d'après les critères de la théorie normative. Evidemment, cette méthode est plus complète que sa cousine descriptive et on pourrait même lui attribuer une supériorité absolue, ne serait-ce du caractère foncièrement évasif de l'intentionnalité...

Dans la conception elstérienne du choix rationnel, le rôle normatif demeure le mode privilégié d'exploitation théorique. La théorie explicative vient en second lieu, car elle est contingente à une théorie normative suffisamment élaborée (IRC : 2, SJ : 1, Harsanyi 1990 : 272). Sans la rejeter complètement, Elster préfère ne pas adopter la conception béhavioriste de la rationalité. Etant donné qu'il croit fermement à la possibilité d'évaluer les désirs et les croyances individuelles, il ne voit pas comment une telle conception pourrait lui être utile (IRC : 2). Aussi, le mécanisme d'association de la rationalité à la régularité ressemble étrangement à une variante de la sélection naturelle en biologie, un mécanisme qui peut fonctionner relativement bien en micro-économie dans un marché libre mais beaucoup plus difficilement dans les autres domaines des sciences sociales (RAT : 2). La définition plus précise de W. Riker et P. Ordeshook (1973 : 13-15) est particulièrement instructive : partant de la prémisse de régularité comme rationalité, ils tentent de discerner l'intentionnel de l'aléatoire dans les comportements individuels. Pour ce faire, ils proposent deux procédures, les préférences révélées (*revealed*) et postulées (*posited*). Dans un contexte donné, la première

---

[6] Il faut toutefois spécifier que le modèle général de rationalité forme toujours la base de la théorie. Les "behavioristes sociaux", biens qu'animés par un désir d'objectivité, seront toujours à la merci d'un concept essentiellement subjectif, l'intentionnalité, ce qui amène Rosenberg à affirmer que le descriptif n'est que le normatif (ou l'explicatif) formalisé à un très haut degré (1988: 65-74).

permet de déduire le but visé par l'individu à partir de l'observation de ses choix et du respect des « règles de logique », qui n'est rien d'autre que le modèle intentionnel complété par l'axiome de transitivité (que nous verrons plus loin), alors que la seconde prédit le choix d'un individu à partir d'un double postulat de son but et des règles de logique. La procédure complète d'analyse descriptive prend alors la forme d'une *description* des buts par les préférences révélées et ensuite d'une *prédiction* de l'action par les préférences postulées utilisant l'*output* des préférences révélées comme buts. Voyons maintenant comment la théorie gère une situation où les individus n'agissent pas en fonction de ce qui a été prédit. Le constat est accablant : « If such choices do not occur, then we are at a loss to discover whether the fault lies in our attribution of logic or our attribution of goals » (p. 14). Etant donné qu'en théorie descriptive il est fort peu recommandé d'évaluer les caractéristiques mentales, les auteurs proposent la stratégie suivante : « (...) it is often thus necessary to move back and forth between methods » (p. 15), soit tenter une nouvelle description des buts. Par conséquent, il n'est pas difficile d'accepter, avec Elster, Harsanyi et d'autres, l'analyse introspective de l'action si elle peut nous permettre d'échapper à ce cercle vicieux.

C'est cette recherche des causes de l'action exigée par la théorie explicative qui va pousser Elster à privilégier la méthode des *mécanismes* en sciences sociales. Un mécanisme est un lien causal spécifique, élaboré à partir d'un événement qui a déjà eu lieu (donc observable), et suffisamment général pour que son pouvoir explicatif puisse persister dans différentes situations (PP : 3-5). C'est donc un théorème plus ou moins *ad hoc*, tout le contraire d'une théorie générale dont Elster qualifie de « rêve illusoire » en sciences sociales (PP : 2). Autre différence marquée avec la théorie générale, le mécanisme peut s'avérer faux dans certaines situations sans se voir infirmé pour autant. La nature précise de la méthode des mécanismes n'est pas claire ; Elster lui-même demeure incapable d'en fournir une définition satisfaisante. Pour la situer épistémologiquement, tout ce que nous avons repose sur ces quelques mots : « It is less than a theory, but a great deal more than a description » (PP : 5). A tout le moins sommes-nous maintenant convaincu de l'incrédulité de Elster face à la théorie descriptive !

Chez Elster les mécanismes servent de support au modèle intentionnel. Ils peuvent opérer aux niveaux de la *causalité psychique* (ou *sub-intentionnelle*) et de la *causalité sociale* (ou *supra-intentionnelle*). Brièvement, la causalité psychique implique la formation « irrationnelle » des désirs et des croyances alors que la causalité sociale concerne les conséquences inattendues de l'action dues à la présence d'autres acteurs, un peu à la manière de la « main invisible » (ETC : 19-20). Elster attribue cette conception intentionnelle/causale de l'être humain à Karl Marx, qu'il résume ainsi dans son livre consacré au grand penseur : « Individuals are as it were caught in the middle, between the psychic causality that shapes their aims and desires and the social causality that thwarts and frustrates them » (MSM : 24).

II - Éléments de la théorie elstérienne

Le système de l'action rationnelle tout juste mentionné, jumelé avec quelques axiomes de cohérence, compose le fondement de la théorie du choix rationnel. Les économistes en quête d'une conception simple et universelle de l'humain ne vont adopter que les axiomes absolument nécessaires. L'essentiel du programme de recherche de Elster se résume à compléter cette théorie *restreinte* du choix rationnel par des postulats plus réalistes - et bien entendu plus complexes - de l'action rationnelle tout en demeurant dans la sphère du choix rationnel, composant ainsi la théorie *étendue*. La distinction restreinte / étendue se situe à la frontière de l'explication externe / interne du comportement individuel : en théorie étendue on se permet une introspection, une évaluation de la nature des désirs et des croyances, ce qui demeure interdit en théorie restreinte.

II.1 - La théorie restreinte -

Comme nous l'avons vu précédemment, l'action rationnelle correspond au choix du meilleur moyen pour arriver à une fin donnée. Bien entendu, ce choix devra compter avec les particularités de la situation présente. Il importe d'abord et avant tout d'éliminer les possibilités irréalisables avant de choisir. Nous avons là une définition un peu plus précise du choix rationnel, que Elster qualifie de la théorie des deux filtres (NB : 13). L'ensemble des possibilités doit passer à travers deux filtres et au bout de ce processus, nous obtenons (idéalement) un choix unique. Le premier filtre est constitué des contraintes externes logiques, physiques et

économiques qui élimine les possibilités irréalisables. Ce jeu réduit se nomme *ensemble faisable* (*feasible set*). Le second filtre servira ensuite à la sélection d'une possibilité parmi celles de l'ensemble faisable. En choix rationnel ce mécanisme est celui de la maximisation de l'utilité, mais comme nous le constaterons plus loin ce n'est pas le seul possible. Il est même possible que le premier filtre soit si contraignant que l'ensemble faisable ne contienne qu'un seul élément ; dans ce cas le second filtre n'est évidemment pas requis. En choix rationnel cette situation est exceptionnelle mais pas impossible ; par contre la théorie structuraliste - bien représentée par la sociologie française et le marxisme classique - postule cette conception du premier filtre comme modèle général d'explication. Le structuralisme diffère radicalement du choix rationnel car il rejette la notion d'action intentionnelle pour la remplacer par un automatisme causé par les forces sociales[7].

Le modèle des filtres ne constitue pas une explication adéquate de l'action, car il considère que l'individu a une connaissance complète, juste et immédiate de l'ensemble faisable alors que dans la réalité ce n'est presque jamais le cas. L'ensemble faisable ne peut causer l'action, mais il peut causer une *croyance* des possibilités. Nous avons donc comme causes de l'action un ensemble de croyances sur l'ensemble faisable ainsi qu'une fonction préférentielle - soit les désirs - permettant de choisir une action parmi d'autres. Nous voilà donc de retour au modèle de Davidson, mais avec un élément supplémentaire (fig. 2).

---

[7] Pierre Bourdieu en est un de ses plus ardents défenseurs. Pour lui, l'action individuelle (*habitus*) est causé par des pulsions (*libido*) entièrement déterminées par le contexte social (*illusio*). Voir Bourdieu (1994: 149-167) pour une critique acerbe des deux postulats fondamentaux de l'utilitarisme qu'il définit ainsi: 1) la "raison consciente" et 2) la réduction de toute motivation en valeur monétaire. Bien que ce dernier n'ait peu d'intérêt car très mal compris au départ, le rejet complet du premier postulat en vaut le détour.

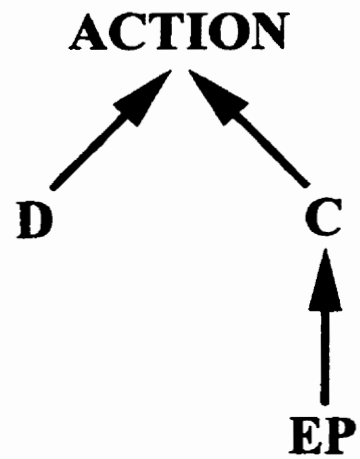


Fig. 2

L'ensemble des possibilités (EP) devient l'ensemble faisable par la médiation des croyances. La simplicité caractérisant toute la théorie restreinte du choix rationnel, le modèle le plus simple des croyances postule l'infaillibilité, que celles-ci soient exactement équivalentes à l'ensemble complet des options réalisables. C'est la thèse de l'information parfaite. A partir de là il n'y a plus rien à redire sur la rationalité des croyances, sauf peut-être l'exigence du respect des lois fondamentales de la logique (SG : 4-5 ; Riker, Ordeshook 1973 : 29-31).

Regardons maintenant du côté des désirs. Ici par contre, la rationalité ne se préoccupe pas de leur origine. Bien qu'il soit reconnu que la socialisation joue un rôle dans leur détermination, le modèle formel de la rationalité considère par mesure de simplicité les désirs comme *sui generis* ; seul un minimum de *cohérence* importe. Voici les quatre axiomes de cohérence généralement reconnus à travers la discipline. D'abord, la *transitivité* des préférences représente le critère minimal de rationalité (SG : 6). Un classement sur un ensemble d'options (*a*, *b*, *c*) est transitif si *a* est préféré à *b*, *b* à *c* et *a* à *c*. Les préférences sur cet ensemble doivent aussi être *complètes*, c'est à dire qu'une préférence (ou une indifférence) est exprimée pour chaque paire d'options. Un autre critère formel est la *continuité*, que Elster définit ainsi : si *a* est préféré à *b*, et que l'une des deux options subit une variation très faible, ceci ne devrait pas suffire à renverser la préférence (SG : 8-9). On retrouve une définition plus précise dans un autre texte : « (...) if a commodity bundle is reduced in one component, then it is always

possible to increase another component so as to leave the consumer equally well off as before the reduction » (RAT : 122). Enfin, un critère d'*égoïsme* est imposé aux paires d'options : l'individu préférera toujours les options qui satisfont le plus son intérêt particulier. Bien que l'égoïsme soit assez près de la réalité dans la modélisation du comportement rationnel en économie, il demeure difficile à définir proprement, nous y reviendrons plus loin. Elster affirme toutefois que l'égoïsme jouit d'un plus grand prestige explicatif que l'altruisme car bien qu'un monde composé uniquement d'égoïstes est logiquement possible, il n'en est pas de même d'un monde parfaitement altruiste. La satisfaction du bienfaiteur dépend des plaisirs égoïstes de ses pairs (NB : 53-54). Toutefois, l'égoïsme n'est jamais une condition nécessaire de la rationalité ; elle joue tout bonnement un rôle de simplification. Ces quatre critères tracent le portrait du « sujet économique » maintes fois décrit par les économistes. Selon Elster, nous n'avons pas absolument besoin de tous ces critères ; seule la prémisse de transitivité suffit, qui est la marque du « sujet rationnel » (SG : 10).

Les désirs peuvent aussi être évalués temporellement ; dans ce cas, des options immédiates sont comparées à d'autres qui se matérialiseront plus tard. Le seul critère de rationalité formelle s'appliquant à ces préférences est la *cohérence temporelle*, qui stipule que les distributions de ressources pour l'avenir doivent être réalisables lorsque celles-ci deviennent effectives (SG : 7, RAT : 124). Comme nous le verrons plus loin, la théorie étendue des désirs permet une application plus étendue de la rationalité intertemporelle. Viennent ensuite les préférences sur des probabilités d'options, dont les axiomes relèvent de la théorie de l'utilité cardinale ; nous y reviendrons.

Gary Becker, un des plus fervents partisans de la théorie restreinte, résume celle-ci comme suit :

> « All human behavior can be viewed as involving participants who maximize their utility from a stable set of preferences and accumulate an optimal amount of information and other inputs in a variety of markets » (Becker 1986 : 119).

La simplicité, voire l'idéalisme de cette conception du choix rationnel se retrouve dans une restriction particulière des désirs (stables) et des croyances (optimales) au nom de la généralisation du modèle d'explication. Becker croit cette fixation des caractéristiques internes de l'individu nécessaires à la recherche empirique ; de cette manière le chercheur ne sera pas porté à réinterpréter une hypothèse infirmée en postulant un changement de préférence ou une information imparfaite, phénomènes impossibles à analyser dans la théorie restreinte (Becker 1986 : 110-112). En fait, cette version du choix rationnel va prospérer ou périr selon la solidité de ces deux prémisses.

Du côté des préférences, stabilité n'est pas synonyme d'immuabilité. En général on postule un lien entre les préférences et l'ensemble faisable de sorte que lorsque cet ensemble change, les désirs doivent s'ajuster en conséquence. C'est le seul mécanisme de transformation admis en théorie restreinte. En termes béhavioristes, seul un changement de stimuli peut provoquer un changement de réponse. Par le fait même, la théorie restreinte ne peut expliquer un changement de préférences alors que l'environnement demeure stable. Ce ne serait qu'une limite acceptable de la théorie si ce phénomène demeurait sporadique et isolé. Mais c'est justement là où de sérieux problèmes surgissent : ce phénomène est relativement *commun*. Prenons le simple exemple d'une préférence exprimée sur des pommes et des oranges. Face à un panier de ces deux fruits, Linda préfère les pommes mais le lendemain, elle penche vers les oranges et pourtant, le panier n'a pas changé. Linda a-t-elle un comportement anormal ? Irrationnel ? L'intuition nous suggère que non alors que la théorie restreinte peut la condamner car elle viole l'axiome de transitivité *ou* l'accepter en postulant des « externalités cachées », un changement trop subtil de l'ensemble des possibilités pour que le chercheur puisse en tenir compte, mais que le sujet a internalisé consciemment ou non (Becker 1986 : 112). Dans le premier cas la théorie rejette un comportement humain apparemment normal tandis que dans le second cas elle élimine le recours à l'axiome de transitivité, et en cela à une conception rationnelle de l'individu. Pire encore, la théorie ne peut, de par sa nature même, distinguer entre une violation de transitivité et une externalité cachée (Rosenberg 1988 : 70). En plus des changements non provoqués de préférences, la théorie restreinte ne peut expliquer des phénomènes tels le comportement traditionaliste (une préférence immuable face à une

transformation de l'ensemble faisable) ou le manque de volonté (préférer A à B et pourtant agi en fonction de B). L'introduction de la notion de *valeur* morale, culturelle, etc. comm déterminant des désirs peut contribuer à régler de telles indéterminations, mais alors nou entrons dans la théorie étendue du choix rationnel.

Les croyances optimales correspondent à une collecte d'information sur le circonstances de la décision contrainte par le coût d'acquisition (Becker 1986 : 111) En d'autres termes, l'optimum est atteint lorsque le coût marginal d'information supplémentair dépasse l'utilité (marginale) contribuant à la clarification de la situation. Le coût d'opportunit du temps va aussi jouer dans la détermination de l'optimum. Ainsi, la théorie restreinte prétenc expliquer les décisions causées par des croyances incomplètes, un phénomène particulièremen répandu dans les domaines où l'environnement se révèle très complexe ; nous n'avons qu'à songer aux décisions des gestionnaires de grandes entreprises par exemple.

Les critiques de la théorie restreinte vont jeter leur dévolu sur la signification réelle du coût marginal d'une « unité » d'information. Selon Sidney Winter (cité dans US : 135), la détermination des coûts et des bénéfices marginaux de l'information demande une information sur ceux-ci que nous ne pouvons logiquement pas posséder, car l'acquisition d'information sur l'information conduit inéluctablement à une régression infinie. Seule l'*intuition* peut trancher ce noeud gordien. Herbert Simon apportera une modification substantielle à la théorie afin de modéliser l'intuition dans la formation des croyances tout en voulant sauvegarder la théorie restreinte : la notion de *satisfaction.* La prémisse de maximisation de l'utilité est alors remplacée par le concept correspondant de satisfaction, la recherche de l'action non pas maximisante, mais satisfaisante dans les circonstances (Simon *et al.* 1992 : 37). Toutefois Simon ne propose pas de méthode de détermination des niveaux de satisfaction (*aspiration levels*) des individus. Nous voilà donc de retour à une conception béhavioriste extra-rationnelle de la décision : lorsqu'un individu ne maximise pas son utilité, c'est donc qu'il satisfait un niveau quelconque d'aspiration seulement connu de lui-même. Les chercheurs peuvent aussi procéder avec des niveaux postulés de satisfaction, un peu à la manière des préférences postulées mentionnées plus haut. Il faut alors reconnaître qu'une telle

détermination demeure très subjective, pouvant se situer n'importe où entre le désintéressement et la maximisation de l'utilité, sans compter la tentation naturelle du chercheur faisant face à une infirmation de son hypothèse de réajuster le niveau postulé plutôt que de se questionner sur la rationalité du geste de son sujet. Une critique semblable à celle portée sur les désirs s'applique ici : la théorie restreinte ne peut faire la différence entre une satisfaction (rationnelle) et une tentative ratée de maximisation (parfois irrationnelle[8]).

Ces critiques à l'endroit de la version restreinte de la théorie du choix rationnel sont très sévères. Alors que Rosenberg considère que l'adoption du postulat de stabilité des préférences prive la théorie de toute prétention au modèle rationnel de l'individu, Elster affirme de son côté que le paradoxe de l'information exclut la notion d'optimalité de la théorie :

> « This means that when we leave the thin theory of rationality, the link between rationality and optimality is completely broken. Rational behavior can be characterized as optimizing only - or at most - with respect to *given* beliefs about the world, but the principles of rationality governing belief acquisition cannot be spelled out in terms of optimization » (SG : 18).

En définitive, la théorie étroite du choix rationnel s'avère très idéaliste et logiquement chancelante (Rosenberg 1992 : 117). Elle n'est utile que dans certains cas spécifiques, comme les situations de décisions hautement stéréotypées où l'attribution *a priori* de désirs et de croyances spécifiques ne pose pas de problèmes sérieux (NB : 35). En général la recherche rationaliste en sciences sociales s'effectue à partir de postulats appartenant à la théorie étendue (Green, Shapiro 1995 : 101). En formalisant *à l'interne* le concept de rationalité, nous élevons notre compréhension de la rationalité des comportements humains. Etant donné l'impondérabilité fondamentale des proposition de nature intensionnelle, Elster ne rejette pas la théorie restreinte du revers de la main, mais confère une primauté certaine aux explications plus substantielles (IRC : 27).

---

[8] Dans l'analyse d'une "mauvaise" décision, il faut faire la part des choses entre un échec causé directement par le paradoxe de l'information optimale et une simple incapacité de jugement. Seule cette dernière est vraiment irrationnelle (Riker, Ordeshook 1973: 30-31).

II.2 - La théorie étendue -

Passer à la théorie étendue du choix rationnel confère certes plus de réalisme, mais le prix à payer en termes de complexité peut s'avérer lourd. Il faudra toujours garder à l'esprit les principes fondateurs de la théorie - l'explication de l'action par le concept de rationalité - afin d'éviter une trop grande expansion des domaines et l'avènement d'une théorie globale et foncièrement inapplicable empiriquement. La théorie étendue devra se contenter de mécanismes de formation des désirs et des croyances nous permettant de distinguer avec suffisamment de clarté le rationnel de l'irrationnel.

Pour Elster, les désirs demeurent en dernière instance une entité *sui generis*, la source même de la décision libre et intentionnelle. Il y donc danger de confusion entre évaluation rationnelle et éthique des désirs. Mais comme les désirs ne sont pas immunisés aux influences externes, l'espoir d'élaborer des critères d'évaluation strictement rationnels persiste. Elster en retient deux particulièrement révélateurs selon lui : l'optimalité et l'autonomie (IRC : 15). Ce premier critère renvoie à l'idéal béhavioriste de correspondance entre les désirs et l'ensemble faisable ; en choix rationnel il devient synonyme de maximisation de l'utilité[9] à l'intérieur de l'ensemble faisable. Aux dires de Elster, « A rational desire is one which is optimally adjusted to the feasible set » (IRC : 15). Deux possibilités d'irrationalité se présentent, les désirs à l'extérieur de l'ensemble faisable et la non-maximisation. Cette dernière ne devrait pas provoquer de controverses ; c'est ce qui confère au choix rationnel sa caractéristique optimisante (Nida-Rümelin 1994 : 125). Pour ce qui est de l'influence de l'ensemble des possibilités, le critère semble trop sévère. Prenons par exemple le cas d'un dissident vivant dans un régime totalitaire et exprimant un désir de liberté et de démocratie. Bien que son désir se situe nettement à l'extérieur des actions possibles, devient-il irrationnel pour autant ? Tenant mordicus à sa conception du bien-fondé des désirs, Elster prétend que

---

[9] Pour ne pas sombrer dans une définition trop matérialiste (voire monétariste) des désirs, Elster emploie fréquemment le terme *bonheur* (*happiness*) plutôt qu'*utilité*. Il ne faut toutefois pas interpréter la théorie elstériene comme une variante de l'utilitarisme à la Bentham.

oui[10] (NB : 40). L'autonomie des désirs signifie *grosso modo* l'absence de forces causales « aveugles » agissant sur ceux-ci, autrement dit les désirs reposent sur « (...) a causal history with which we can identify ourselves » (IRC : 15). L'autonomie demeure, selon l'aveu même de Elster, un mécanisme extrêmement vague et foncièrement insaisissable. Il ne considère pas sa théorie étendue des désirs comme une amélioration suffisante de la théorie restreinte, mais il demeure confiant que cela représente un pas dans la bonne direction (IRC : 14).

La théorie étendue promet aussi de retirer un autre bâton des roues de la théorie restreinte : le mécanisme de changement rationnel des préférences sous un ensemble faisable constant (Follesdal 1979 : 334). Ces mécanismes se révèlent multiples. D'abord, l'individu possède la capacité de réfléchir sur ses désirs et d'en modifier la nature, il a alors recours aux *méta-préférences*. Ce phénomène, que Elster appelle *intentional character planning*, peut expliquer la formation de préférences mais n'en garantit pas la rationalité car les mêmes difficultés qu'à l'origine ne font que se transposer à un niveau supérieur (SG : 21). L'*impulsion* représente un second mécanisme, une force mentale inconsciente recherchant le plaisir immédiat ; sa nature précise reste toutefois à déterminer (SG : 24-25). La voie la plus prometteuse demeure le concept de *valeur*, un ensemble de principes éthiques et moraux influençant les désirs et guidant l'individu dans ses choix. Plusieurs théoriciens ont proposé une théorie des valeurs dans l'explication des désirs, notamment Nida-Rümelin (1994), Little (1991), Hirschman (1985) et Brunner (1987). Harsanyi (1990) y voit même le complément idéal à la théorie des jeux. Les valeurs diffèrent des désirs en ce qu'elles sont indépendantes de l'ensemble faisable et qu'aucun critère de rationalité, sauf peut-être l'exigence de coexistence logique de plusieurs valeurs, ne s'y applique. Ce troisième membre intensionnel de l'esprit humain pourrait loger le concept de maximisation de l'utilité (Sen 1982 : 93-94) ou de méta-préférence (Hirschman 1985 : 8-9). Les valeurs, contrairement aux désirs, sont de

[10] La controverse s'accentue grâce à une conception trop vague de la nature de l'ensemble faisable. Si l'ensemble n'inclut que le présent immédiat, Elster a probablement raison; mais en y rajoutant l'espérance, soit une évaluation des probabilités d'un changement des opportunités, le comportement du dissident se rapproche beaucoup du rationnel. Prenons l'exemple d'un dissident soviétique: ses désirs seraient irrationnels durant le règne communiste, mais à un niveau certainement différent sous Staline que sous Gorbatchev.

nature délibérative - elles sont sujettes à changements simplement par la réflexion. Nous avo
là une explication plausible d'un changement de préférences.

Selon Elster, nous pouvons définir la liberté selon deux extrêmes : la satisfaction d
désirs subjectifs d'un côté (liberté active), le nombre d'opportunités substantielles à not
disposition de l'autre (liberté passive) (SG : 127). Bien entendu, ces deux conceptions peuve
entrer en conflit dans le cas d'un désir subjectif irréalisable. L'individu peut alors limiter s
désirs en fonction de ce qui est possible, selon une adaptation de type intentionnell
Il peut aussi faire primer la première conception et proclamer tous les désirs qui lui passe par
tête ; nous serions alors conduits à affirmer qu'un individu n'ayant que quelques désirs est plu
libre que celui plus ambitieux, ce qui n'a pas beaucoup de sens. Une préférence « libre » n
devrait être ni adaptive, ni infinie mais plutôt « autonome », soit une combinaison des deu
La définition de Elster de la liberté va comme suit : « (...) the degree of freedom depends on th
number and importance of the things that one (i) is free to do and (ii) autonomously wants t
do » (SG : 128). Le point (i) concerne le jeu des possibilités tel qu'existant dans la sociét
tandis que le second point concerne les préférences non adaptives. La définition d'un dés
rationnel devient alors une affaire de juste médiation entre l'autonomie et l'adaptation a
possible.

Le premier critère de rationalité des croyances concerne leur relation avec l'ensembl
faisable. Alors que la théorie restreinte demande une parfaite correspondance entre les deux, l
théorie étendue se révèle moins sévère en permettant les approximations, notamment par l
principe de satisfaction. Toutefois, la théorie exclut la simple *coïncidence* de la définition de l
rationalité. Il doit être démontré que l'ensemble faisable a causé la croyance (SG : 16)
Par exemple, on demande à deux personnes de nommer l'équipe gagnante du match de hocke
de la veille et elles ont tous deux raison, mais un interrogatoire révèle que la seconde ne savai
pas la réponse et avait choisi un club au hasard. Bien que les deux jouissent de croyance
correspondant au réel, seul le premier individu voyait ses croyances *fondées* sur le réel.

Nous pouvons évaluer les croyances par rapport au réel, c'est à dire l'information disponible. C'est ce que Elster nomme d'ailleurs la condition de « well-grounded beliefs » (RAT : 113). Il ne s'agit pas ici de confronter les croyances d'un individu avec le réel (autrement la croyance en Dieu serait toujours irrationnelle), mais plutôt d'examiner les liens entre l'information et les croyances (SG : 16). La question est de savoir si les croyances ont été suffisamment influencées par l'information disponible. L'irrationalité peut survenir aux deux extrêmes possibles : une cueillette d'information nulle ou infinie. La rationalité substantielle des croyances devient ainsi, dans les termes de Elster, une affaire de *jugement*, qu'il définit par « (...) the capacity to synthetize vast and diffuse information that more or less clearly bears on the problem at hand, in such a way that no element or set of elements is given undue importance » (SG : 16). Le jugement implique les croyances approximatives sur la situation et indique un certain lien causal réciproque des croyances vers l'information. Cette définition pour le moins ambiguë réfère au mécanisme d'acquisition des croyances, où il est impossible de connaître à l'avance la quantité optimale d'information nécessaire (SG : 18). En second lieu il faut s'assurer que l'information ait été proprement considérée. En somme, les phénomènes d'*illusion* et de *distorsion* de l'information peuvent causer des croyances irrationnelles (SG : 24, RAT : 113). Les limites d'acquisition d'information relèvent du domaine des désirs pour la simple raison que celles-ci ne peuvent être définies autrement. Si nous pouvions calculer à l'avance le coût de l'information à la marge, nous aurions une base objective pour l'établissement de limites et par le fait même il serait possible d'aspirer au niveau optimal d'information, mais comme nous l'avons vu plus haut ce n'est pas le cas. Les croyances agissent sur l'accumulation d'information par un mécanisme de *feed-back* : la recherche d'information nouvelle dépend de la quantité et de la qualité de celle déjà accumulée (IRC : 14). Incorporant la notion d'approximation, le schéma de l'action rationnelle prend maintenant la forme de la fig. 3, fréquemment reproduite dans le corpus d'Elster.

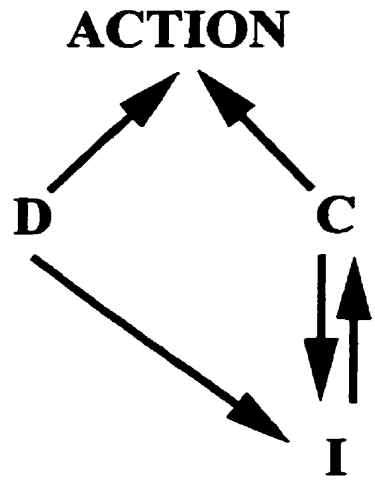


Fig. 3

La rationalité exclut aussi la possibilité pour un individu de générer lui-même ses croyances. « Vouloir croire » constitue un acte non seulement irrationnel, mais impossible : « (...) one cannot - conceptually cannot - at one and the same time believe something and believe that the belief has been adopted on noncognitive grounds » (SJ : 7).

Nous remarquons que les désirs et les croyances ne sont pas reliés dans le modèle de Elster. Ceci signifie qu'en aucun cas les désirs doivent affecter les croyances (sauf indirectement, dans le processus d'accumulation d'information), et vice versa. Aussi, ces deux éléments sont conjointement et également nécessaires à la prise de décision. Il existe par conséquent quatre instances d'irrationalité dans le processus de décision, que nous allons examiner à l'instant.

Le premier cas implique la domination[11] des désirs sur les croyances ou la réduction de dissonances, cesser de désirer l'inaccessible. C'est le phénomène « *sour grapes* » analysé par

[11] Ici, les termes "domination" et "influence" auront des significations particulières. On dira qu'un élément "domine" lorsqu'il a plus de poids que l'autre dans la prise de décision, alors qu'il "influence" l'autre s'il détermine directement sa composition. Dans le cas de l'influence, on trace une flèche entre les désirs et les croyances.

Elster dans l'ouvrage du même nom. Les désirs impulsifs ou trop axés sur la satisfaction d plaisirs immédiats se retrouvent dans cette catégorie, car ils ignorent la croyance des besoi pour l'avenir (NB : 22), comme dans la fable de la cigale et de la fourmi. En deuxième lieu no avons la domination des croyances sur les désirs, qui peut prendre la forme d'une trop gran confiance dans l'information disponible. Souvent, les conseils des firmes de courtage so suffisants pour pousser un investisseur à vendre ses titres, même si la compagnie en questi semble en bonne santé. Dans un même ordre d'idées, les croyances idéologiques ou religieus peuvent facilement prendre le dessus sur les désirs. Troisièmement, les désirs peuvent influenc les croyances, communément appelé les « voeux pieux ». Généralement, cette for d'irrationalité survient lorsque les désirs censurent une partie de l'information qui empêchera ceux-ci de se réaliser. L'individu réalise alors (ou tente de réaliser) ses désirs sans êt pleinement conscient de leurs conséquences. Le phénomène des voeux pieux ouvre la voie une influence déraisonnée des désirs sur le niveau d'information à accumule La volonté irrationnelle de vouloir accumuler de l'information jusqu'à ce que celle-ci no donne raison peut très bien situer le niveau d'information en-deça ou au-delà des limit acceptables définies au préalable (IRC : 20-21). Cette discrimination de l'information constitu un obstacle sérieux au bon fonctionnement du modèle rationaliste. Le dernier lien illégitime e l'influence des croyances sur les désirs, ou la reformulation des désirs en fonction d croyances. L'illusion fiscale représente un tel phénomène irrationnel : un individu n'acceptera jamais de payer de sa poche telle réforme gouvernementale, mais il accepte lorsque l'État l promet qu'il se financera autrement. En définitive, l'individu déboursera autant, mais de faço indirecte par les taxes (SJ : 20-24). La publicité et la propagande opèrent de la même façon.

Ce catalogue non exhaustif de mécanismes causaux fut élaboré par Elster dans l'espoi de pousser plus loin le domaine d'application de la théorie du choix rationne L'introspection exigée par la théorie étendue fait passer le statut de la théorie d'outil empiriqu à une conception particulière de l'individu-maximiseur se rapprochant de la philosophi Elster résume son pari comme suit : « (...) between the thin theory of the rational and the fu theory of the true and the good there is a room and need for a broad theory of the rational. (SG : 15 ; aussi Green, Shapiro 1995 : 123). Par la théorie étendue Elster veut aussi éviter qu

le choix rationnel sombre dans la tautologie et la vacuité, en s'efforçant de donner forme a rationnel et à l'irrationnel, un élément manifestement manquant dans le modèle béhavioriste.

III - Limites de la théorie du choix rationnel

Il semble que la classification rationnelle - irrationnelle du comportement humai permette à la théorie du choix rationnel de s'exprimer sur toutes les situations d'action Bien entendu, ce ne sera pas le cas. D'abord, la théorie n'est pas à l'abri de contradiction internes. La relation entre croyances et information en est un exemple frappant, mais il n'est pa seul ; c'est ce que nous verrons en analysant les possibilités d'indétermination de la théorie Les qualificatifs « rationnel » et « irrationnel » suscitent une confusion péjorative avec de termes comme « sensé » et « insensé », la définition d'une action intelligente ou stupide.. Les cas isolé d'irrationalité se prêtent sans trop de problèmes à ce jugement Un problème plus sérieux survient lorsque le non-respect des principes de rationalité s généralise. Un doute s'installe alors concernant la validité de la théorie, elle peut s'avére inadéquate à l'explication. Aussi, les axiomes particuliers de la théorie rendent particulièremen difficile l'analyse de certains phénomènes particuliers comme l'influence des forces sociales su le comportement humain par exemple. Nous énumérerons à la fin de ce chapitre quelque alternatives pratiques à la théorie du choix rationnel.

III.1 - Indétermination et inadéquation -

Le modèle de la rationalité individuelle que nous venons d'élaborer indique la formation d'une décision (ou d'une action) dans le cadre des « lois » de la théorie du choix rationnel. Bien entendu, ce modèle ne peut prétendre à l'explication universelle. Cette section sera consacrée aux défaillances possibles du modèle rationnel. D'abord, il existe certaine situations dans lesquelles le modèle ne pourra générer de décision à partir de ses déterminants. Ensuite, même si le modèle nous fournit une décision, il demeure possible que l'individu ne respecte pas ses prescriptions. Le premier cas se nomme indétermination ; le second, inadéquation. Etudions ces défaillances dans l'ordre.

L'indétermination survient en théorie du choix rationnel lorsque l'*output* du modèle cesse de fournir un résultat unique pour chaque situation donnée. La théorie s'en trouve alors affaiblie, mais pas inutile. en particulier si elle peut toujours exclure certaines alternatives de l'ensemble possible de solutions (SJ : 8). Deux types d'indétermination sont possibles ; lorsque le résultat n'est pas unique, il peut être soit multiple ou non-existant. Aussi, le problème peut survenir à n'importe quel niveau du modèle. Elster fait mention de quatre cas particuliers d'indétermination, que nous allons à présent examiner.

Tout d'abord, la théorie peut suggérer plusieurs actions pour une seule et même situation. C'est le cas notamment lorsque l'individu se retrouve devant un choix où il ne peut comparer les alternatives, comme par exemple avoir à choisir entre deux verres d'eau exactement semblables. Selon Elster, ceci ne constitue pas une faillite du choix rationnel car ce type de choix ne concerne même pas la théorie : « If two options are equally and maximally good, we don't need a theory to tell us how to choose between them » (IRC : 17). Comme nous le verrons au chapitre suivant, la théorie des jeux peut exprimer une indifférence entre plusieurs actions. Les stratégies dites « mixtes » constituent un moyen d'éviter la multiplicité des actions prescrites en les réunissant dans un calcul probabiliste unique. Par exemple, si la théorie révèle une indifférence entre *a* et *b*, elle aura à fournir deux solutions mais si elle les combine, par exemple, en « 50% *a*, 50% *b* » cette solution a la propriété d'être unique. Selon Elster, la stratégie mixte ne cadre pas dans un modèle rationaliste car elle remplace l'intention d'agir par un mécanisme (causal) aléatoire[12] (IRC : 18, NB : 106-107).

La théorie du choix rationnel peut aussi manquer à sa tâche et se révéler incapable de nous fournir une action optimale. Cette faillite peut survenir à trois niveaux, soit à la formation de la décision et des croyances, et à la cueillette d'information. Nous avons vu précédemment que pour être formellement rationnel, l'ordre de préférence doit être complet. Les théories restreintes du choix rationnel ayant recours aux fonctions d'utilité ne peuvent

[12] Cette critique des stratégies mixtes se révèle quelque peu ambiguë car l'agent conserve l'intention de se servir du mécanisme aléatoire. Mais demeure-t-il aussi *responsable* de l'action que celui qui n'a pas eu recours à la chance?

opérer lorsque cet ordre est incomplet, ce qui survient lorsque certaines paires d'alternatives sont incomparables. La définition de l'incomparabilité n'est toutefois pas elle-même formelle. Bien que le choix d'une voiture soit parfaitement - quoique difficilement - comparable, le choix d'un(e) époux(se) ou d'une carrière ne l'est pas car l'individu ne peut avoir en sa possession toute l'information nécessaire à un choix éclairé[13]. Les coûts et les bénéfices, surtout futurs, sont ici incalculables et l'expérience de l'individu dans le domaine fait défaut. Elster n'exige pas un ordre complet des préférences dans son modèle ; nous savons qu'il n'exige que la transitivité des préférences. Pour lui, la recherche absolue en tout temps d'un ordre complet est une forme d'hyper-rationalité (SJ : 9-10). Un autre critère de sélection doit alors remplacer le choix rationnel. Ce pourrait être par exemple la satisfaction ou même le hasard, mais toujours est-il que la théorie du choix rationnel demeurera muette sur la meilleure alternative à adopter (NB : 33).

L'action peut être impossible à déterminer à cause de croyances rationnelles inexistantes. C'est le cas lors de choix en situation d'incertitude. Ici, l'individu n'est pas en mesure d'appliquer une utilité aux alternatives, ni même une probabilité d'utilité[14]. Selon le modèle de l'action rationnelle, l'agent ne peut alors dériver l'action de son ordre de préférences, même complet et proprement conçu. La théorie du choix rationnel ne pourra générer de solution ; elle pourra seulement éliminer certaines alternatives selon un critère très vague : « Sometimes we are able to dismiss an option in the presence of another that, regardless of which state of the world obtains, has better consequences » (SJ : 11). Ce critère s'applique dans les cas où l'individu peut établir diverses valeurs d'utilité aux alternatives dans de multiples « univers parallèles » dont il ne sait absolument pas lequel surviendra. Celui-ci peut alors appliquer les principes de « maximin » ou de « maximax » dans sa prise de décision. Deux problèmes persistent toujours : d'abord, la stratégie « maximin » est aussi rationnelle que la

[13] La difficulté conceptuelle des comparaisons interpersonnelles d'utilité peut aussi représenter un problème d'incomparabilité (SJ: 9).

[14] Dans les situations de choix risqué - pleinement compatibles avec la théorie du choix rationnel - l'individu peut substituer l'utilité par une probabilité d'utilité. La théorie demande alors de maximiser l'utilité espérée.

« maximax » et il est impossible de trancher objectivement entre les deux. Ensuite, le modèl ainsi perfectionné peut très bien fournir plusieurs alternatives optimales.

L'interaction stratégique peut aussi empêcher la formation de croyances rationnelles En théorie des jeux, le phénomène d'optimum multiple dans une situation symétriqu (identique pour chaque joueur) nous rend parfaitement ignorants du choix de l'autre Sans coopération, la solution d'un tel jeu relève du hasard (SJ : 13, IRC : 19), nous y reviendrons. La dernière forme d'indétermination concerne un phénomène que nous avons déjà abordé dans les pages précédentes, l'impossibilité d'accumuler un niveau optimal d'information On ne peut espérer atteindre qu'un niveau « acceptable » d'information respectant certaine contraintes définies par une évaluation subjective de la situation.

Même en considérant la théorie du choix rationnel comme déterminée, c'est à dire qu'elle fournit une prescription unique pour une situation donnée, il demeure possible pour l'individu de violer les principes de la rationalité. La théorie souffre alors d'inadéquation ; ses solutions ne sont pas appliquées par les individus, ils se comportent alors de façon irrationnelle. L'inadéquation peut survenir à tous les niveaux du modèle rationaliste. Contrairement aux indéterminations, elle se matérialise à travers des liens causaux d'agent « fautifs ». En plus des mécanismes affectant parfois les désirs et les croyances dont nous avons fait état dans la théorie étendue, l'irrationalité peut survenir entre l'ordre formé de préférences et l'action primitive.

Tout en haut du modèle, au niveau de l'action, l'inadéquation survient d'abord et avant tout dans les cas de faiblesse de la volonté (*weakness of will*) : l'agent n'utilise qu'une partie de ses désirs dans la formation de ses préférences et de son action, alors qu'il existe consciemment en lui d'autres désirs suffisants à renverser sa décision. Qui plus est, l'action sous-optimale est intentionnellement, mais non rationnellement, entreprise : « When the weaker desires win out, it must be because they are in some sense stronger - not stronger as reasons, but stronger as sheer

psychic turbulence » (NB : 36)[15]. Ce phénomène viole une des condition de base de la rationalité, car les causes de l'action n'en sont plus les raisons (Davidson 1980 : 39-42). Le problème devient d'autant plus sérieux pour le choix rationnel que la faiblesse de la volonté tend à démontrer une dichotomie entre ordre (mental) des préférences et ordre (psycho-physiologique) des conséquences plus ou moins sensuellement agréables de l'action (Hindess 1988 : 51-54). Le calvinisme (*everyday calvinism*) contrevient aussi à la rationalité de l'action. Ici, l'individu tente d'agir sur les symptômes d'une cause en croyant à tort changer la cause, comme par exemple la personne qui cache son problème de drogue ou d'alcool en croyant s'en sortir ainsi. Certaines formes de superstition peuvent être ainsi comprises, comme chez l'étudiant qui réussit son premier examen alors qu'il portait une casquette bleue et qui par la suite arbore la même à tous les examens, croyant que ce geste lui apportera le succès (CS : 196-198). Enfin, la recherche d'états contingents (*self-defeating actions*), aussi appelés « excès de volonté », comme volontairement essayer d'oublier, de croire, etc., représente une autre forme importante d'irrationalité (IRC : 20-21, SG : chap. 2).

Ces limites internes de la théorie du choix rationnel délimitent avec plus de précision son champ privilégié d'application pour fins normatives et explicatives. Les phénomènes d'indétermination montre que même lorsque le *sujet* est rationnel, son *environnement* peut empêcher la théorie d'opérer correctement. Dans ces circonstances, la théorie n'est toutefois pas à rejeter : « Non-existence and non-unicity create difficulties for rational choice, but they do not totally invalidate it » (IRC : 19). L'utilité de la théorie persiste lorsque, faute de pouvoir cerner une action optimale, elle élimine certaines possibilités clairement irrationnelles. La leçon essentielle à retenir selon Elster est de reconnaître cette indétermination et de savoir quand abandonner la rationalité pour un autre mécanisme de choix. Une confiance aveugle en l'universalité du choix rationnel constitue en soi une irrationalité ! (NB : 36, SJ : vii) L'inadéquation va porter un coup beaucoup plus rude à la théorie. En admettant que les diverses formes d'irrationalité sont non seulement possibles mais bien présentes dans le

[15] Le phénomène d'incontinence demeure controversé en philosophie de l'action. La documentation sur le sujet est riche et une étude approfondie ne cadre pas dans les objectifs de cet exposé. Pour une critique élaborée de la conception de Elster, lire Rorty (1980).

comportement habituel de l'humain, il faudra trouver des alternatives au choix rationnel comme explication générale du comportement. La plupart des théoriciens actuels du choix rationnel ne sont pas aussi optimistes que Riker et Ordeshook, qui eux prétendent que l'inadéquation résulte le plus souvent d'erreurs de calcul et que « erreur n'est pas stupidité » (Riker, Ordeshook 1973 : 31). Les théoriciens vont alors s'ouvrir à d'autres domaines des sciences humaines pour tenter de sauver la théorie du choix rationnel.

III.2 - Alternatives à la théorie du choix rationnel -

Depuis les années '70 les courants de pensée en théorie du choix rationnel remettent sans cesse en question l'orthodoxie que représente la version restreinte/béhavioriste, grâce surtout à des publications comme Sen (1982) qui remit en question la maximisation de l'utilité et Ullman-Margalit (1977) qui proposa en force détails une théorie des normes sociales fondée sur la théorie des jeux. La décennie suivante verra une explosion de motivations supplémentaires ou substitutives à la rationalité instrumentale, dont Elster fut l'un des auteurs les plus prolifiques. Il importe de faire la distinction ici entre *mécanismes* causaux expliquant des comportements irrationnels - le sujet de la théorie étendue - et *théories* alternatives extra-rationnelles servant à expliquer des comportements insaisissables par une simple dichotomie rationalité - irrationalité. C'est ce que nous nous proposons maintenant d'analyser.

Un survol des alternatives les plus fréquemment proposées dans la littérature nous permet de dégager deux tendances majeures : le choix rationnel aurait besoin de théories issues de la *psychologie* et de la *sociologie*. Les avenues explorées par Elster abondent dans le même sens. Il empruntera une typologie de Tocqueville en classifiant l'ensemble des motivations de l'action en intérêts, passions et normes sociales (PP : 9 ; RESN : 21) correspondant aux domaines rationnel, psychologique et sociologique des sciences de l'homme. Elster a consacré quelques textes à l'étude des passions, ou *émotions* pour employer le terme qu'il privilégie. Ses positions sur le rôle des émotions dans la détermination de l'action individuelle manque de la clarté et de la concision de ses autres écrits ; son aveu de ne pas vouloir s'engager dans la psychologie, un domaine qu'il ne maîtrise pas (SJ : 30-31) prend ici une intrigante

signification[16]. En résumé, il situe les émotions dans le modèle d'action rationnelle comme déterminant des désirs, d'une manière étrangement semblable au lien entre information et croyances (fig. 4). « Agir sur le coup de l'émotion » résume assez bien l'influence des émotions sur les désirs. Le lien inverse représente la satisfaction ou la frustration d'un désir réalisé ou non. Les émotions ont des croyances particulières comme source, d'où la relation croyances-émotions (RESN : 34-35). Elles influencent aussi la formation des buts de l'action, un peu de la manière suivante : « Emotions provide a meaning and sense of direction to life, but they also prevent us from going steadily in that direction » (NB : 70). Enfin, elles pourraient venir compléter la définition de l'utilité (Almond 1991 : 49).

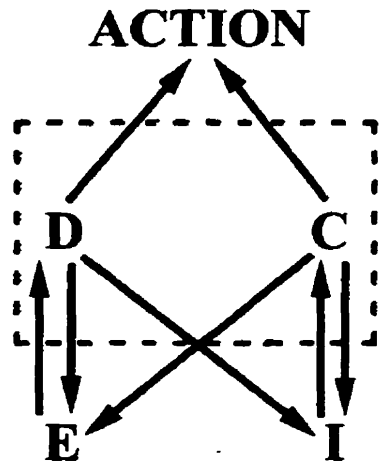


Fig. 4
Tiré de (RESN : 35)

L'action non-instrumentale - poser un geste simplement pour le plaisir de l'acte - défie la conception de l'individu maximiseur. Hirschman (1985 : 19-20) se sert des notions de *striving* et de *savoring* pour expliquer l'activisme politique chez certaines personnes et prétend qu'elles peuvent entre autres solutionner le paradoxe du vote. Pour Elster, il est clair que la jouissance d'une action dépend de son caractère substantiel : « It can indeed be highly

[16] Surtout que "Solomonic Judgements" (1989) fut publié quatre ans après "Sadder but Wiser" (1985), la première élaboration de sa théorie des émotions. Son aveu d'impuissance ne le résigna pas à l'abandon, car il revint à la charge dans "Rationality, Emotions and Social Norms" (1994) qui reprend essentiellement les mêmes idées.

satisfactory to engage in political work, but only on the condition that the work is defined by serious purpose which goes beyond that of achieving this satisfaction » (MF : 121). Cett définition découle de la théorie des états contingents (*by-products*)[17] : l'action politiqu entreprise dans le but d'une jouissance participative tournera rapidement court. Pour le momer nous pouvons nous passer d'une théorie non instrumentale car de telles actions peuvent êtr reclassifiées soit comme subtilement maximisante (d'une manière que le chercheur n'a pa remarqué) ou comme non-maximisante et émanant de désirs formés irrationnellement[18]. Si le buts de l'action sont fixés l'individu peut, en plus de rechercher la réalisation de ses buts profiter des plaisirs de la participation. Cette forme d'explication par l'activisme sera élaboré au prochain chapitre.

De la sociologie nous provient une autre conception de l'action non instrumentale d nature beaucoup moins controversée. Il s'agit de la *socialisation* de l'individu par l'éducation la culture ou autres mécanismes semblables. Les normes sociales, que nous élaborerons a troisième chapitre, font partie de cette catégorie. Les « forces sociales », contrairement à l rationalité, provoquent des actions sans en élaborer les conséquences ; l'individu agit parc qu' « il le faut » ou s'abstient parce que « ça ne se fait pas ». La simple observation de notr quotidien suffit à nous convaincre de l'étendue de son influence. Elles entrent dans le modèl rationnel à divers endroits, comme la formation des préférences (Kolm 1986 : 93), l détermination de l'utilité ou de l'échelle des valeurs (Almond 1991 : 48), le contenu d l'ensemble faisable (excluant les alternatives socialement inacceptables), l'identité de participants en théorie des jeux (lire Shepsle 1989 : 135-36 ; Johnson 1991 : 121-22 ; NB 147-158 sur le « néo-institutionnalisme »). Les théories sociologiques mettent en génér l'emphase sur l'ensemble faisable plutôt que les préférences ; elles ont tendance à considér

---

[17] C'est le sujet du second chapitre de "Sour Grapes". Un *by-product* est un état qui peut être visé intentionnellement mais qui de par sa nature même demeure impossible à atteindre de cette façon; vouloir oublier par exemple.

[18] Williams et Findlay (1981: 19) soutiennent qu'une théorie remplaçant la maximisation de l'utilité par un "besoin d'accomplissement" serait "anti-rationnelle" et plus plausible dans l'explication du comportement. Malheureusement, ils attaquent une vision strictement monétariste du choix rationnel, où l'utilité s'exprime toujours en argent. Bien entendu, rien n'est plus faux car dans toutes les versions de la théorie, les sujets sont libres de déterminer eux-mêmes les utilités à maximiser, en particulier chez Becker (1986).

l'individu comme moins doué d'intentionnalité (et de liberté) que le choix rationnel leu confère. Les préférences deviennent l'affaire de classes sociales et d'identité de groupe (US 138-39). Donc, les explications sociologiques et rationnelles d'un même phénomène vont for probablement varier. Nous ne nous étendrons pas plus sur le sujet des normes sociales ici, nou y reviendrons en détail plus tard.

La pression est forte pour compléter la théorie du choix rationnel par des théorie psychologiques et sociologiques. Il faut se rendre à l'évidence, l' « économisme » pur et dur e sciences sociales - pour employer un mot de Almond (1991 : 48) - n'est plus à la mode Ses résultats concrets demeurent insatisfaisants et la théorie étroite se révèle trop simpliste pa rapport au réel (Bunge 1990 : 5 ; Green, Shapiro 1995). Parfois les critiques se font incisives « Les 'ordres de préférences' individuels (ou 'fonctions d'utilité', ou autres notions indiquées sont donc la haute muraille derrière laquelle les économistes protègent leur paresse à apprendr de la psychologie (et de la sociologie) » (Kolm 1986 : 93). Les diverses théories en science sociales doivent reconnaître leurs propres limites :

> « Les hypothèses qui découlent de la théorie des choix rationnels seraient plus pénétrantes si la distinction entre l'action rationnelle et les autres modes de comportement était plus claire, et les tests empiriques seraient plus convaincants et plus instructifs s'ils étaient destinés à explorer les limites de ce que la théorie des choix rationnels peut expliquer » (Green, Shapiro 1995 : 126).

Deux derniers points concernant cet idéal d'harmonie en sciences sociales. D'abord, il n'est pas question pour les rationalistes d'abandonner la primauté épistémologique de la théorie du choix rationnel. Du côté normatif, la rationalité a le statut d'un principe de charité : la communication repose sur un préjugé en faveur de la rationalité de son interlocuteur (Davidson 1980). Il est futile de tenter d'exposer sa pensée à un autre si l'on ne le considère pas comme étant doué des mêmes capacités de rationalité que nous. Elle a aussi un autre statut, celui de norme, de règle de conduite : « First and foremost, rationality is privileged because we want to be rational » (SJ : 28). La rationalité est aussi privilégiée parmi les théories explicatives, car elle ne repose sur aucune autre théorie du comportement, alors que beaucoup de théories

dépendent d'un concept de rationalité (SJ : 29). Le second point concerne une caractéristique commune nécessaire avant toute tentative d'unification : les théories doivent relever de l'individualisme méthodologique (Kolm 1986 : 94-95). Le modèle rationaliste n'est pas fait pour être appliqué à des « acteurs collectifs » (SJ : 175-81).

Les voies s'ouvrant aux théoriciens une fois la « haute muraille » démolie sont riches et multiples. Un bon nombre d'entre eux vont contribuer au développement d'une théorie des normes sociales liée au concept de rationalité collective modélisée en théorie des jeux. Elster y fonde ses espoirs : « Among the alternatives to rational-choice theory, the (as yet undeveloped) theory of social norms holds out much promise. It is radically different from rational-choice theory, whereas the other alternatives are largely variations on the same consequentialist theme » (SJ : 35). Et pour conclure, voici la vision de l'avenir des sciences sociales de Elster, très révélatrice de ses nombreux combats :

> « Neoclassical economics will be dethroned if and when satisficing theory and psychology join forces to produce a simple and robust explanation of aspiration levels, or sociological theory comes up with a simple and robust theory of the relation between social norms and instrumental rationality » (US : 136).

IV - Conclusion

Ce chapitre fut un bref aperçu d'un sujet très vaste, la théorie du choix rationnel et ses limites explicatives. Suivant la philosophie de Elster et de nombreux théoriciens de ces dernières années, nous en sommes venus à la conclusion que l'entreprise simplificatrice de la théorie orthodoxe ou restreinte du choix rationnel laisse d'importantes questions en suspens. Au stade actuel de développement de la théorie, ces questions ne peuvent plus être ignorées et il semble inévitable que l'on sorte des axiomes du choix rationnel pour espérer trouver des réponses satisfaisantes.

Les phénomènes les plus importants résistant à l'explication restreinte sont la formation des désirs, le changement de préférences sous un ensemble faisable stable et l'accumulation d'information. A un niveau plus fondamental, la théorie restreinte adopte une définition

tautologique de la rationalité en ne reconnaissant pas suffisamment le comportement authentiquement irrationnel. La théorie étendue se propose de corriger ces aberrations en incorporant des mécanismes extra-rationnels à la détermination de l'action. D'abord, les désirs doivent résulter d'un équilibre réflexif entre la maximisation de l'ensemble faisable et la liberté individuelle d'action. De l'aveu même d'Elster toutefois, ce concept demeure peu convaincant. Ensuite, le problème de l'explication des changements de préférences peut être résolu par les méta-préférences, le comportement impulsif ou la notion de valeur. Le paradoxe de l'information optimale représente l'obstacle le plus fabuleux pour la théorie restreinte. Pour le contourner, la théorie étendue propose la satisfaction ou le jugement, deux mécanismes assez mal définis pour l'instant mais hautement prometteurs. Pour ce qui est du concept substantiel de l'irrationalité, Elster le situe dans les liens illégitimes entre désirs et croyances que l'on rencontre parfois chez l'individu.

L'élaboration de mécanismes extra-rationnels permet certes de sauver la théorie du choix rationnel, mais ils demeurent impuissants devant la double menace d'indétermination et d'inadéquation. Bien que l'indétermination ne soit pas un problème vraiment sérieux, l'inadéquation constitue une raison suffisante pour rejeter la théorie. A ce moment il faut se tourner vers d'autres théories susceptibles de fournir des alternatives viables et compatibles avec l'individualisme méthodologique du choix rationnel. Les deux champs d'études fréquemment mentionnés par les théoriciens - Elster compris - sont la psychologie et la sociologie. Le premier pourrait servir à définir les composantes du sujet rationnel, soit les désirs et les croyances, alors que le second permet l'analyse d'un phénomène primordial et pourtant largement ignoré par le choix rationnel, l'influence des forces sociales sur le comportement individuel. La théorie des jeux, un pur produit du choix rationnel, nous sert à expliquer une partie de la causalité sociale, l'interaction stratégique. Pour le reste des forces sociales, nous devons laisser de côté le choix rationnel.

Comme nous le verrons au prochain chapitre, le Dilemme du Prisonnier représente le cas le plus célèbre d'inadéquation de la théorie : ses prédictions ne sont pas toujours respectées dans la réalité. Par exemple, dans l'étude du vote populaire en régime démocratique, ce jeu

prédit une participation extrêmement faible des électeurs, ce qui n'est pas le cas
L'étude de M. Olson fit implicitement référence au Dilemme pour affirmer que dans de
groupes de bonne dimension, les biens publics ne sont jamais produits, une autre conclusio
infirmée par la réalité. Au quatrième chapitre, nous tenterons d'expliquer ces phénomènes pa
des théories alternatives, dont celle privilégiée par Elster, les normes sociales.

CHAPITRE 3

SOLUTIONS RATIONNELLES AU DILEMME DU PRISONNIER.

Avant de procéder aux propositions de réforme de la théorie du choix rationnel de Elster, il serait de bon usage de donner la chance au coureur et examiner minutieusement comment la théorie aborde le problème de l'action collective. L'outil conceptuel privilégié pour ce genre d'exercice demeure la théorie des jeux, qui correspond au choix rationnel appliqué aux situations d'interaction entre individus rationnels. Nous débuterons par une introduction essentielles aux prémisses sous-tendant la théorie des jeux, notamment le concept d'utilité et les représentations des possibilités sous forme de matrice ou de graphique. Viendra ensuite l'analyse proprement dite de l'action collective idéalement représentée sous la forme d'un jeu nommé le Dilemme du Prisonnier. Le but ultime de ce chapitre est de déterminer si la théorie du choix rationnel peut expliquer avec satisfaction les instances d'action collective que nous observons dans la réalité. Si tel est le cas, alors la théorie se révélera adéquate pour cette problématique particulière ; sinon il faudra rejoindre Elster dans sa volonté à compléter ou remplacer la théorie par quelque chose de mieux.

I - La théorie de l'utilité

I.1 - Utilité et rationalité -

D'après le modèle davidsonien de l'action, l'individu doit arrêter son choix sur deux éléments de l'action, le but visé (l'intention) et le moyen d'y parvenir (l'action). La théorie moderne de l'utilité s'intéresse particulièrement au premier. Elle permet de formuler une *fonction d'utilité* permettant au sujet de choisir une finalité appropriée par la *maximisation de l'utilité*, soit le choix de l'option ayant la valeur d'utilité la plus élevée.

La distinction entre maximisation de l'utilité et maximisation de l'efficacité est clairement exprimée chez Elster :

> « First, to maximize utility is not to engage in the carrying out of a plan, choosing the best means to realize an independently defined end. In the modern theory of utility, it is essentially a short-hand for preferences, and implies nothing about more or less pleasurable mental states that could be seen as the goal of behaviour » (SG : 9).

La dernière phrase peut porter à confusion, car elle semble indiquer que l'utilité ne pe correspondre à un but de l'action. En fait, c'est une condamnation de la maximisati *intentionnelle* de l'utilité. En choix rationnel, l'utilité est toujours définie *ex ante* d'après le b visé par l'individu. Celui-ci vise évidemment un but quelconque, mais il doit être substantiel décrit en termes appropriés à la situation. Prenons l'exemple de Pierre qui, posté devant u distributrice de jus de fruits, doit choisir parmi différentes saveurs. Il arrête son choix s « raisin », et il appuie sur le bouton correspondant. Ses croyances sont rapidement formées p l'éventail des choix de saveurs et sa connaissance générale sur le fonctionnement de la machin Ensuite il analyse les saveurs selon divers critères et choisit le raisin, le domaine des désir A partir de cette sous-décision, il analyse les moyens possibles pour obtenir la bouteille de j de raisin et décide d'appuyer sur le bouton « raisin » (action intentionnelle primitive déclenchant un mécanisme (causal ordinaire) à l'intérieur de la machine lui conférant en bout ligne l'objet de son désir. Quel est l'intention de Pierre ? Qu'a-t-il maximisé au juste ? Il devra exprimer son intention en des termes du genre « Je voulais un jus de raisin », et non « Je voula maximiser mon bonheur, et celui-ci passe par l'obtention d'un jus de raisin ». La théorie d choix rationnel, dans sa recherche des raisons de l'action, ne demande pas plus de précisio que le premier énoncé de Pierre. A partir de celui-ci, on sait que Pierre maximise son utilit mais ce n'est pas le but (conscient, intentionnel) de son action. Le choix rationnel ne s'intéres pas à la nature ou au contenu de l'utilité. La recherche du bonheur est, dans les termes d Elster, un sous-produit (*by-product*) de l'action intentionnelle (SG : 9-10). En fait, l'explicatio par la maximisation de l'utilité est tautologique car l'option jouissant de la plus haute valeu d'utilité est tout simplement l'option *déjà préférée* par dessus tout autre par l'individu (Luc Raiffa 1957 : 16). Cette tautologie n'est toutefois pas vide de sens ; elle demeure la clé de l représentation des paiements en théorie des jeux (Binmore 1994). En fait, le phénomène d

faiblesse de la volonté (*weakness of will*) concerne justement une action où l'agent ne maximis pas son utilité.

I.2 - Utilité ordinale et cardinale -

L'élaboration de la fonction d'utilité s'effectue à partir des préférences du suje On demande à l'individu d'exprimer sa préférence sur toutes les paires possibles d'alternative reliées à la situation. En somme, pour une situation S comprenant les alternative $S = [a_1, a_2, ..., a_n]$, l'ordre de préférence doit respecter les axiomes[1] suivants :

1) Transitivité : si a_1 P a_2 et a_2 P a_3, alors a_1 P a_3.

2) Totalité : Pour tout $i, j \leq n$, soit a) a_i P a_j, b) a_j P a_i ou c) a_i I a_j.

Il est possible d'attribuer un nombre, ou valeur d'utilité, à chacune des alternatives d façon à ce qu'une préférence plus élevée ait une utilité plus élevée, c'est à dir $u(a_i) > u(a_j)$ ssi a_i P a_j. Ces deux axiomes définissent ce qu'il est convenu d'appeler l'utilit *ordinale* car ces nombres ne représentent aucune intensité de préférences, mais seulement u ordre particulier. La seule opération mathématique possible sur ceux-ci est la relation d grandeur (<, >, =). On peut attribuer n'importe quels nombres aux alternatives, pourvu qu l'ordre numérique respecte l'ordre des préférences (Luce, Raiffa 1957 : 25-26) L'utilité ordinale constitue l'instrument de choix dans le domaine de la décision en parfait certitude ; l'action rationnelle y étant définie par la maximisation de l'utilité. Ici, l'utilit n'exprime rien, sauf l'ordre de préférences. Serait-il possible de conférer un caractère plu substantiel à ces nombres, afin qu'ils puissent représenter une « quantité » d'utilité ? Le passag à un environnement à risques permet ce genre d'utilité.

John Von Neumann et Oskar Morgenstern, les précurseurs de la théorie des jeux en sciences sociales, ont démontré qu'une utilité *cardinale*, exprimant une intensité de préférences pouvait être déduite à partir de l'utilité ordinale simplement en ajoutant un élément de risque au processus de décision (Von Neumann, Morgenstern 1953 : 18). Le problème se posant à l'individu ressemblerait à ceci : en supposant que a_1 P a_2 P a_3, préféreriez-vous a_2 à une loterie

[1] Les symboles utilisés seront: P pour "préféré à", I pour "indifférent".

où vous auriez 50% des chances d'obtenir a_1 et 50% d'obtenir a_3 ? La question se révèle très pertinente, car si le sujet préfère a_2, c'est qu'il chérit davantage la préférence a_2 P a_3 que a_1 P a_2, et vice versa. En variant les probabilités reliées à a_1 et a_3, il est possible d'en arriver à une probabilité p ($0 \leq p \leq 1$) tel que a_2 I [p a_1, (1-p) a_3]. L'utilité de a_2 s'exprime ainsi : $u(a_2) = pu(a_1) + (1-p)u(a_3)$. En supposant $u(a_1) = 1$ et $u(a_3) = 0$[2], alors $u(a_2) = p$. Voici les axiomes supplémentaires concernant l'utilité cardinale :

3) réduction des loteries complexes : une « loterie de loteries » peut toujours être reformulée en loterie simple.

4) continuité : pour n alternatives, il existe toujours un p tel que a_i I [p a_1, (1-p) a_n].

5) substitution : dans l'ordre de préférences, il est toujours possible de remplacer a_i par son équivalent [p a_1, (1-p) a_n].

6) monotonicité : [p a_1, (1-p) a_n] P [p' a_1, (1-p') a_n] ssi $p > p'$.

Une fonction d'utilité respectant les axiomes 1 à 6 porte dans la littérature rationaliste le nom de *fonction d'utilité de Von Neumann et Morgenstern* (ou fonction d'utilité VNM)[3]. Ces axiomes permettent aussi le calcul d'une valeur d'utilité pour une loterie. Ainsi, il devient possible, à partir de l'utilité cardinale, de déterminer quelle loterie d'alternatives l'individu devrait préférer. L'utilité d'une loterie se calcule ainsi : $u(L) = p_1u(a_1) + p_2u(a_2) + \ldots + p_nu(a_n)$. Les utilités $u(L)$ sont toutefois ordinales ; elles n'expriment pas d'intensité de préférences entre les loteries.

I.3 - Intensité de préférences -

La fonction d'utilité VNM représente un outil très puissant en théorie des jeux - et en théorie de la décision en général - car ainsi l'utilité associée à une option exprime une certaine quantité. Mais qu'est-ce que ces nombres peuvent nous indiquer sur la structure de préférences d'un individu ? D'abord, ils nous renseignent sur sa propension au risque.

---

[2] Il est de bon usage en théorie de l'utilité de fixer l'utilité de l'alternative la plus prisée à 1 et de la moins prisée à 0. Il ne faut pas oublier que l'utilité cardinale n'est qu'une valeur relative servant à comparer les préférences entres elles.

[3] En fait, seuls les axiomes 1 et 2, ainsi que la notion de risque (l'introduction de la variable p) sont nécessaires. Les axiomes 3 à 6 peuvent être déduits à partir de ces trois éléments.

Reprenons le dilemme de Von Neumann et Morgenstern tel que décrit plus haut, mais ce fois-ci avec des valeurs monétaires : en supposant a_1 = 10$, a_2 = 5$ et a_3 = 0$ et u(a_i) = a préféreriez-vous a_2 à une loterie (50% a_1, 50% a_3) ? Normalement, la relation en est u d'indifférence, car 5$ = 0.5 x 10$ + 0.5 x 0$. Si l'individu préfère a_2 avec certitude, c'est dc que l'utilité du gain espéré est plus grande que l'utilité espérée de la loterie, autrement d u(a_2) > 0.5 u(a_1) + 0.5 u(a_3) ou encore u(0.5 a_1, 0.5 a_3) > 0.5 u(a_1) + 0.5 u(a Cette formule modélise une crainte envers le risque. A l'inverse, si l'individu préfère la loter c'est qu'il favorise le risque (Ordeshook 1986 : 44-45). Il est important de spécifier qu'u fonction d'utilité VNM ne s'exprime que sur les *options* et non sur la nature des loteri Une préférence sur tel ou tel type de loteries est considérée comme irrationnelle et viole mêr dans certains cas le troisième axiome[5] (Binmore 1994 : 261).

L'utilité cardinale nous renseigne aussi sur l'intensité des préférences propreme dite. Le principe est simple : a_1 P a_2 est plus intense que a_3 P a_4 si u(a_1) - u(a_2) > u(a_3) - u(a (Binmore 1994 : 277-78). Deux conditions se révèlent primordiales toutefo D'abord, les options comparées doivent naturellement faire partie du même ensemble S résultats d'une action. On ne peut donc pas comparer les préférences de deux actio complètement divergentes. Ensuite, on ne permet que les comparaisons intrapersonnelles, c'e à dire l'évaluation de préférences *chez le seul et même individu*. Cette condition ne nie p expressément la possibilité de comparaisons interpersonnelles - sujet controversé en théorie la décision s'il en est un - mais permet d'éviter toute une série de précisions dont nous pouvo nous passer car il ne sera jamais question de telles comparaisons dans le cadre de ce travail.

---

[4] C'est à dire, l'utilité de l'argent est linéaire avec l'argent, ce qui n'est généralement pas le cas dans la réalité o le taux de croissance de l'utilité de l'argent tend à diminuer alors que la somme d'argent augmente. Cette simplification ne sert qu'à clarifier le concept d'utilité cardinale.

[5] Par exemple, le choix entre p $a1$, (1-p) $a2$ à la roulette ou aux dés. L'axiome 3 est trangressé lorsque l'individ préfère une loterie complexe à son équivalent simplifié pour l'aspect plus "palpitant" de ce premier.

II - La théorie des jeux

II.1 - Notions de base -

La maximisation de notre utilité serait excessivement facilitée si l'on était seul au monde, mais ce n'est évidemment pas le cas ! Un outil très pertinent en théorie de la décision, la théorie des jeux, propose un cadre d'analyse du problème de maximisation impliquant au moins deux individus en situation d'interaction. R.D. Luce et H. Raiffa élaborent ainsi la portée de la théorie :

> « Intituively, the problem of conflict of interests is for each participant a problem of individual decision making under a mixture of risk and uncertainty, the uncertainty arising from his ignorance as to what the others will do. In game theory one attempts to idealize this problem in such a way as to transform it into interacting problems of individual decision making under risk » (Luce, Raiffa 1957 : 14).

L'interaction qui nous intéresse ici met en cause trois principes fondamentaux. D'abord, la récompense de chacun dépend de la récompense de tous ; c'est la modélisation de l'égoïsme ou de son opposé, l'altruisme[6]. En second lieu, la récompense de chacun dépend des choix de tous. Enfin, le choix de chacun dépend des choix de tous, car chacun doit calculer sa meilleure stratégie par rapport à ce qu'il s'attend de la part de ses « adversaires » (RAT : 119)[7]. Nous avons déjà vu que les décisions risquées appellent la formation d'utilités cardinales. Celles-ci serviront à la représentation dite « normale » ou « stratégique » du jeu, dont en voici un exemple :

---

[6] Ce principe peut aussi signifier une recherche de redistribution dans un jeu coopératif comme une négociation syndicale par exemple.

[7] La quatrième possibilité, que le choix de chacun dépende de la récompense de tous, ne fait pas partie du domaine de la théorie des jeux; elle prendrait la forme de la socialisation ou du comformisme (Leca 1985: 108; ETC: 77; MSM: 10). Elle pourrait aussi très bien correspondre aux normes sociales, notamment la norme kantienne, comme nous le verrons dans la section suivante.

	C	D
C	1,3	3,4
D	2,1	4,2

Fig. 5

Ici, deux individus s'affrontent : le joueur des *lignes* (désormais appelé joueur A) e celui des *colonnes* (joueur B). Ils font face à un choix binaire, soit collaborer (C) ou faire défection (D)[8]. Les nombres à la croisée des choix possibles représentent les *paiements* exprimés en valeurs d'utilité cardinale sous la forme (A, B). La « solution » de ce jeu es évidente : A et B font défection. Comment les joueurs en sont-ils arrivés à ce résultat ? Simplement en comparant la valeur des alternatives. Pour le joueur A, choisir C lui donne soit 1 ou 3 « points d'utilités » alors que D lui en confère 2 ou 4, dependamment du choix de B. Nous constatons que D lui donnera nécessairement plus d'utilité que C peu importe ce que son adversaire choisit. Un tel choix est ainsi *dominant* pour lui, et en supposant qu'il cherche à maximiser son utilité, il est donc forcé de jouer une telle option. Le même raisonnement s'applique pour son adversaire. Donc, la défection mutuelle est obligatoire et la solution du jeu se situe à (D, D) dont les paiements sont de 4 pour A et 2 pour B.

Le résultat (D, D) comporte plusieurs propriétés. D'abord, il est *Pareto-optimal*, ce qui signifie qu'aucune autre possibilité ne pourrait améliorer le sort d'un joueur sans pénaliser l'autre. A l'inverse, la situation (D, C) n'est pas optimale au sens de Pareto car les joueurs pourraient augmenter leurs gains en passant à (C, D) ou (D, D). Ensuite, la solution est en *équilibre de Nash*, car D constitue la meilleure réponse de la part de A sachant que son adversaire va jouer D, et vice-versa. Un jeu peut comporter plusieurs équilibres de Nash, comme le démontre cet autre exemple :

---

[8] Comme nous le verrons un peu plus loin, cette notation relève du dilemme du prisonnier. Elle n'est pas tout à fait adéquate en tout temps, mais elle nous fournit tout de même une représentation simple et pratique.

	C	D
C	2,2	1,1
D	1,1	3,3

Fig. 6

Malgré ses deux équilibres (C, C) et (D, D), la solution de ce jeu est unique et se trouve en (D, D). Il en est ainsi car ce jeu est *coopératif*, c'est à dire que les participants peuvent convenir en début de partie de faire mutuellement défection. Une fois l'engagement pris, il demeure rationnel pour chacun de le respecter, car le gain maximum (3, 3) Pareto-optimal est ainsi assuré. Sans coopération, ce jeu demeure sans solution car il n'existe pas de stratégies pures dominantes. Une stratégie *pure* consiste en un choix simple parmi les options présentées à l'individu. Si l'on admet les stratégies *mixtes*, c'est à dire un choix aléatoire entre les options, on peut en arriver à une solution. Une stratégie mixte *individuelle* en choix binaire se présente sous la forme (p C, (1-p) D), avec $0 < p < 1$. On peut aussi faire appel au hasard dans le choix d'un résultat, dans le cadre d'une stratégie mixte *commune*. Les joueurs sélectionnent les cases du jeu qui les intéressent mutuellement et entreprennent d'en retenir une à l'aide d'un mécanisme aléatoire accepté par tous. Ceux-ci s'engagent à respecter la proposition du mécanisme et à jouer la stratégie demandée. Le théorème d'équilibre de Nash prouve que pour tout jeu doté d'un nombre fini de solutions et de joueurs, il existe au moins un tel équilibre. Malheureusement, ce théorème ne fait que prouver l'existence de l'équilibre ; il ne nous renseigne aucunement sur la détermination des stratégies pour y parvenir (Rapoport 1966 : 74-77). Souvent, le calcul des probabilités portant sur les stratégies mixtes en équilibre est extrêmement complexe et relève dans certains cas de la simple intuition et de l' « essai et erreur » (Luce, Raiffa 1957 : 424-25).

La théorie des jeux admet aussi des situations à plus de deux joueurs, appelées *jeux à n-personnes*. Evidemment, la représentation normale ne peut adéquatement accommoder plus de deux joueurs. Les jeux à n-personnes étant particulièrement complexes - bien au delà des intentions de ce texte - nous nous contenterons d'un modèle simplifié, la représentation de Schelling (d'après Schelling 1978 : 213-243). Celle-ci exige toutefois que la matrice des paiements soit la même pour tous, formant ainsi un jeu *symétrique*. Par exemple, le jeu de la fig. 6 est symétrique tandis que le précédent ne l'est pas. Reprenons justement cet exemple (fig. 7).

	C	D
C	2,2	1,1
D	1,1	3,3

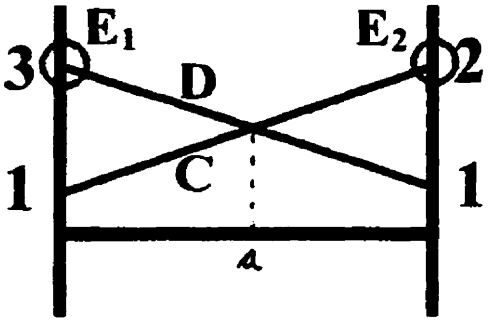


Fig. 7

Ici, l'axe horizontale représente la population de joueurs autre que le joueur actuel qui coopèrent (stratégie C), de 0 à n-1. Les valeurs d'utilité figurent sur la verticale. Les deux courbes se croisant représentent le gain d'utilité possible pour chacune des stratégies pures C et D accessibles au joueur actuel selon le nombre de coopérants. La courbe D part de l'utilité de la défection universelle (D, D) vers la défection unilatérale (D, C), tandis que la courbe C va de la coopération unilatérale (C, D) vers la coopération universelle (C, C). Alors que la solution au jeu à deux personnes se situe à (D, D), la variante à n personnes nous révèle une autre possibilité. En autant que la collaboration se limite à une certaine fraction m de la population ($m < a$), la courbe D demeure toujours la plus payante pour un individu. Chacun sera donc rationnellement porté à faire défection, et ce groupe atteindra rapidement l'équilibre E_1. Mais si, d'une manière quelconque, la coopération dépasse le seuil a, cette stratégie n'est plus avantageuse et tous s'adonneront à la stratégie C, en route vers un nouvel équilibre E_2, équivalent à (C, C) en forme normale et conférant 2 utilités à chacun.

Quelles conclusions faut-il en tirer ? Simplement que la forme normale nous indiquait que (D, D) étant la solution, il valait toujours mieux que tous choisissent D, mais ce n'était pas une stratégie dominante. Quoique la version à deux joueurs ne pose pas de problèmes, il en est autrement de la version à *n* joueurs, où un nombre suffisamment grand de coopérants remet en question la pertinence du choix de D. Le second équilibre de Nash fait donc finalement partie de la solution. Les caractéristiques de la forme normale transposée vers le graphique de Schelling vont comme suit :

- Stratégie dominante : une courbe domine l'autre lorsque tous ses points se situent au-dessus de celle-ci.
- Equilibre de Nash : ici, représente une situation où le nombre *m* de coopérants demeure fixe. A l'atteinte d'un tel équilibre, le choix de l'individu ne peut changer les paiements offerts aux participants[9].

II.2 - Le dilemme du prisonnier -

Le plus connu et le plus décortiqué de tous les jeux est sans contredit le dilemme du prisonnier (ci-après nommé DP). Il représente le cas paradigmatique de la faillite de la rationalité. En cherchant à maximiser leur utilités, les participants au DP ne peuvent qu'atteindre un résultat qui n'est pas Pareto-optimal. Il existe en effet une autre situation de jeu plus payante pour tous, mais qui demeure inatteignable rationnellement.

La légende du dilemme du prisonnier va comme suit : la police parvient à mettre la main au collet de deux présumés complices d'un même crime. Ils subissent chacun un interrogatoire à huis clos, isolé de l'autre, où on leur demande de dénoncer leur partenaire. Si un suspect réussit à dénoncer l'autre sans se faire dénoncer soi-même, il retrouvera immédiatement sa liberté tandis que l'autre purgera une peine maximale de dix ans. Si les deux confessent simultanément, ils passeront cinq ans derrière les barreaux, tandis que si les deux se taisent, ils subiront une peine exemplaire d'un an. En attribuant arbitrairement 0 au résultat le moins utile

[9] Schelling (1978: 226) a démontré que le point *a* n'est pas un "point" proprement dit, mais bien une "zone" large d'une unité de population. Donc, lorsque le nombre de coopérants se stabilise à *a*, possible si *a* est en équilibre de Nash, un individu ne peut à lui seul déséquilibrer le jeu.

(10 ans de prison), 1 au plus utile (la liberté) et 0.1 et 0.9 aux deux autres (respectivement, 5 ans et 1 an), nous obtenons le jeu suivant :

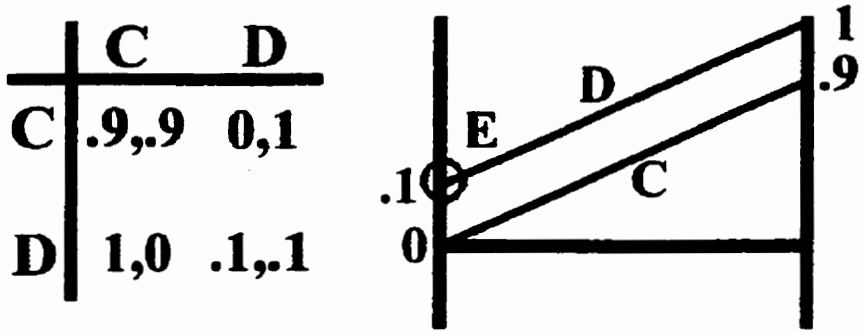


Fig. 8

Chacun des suspects doit maintenant choisir entre se taire (coopérer) ou dénoncer son confrère (faire défection). Une analyse technique du jeu nous permettra d'entrevoir les enjeux. D'abord, le DP se révèle être un jeu symétrique. Il n'arbore qu'un seul équilibre de Nash, la situation (D, D). La stratégie de défection dominant la coopération, la solution se retrouve donc à (D, D), conférant 0.1 utilité à chacun. Cette solution n'est toutefois pas Pareto-optimale, car le résultat (C, C) donnant 0.9 utilité à chacun est préférable pour tous. Tout l'intérêt des sciences sociales pour le DP pourrait se résumer à la question suivante : comment atteindre le résultat coopératif ? A première vue, il semblerait qu'une simple entente « motus et bouche cousue » entre les suspects de notre histoire suffirait à garantir un résultat (C, C). Malheureusement pour nos deux comparses, le DP est un jeu non coopératif. Voici pourquoi : supposons qu'effectivement, les suspects se soient mutuellement promis de jouer C à l'intérieur de leur isoloir. Le suspect A, sachant que son confrère coopérera, sera naturellement porté à jouer D, gagnant ainsi sa liberté (et un maximum d'utilité). Même s'il tient mordicus à coopérer, un simple effort de réflexion de sa part lui démontrera que l'autre aura tout intérêt à profiter de la situation et à faire défection ; il prendrait alors 10 ans pour sa gueule, le pire résultat possible. La situation (C, C), bien qu'envisageable, est extrêmement instable et strictement inatteignable dans les règles de base de la théorie des jeux, qui considère la maximisation de l'utilité comme le but ultime de tout joueur.

Les valeurs d'utilités exprimées à la fig. 8 sont arbitraires. Cela signifie qu'il existe une infinité de variantes du DP. En fait, tout jeu en forme normale équivalent à celui de la fig. 9 et respectant l'ordre $a < b < c < d$ se qualifie comme un DP. En représentation de Schelling il suffit simplement que la courbe D domine complètement la courbe C, que le point à l'extrême droite de C soit plus haut que l'extrémité gauche de D et que les deux courbes exhibent une pente positive.

	C	D
C	c,c	a,d
D	d,a	b,b

Fig. 9

Il existe un cas spécial de DP, nommé DP « étiré » (Sobel 1991 : 34), lorsque $.5(a + d) > b$. Dans ce cas, la stratégie mixte commune [.5(C,D), .5(D,C)] confère une utilité supérieure à (C,C). De plus, si $.25(a + b + c + d) > b$, la stratégie mixte individuelle [.5(C), .5(D)] pour chacun devient plus payante que (C,C). Il ne faudrait pas en conclure toutefois que de telles stratégies seraient nécessairement retenues par les joueurs ; nous y reviendrons plus loin.

III - Solutions rationnelles au Dilemme du Prisonnier

Le coeur de ce chapitre se veut une analyse exhaustive du Dilemme du Prisonnier en utilisant les outils conceptuels de la théorie des jeux. Tout jeu peut prendre plusieurs formes qualitativement distinctes selon le nombre de joueurs et le nombre de répétitions du jeu. D'abord, les jeux à plus de deux joueurs - par opposition au jeu à deux - introduit les notions de solutions semi-coopératives où une fraction seulement des participants coopèrent. La répétition du jeu - par opposition au jeu à un coup - permet aux joueurs d'adopter des stratégies conditionnelles, soit réagir en fonction des actions antérieures de son ou ses adversaires. En tout cela nous donne quatre types de jeux que nous étudierons séparément, chacun comportant ses problèmes et ses concepts particuliers de solution.

Avant de débuter l'analyse, voici un bref survol des théorèmes qui seront employés. Dans le jeu à deux joueurs et un seule instance, le postulat des joueurs symétriquement rationnels sera testé et nous critiquerons l'exposé de Gauthier sur la possibilité de coopération avant le match. La seconde partie concerne le jeu à plus de deux participants toujours joué une seule fois ; il sera question de l'analyse de Schelling des seuils critiques de coopération ainsi que de la proposition originale de Hardin de permettre aux joueurs de voter à l'avance pour la solution préférée. En troisième lieu viendra le DP à deux joueurs mais à plusieurs rondes. Nous y ferons la distinction entre répétition finie, infinie et à durée aléatoire. Les méthodes d'analyse seront l'induction à rebours, les équilibres avec utilité temporellement décroissante de Taylor et les tournois simulés d'Axelrod. Pour terminer, nous nous pencherons sur le jeu le plus intéressant, celui à plusieurs joueurs à répétition. Nous tenterons là un amalgame des théorèmes valides chez les formes moins complexes du DP en y ajoutant l'analyse coût/bénéfice de Taylor et de Schelling. Cet exercice devrait nous fournir les instances où l'équilibre coopératif constitue une solution rationellement atteignable.

III.1 - Deux joueurs, un coup.

Nous avons vu précédemment que la seule solution possible au DP tel que représentée à la fig. 9 est la défection de la part de chacun des joueurs, amenant le résultat (D,D). Et pourtant, un doute lancinant hante toujours nos esprits : le résultat (C,C), préférable pour tous, devrait être à la portée de joueurs « intelligents » ! Mais intelligent rime-t-il avec rationnel ? Le concept même de rationalité individuelle nous conduit vers une avenue intéressante de recherche. En théorie des jeux, les joueurs sont considérés également rationnels. S'ils participent à un jeu symétrique, le principe de rationalité doit donc *nécessairement* prescrire la même stratégie à chacun ; il ne peut pas prescrire deux stratégies différentes lorsqu'autant les joueurs que les paiements sont identiques[10] . Ceci restreint les possibilités à la diagonale nord-ouest du DP, soit les cases (C,C) et (D,D). Etant donné que la première domine l'autre au sens de Pareto, il devient *rationnel de coopérer* pour chacun. Cet argument a

[10] Les joueurs sont identiques uniquement par leur caractère rationnel. Dans un jeu où les utilités ont déjà été déterminées, il ne reste plus que le principe de maximisation d'utilité qui puisse les "unir". Donc, deux maximiseurs sont identiques du point de vue du choix d'une stratégie.

principalement été développé par Lawrence H. Davis, qui y a vu une résolution du « paradoxe » fondamental du DP, le choix rationnel d'un état Pareto-inférieur à un autre (Davis 1977 : 319).

Le véritable paradoxe devient le suivant : comment le *même* principe de maximisation d'utilité peut-il simultanément prescrire la coopération et la défection ? Certainement, le théorème de Davis doit être faux, ou alors il faudra repenser sérieusement le principe de dominance. D'abord, Davis nous signale qu'il se sert d'une conception plutôt radicale de la rationalité - l'agent ici ne se trompe jamais, autant sur lui-même que sur les intentions de son adversaire[11] (Davis 1977 : 322). Tout en respectant les prémisses de Davis, il est possible d'infirmer ses propos. D'abord, le principe de symétrie nous indique que si le joueur A choisit la coopération, alors B fera de même et vice-versa. Appliquons l'effet accordéon[12] à cette proposition (Pettit 1986 : 183) :

- Si A choisit x, alors x est rationnel pour A (rationalité absolue des joueurs)
- Si x est rationnel pour A, alors il l'est aussi pour B (symétrie)
- Si x est rationnel pour B, alors B choisit x.

Binmore rejette d'emblée la premier point comme étant contraire à la rationalité : « An action is not rational because it is chosen by a rational person. On the contrary, a person is said to be rational because he only chooses rational actions » (Binmore 1994 : 207), ce qui interdit à A de jouer C. Mais laissons la chance au coureur et acceptons telle quelle la première proposition. Remplaçons x par C, la coopération. Nous pourrions alors combiner les deux premières propositions de la façon suivante : si A choisit C, alors il est rationnel pour B de choisir C. Cette proposition est évidemment fausse : il n'est *pas* rationnel de jouer C en réponse à C dans un DP (Pettit 1986 : 183). Même si (C,C) domine (D,D) dans le jeu restreint de Davis,

---

[11] L'auteur avoue que toute conception plus "faible" de la rationalité infirmerait son théorème. Ce qui prouve une fois de plus que l'équilibre (C,C) - si équilibre il y a - est extrêmement volatile. Selon l'analyse traditionnelle du DP, au moindre petit doute de la part de l'un des joueurs le résultat (D,D) s'en suivra. Davis, dans une tentative de conséquence avec lui-même, prétend qu'en cas de doute, le résultat sera *indéterminé*: " (...) then it is by no means so obvious what he should do" (Davis 1977: 322).

[12] L'effet accordéon permet la manipulation de phrases décrivant les actions en les "étirant" pour inclure un maximum de liens causaux ou en les "réduisant" au minimum pour ne décrire que l'action et sa conclusion (Davidson 1980: 53-54).

force nous est de constater que le sous-ensemble des résultats *possibles* ne comprend que (D,D) (Binmore 1994 : 205). Le « paradoxe » de Davis s'estompe, car la rationalité n'exige pas que les joueurs atteignent un résultat mutuellement payant[13], mais seulement que chacun maximise son utilité (Sowden 1983 : 351).

L'obstacle principal à l'atteinte du résultat (C,C) réside dans le fait que toute entente de collaboration entre les joueurs demeure trop instable pour constituer une raison efficace d'action. David Gauthier (1990) introduit la notion de *métastratégie*, le choix rationnel d'une disposition particulière d'action dans le DP. En langage informatique, la métastratégie serait un algorithme préétabli régissant toutes les circonstances possibles d'une situation donnée. Une fois que le joueur a choisi sa métastratégie il est alors lié, comme un ordinateur, à l'exécution de celle-ci lorsque la situation (ici, le DP) se présente. La métastratégie naturelle en choix rationnel est la « maximisation brute » (*straightforward maximization*, SM) qui prend la forme suivante : faire défection peu importe le choix de l'adversaire. Le coeur de la solution de Gauthier réside dans une métastratégie visant à l'atteinte du résultat coopératif. Celle-ci se nomme « maximisation contrainte » (*constrained maximization*, CM) : si l'adversaire adopte CM, collaborer ; s'il choisit SM, faire défection. Pour fonctionner, l'individu CM doit alors jouir de la possibilité d'*afficher* publiquement son (méta-) choix et de *détecter* celui de son adversaire. Si les joueurs peuvent détecter sans failles la métastratégie de l'autre en tout temps, la solution du « méta-jeu » devient (CM, CM) qui se transpose en (C, C) dans le DP. Mais Gauthier rejette cette possibilité, qu'il nomme la *transparence*, parce que trop idéaliste (Gauthier 1990 : 315-24).

[13] De toute façon, qu'est-ce qu'un "résultat mutuellement payant"? Un résultat Pareto-optimal? Il y en a trois dans le DP: (C,C), (D,C) et (C,D). L'obsession que certains vouent à (C,C) dans le jeu à une seule itération provient du fait que celui-ci représente une distribution égalitaire et que les joueurs ne s'entendraient jamais sur (D,C) ou (C,D), du moins dans le DP non étiré. Mais ces questions d'ententes sont absolument non avenantes dans le contexte d'un jeu non coopératif. Comme nous le verrons plus tard, (C,C) devient intéressant dans le cas du DP répété indéfiniment, où une forme subtile de négociation fait son apparition.

Par souci de réalisme, l'auteur a alors recours à des acteurs *translucides* doués d'une capacité de détection imparfaite. Pour modéliser un environnement translucide, il a recours à une population d'individus[14] dont une fraction (r, $0<r<1$) adopte CM et l'autre ($1-r$) adopte SM. Lorsque deux CM se rencontrent, la probabilité qu'ils se reconnaissent est de p, la probabilité de *succès coopératif*. A l'inverse, si un SM reconnaît un CM alors que celui-ci se trompe sur son adversaire, le SM peut alors *exploiter* le CM ; cette situation survient avec la probabilité q (Gauthier 1990 : 325). A la suite de savants calculs que nous ne reproduirons pas ici, Gauthier en vient à la conclusion qu'il est rationnel de choisir CM lorsque p et r sont suffisamment grands et q suffisamment petit (Gauthier 1990 : 326). Par conséquent, il devient rationnel de coopérer dans le DP.

Plusieurs ont critiqué le modèle de Gauthier comme étant hautement idéaliste, donc trop détaché de la réalité. Il y a d'abord les arguments concernant le pouvoir réel d'une métastratégie sur l'action. En d'autres termes, le choix d'un algorithme de décision peut-il *absolument* déterminer une décision ultérieure ? La question relève de la philosophie de l'action : il ne peut y avoir de lien causal entre une intention et une action lorsqu'il existe un délai de temps entre les deux (Davidson 1980 : 99 ; Smith 1991 : 235-36). Une explication alternative serait l'engagement volontaire irrévocable (*self-binding*) de la part des joueurs CM. Elster (US : 42) pose comme condition essentielle le déclenchement d'un événement externe qui agira causalement sur l'individu au moment de l'action, ce qui ne semble pas être le cas ici. L'idée d'un mécanisme « interne » d'engagement irrévocable a été rejetée par Nelson (1988 : 157-58) et Bicchieri (1993 : 208-9). Les métastratégies demeurent tout de même largement efficaces chez la plupart des individus ; il est indéniable qu'il existe une norme interne[15] favorisant le respect d'un engagement antérieur.

---

[14] Un nombre d'individus supérieur à 2 n'implique pas ici les jeux à n-personnes. Le jeu demeure à deux personnes, mais la nature des adversaires peut prendre diverses caractéristiques dépendamment de la distribution de la population.

[15] La norme externe (sociale) fera l'objet du chapitre 3. L'idée ici est que l'individu peut apprécier la valeur *intrinsèque* d'une promesse réalisée, indépendamment du jugement de ses pairs.

Mais même en admettant la pertinence des métastratégies, de sérieuses questions demeurent en suspens. Tout d'abord, Gauthier précise que le choix d'une métastratégie constitue une décision paramétrique risquée[16] (Gauthier 1990 : 350). L'individu fonde son choix de CM ou SM uniquement sur les valeurs *p*, *q* et *r*. La situation devient paradoxale lorsque l'on considère que les autres joueurs doivent nécessairement prendre la même décision dans des circonstances semblables, ce qui soulève de sérieux doutes sur la validité de *r* : comment peut-on connaître la composition CM/SM d'une population si les membres choisissent simultanément ? Cette variable est manifestement influencée par les décisions de tous, qui en retour influencent ces mêmes décideurs. Nous sommes alors en présence d'un environnement stratégique où la variable *r* perd toute signification (Nelson 1988 : 160 ; Franssen 1994 : 260 ; Danielson 1991 : 302). Le passage à un tel environnement peut soulever le problème de la décision conditionnelle simultanée (je collabore si vous collaborez, mais vous collaborez seulement si je collabore...), mais la logique moderne - dont je suis ignorant - prouve aisément qu'il n'y a aucune régression infinie et que la décision peut aboutir (Binmore 1994 : 181 ; Danielson 1991 : 310).

L'aspect le plus crucial du modèle de Gauthier se retrouve dans la confiance réciproque absolue des joueurs. Lorsque deux CM se rencontrent, la collaboration s'en suit immédiatement et ceci n'a rien à voir avec les facteurs de risque chez des adversaires translucides. De fait, même lorsqu'un CM croit faussement faire face à un CM alors qu'il est en réalité un SM, ce premier collabore quand même. Introduisons dans le modèle un « faux CM » (FCM), un SM qui tente de se faire passer pour un CM et admettons qu'il excelle dans l'art du déguisement. Nous observons que contre d'autres SM, le résultat sera (D, D) dans le DP alors que face aux CM, il les exploite (D, C). Deux FCM s'affrontant est stratégiquement équivalent à deux SM. La possibilité d'exploitation fait de FCM la métastratégie dominante (Rainbolt 1989 : 218-19). Celui-ci précise toutefois qu'une prémisse d'égalité absolue de la capacité de « déguisement » et de détection de tous les joueurs sauve les conclusions de

[16] Un environnement est *paramétrique* lorsque les variables (incluant les décisions des autres individus) entrant dans la décision sont fixes. Il devient *stratégique* lorsque la prise de décision a une influence sur ces même variables.

Gauthier, mais prétend que c'est trop demander. Gauthier empêche l'exploitation de masse de se généraliser en précisant qu'au-delà d'un certain seuil d'exploitation (q), la métastratégie CM n'est plus rationnelle[17]. Mais qu'en est-il du jeu le plus simple, impliquant deux joueurs ne se connaissant pas jouant une seule partie du DP ? Il semble que la seule défense possible soit l'honnêteté parfaite (Danielson 1991 : 312), argument ressemblant étrangement à celui de Davis mentionné plus haut. Toutefois, cette prémisse est si contraignante dans l'explication de l'agir humain qu'elle doit être rejetée (Franssen 1994 : 256-57 ; Smith 1991 : 236). Harsanyi (1987 : 343) prétend que l'honnêteté est en fait une *norme* qui précède la théorie rationnelle, ce qui est incompatible avec le projet de Gauthier, élaborer un fondement rationnel aux normes (Gauthier 1987 : 2-3 ; 1990 : 330). Comme nous le verrons au chapitre suivant, les normes peuvent effectivement expliquer la coopération dans le DP.

III.2 - Plusieurs joueurs, un coup.

Bien que la théorie des jeux à deux participants soit riche en implications en sciences sociales, il est bien entendu qu'elle fait preuve d'un profond irréalisme car les situations d'interactions sociales impliquent la plupart du temps plus de deux individus. Il devient alors nécessaire de passer à une théorie générale des jeux sans limite de joueurs, mieux connu dans la littérature sous le terme *jeux à n-personnes*. Aux premiers balbutiements de cette généralisation, R.D. Luce et H. Raiffa commentaient :

> « Nonetheless, it is the *n*-person theory which must be of greater interest in sociology and economics. It is here, more than in two-person theory, that game theory as a part of social science, though not as a part of mathematics, will stand or fall » (Luce, Raiffa 1957 : 157).

L'intérêt du jeu à n-personnes pour les sciences sociales démarra au début des années soixante-dix, quelques années après la parution de l'ouvrage de M. Olson, *The Logic of Collective Action*. Sa thèse bouleversa la théorie traditionnelle des organisations, en affirmant que l'individu dans un groupe, en recherchant la maximisation de son intérêt particulier, ne

---

[17] Et étant donné que FCM est dominant, CM est rapidement évincé pour donner une population entièrement composée de SM ($r = 0$). Gauthier ne se prononce pas sur cette perversion, qui tourne de toute façon sur une conception stratégique de prise de décision où r perd toute signification.

réalisera pas l'intérêt du groupe lorsque celui-ci consiste à la production d'un bien public. Olson lui-même ne fit aucune référence à la théorie des jeux - laquelle il croyait incapable de généralisations abstraites - malgré les ressemblances au niveau des postulats de base de l'action collective. R. Hardin (1971) fut le premier à associer les propositions d'Olson à la théorie des jeux, particulièrement au Dilemme du Prisonnier (Landry 1993 : 46), ce qui engendra une activité intense de la part des chercheurs autour de la question de l'action collective, activité qui n'a pas cessé depuis.

Les paramètres de l'action collective doivent respecter certains critères pour pouvoir être représentés comme un DP sur un graphique de Schelling. En résumé, nous avons une situation d'action collective sous la forme d'un DP lorsque :

a) Le bien à produire est « public », dans le sens de « non exclusif » ; aucun membre du groupe ne peut être exclu de sa consommation.

b) Les individus font face à un choix binaire, le même pour tous : contribuer ou ne pas contribuer.

c) Ceux-ci sont parfaitement libres de contribuer ou non, et leur choix s'effectue indépendamment des autres (Landry 1993 : 49, CS : 24).

d) Les paiements (ou utilités[18]) associés aux divers choix sont les mêmes pour tous. Autrement dit, chacun reçoit une part égale du bien public et les coûts de contribution sont les mêmes (Schelling 1978 : 218)[19].

---

[18] Par souci de simplicité, l'utilité d'une part de bien public est linéaire avec sa valeur objective (argent, bénéfices, ...). L'utilité varie seulement avec la quantité de biens produits.

[19] Techniquement, cette condition n'est pas nécessaire, mais elle simplifie grandement l'analyse. Elle ne fait qu'assurer que les courbes demeurent équidistantes. En fait, les conditions d'existence du DP sont respectées tant que son équivalent en forme normale demeure un DP et que tous les points de la courbe D se situent au-dessus de C.

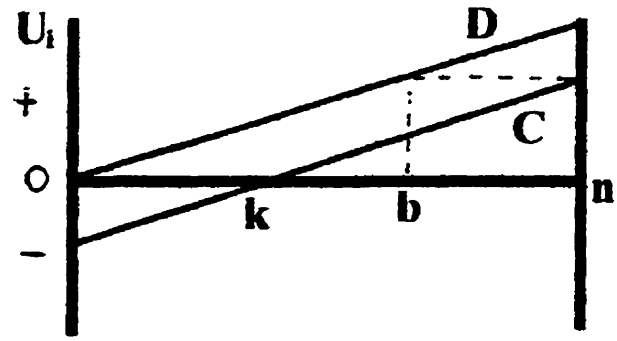


Fig. 10

En observant un DP de Schelling comme celui de la figure 10, on en vient à une conclusion immédiate : il est *toujours* préférable pour un individu de faire défection, et ce peu importe le nombre de coopérants. Toutefois, le graphique nous fournit certaines informations supplémentaires, notamment les niveaux *k* de coalition viable et *b* de coalition Pareto-optimale. Le nombre *k* indique le nombre minimum de coopérants nécessaires afin que la part individuelle du bien public produit dépasse le coût de contribution. C'est seulement à partir de *k* contribuables que l'action collective sera couronnée de succès, bien qu'il existe différents degrés de réussite. Le point *b* représente le nombre minimum de coopérants à partir duquel les resquilleurs reçoivent un meilleur paiement que la coopération universelle. A partir de *b* ceux-ci n'ont aucun avantage à se rallier à la coopération. En considérant N comme représentant le nombre d'individus dans le groupe, nous avons les conditions supplémentaires suivantes :

e) Le bien public peut être produit efficacement lorsqu'au moins *k* individus contribuent, $k <= N$ (Schelling 1973 : 218).

f) Le coût de contribution doit être supérieur à la part individuelle du bien public partiellement produit à l'aide d'un seul contribuable, sinon le jeu cessera d'être un DP.

Lorsqu'un groupe d'individus désire produire un bien public, leur situation peut être modélisée comme un DP à n-personnes lorsque les conditions ci-dessus sont toutes respectées. La solution alors n'est pas différente que celle du DP à deux joueurs : la défection constitue une stratégie dominante pour chacun. De légères modifications apportées à certaines conditions pourraient possiblement sortir les participants du cercle vicieux de la défection rationnelle.

Procédons à l'examen de celles-ci. D'abord, les conditions a), b) et c) déterminent les paramètres essentiels du jeu d'action collective. Elles ne peuvent être modifiées sans affecter la nature même de la situation d'analyse. Il en est tout autrement des autres conditions. Nous pourrions considérer contrairement à la condition d) que les paiements et les coûts ne soient pas les mêmes pour tous. Si le ratio coût-bénéfice d'un individu en particulier est très faible par rapport aux autres, il pourrait à lui seul fournir le bien public. Par exemple, un groupe de l'âge d'or organisent une levée de fonds pour se procurer un autocar. Un membre particulièrement riche pourrait considérer la « désutilité » du coût du véhicule inférieure à l'utilité des promenades avec les amis le dimanche après-midi et ainsi fournir seul le bien public. Le jeu devient alors trivial. La condition e) stipule que la dimension de la coalition viable ne peut dépasser celle de la population totale du groupe, sinon le règlement (C, C, ..., C) aurait une utilité inférieure à (D, D, ..., D), ce qui rendrait ce dernier non seulement en équilibre mais aussi Pareto-optimal. Personne, à juste raison, n'envisagerait la collaboration dans ce nouveau jeu trivial.

Le non-respect de la dernière condition renvoie à une situation similaire à celle du contribuable aisé où, dans une certaine mesure, $k = 1$ car un seul individu peut efficacement produire le bien public. La particularité ici est que n'importe quel membre du groupe peut à lui seul produire le bien (CS : 27). La stratégie de défection n'est plus dominante, comme le démontre ce jeu « individu vs. collectivité » (fig. 11).

		Collectivité C	Collectivité D
Individu	C	3,3	1,4
	D	4,1	0,0

Fig. 11

Cette matrice représente le jeu de la Poule Mouillée. Il possède deux équilibres de Nash en stratégie pure, (D, C) et (C, D). Lequel des deux équilibres sera atteint relève de la coordination entre les parties (Landry 1993 : 79-82). Nous n'entrerons pas dans les détails d'un jeu de la Poule Mouillée à n-personnes, nous avions seulement besoin de prouver que le jeu de l'action collective se transforme lorsque la condition f) n'est pas respectée[20] .

Une tentative de solution rationnelle autre que la défection universelle fut proposée par R. Hardin dans un article au titre évocateur, *Collective Action as an Agreeable n-Prisoners' Dilemma* (1971). L'élément principal de sa preuve est le choix de Condorcet : *a* est un choix de Condorcet fort si *a* P *b* pour un nombre *m* d'individus supérieur à la moitié de la population totale N du groupe, sinon le choix est faible si plus d'individus adhèrent à *a* P *b* qu'à *b* P *a* (Hardin 1971 : 477), la différence entre les deux dépendant du nombre d'individus pour lesquels *a* I *b*. Son théorème est le suivant : dans un DP à n-personnes, la coopération universelle (C_N) constitue un choix de Condorcet fort sauf si N est pair et que les coûts individuels de participation atteignent la moitié des bénéfices, alors C_N devient un choix de Condorcet faible (Hardin 1971 : 478).

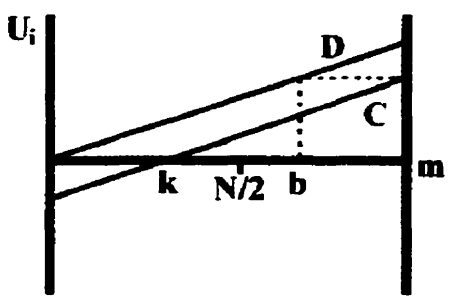


Fig. 12a

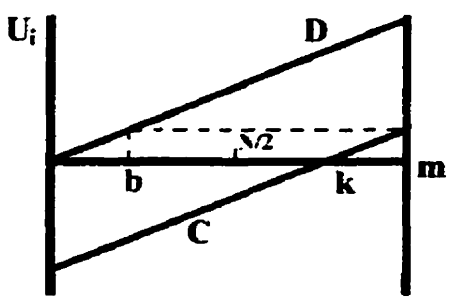


Fig. 12b

Dans la démonstration de Hardin, tous les participants doivent se prononcer sur la question suivante : préférez-vous la coopération universelle à votre situation actuelle ? Autrement dit, pour chaque individu *i* d'une population N, les collaborateurs se prononcent sur C_N P C_i et les resquilleurs sur C_N P D_i. Hardin impose aussi la condition du minimax : les situations où l'utilité du collaborateur se situe en-deçà de l'utilité de la défection universelle

[20] Voir Hovi (1986) pour une étude des différents jeux autre que le DP dans les situations d'action collective.

sont irréalisables car ceux-ci auront vite fait de renverser leur décision et de revenir à la défection universelle. La valeur d'utilité 0 représente le minimax garanti pour chacun. Sur le graphe de Schelling, les valeurs de U(*i*) tel que $0 < i < k$ se situent sous le minimax. De ces prémisses nous pouvons tirer les conclusions suivantes :

1) Les coopérants votent toujours pour la coopération universelle, donc :

1a) Si les coopérants représentent plus de la moitié de la population ($i > N/2$), C_N deviendra un choix de Condorcet fort.

2) Les resquilleurs ne votent pour la coopération universelle que lorsque leur utilité y est inférieure ($i < b$) ce qui fait apparaître deux cas :

2a) $b > N/2$ (fig. 12a) - Lorsque $i > b$, $i > N/2$ aussi et la conclusion 1a s'applique.

2b) $b < N/2$ (fig. 12b) - Dans tout graphe de Schelling symétrique, si $b < N/2$ alors $k > N/2$, donc $b < k$ et b se situant en-deçà du minimax la situation est impossible.

Le théorème de Hardin est ainsi prouvé[21].

Il est tout de même frappant de constater que certaines possibilités du DP à n-personnes sont irréalisables. Hardin retient le concept de minimax comme condition de faisabilité : « Hence, none of these outcomes is realizable, i.e. they would require that some player willingly recline below his minimax, *as few of us are wont to do* » (Hardin 1971 : 477, italiques rajoutées). Même s'il est vrai que « peu d'entre nous » accepteraient un règlement moins payant que la défection universelle, il ne s'en suit pas que tous les points sous le minimax sont des choix irrationnels alors qu'au-dessus, la rationalité règne. Comme le mentionne Elster (IRC : 7), le minimax tout comme ses semblables - maximin, maximax, etc. - peuvent être *conformes* aux principes de la rationalité sans toutefois être *recommandés* par lui. En fait, la domination de la stratégie de défection démontre que le seul choix rationnel est D pour chacun, par conséquent la seule situation rationnellement possible est la défection universelle.

[21] Sauf pour le cas spécial du choix de Condorcet faible qui demande des calculs plus complexes.

En dernier lieu, la question posée aux participants, C_N P C_i ou C_N P D_i, n'est pa pertinente car il existe une autre question qui fait l'unanimité dans tous les cas : préférez-vou être le seul resquilleur dans un univers de coopération quasi-universelle ? Même si l'on accept le théorème de Hardin et la formulation de la question, la réponse anticipée C_N P C_i pou n'importe quel i <= N n'engendra pas la coopération universelle à elle seule. Ce ne serai qu'une légitimation du point C_N comme « solution rationnelle » qu'il faudrait atteindre pa d'autres moyens. Hardin (1971 : 479) suggère la coercition, mais ceci violerait la condition c de l'action collective comme DP, la liberté d'action.

III.3 - Deux joueurs, plusieurs coups

L'une des principales causes de l'impossibilité pour des participants à un Dilemme d Prisonnier à atteindre le résultat coopératif Pareto-optimal est que la théorie suppose de individus ne se connaissant que par les attributs de la rationalité, qui signifie en gros l'axiom de maximisation de l'utilité. Bien que le résultat coopératif soit logiquement possible, l moindre doute (dont nous avons prouvé qu'il était plus que raisonnable) entretenu sur so adversaire forcera un joueur à adopter sa stratégie dominante, soit la défection. Ce informations se trouvent qualitativement modifiées lorsque les joueurs s'affronten mutuellement plusieurs fois au même jeu. On dira alors qu'ils participent à un « tournoi ». A tout moment du tournoi, les joueurs ont à leur disposition une source inestimabl d'information : le comportement de l'adversaire dans les parties précédentes. Mais l'innovatio la plus significative des tournois, en relation avec la précédente, est le concept de réciprocité. En effet, un joueur insatisfait peut punir son adversaire lors des parties subséquentes et comm l'action punitive cadre parfaitement à l'intérieur de la théorie des jeux, elle peut constituer un forme de menace crédible, comme on en rencontre dans les jeux coopératifs. Ces deux principe fondamentaux de la théorie des jeux répétés amènent un nouveau type de stratégie : le stratégies contingentes. Une stratégie est dite contingente si elle prend en considération le coups précédents de l'adversaire dans son calcul du meilleur coup pour la partie actuelle. Un stratégie contingente simple pourrait être « Jouer D si l'adversaire a joué D dans la partie précédente, sinon jouer C », alors qu'une stratégie ordinaire (non-contingente) pourrait se lire « Toujours jouer C » ou « Jouer C pour les dix premiers coups, toujours jouer D ensuite », etc.

La valeur d'utilité de la stratégie correspond à la somme des utilités des résultats obtenus à chaque partie du tournoi[22].

Nous débuterons par l'étude des tournois finis, ou à nombre fixe de parties connu par les joueurs, du Dilemme du Prisonnier. Une représentation en forme normale d'un tournoi est parfaitement possible, quoiqu'elle peut s'avérer rapidement très complexe. Procédons par le plus simple des tournois, celui qui ne comporte que deux parties et ne faisant usage que des quatre stratégies pures non contingentes possibles (fig. 13). Les utilités de chaque partie sont celles d'un DP générique (fig. 9) avec a = 4, b = 3, c = 2 et d = 1. Ce jeu comporte un seul équilibre de Nash en (DD, DD) conférant 4 utilités à chacun. La stratégie DD est dominante pour chacun, mais son résultat est dominé au sens de Pareto par toutes les stratégies n'employant DD de la part d'aucun des joueurs. Les conclusions ressemblent en tous points à celles que nous tirions du DP simple à la section III.1. Pour une analyse vraiment significative, il nous faudra employer les stratégies contingentes.

	CC	CD	DC	DD
CC	6,6	4,7	4,7	2,8
CD	7,4	5,5	5,5	3,6
DC	7,4	5,5	5,5	3,6
DD	8,2	6,3	6,3	4,4

Fig. 13

[22] Comme le précise Luce et Raiffa (1957: 97-98), la théorie Von Neumann-Morgenstern de l'utilité ne justifie pas un tel calcul, mais elle constitue quand même une interprétation assez simple et réaliste de l'utilité d'un jeu répété.

En plus des quatre stratégies non contingentes, il existe quatre autres stratégies contingentes (Hardin 1982 : 165), toutes du type « Jouer x au premier tour. Au second tour, si l'adversaire a joué C précédemment alors jouer y, sinon jouer non-y », qui sera noté (x [y]). Les quatre stratégies seront donc (C [C]), (C [D]), (D [C]) et (D [D]). Nous ne représenterons pas la matrice 8 x 8 nécessaire à l'analyse technique de ce jeu. Il existe encore un équilibre en (DD, DD) mais cette fois-ci l'équilibre est instable, car la stratégie (D [C]) est équivalente à (DD) en réponse à (DD) ou à (D [C]). Donc, chaque joueur est indifférent entre (DD) et (D [C]). Les résultats des quatre combinaisons possibles ont toutes la même valeur. De plus, ces deux stratégies exhibent la propriété du maximin. Le problème réside dans le fait qu'aucun de ces règlements ne soit en équilibre de Nash, aucune de ces stratégies n'est donc dominante. En effet, même si (D [C]) est une bonne réplique à (DD), la meilleure réplique à (D [C]) est soit (CD) ou (C [C]). La solution du jeu en stratégie pure est indéterminée.

La solution d'un tournoi fini peut se calculer à partir des principes de stratégies rationnelles de Kohlberg et Mertens (Binmore 1994 : 192), soit l'induction à rebours (*backwards induction*) et l'élimination successive des stratégies dominées. Par l'induction à rebours, nous débutons par l'analyse de la dernière partie et nous revenons progressivement vers le début du tournoi, où la solution devrait se révéler. Dans un DP répété, l'absence de stratégies à priori dominantes permet d'envisager des stratégies coopératives, semi-coopératives ou conditionnellement coopératives. Le coopérant peut toujours punir le resquilleur en faisant lui-même défection à la prochaine partie[23]. A la dernière partie, la défection devient dominante : l'adversaire n'a plus les moyens de ramener le resquilleur à la raison. Si le règlement de la dernière partie est déjà déterminé par (D,D), l'avant-dernière partie devient effectivement la dernière ; alors le même raisonnement s'applique et, de fil en aiguille, on découvre que la stratégie (DDDD...) est dominante pour tous dès le début du tournoi.

[23] Les détails de ce genre de stratégie seront élaborés plus loin. Il s'agit ici de considérer la possibilité de coopération sous la menace de défection dans le futur.

Par contre, lorsque les joueurs ne connaissent pas la durée du tournoi, ils ne peuven évidemment appliquer le raisonnement à rebours. Ces situations de jeu répété *indéfinimen* permet une certaine forme de coopération. Supposons d'abord un tournoi infini Dans un tel tournoi, l'utilité est calculée non pas sur la somme des jeux, mais sur la moyenne par jeu. Existe-t-il alors un équilibre de Nash autre que la défection universelle de la part de chacun des joueurs ? En fait, il existe une infinité de stratégies en équilibre Toute stratégie collectivement stable[24] et donnant une moyenne par jeu égal ou supérieur à l'utilité de (D, D) à chacun des joueurs est en équilibre de Nash (Binmore 1994 : 115-6) Nul besoin de préciser que nombre de stratégies partiellement coopératives se situent dans cet ensemble de possibilités. La plus célèbre d'entre elles est sans doute celle qui fut couronnée par R. Axelrod championne d'un important tournoi informatique, la stratégie « tit fot tat » (TFT) C'est une stratégie conditionnelle qui coopère toujours au premier tour, et qui aux tours subséquents imite simplement le comportement précédent de l'adversaire (C ou D). La TFT est en équilibre de Nash avec elle-même, voici pourquoi : deux joueurs utilisant TFT gagnent 3 utilités en moyenne par tour. Si l'un d'eux décide de passer à la défection universelle, il gagnera 4 utilités au moment de la défection et 2 utilités par tour par la suite. Donc, la moyenne par tour tendra vers 2 utilités. Dans un tournoi infini rien ne justifie la déviation d'un équilibre (TFT, TFT) qui incidemment paye la même somme que la coopération universelle !

Malheureusement, cette découverte ne solutionne pas le DP car les tournois infinis se révèlent plutôt irréalistes, donc largement inintéressants pour le chercheur empirique. Aussi est-il convenu d'employer les tournois à durée indéfinie gérés par une probabilité w qu'il se poursuive à chaque tour (ou inversement, une probabilité $1-w$ qu'il se termine à un tour donné). La durée espérée d'un tel tournoi est de $1/(1-w)$. La notion de temps introduit aussi le principe du poids décroissant de l'avenir. En théorie des jeux, cela se traduit par des valeurs d'utilités plus petites pour les tours subséquents au tour présent. Tous les facteurs de

[24] Si tous les joueurs adoptent la stratégie S1, un ou plusieurs joueurs employant une stratégie S2 peuvent *envahir* le groupe si leur utilité est supérieure à celle du groupe, soit U(S2) vs. S1 > U(S1) vs. S1. Une stratégie est *collectivement stable* si aucune autre stratégie ne peut l'envahir (Axelrod 1990: 301).

décroissance se valent, mais nous utiliserons le décroissement logarithmique par convenance et conformité à la littérature rationaliste. Le facteur d'escompte e_i ($0 < e_i < 1$), un par joueur, influence les utilités du tour n (le tour présent étant zéro) par le multiplicateur e_i^n.

Le principe d'utilité décroissante impose de nouvelles limites à la coopération. D'abord, il a été prouvé par Taylor (1987 : 69) que l'équilibre de TFT avec elle-même n'est possible que si $e_i >= \max((a-b)/(a-c), (a-b)/(b-d))$ pour les deux joueurs. Lorsque $e_i < (a-b)/(a-c)$, la défection universelle est plus payante pour le joueur i contre la stratégie TFT, et lorsque $e_i < (a-b)/(b-d)$ c'est la stratégie « Nasty tit-for-tat » (TFT ') - semblable à TFT sauf qu'elle fait défection au premier tour - qui devient la meilleure réponse à TFT, conduisant à l'équilibre (TFT, TFT ') conférant en moyenne $(a+d)/2$ à chacun ce qui est toujours inférieur à b dans un DP non-étiré (Taylor 1987 : 66-68)[25]. Donc, le facteur d'escompte de chacun des participants doit être suffisamment élevé afin qu'une stratégie coopérative soit envisageable.

Le tournoi indéfini, géré par la probabilité de continuité w, bien que plus réaliste, se rapproche du tournoi fini, avec les conséquences que l'on connaît. La stratégie TFT n'y est plus aussi inattaquable que dans un tournoi infini. En fait, tout comme dans un tournoi fini, elle est battue par sa consoeur qui se comporte comme une TFT sauf qu'elle fait automatiquement défection au dernier tour (TFT-1). Etant donné que la durée réelle d'un tournoi indéfini demeure inconnue à l'avance, le succès de TFT-1 relève du hasard mais l'important c'est qu'il *existe* une stratégie préférable à TFT. Donc, TFT ne peut plus être en équilibre de Nash avec elle-même. Sober (1992 : 134-9) met l'accent sur la nature circonstancielle et non intentionnelle de TFT-1, alors qu'une étude intentionnelle demeure possible car l'individu connaît la durée *espérée* du tournoi et est donc en mesure de « miser » sur un gain avec TFT-1 contre un adversaire jouant TFT. La possibilité de TFT-1 empêche en fait tout équilibre de Nash car cette stratégie peut être battue par TFT-2, elle-même vulnérable à TFT-3, etc. jusqu'à TFT-n (n = durée du tournoi) stratégiquement équivalent à la défection universelle. Mais la défection universelle cède devant TFT ! L'ordre de préférence des stratégies TFT-x ($0 <= x <= n$)

---

[25] Les utilités moyennes telles qu'exprimées ici ne reflètent pas l'utilité décroissante, pour des raisons de simplicité. A titre de comparaison, la valeur de b est en réalité $be_i/(1-e_i)$ et $(a+d)/2$ devient $(a+de_i)e_i/(1-e^2_i)$.

devient donc intransitive (Sober 1992 : 133), et conséquemment, il n'y a plus de stratégie dominante.

Bien qu'il existe en théorie une multitude d'équilibres de Nash dans le DP répété indéfiniment, il faut se rappeler que la popularité de TFT n'est due qu'à sa victoire dans le tournoi d'Axelrod et à sa simplicité. Ce fameux tournoi fut contesté par quelques auteurs, notamment Binmore qui ne concède pas la victoire à TFT. Le tournoi d'Axelrod fut *évolutif*, c'est à dire qu'il employa une population de personnages (fictifs, bien sûr) utilisant l'une ou l'autre des stratégies soumises à l'expérimentation. Pour éviter un DP à n-personnes, les interactions entre individus furent limités à la paire. Le principe du tournoi se résume ainsi : les personnages moins performants sont graduellement éliminés et à la fin du tournoi (à durée fixée par Axelrod mais inconnue des participants) le groupe majoritaire parmi les survivants est déclaré gagnant. Binmore (1994 : 200) voit comme véritable gagnant dans ce tournoi une stratégie mixte composée de toutes les stratégies survivantes en proportion de leur importance dans la population finale[26]. Aussi, différentes expériences peuvent mener à différents résultats[27]. Par contre, l'ensemble des expériences conclues jusqu'ici nous permettent d'affirmer que peu importe le « gagnant », il s'agit presque toujours d'une stratégie initialement coopérative, c'est à dire qui amorce le tournoi par la coopération (Binmore 1994 : 202 ; Axelrod 1988 : 1388-9).

III.4 - Plusieurs joueurs, plusieurs coups

L'intérêt de la science politique pour le Dilemme du Prisonnier prend tout son sens lorsque le jeu comprend plus de deux participants et que ceux-ci interagissent continuellement. Nous entrons alors dans les cas de production continue d'un bien public par une collectivité. Il est toutefois surprenant de constater que l'analyse théorique de tels jeux n'est pas tellement

---

[26] Par exemple, si à la fin du tounoi la stratégie x représente 75% de la population et y 25%, le gagnant ne devrait pas être x (comme le propose Axelrod), mais bien la métastratégie mixte (.75 x, .25 y).

[27] Le tournoi de B. Linster (cité dans Binmore 1994: 200), employant 26 stratégies - dont TFT - plutôt que 63 pour Axelrod, donna la stratégie "Coopérer jusqu'à ce que l'adversaire fasse défection; adopter alors la défection universelle" - une version très revencharde de TFT - majoritaire dans la population restante.

développée[28]. Notre analyse portera donc sur un amalgame de théorèmes s'appliquant aux jeux à n-personnes (III.2) et aux jeux répétés à deux joueurs (III.3).

L'étude la plus complète à ce jour - et la plus citée - demeure celle de M. Taylor (1987), qui transpose ses calculs du jeu répété à deux joueurs vers une multitude de joueurs. Il en vient à la conclusion que la stratégie TFT, employée par un certain nombre d'individus, permet un certain niveau de coopération rationnelle. L'exposé du problème fait référence au courbes de Schelling : alors que dans le cas d'un jeu à n-personnes pratiqué une seule fois, le point k assurant une coalition viable est inatteignable rationnellement, l'utilisation de la stratégie TFT permet au groupe de collaborateurs de dépasser ce niveau critique de participation.

Dans une situation d'action collective, les individus d'un groupe doivent produire un bien public. La contribution est volontaire pour chacun, et elle est coûteuse. Le niveau de production du bien est calculé de façon linéaire selon le nombre de personnes collaborant à sa création.

Au niveau formel :

N = nombre d'individus dans le groupe

V = valeur totale du bien public

V_i = valeur du bien public pour l'individu i

T_i = coût de participation pour l'individu i

Ainsi, si N individus s'unissent pour produire un bien de valeur V, ils en récolteront chacun un bénéfice brut de $V_i = V/N$, ou un bénéfice net de $V_i - T_i$. Evidemment, l'entreprise doit être profitable pour chacun, sinon l'action collective ne décollera jamais ; il s'agit donc d'exiger que $V_i > T_i$.

[28] D'autres champs de recherche semble être préconisés pour ce type de situation, notamment la théorie des négociations (concernant les jeux coopératifs) et la théorie du choix social.

a) DP à deux joueurs : individu vs. collectivité.

Cette première représentation de l'action collective comme un DP place un individu contre le reste du groupe dans une matrice de jeu conventionnelle à deux joueurs (fig. 14). Rappelons que cette matrice constitue un DP en autant que $a < b < c < d$. Ici, l'individu doit choisir entre la collaboration à la production du bien public (C) ou le resquillage (D), alors que la collectivité (moins l'individu en question) se retrouve devant un choix similaire : elle doit coopérer ou faire défection en bloc.

		Collectivité C	Collectivité D
Individu	C	c,c	a,d
	D	d,a	b,b

Fig. 14

Si tous mettent la main à la pâte, ils obtiendront le règlement (C, C), soit $V_i - T_i$. Inversement, si tous se dérobent (D, D), aucun bien ne sera produit, avec un résultat de 0. Lorsqu'un individu refuse de collaborer alors que tous les autres participent à la production, le bénéfice du resquilleur sera de $((N-1)/N)\ V_i$, alors que les autres recevront chacun $(((N-1)/N)\ V_i) - T_i$. Enfin, si l'individu contribue seul au bien public (C, D), il en récoltera $((1/N)\ V_i) - T_i$, et les autres $(1/N)\ V_i$. La structure du DP pour le joueur individuel requiert donc les inégalités suivantes :

$$\frac{(N-1)}{N} V_i > V_i - T_i > 0 > \frac{V_i}{N} - T_i$$

1ère inégalité : Pour que $((N-1)/N)\ V_i > V_i - T_i$, il s'agit que

$T_i > (1-((N-1)/N))\ V_i$, qui se simplifie par

$T_i > V_i/N$

2ème inégalité : Celle-ci est déjà résolue, car la condition préalable $V_i > T_i$ rend nécessairement vraie l'inégalité $V_i - T_i > 0$.

3ème inégalité : Pour que $0 > (V_i/N) - T_i$, il s'agit que $T_i > V_i/N$, ce qui revient à la première inégalité.

Conclusion : Le coût de participation doit respecter $V_i > T_i > V_i/N$ afin que le jeu d'individu vs. collectivité soit considéré comme un DP.

Il existe donc une relation entre le nombre d'individus formant le groupe et les coûts individuels de production. Cette importante relation est la suivante : le bénéfice encouru par un seul contribuable doit être inférieur à ce qu'il lui en coûte. Si cette condition n'est pas respectée - autrement dit si $V_i/N >= T_i$ - il devient rationnel pour un individu de fournir seul le bien public[29], car ses bénéfices dépassent ses coûts ; la matrice cesse alors d'être un DP.

En tenant compte du choix d'*un seul* individu dans le DP « individu vs. collectivité », il est possible de connaître les choix de *tous* les membres du groupe, simplement en les plaçant tour à tour dans la peau du joueur des lignes. Etant tous également rationnels, ils tenteront tous de maximiser leur utilité et choisiront de faire défection. Résultat : aucun bien n'est produit, c'est la faillite de l'action collective.

b) DP à n-personnes - analyse coût/bénéfice

Le jeu individu vs. collectivité souffre d'une grave lacune, car il ne permet pas l'analyse de la coopération partielle : le joueur « collectivité » doit soit collaborer, soit faire défection en bloc. La théorie des jeux à n-personnes permet de contourner cet inconvénient, en élaborant une matrice « joueur vs. joueur » comportant autant de dimensions que de participants. Celle-ci peut toutefois devenir rapidement très complexe : un DP à dix joueurs ne comporte pas moins de 1024 cellules ! Et nous ne commenterons pas sur la représentation en forme normale d'un tel jeu... C'est ici que la représentation de Schelling devient hautement salutaire. La fig. 10 montre un DP générique selon Schelling. La défection universelle ne générant aucun

[29] En langage olsonnien, le groupe est "privilégié" (Hardin 1982: 28).

bénéfice, l'extrémité gauche de la courbe D est fixée à zéro. La valeur de la courbe correspond au nombre de coopérants : plus il est élevé, plus le bien produit est grand, et plus le bénéfice du resquilleur est élevé. La courbe correspond à (m/N) V_i, m représentant le nombre de collaborateurs (m <= N-1). La courbe C doit, dans la logique du DP, demeurer en-dessous de D. L'écart entre les courbes est fixe et correspond au coût de contribution, T_i. La valeur de la courbe C est donc de (m/N) V_i - T_i.

Bien qu'il soit évident que face à une telle situation, un individu rationnel fera toujours défection, il existe deux valeurs de m, k et b, dont les propriétés ouvrent la voie à de nouvelles possibilités de solution au DP. Le point k se situe sur la courbe C, vis-à-vis l'extrémité gauche de D (fixée arbitrairement à zéro). A partir de ce point, la coopération - partielle du moins - devient plus payante que la défection universelle. C'est le nombre minimum de collaborateurs nécessaire pour une « coalition viable » (Schelling 1978 : 221). Sa valeur précise est de :

(k/N) V_i - T_i = 0 ou simplifié,

$k = (T_i/V_i)$ N

L'autre point d'intérêt, b, représente le niveau critique d'optimum de Pareto ; il se situe sur la courbe D, au niveau de l'extrémité droite de C. Lorsque $m < b$, la coopération universelle profite à tout le monde, alors que si $m > b$, la coopération universelle n'améliore le sort que des seuls membres de la coalition de production du bien public (CS : 28-29). Lorsque la dimension de la coalition dépasse b, il devient encore plus difficile de convaincre les resquilleurs de joindre le mouvement. La valeur de b est de :

(b/N) $V_i = V_i - T_i$ ou,

$b = ((V_i - T_i)/V_i)$ N

Une analyse des coûts de participation nous permet de tirer quelques conclusions sommaires. Nous remarquons que plus T_i est petit, k tend vers zéro alors que b tend vers N, ce qui signifie que plus les coûts sont minimes par rapport aux gains, plus la coalition viable est facile à atteindre et plus la coalition Pareto-optimale comprend d'individus. Le nombre de resquilleurs (quasi-) irréductibles ((N-1) - b) devient très petit. Lorsque T_i = 50% V_i, nous avons $k = b$, la coalition viable prend la dimension (N-1)/2. Avec un grand T_i dépassant 50% V_i, les tendances s'inversent : b devient plus petit que k, la possibilité d'une

coalition viable est remise en question. Ces conclusions ne suggèrent que des « possibilités dynamiques » (CS : 29) ; les valeurs *k* et *b* ne schématisent en aucun cas des équilibres ou des solutions rationnelles au DP, et la rationalité individuelle immédiate recommande toujours la même chose : faire défection en tout temps.

c) DP à n-personnes - stratégie TFT

Nous étudierons les possibilités d'équilibre dans une population composée de N individus, chacun ayant le choix entre les trois stratégies suivantes : coopération inconditionnelle (C^{∞}), défection inconditionnelle (D^{∞}) et coopération conditionnelle, d'après la stratégie « tit for tat » et symbolisée par TFT_a, *a* représentant le nombre de coopérants requis au jeu précédent pour que le joueur poursuive sa coopération. Si ce nombre n'est pas atteint, l'individu fait immédiatement défection. Le nombre d'individus adoptant ces stratégies est représenté par N_C, N_D et N_{TFT} ; leur somme équivalent à N. Le nombre de coopérants (conditionnels ou non) est représenté par $M = N_C + N_{TFT}$.

Tout d'abord, la valeur de *a* chez les adeptes de TFT_a doit nécessairement respecter $a = M-1$, soit tous les autres coopérants excluant le joueur lui-même, car avec $a < M-1$ un certain resquillage devient possible sans pour autant compromettre la coopération partielle. A ce moment, tous les joueurs ont intérêt à faire défection dans l'espoir de combler les quelques postes disponibles de resquilleur et l'équilibre devient D^{∞} pour tous - en fait, c'est un DP à l'intérieur d'un DP ! Dans le cas de $a > M-1$ il ne peut y avoir assez de coopérants pour satisfaire l'inégalité, donc la défection universelle s'ensuit aussitôt.

Comme pour le DP répété à deux joueurs, l'équilibre dépend du facteur d'escompte e_i, lequel doit demeurer suffisamment élevé pour *tous les coopérants* (M)[30] . La condition générale d'équilibre est la suivante :

$$e_i >= \frac{D(M-1) - C(M-1)}{D(M-1) - D(N_C)}$$

Condition pour un groupe constitué de D^{∞}, C^{∞} et TFT_a.

[30] Les resquilleurs inconditionnels se situent toujours en équilibre peu importe leur facteur d'escompte.

En supposant la monotonicité des deux courbes, l'écart entre celles-ci - le numérateur de la fraction - demeure constant peu importe la valeur de M. Puisque $0 >= e_i >= 1$, le numérateur doit demeurer égal ou inférieur au dénominateur sinon l'inégalité sera fausse. Il faut donc que $D(M-1) - C(M-1) <= D(M-1) - D(N_C)$ qui se simplifie en $D(N_C) <= C(M-1)$. En clair, le niveau minimum de resquillage doit demeurer inférieur au bénéfice maximum de la coopération. Etant donné un numérateur fixe, seule une variation du dénominateur peut influencer sur la réalisation de l'inégalité : plus il sera grand, plus l'ensemble des valeurs de e sera grand. La maximisation de l'écart $M-1 - N_C$ devient déterminant : il correspond en fait à $N_{TFT} - 1$. Conclusion : plus le nombre d'adeptes du TFT_a est élevé, plus il y a de chances que l'équilibre soit atteint.

Lorsque les coopérants inconditionnels disparaissent, le revenu minimum du resquilleur n'est plus assuré. L'équation se lit alors :

$$e_i >= \frac{D(M-1) - C(M-1)}{D(M-1) - D(0)}$$ Condition pour un groupe constitué de D^{∞} et TFT_a.

La première condition, $D(0) <= C(M-1)$, correspond au concept de « coalition viable » de Schelling (1978 : 221) : une telle coalition existe lorsque le nombre de coopérants dépasse k ($M > k$). La seconde, que $D(M-1) - D(0)$ soit le plus élevé possible, signifie simplement que $M-1$ doit être grand ou, plus pertinemment, qu'il existe un nombre maximum de resquilleurs inconditionnels ($N - (M-1)$) au-delà de lequel un équilibre partiellement coopératif devient impossible à atteindre.

La dernière possibilité concerne la présence uniquement d'adeptes du TFT_a. Ces individus peuvent atteindre un équilibre stratégiquement équivalent à la coopération universelle si :

$$e_i >= \frac{D(N-1) - C(N-1)}{D(N-1) - D(0)}$$ Condition pour un groupe constitué de TFT_a.

La condition D(0) <= C(N-1) est nécessairement vraie dans le contexte d'un DP : la coopération universelle est plus payante pour chacun que la défection universelle. Comme il n'y a aucune variable manipulable dans cette équation, le succès de l'entreprise dépendra uniquement de la pente des courbes, donc des utilités du jeu en forme normale. Plus la coopération sera payante par rapport à la défection, plus la plage des valeurs possibles de e_i sera grande.

En résumé, le problème soulevé par Taylor est que les conditions d'une coopération (même partielle) sont encore plus sévères que pour le jeu à deux joueurs. Considérons d'abord la situation où tous les participants utilisent TFT. Chacun doit alors se soumettre aux contraintes suivantes :

C1 - La stratégie TFT doit être très sévère : un joueur doit passer à la défection du moment *qu'un seul joueur* fait défection.

C2 - Chaque individu doit respecter le facteur d'escompte minimum imposé par la situation particulière.

Nous constatons alors que plus il y a de joueurs, plus ces contraintes ont des chances de ne pas être respectées. Taylor précise qu'un certain nombre d'individus ayant adopté une ou l'autre des stratégies de coopération ou de défection universelles peuvent cohabiter avec les TFT, pourvu que ces groupes ne soient pas trop importants par rapport au nombre de TFT, ce qui nous amène à une troisième contrainte :

C3 - Une présence minimale de TFT est nécessaire.

Le théorème de Taylor accepte la présence de resquilleurs inconditionnels puisque la « coalition viable » de coopérants peut se situer en-deçà de la population totale. Il peut survenir un sérieux problème à ce niveau, celui de déterminer qui collaborera et qui pourra « se la couler douce » en resquillant. Avec différentes compositions des deux groupes, nous nous retrouvons avec une multitude d'équilibre de Nash où un individu particulier est tantôt un TFT, tantôt un resquilleur et bien entendu, il préférera ce dernier. Cette nouvelle situation de jeu est en fait une Poule Mouillée à n-personnes dont les conclusions demeurent imprécises (Taylor 1987 : 104 ; Jankowski 1990 : 449-58).

Bien entendu, nous pouvons adapter la stratégie TFT-n aux jeux itérés à n-personnes La conclusion pourrait être que peu importe l'équilibre adopté, les joueurs feront défection un peu avant la fin probable du tournoi. En fait, par la contrainte C1, le premier à faire ains défection entraînera tous les autres à faire de même. Les choses se compliquent encore plus s nous considérons non pas une probabilité *collective* (que le jeu se poursuive à chaque tour) mais une probabilité *individuelle* qu'un joueur demeure dans la communauté. Nous aurions alors des possibilités de défections à n'importe quel moment. Je ne me propose pas d'élaborer sur le sujet mais voilà une avenue intéressante pour la recherche sur l'action collective en théorie des jeux[31].

IV - Conclusion

L'analyse technique du Dilemme du Prisonnier nous permet de dégager certaines conclusions concernant l'explication rationnelle de l'action collective. D'abord il semble que si le jeu n'est pas répété, la coopération ne constitue jamais une stratégie rationnelle. Nous avons établi que les quelques auteurs qui ont tenté de justifier la coopération dans de telles conditions voyaient leur théorèmes reposer sur de fausses conceptions de la rationalité. Elster en vient à un même constat : « If (i) the game is played only once, (ii) the actors are solely motivated by the payoff in the matrix and (iii) they behave rationally, collective action *must* fail » (MSM : 360).

L'espoir renaît lorsque le jeu est répété plusieurs fois. La possibilité de stratégies conditionnelles - dont la plus populaire, TFT - permet l'équilibre coopératif sous certaines contraintes. La première concerne la durée du tournoi. S'il est *infini*, la stratégie TFT devient rationnelle et la coopération s'en suit. Toutefois, ce type d'interaction n'est pas très réaliste. La rationalité ne justifie pas TFT dans un tournoi *fini*, à cause du phénomène d'induction à rebours. Reste le tournoi *indéfini*, qui a l'avantage de se situer beaucoup plus près de la réalité : toute chose a une fin, mais nous ne savons pas quand ça arrivera. Ici, en autant que les acteurs

[31] Le postulat des probabilités individuelles n'est pas nécessairement destructif pour l'équilibre (semi-) coopératif. Prenons l'exemple d'une association étudiante permettant la contribution financière volontaire pour des activités parascolaires. Les étudiants approchant la fin de leurs études cessent de contribuer, mais ils seront remplacés par de nouveaux arrivants. L'équilibre peut alors perdurer.

accordent une certaine importance aux utilités futures (la seconde contrainte) , l'équilibre TFT devient possible mais pas automatique. C'est alors qu'une troisième contrainte entre en scène : le nombre de participants. A deux joueurs, la coopération demeure relativement facile en autant que le facteur d'escompte demeure dans les limites prescrites. Lorsque l'on passe au jeu à plus de deux joueurs, les choses se compliquent grandement. La seconde contrainte devient plus difficile à respecter car tous les coopérants (TFT et coopérants inconditionnels) doivent s'y conformer. Aussi, la réussite de l'entreprise dépendra de la distribution de la population selon les stratégies possibles. Nous n'en avons étudié que trois, mais les possibilités sont bien sûr infinies... En définitive, la solution coopérative au DP est prescrite par la théorie des jeux lorsque a) le jeu est répété et les joueurs ne connaissent pas la fin du tournoi, b) ils ont une certaine préférence pour les gains futurs et c) il ne se trouve pas trop de resquilleurs dans le groupe. Le non-respect d'une de ces prescriptions entraînera automatiquement l'échec total par la défection universelle. Voilà qui est bien contraignant comme condition de coopération et qui ne correspond certainement pas aux situations concrètes où la coopération semble se réaliser beaucoup plus fréquemment. Peut-être trouverons-nous une meilleure formulation en regardant à l'extérieur de la théorie du choix rationnel...

CHAPITRE 4

MOTIVATIONS EXTRA-RATIONNELLES ET SOLUTIONS ALTERNATIVES AU DILEMME DU PRISONNIER.

Ayant jugé insatisfaisantes les conclusions de la théorie des jeux concernant l'explication de l'action collective, Elster va se tourner vers les normes sociales dans l'espoir d'y trouver une élaboration plus juste. L'une de ses convictions les plus profondes est que la théorie du choix rationnel comporte des limites explicatives ; c'est pour cette raison qu'il ne cherche pas à redéfinir le modèle rationnel afin de combler ses lacunes propres.

Ses attentes face à une théorie des normes sociales se présentent en deux volets. Premièrement, la théorie devra se distinguer le plus possible du modèle rationaliste tout en conservant la prémisse de l'individualisme méthodologique. Pour ce faire, il abordera le problème des normes sous les angles non conséquentialiste et non intentionnel. En second lieu, il introduira la notion vitale - et largement absente du choix rationnel - d'influence de la société sur l'individu en se servant principalement des outils théoriques de la sociologie. Les normes peuvent aussi être analysées par la philosophie morale ou éthique mais Elster ne prêtera que peu d'attention à ces aspects.

Dans ce chapitre, il ne sera pas seulement question des normes sociales comme alternatives au choix rationnel. Respectant sa trichotomie intérêts - passions - normes sociales, Elster se servira de quelques motivations psychologiques comme solutions possibles au problème de l'action collective. Il faut garder à l'esprit que Elster privilégie les normes sociales avant la psychologie, et que le choix rationnel - lorsqu'il est fonctionnel - doit toujours précéder ces deux théories. Nous débuterons l'étude des solutions normatives au Dilemme du Prisonnier par une définition des normes sociales, autant du point de vue individuel que social. Ensuite nous passerons aux concepts de Elster et aux idées d'autres théoriciens de la même veine.

Nous serons alors en possession de tous les éléments pour élaborer une typologie des motivations extra-rationnelles et finalement, une nouvelle tentative de solution au DP.

I - Qu'est-ce qu'une norme sociale ?

Pour son analyse des normes sociales, Elster s'impose un objectif épistémologique exigeant : il désire élaborer une théorie des normes qui viendra remplacer la théorie du choix rationnel là où cette dernière fait preuve d'inadéquation. Bien que l'effort soit louable, les choses ne se présentent pas aussi simplement et Elster lui-même sera forcé de l'admettre. La frontière entre rationalité et normes sera plus floue que prévu.

Dans cette partie nous tenterons de définir les normes sociales au niveau individuel et collectif. Pour ce premier objectif nous situerons les normes sociales parmi d'autres normes affectant l'individu et nous réviserons le modèle intentionnel pour tenter d'y situer le comportement normatif. Ensuite nous passerons aux rapports entre normes sociales et collectivité ; il sera question des origines sociales des normes ainsi que des effets potentiellement bénéfiques des normes sur la société.

I.1 - Normes sociales et individu -

Chez Elster, l'action normative est définie en opposition à l'action rationnelle ; une norme est un « commandement » psychique, relié aux émotions, qui pousse l'individu à agir compulsivement, sans but réfléchi. La norme garde toutefois une parenté essentielle avec la rationalité : toutes deux représentent des motivations individuelles de l'action. Pour qu'une norme puisse être qualifiée de *sociale*, elle doit être partagée par d'autres personnes et son application sujette à l'approbation ou à la désapprobation de ceux-ci (RESN : 24, CS : 98-99). Cette définition minimaliste demande que l'on prête un intérêt particulier aux sources possibles de confusion qu'elle comporte.

En premier lieu, la norme sociale ne doit pas s'attarder aux conséquences des actes prescrits, comme par exemple les normes morales qui attachent beaucoup d'importance aux conséquences de certains comportements, telles les recommandations émanant de l'utilitarisme

de la règle (Kymlicka 1990 : 27-28). Les normes sociales s'identifient au passé, généralement par la tradition, mais il ne faudrait pas accorder trop d'importance à cette distinction. Il sera toujours possible de rattacher une évaluation des conséquences à n'importe quelle norme, ne serait-ce qu'une vague conception du Bien et du Mal. Peut-être le plus important à retenir ici est que dans le cas de la moralité utilitariste, les conséquences jouent un rôle primordial dans l'élaboration de la norme alors que pour les normes sociales, celles-ci sont moins déterminantes (RESN : 24, CS : 101).

Ensuite, l'explication des normes sociales ne peut être réduite à l'intérêt particulier, ce qui exclut les conventions sociales. Par exemple, conduire sa voiture à droite de la route pourrait être considéré comme une norme sociale : elle est partagée par les autres automobilistes et elle est sujette à sanctions. Ce qui la diffère toutefois d'une véritable norme sociale est le fait qu'il est individuellement rationnel de respecter la norme[1]. Lorsque l'on part de zéro (la « situation originale » sans routes ni voitures), la décision de conduire à droite ou à gauche devient un jeu d'assurance où la rationalité est impuissante à déterminer quoi que ce soit. La décision devient alors une authentique norme basée sur la délibération. Imaginons maintenant un parfait étranger voulant circuler sur ces voies. Quel côté choisira-t-il ? Cette décision relève alors exclusivement de la rationalité, qui dictera de suivre le même côté que tous les autres. Rappelons que pour Elster, la décision normative doit s'opposer aux principes de la rationalité si elle veut conserver un pouvoir explicatif propre (CS : 101-102). Les « normes d'obligation » de Ullman-Margalit (1977 : 12-13) permet aussi une distinction entre normes sociales (selon Elster) et conventions. Une norme d'obligation répond à trois caractéristiques, en plus de celle implicite d'être partagée par d'autres : une menace de sanctions, une certaine utilité sociale et une contrainte à l'intérêt particulier. La norme « PD »[2]

---

[1] La distinction entre intérêt rationnel et normes sociales comporte une importante zone grise où une norme sociale *s'apparente* à l'intérêt particulier, dans le sens faible que l'individu ne perçoit pas de conflits majeurs entre la norme et ses intérêts (Margolis 1990a: 823). Dans une telle situation, la norme demeure authentiquement sociale (au sens de Elster) car ici la coïncidence est non universelle et purement *ad hoc*.

[2] Ce type de normes peut être considéré comme un sous-ensemble appartenant aux normes sociales de Elster. Elles ont la caractéristique d'être intimement liées à un DP en particulier, et ce de deux directions. Ces normes émergent d'une situation de type DP et servent exclusivement à y maintenir l'équilibre coopératif.

de Ullman-Margalit correspond aux normes d'obligation, alors que la convention sociale n'y correspond pas du tout. Chez Elster, la norme sociale respecte la première et la troisième caractéristique, mais pas la seconde ; il n'est pas nécessaire qu'une norme sociale ait une quelconque utilité sociale. Il existe même des normes sociales opérant au détriment du bien-être collectif, comme les normes d'honneur et de revanche (NR : 876-880).

La menace de sanctions face au non-respect d'une norme sociale pourrait motiver l'individu à adhérer à la norme par pur intérêt personnel, soit éviter la persécution. A ceci Elster apporte deux réponses. On peut d'abord admettre que l'individu considère la punition dans un calcul rationnel. Mais alors, qu'est-ce qui motive les « punisseurs » ? A première vue, ils seraient eux aussi sujets aux reproches d'autres individus s'ils évitent de châtier convenablement le fautif, mais alors nous sombrons dans une régression infinie. Selon Elster, cette régression ne va rarement très loin car plus on avance dans la chaîne punitive, plus la chance d'être perçu comme un violeur de (méta) normes est mince (RESN : 30). Si le groupe est suffisamment important, il s'agit que les individus affichent une *disposition* à punir[3] pour maintenir chacun dans le rang (Pettit 1990 : 739). Souvent aussi, cette peur d'être mal jugé se retrouve intériorisée chez l'individu de telle sorte qu'il ne transgressera pas de normes sociales même si personne ne le regarde ; les contraintes externes se transforment en préférences (Lindbeck 1996 : 11). On retrouve cette distinction chez Ullman-Margalit (1977 : 33-41) dans son traitement du « dilemme du soldat »[4] en élaborant deux types de solutions se qualifiant comme normes « PD », la discipline et l'honneur. Cette première concerne le respect des ordres supérieurs sous crainte de représailles sévères alors que pour la seconde la sanction est intériorisée. Elle précise alors que l'honneur est une solution préférable à la discipline car les intentions des compatriotes ne sont pas prises en considération. Un soldat discipliné cessera

---

[3] Cette disposition doit reposer sur des bases normatives. Les individus doivent croire fermement qu'ils risquent le châtiment s'ils transgressent la norme. Cette croyance devient sans fondements si tous les membres de la société (punisseurs inclus) agissent de manière parfaitement rationnelle car comme il a été démontré plus haut, il n'est pas dans l'intérêt (individuel, égoïste) d'un punisseur d'exécuter ses menaces.

[4] Ce jeu est exactement équivalent au DP: un soldat sur la ligne de front fait face à deux alternatives, combattre ou déserter. Si tous les soldats demeurent à leurs postes ils vaincront l'ennemi mais si un certain nombre font défection, ceux-ci survivront alors que les "braves" périront.

d'obéir lorsqu'un certain nombre de membres de son unité auront cessé eux aussi (un jeu d'assurance), alors qu'un soldat « noble » demeurera au front sans regarder autour de lui (un jeu trivial). L'honneur se qualifie comme norme sociale elstérienne, mais pas la discipline qui n'est en fait qu'une réaction rationnelle face à un système punitif efficace et connu de tous. La hiérarchie militaire garantit que les punisseurs exécuteront leur tâche, contrairement aux sanctions informelles rencontrées dans la vie de tous les jours. Dans le cas de l'honneur et autres normes semblables, le coût de transgression devient beaucoup plus que la quantité de désapprobation venant de l'extérieur. Il inclut aussi une réaction émotive négative résultant de la proche parenté entre émotions et normes[5]. Ce caractère profondément émotif des normes leur confère toute leur force, au-delà du calcul rationnel (CS : 99-100, RESN : 31). Une telle émotion pourrait certainement être la recherche d'approbation des autres, d'où la notion d'identité de groupe comme force pouvant soutenir les normes sociales (Bergstein 1985 : 118, Hirschman 1985 : 14). Chez Bergstein le groupe remplit deux fonctions : source de sanctions et lieu d'identification. L'identité demeure en dernière instance un « état contingent » (*by-product*) (Pettit 1990 : 741) dont Elster (SG) fait largement mention en spécifiant que ces états ne son pas atteignables intentionnellement.

Même si l'action normative se soustrait généralement aux canons de la rationalité, il ne s'en suit pas qu'elle soit de nature absolument non intentionnelle[6]. On peut considérer les normes sociales comme contraintes à l'intérêt particulier (par ex. ne pas être excessivement bruyant tard le soir) mais on peut aussi manipuler les normes selon ses intérêts. Certaines normes conditionnelles du type « l'action x doit entraîner l'action y » sont rationnellement manipulables dans le sens qu'il s'agit d'éviter x pour se soustraire à la norme ; si le dernier à sortir de la salle doit fermer les fenêtres par exemple, il s'en suivra que certains se précipiteront vers la sortie (RESN : 28). Aussi, il peut s'avérer rationnel de paraître irrationnel : faire croire à ses adversaires que l'on est lié à un code d'honneur strict peut contribuer à calmer

---

[5] Pour en savoir plus sur la théorie des émotions de Elster, lire RESN, SBW et NB ch. 7.

[6] Sandven (1995: 311) prétend que pour Elster, l'action normative a une origine directement causale, ne laissant aucune place à l'intentionnalité. Cette affirmation se révèle fausse, car Elster prétend à maintes reprises (nous allons le voir) que l'intentionnalité de l'acteur peut jouer un rôle important dans ce type d'action.

leurs ardeurs. Cependant, la norme d'honneur doit être présente dans cette société, et la seule preuve demeure l'existence d'individus respectant fidèlement la norme. Ainsi, tous ne peuvent faire semblant afin d'en tirer quelque avantage car alors la norme perd tout son pouvoir de motivation qui, rappelons-le, réside largement sur la disposition à la sanction de la part de ses pairs (NR : 875-76). La dichotomie rationalité - normes n'est pas aussi absolue qu'on pourrait imaginer, car l'intentionnalité peut agir directement sur les normes. Ceci nous conduit vers la notion de liberté : les normes ne sont pas des prescriptions obligatoires, leur pouvoir émotif est significatif mais il est toujours possible pour un individu de les violer s'il est prêt à en subir les conséquences. La violation d'une norme semble être l'exception confirmant la règle que les normes sont inviolables. Une norme constamment violée ne peut évidemment plus porter le qualificatif de « norme ». Donc, la violation occasionnelle demeure contingente à l'adhésion générale à la norme (NR : 878-80).

Il existe deux schémas possibles d'influence des normes sur l'action. Le premier est la norme comme contrainte à l'action (fig. 15). Ici, la norme agit comme un filtre à la décision rationnelle et la seule façon intentionnelle d'éviter la norme est de redéfinir la situation de décision (*framing*). Elster supporte plutôt le schéma de la fig. 16 qui permet une certaine interaction entre la prescription normative et les préférences individuelles (RMCA : 153). Le problème majeur rencontré par ce modèle est la possibilité de confusion entre préférences et normes comme sources de l'action. Comme la théorie du choix rationnel se propose d'examiner les préférences à partir de l'action, il devient impossible - du moins dans la version orthodoxe de la théorie - de distinguer entre les deux. En effet, toute tentative de distinction entre normes et préférences se révèle rapidement très complexe. A la lecture de Elster (SG : 117-19, NB : 40) et de Griffith et Goldfarb (1991 : 65-66) on peut affirmer de façon simpliste que les désirs ont une origine interne pouvant être extérieurement modifiée - soit par la socialisation ou encore par le *character planning* de Elster - et que les normes constituent des sources causales d'action provenant de l'extérieur tout en étant sujettes à des modifications intentionnelles à l'interne[7]. Notamment, les normes sociales doivent respecter un minimum d'intérêt particulier

[7] Griffith et Goldfarb affirment que les valeurs morales (servant pour eux de base aux normes) diffèrent des simples préférences par leur caractère intentionnellement transformable, tandis que Elster prétend que même les

(Pettit 1990 : 726, Bicchieri 1993 : 231), du moins à long terme (Williams 1988 : 11), sinon il s'en suivra une déception qui conduira l'individu à agir rationnellement contre celles-ci. En résumé, les normes chez Elster constituent de forts incitatifs à l'action mais elles peuvent exceptionnellement être manipulées ou surpassées par les intentions du sujet.

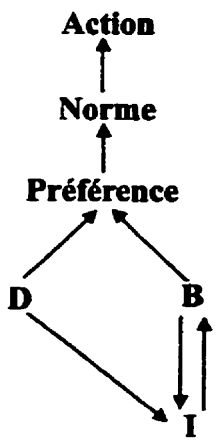


Fig. 15

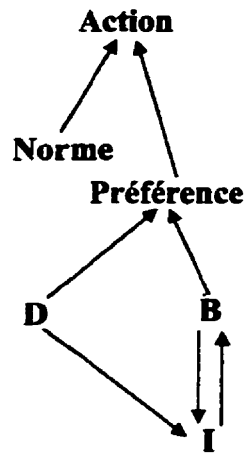


Fig. 16

I.2 - Normes sociales et collectivité -

Passons maintenant de la rationalité individuelle à la rationalité collective. Il a été démontré que les normes ne s'expliquent pas par le bénéfice accordé à l'individu. En est-il autrement en ce qui a trait au bénéfice du groupe ? Plusieurs normes contribuent manifestement au bien être de la collectivité, comme les conventions sociales et les normes contribuant à éliminer le resquillage ; il devient alors tentant d'y voir l'explication de l'origine ou de la persistance de celles-ci. En sciences sociales, le fonctionnalisme d'inspiration darwinienne se penche sur ces possibilités.

---

préférences peuvent être modifiées de la sorte, mais ce processus de *character planning* est plus compliqué à réaliser car la méthode doit être indirecte, notamment par l'engagement (*precommitment*) (US: 39-46).

Le fonctionnalisme explique un événement par ses effets bénéfiques. Le paradigme d l'explication fonctionnaliste se retrouve en biologie, où les caractéristiques des organisme vivants sont reliés à l'accroissement de sa capacité de reproduction. Les diverses mutation aléatoires que subissent les organismes répondent à un mécanisme de sélection naturelle : le mutations bénéfiques contribuent à la survie de l'espèce qui reproduit alors celles-ci dans le gènes de ses enfants alors que les organismes victimes de mutations moins réussies périclitent e disparaissent. Les normes sociales s'expliquerait d'une manière semblable Les sociétés respectant des normes bénéfiques auraient plus de chance de survivre et de s propager que celles aux normes destructives. Elster soutient qu'un tel mécanisme d'explicatio est possible en sciences sociales, mais ses applications demeurent très limitées. En science sociales, il n'existe pas de principe explicatif général semblable à la « loi de la jungle » e biologie ; il faut spécifier la nature de celle-ci pour chaque analyse fonctionnaliste Le concept même de « reproduction » se révèle difficilement exportable vers l'analyse sociale ; une société bien adaptée ne laisse pas plus de copies d'elle-même que les autres[8] (Rosenberg 1988 : 136-37). Aussi, l'explication en termes de bénéfices demeure très vague, il faut encore une fois préciser dans chaque cas ce que l'on entend par « bénéfices ». Une explication par les bénéfices collectifs dérive dangereusement de la théorie du choix rationnel vers un collectivisme méthodologique où le groupe maximise une utilité propre, indépendant (du moins directement) des utilités de ses membres. L'alternative la plus intéressante, proposée par Howard Margolis, consiste à considérer l'équilibre social créé par un respect universel d'une norme comme un bien collectif : « (...) what makes the norm socially valued is not some functional contribution to social efficiency but only the looser propensity to social coordination for its own sake » (Margolis 1990a : 834). Cette préférence pour la stabilité sociale provient d'un mécanisme darwinien selon lequel les sociétés bien ordonnées survivent plus longtemps que les sociétés chaotiques (Margolis 1990a : 830-834 ; 1990b : 248). Le lien entre utilité individuelle et sociale devient alors facile à obtenir : il est dans l'intérêt de chacun de prévenir le désordre parmi le

[8] Gibbard (1985: 18) considère l'intériorisation des normes comme un mécanisme d'adaptation biologique propre aux créatures sociales servant à augmenter les chances de survie. Ainsi les normes sont associées aux rituels (de combat, de séduction, etc.) des animaux. Bien que ce modèle résoudrait le problème de l'utilité collective, il est difficile de croire que les normes soient innées chez l'humain et non acquises. Elster (CS: 139) précise que toutes les normes n'apportent pas de bénéfices, autant à l'individu qu'à la collectivité.

groupe auquel il appartient. Toute autre explication par l'utilité sociale se bute d'une manière ou un autre au problème du resquillage : à la marge, un individu préférera la défection si le gain personnel excède la perte sociale et que le risque de détection est faible (Griffith, Goldfarb 1991 : 74). Il faut donc être très prudent dans l'utilisation de l'explication fonctionnaliste en sciences sociales (ETC : 49-68).

Une autre explication possible de l'origine des normes sociales réside dans le contexte culturel. Ici le mécanisme devient l'apprentissage de la part des individus ; on parle alors soit d'éducation[9] ou de béhaviorisme. Les valeurs partagées par la société sont bien entendu primordiales dans la détermination des normes sociales en vigueur (Bergstein 1985 : 118 parle de « considerable significance »). Par exemple, l'individualisme libéral caractéristique de nos sociétés occidentales conduit à des normes politiques solides comme le devoir de l'électeur (permettant de résoudre le paradoxe du vote en choix rationnel) et une absence relative de normes économiques, où le laissez-faire et l'enrichissement sans gêne prédominent (Bergstein 1985 : 130 ; Mueller 1986 : 10). Marx concevait aussi les normes - le système judiciaire entre autres - comme dépendantes du mode de production économique particulier de la société. Un problème avec cette approche réside dans le concept même de culture : si elle se pose en termes de normes sociales, nous nous enfonçons alors dans un cercle vicieux (Ullman-Margalit 1977 : 9). Pour reprendre l'exemple ci-haut, le libéralisme reposerait lui-même sur les règles et les normes qu'il contribue à maintenir par la socialisation - un autre thème qui n'est pas étranger au marxisme.

Une dernière alternative consiste à étudier non pas l'origine historique ou fonctionnaliste des normes, mais d'en étudier les conditions d'existence générale. Cette méthode de travail permet à Ullman-Margalit d'affirmer par exemple que les situations de jeu correspondant au PD contribuent à l'émergence de normes d'obligation (ou normes « DP ») (Ullman-Margalit 1977 : 8, 12-13). Les normes peuvent servir de supplément à la théorie du choix rationnel lorsque les individus ne peuvent formuler aisément de décisions en raison d'un

---

[9] Mueller prétend entre autres que la solution coopérative du DP se rencontre fréquemment parce que les individus apprennent les vertus de la coopération dès leur plus jeune âge (Mueller 1986: 5).

manque d'informations fiables. Lors de situations critiques où une décision doit être prise peu importe si elle ne correspond pas au meilleur moyen d'en arriver à une fin donnée, un mécanisme alternatif d'action comme l'habitude - une norme personnelle (Skyrms 1990 : 126)[10] - ou le code de comportement ou « loi non-écrite » - une norme sociale (Bergstein 1985 : 114 ; Griffith, Goldfarb 1991 : 49) - sert souvent à trancher les noeuds. Le terme « avoir le sens du jeu », utilisé par Bourdieu (1994) résume à merveille le mécanisme d'action à l'étude ici[11]. Les sports d'équipe débordent de ces codes permettant aux joueurs de réagir convenablement aux situations peu communes où le temps de réaction prend un aspect primordial. On remarque aussi le phénomène inverse, plus les paramètres de choix sont clairs et précis, moins la portée des normes est grande (Taylor 1987 : 111). Elster reconnaît la faillite du choix rationnel en situation de grande incertitude mais il demeure réticent à attribuer un rôle significatif aux normes sociales. Pour des raisons épistémologiques, les normes doivent se détacher le plus possible du choix rationnel, et non seulement les compléter (NB : 35-36). Les situations d'information asymétrique entre les individus ou les groupes, comme par exemple dans les domaines médical et légal, pourraient toutefois correspondre à l'idéal d'Elster. L'information asymétrique ouvre toute grande la porte à une exploitation *rationnelle* - parfaitement compatible avec le choix rationnel - de l' « ignorant » par le « savant » ; certaines normes contre l'exploitation peuvent alors apparaître, comme les codes de déontologie des organismes professionnels[12] (Bergstein 1985 : 114 ; Griffith, Goldfarb 1991 : 49).

---

[10] Il n'est pas certain que cela soit le point de vue de Skyrms, car plus loin il affirme que "If all payoffs have been taken into account and it is still evident in a particular case that acting against the habit has higher expected utility than acting in accordance with it, then the rational decisionmaker will prefer to act against the habit" (1990: 129). Comment peut-on calculer l'utilité espérée complète (*all payoffs have been taken into account)* d'une situation permettant d'évaluer la valeur d'une habitude qui existe précisément parce que le calcul d'utilité se révèle incomplet? L'auteur nage ici en pleine contradiction logique.

[11] Bourdieu toutefois n'endosserait certainement pas cette utilisation de cette expression. Même si le "sens du jeu" constitue une authentique formule de décision en cas d'incertitude, il va en fait beaucoup plus loin en affirmant qu'il est *impossible* pour quiconque d'appliquer avec régularité le processus rationnel de décision pour diverses raisons, notamment l'impossibilité pour un individu de viser intentionnellement des fins spécifiques et l'irréductibilité des intérêts en valeurs d'utilité. Cette critique acerbe du choix rationnel fut l'objet de fréquentes réfutations par Elster, voir entre autre NB: 14-15, IRC: 22-23, ETC: 85.

[12] Pour une analyse détaillée des codes de déontologie dans une perspective rationaliste, lire Bergstein (1985: 122-127).

Les normes sociales ne se laissent pas facilement conceptualiser. Comparativement à la théorie du choix rationnel, les tentatives de définition des normes apparaissent floues et peu convaincantes. L'essentiel à retenir demeure la « fonction » de la norme comme agent causal de l'action individuelle, sujette à l'intervention limitée de l'intentionnalité. De cette façon les normes peuvent servir de complément utile au choix rationnel. Cette citation de Elster résume admirablement toute la difficulté du projet :

> « There can be little doubt that the appeal to norms sometimes enables us to explain why rational choice fails. The appeal tends, however, to be *ad hoc* and *ex post facto*. There does not exist a robust, well confirmed theory specifying the conditions under which, and the limits within which, norms override rationality. In particular, we do not know what determines when norms remain strong and stable and when they yield to the pressure of self-interest » (IRC : 24).

L'origine des normes constitue beaucoup plus une question empirique de « cas par cas » que l'objet d'une théorie générale (Griffith, Goldfarb 1991 : 65). Dans ce travail l'origine nous importe peu de toutes façons. L'important est de se concentrer sur les effets des normes sociales sur les situations problématique d'action collective.

II - Les normes comme alternative au choix rationnel

II.1 - Les « motivations mixtes » de Elster -

Nous avons conclu précédemment que des individus pleinement rationnels peuvent en arriver à la coopération dans le DP sous certaines conditions assez contraignantes. Par contre, certaines motivations extra-rationnelles se révèlent particulièrement bien adaptées pour l'analyse du DP. Elster en retient quatre : la norme kantienne, l'utilitarisme, le participationnisme et le devoir communautaire.

Le terme « kantien » ne désigne pas des disciples de Kant, mais bien un type d'individus répondant à un impératif particulier de Kant : « (...) one cannot will X if the notion of all doing X harbours a logical or pragmatical contradiction » (CS : 192). La norme dérivée de cet

impératif a toutefois une application plus large, elle serait plutôt du genre « un individu ne peut entreprendre l'action X si le fait que tous entreprennent X apporte des pénalités à tous », admirablement résumée par Elster par l'interrogation « What if everybody did that ? » (CS : 193). En termes de la théorie des jeux, le kantien compare l'utilité des stratégies D^{∞} et C^{∞} et prend une décision de coopérer ou non sans se soucier des autres alternatives. Donc, les kantiens n'ont pas de raisons de ne pas coopérer dans un DP, car la défection universelle y est par définition pire pour tous que la coopération universelle. La norme kantienne peut toutefois générer des situations sous-optimales. Dans le cas où l'utilité pour tous serait classée dans cet ordre : U = (coopération partielle, coopération universelle, défection universelle), ceux-ci préféreraient la coopération universelle et ainsi n'arriveraient pas à atteindre le maximum d'utilité.

Il serait faux de croire que les kantiens participent aveuglément à la production de biens publics, pour la même raison que le pouvoir des normes sur le *psyché* n'est pas absolu. Le kantien refusera de coopérer - avec raison - si son coût personnel est trop grand par rapport à ce qu'il peut endosser[13]. Cette contrainte sera affectée par la « valeur » de la coopération universelle, définie par la différence entre $U(C^{\infty})$ et $U(D^{\infty})$, ou C(N) et D(0) sur le graphe de Schelling. Plus cette valeur sera petite, plus le kantien tiendra compte de ses coûts de coopération (CS : 192-3). Par contre, il est typiquement insensible au nombre probable de participants, ce qui peut amener de graves problèmes pour la collectivité comme le démontre la fig. 17. où l'action de n kantiens ($n < A$) plonge l'utilité de tous vers un résultat pire que D^{∞}, la situation originale. Cette propension que Elster nomme « égoïsme moral » démontre que la norme kantienne n'a pas que des bons côtés dans le problème d'action collective.

[13] Ce qui les différencie des individus pleinement rationnels, qui refusent de participer si leur coût est supérieur aux bénéfices.

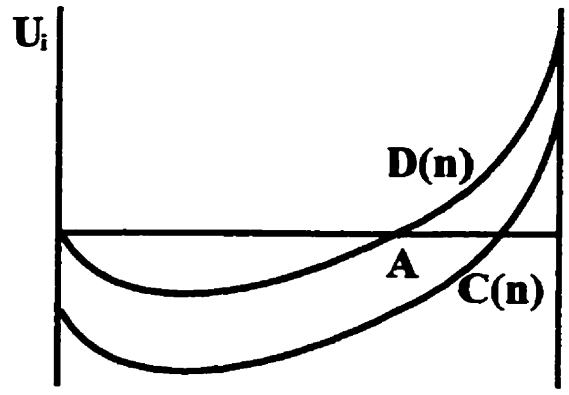


Fig. 17

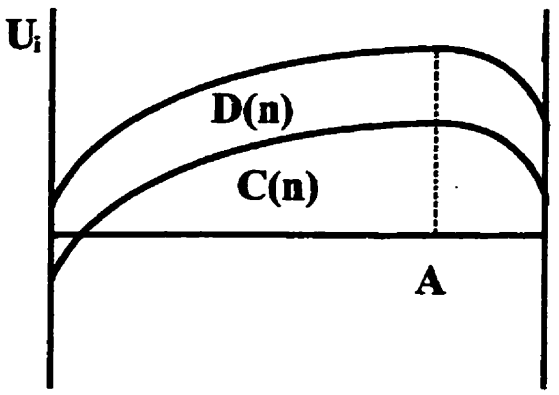


Fig. 18

Même avec une population entièrement composée de kantiens, un résultat sous-optima peut survenir. La fig. 18 représente une situation où le niveau optimal de participation se situ un peu en deçà de la coopération universelle. Une solution envisageable consiste à reformuler l situation de choix : en plus de C^{∞} et de D^{∞}, pourquoi pas y rajouter une option de type C_A coopérer ssi *A* personnes ou moins coopèrent, faire défection sinon[14] (Barry 1985 : 157) Tout d'abord, Elster prétend qu'une considération du nombre optimal de participants dans l situation de choix ne respecte pas la propriété non-conséquentialiste du kantisme (CS : 194) Aussi, ce « kantisme sophistiqué » ne précise aucunement qui collaborera et qui pourr resquiller. A ceci, Barry propose une loterie attribuant à chacun son rôle, mais cette solutio définit la norme kantienne de manière trop spécifique pour être d'un quelconque intérê (Griffin 1985 : 116). Le kantianisme doit demeurer un mécanisme de décision applicable à un grande variété de situations sociales. Et sans loterie, il est difficile de voir comment les kantien sophistiqués arriveront à un équilibre optimal, particulièrement si tous doivent prendre leu décision de coopérer ou non simultanément[15].

[14] Ou encore, l'emploi de stratégies mixtes.

[15] Si l'on permet une décision séquentielle (le premier joueur décide, ensuite le second, ...), le problème se déplace vers la constitution de l'ordre de décision, car les derniers à choisir soutiront plus de l'arrangement que les premiers, à cause du resquillage au-delà de A.

Selon Elster, il faut voir dans la norme kantienne une irrationalité fondamentale qu'il nomme « pensée magique » et qu'il pose comme suit : « (...) the belief that by acting on the symptoms one can also change the cause » (CS : 196). Dans le contexte de l'action collective, ce principe prétend que l'action individuelle de coopération peut *causer* le reste du groupe à faire de même, par l'argument de symétrie des acteurs rationnels que nous avons analysé dans le cadre du DP à deux participants joué qu'une seule fois. Rappelons que l'argument de symétrie suppose que si un individu prend une décision, il s'en suivra que son semblable fera nécessairement de même puisqu'ils sont tous deux pleinement rationnels. Cette « illusion socialement bénéfique » (RMCA : 145) ou motivation « logically defective but psychologically effective » (Ullman-Margalit 1977 : 58) peut se justifier dans de petits groupes aux liens interpersonnels étroits : « The more people are like oneself, the more plausibly (...) one can infer that they will behave like oneself » (CS : 208). Ici, une irrationalité manifeste peut actuellement contribuer à la solution au DP.

L'utilitarisme s'apparente à l'altruisme dans le sens qu'il considère l'utilité des autres dans la prise de décision, avec la particularité que l'utilitariste accorde exactement la même importance à tous, lui-même inclus. Les utilitaristes ne contribuent au bien public que si leur coopération augmentera l'utilité de tous. En fait, un tel individu compare ses coûts personnels de coopération avec les bénéfices de tous (CS : 48). Les utilitaristes ne participent pas à l'action collective lorsqu'ils se trouvent sur la pente décroissante ou nulle de l'utilité collective (CS : 203). Une exception cependant couvre les cas où la pente décroissante se situe à gauche d'un graphique de Schelling et n'est pas trop longue, tout comme la fig. 17 plus haut. En d'autres termes, si une faible participation (en nombres de coopérants) entraîne une perte nette pour tous mais que cette perte peut être rapidement comblée par l'ajout d'une poignée de coopérants, alors les utilitaristes devraient considérer leur situation comme un jeu d'assurance à l'intérieur du DP s'ils se trouvent en nombre suffisant pour redresser la courbe à eux seuls. L'information et la coordination des efforts assureraient la collaboration de tous les utilitaristes. En situation d'information imparfaite il peut arriver que l'utilitariste ne puisse savoir où il se

trouve sur la courbe, ni même à quoi la courbe ressemble[16]. Dans ce cas une forme d'éthique utilitariste préconiserait la coopération, car la coopération universelle étant toujours supérieure à la défection universelle, un déplacement du niveau de coopération vers la droite sur un graphique de Schelling doit être, en général, bénéfique pour tous. Cette formulation de la norme utilitariste en situation d'ignorance partielle ressemble énormément à la norme kantienne (RMCA : 150).

Il existe un type d'individu dont le calcul d'utilité ne se limite pas aux valeurs représentées dans la matrice de jeu. Une seconde source d'utilité est à considérer, les bénéfices de participation. Ces individus - appelons-les « activistes » - recherchent à la fois la maximisation de l'utilité « classique » et le plaisir de l'implication dans un projet collectif. Ce concept d'utilité multiple ne doit pas être pris à la légère. Elster prend bien soin de préciser que la maximisation classique doit demeurer prioritaire à l'utilité de participation (RMCA : 147 ; MF : 121). Cette dernière dépend de la première au sens que si l'action collective échoue, le bénéfice de participation deviendra immédiatement nul. Elster distingue parmi deux niveaux de participation : les activistes d'élite motivés par le désir de fonder un mouvement et d'y être parmi les personnalités importantes et les activistes de masse, motivés eux par le désir de mettre l'épaule à la roue de projets d'envergure[17]. En résumé, ces premiers coopèrent au début du mouvement et ces derniers ne prennent part que plus tard, lorsque suffisamment de gens auront joint le mouvement (CS : 203-4).

Les individus motivés par le devoir communautaire (*fairness*) ou par l'identité au sein du groupe (Hirschman 1985 : 14-15 ; Bergstein 1985 : 118) prennent leur décision à partir d'une seule variable, la coopération (ou défection) immédiate des autres au moment de se joindre : « (...) cooperate if and only everybody else, or a substantial number of others, cooperate »

---

[16] Au minimum, il doit savoir qu'il se trouve dans un DP.

[17] On rencontre souvent ces deux types dans l'entreprise privée. Il y a des entrepreneurs qui ne font que démarrer des entreprises qu'ils revendent aussitôt qu'elles deviennent autosuffisantes et il y en a d'autres qui préfèrent acquérir des entreprises bien établies pour les mener à bon port.

(CS : 187). Le seuil critique de coopération de la part des autres varie grandement pour chaqu individu. Ces individus ne considèrent jamais les bénéfices de participation ; ils ne font qu suivre le troupeau. Par contre, ils peuvent très bien inclure les coûts directs de coopération dan leur prise de décision[18]. On pourrait croire que la stratégie TFT correspond à la norme d devoir, car elle aussi dépend uniquement de la coopération des autres, mais il n'en est pas ainsi La principale différence réside dans le fait que pour la TFT, ce qui compte c'est la coopératio *antérieure*, c'est à dire à la ronde précédente du tournoi. Lorsqu'un système TFT se situe e équilibre, la coopération future devient *anticipée*. La norme de devoir considère plutôt l coopération *actuelle* et ne possède aucun mécanisme permettant d'anticiper les comportement futurs des membres du groupe. Une norme n'est pas une stratégie, et de fait ne s'encombre pa de notions d'équilibre (CS : 188). Elle peut toutefois contribuer à *maintenir* un équilibre comme nous le verrons plus loin.

II.2 - Alternatives au modèle elstérien -

On s'en doute, outre ces quatre motivations il existe d'autres alternatives au choi rationnel que Elster n'a qu'effleuré au passage. La littérature rationaliste étendue se penchera sur trois motivations s'adaptant assez bien au choix rationnel, permettant l'élaboration d *supports* à la théorie plutôt que d'*alternatives* pures et simples comme le souhaite Elster. Ce motivations particulières sont l'altruisme, la confiance et une norme fondée sur la célèbre stratégie TFT.

Dans la théorie du choix rationnel, l'altruisme est généralement défini par l'inclusion d'utilités d'autres individus dans sa propre fonction d'utilité. La motivation altruiste est alors considérée comme une externalité, elle ne comporte donc que très peu de difficultés supplémentaires au choix rationnel (Sen 1982 : 92-93 ; Brennan 1991 : 87). En théorie formelle, il s'agit simplement d'additionner sa propre utilité avec celles des autres. Pour plus de réalisme, ces utilités doivent être pondérées ; la variation de la pondération permet d'obtenir

[18] En fait cette précision devient redondante, car il a été spécifié que chez Elster, toute prescription normative peut être rejetée au nom de la rationalité dans des situations exceptionnelles. Le coût prohibitif de participation correspond parfaitement à cette condition.

différents degrés d'altruisme. Il est à noter que la pondération dépend d'une certaine manière du profit potentiel du resquillage ; plus il est élevé, plus l'individu penchera vers l'égoïsme (Mueller 1986 : 12). Suivant M. Taylor (1987 : 113-14) et D. Mueller (1986 : 7) [19] nous adopterons l'équation suivante :

$U_A = W_{A0}\, U_0 + W(U_1 + U_2 + \ldots + U_n)$

U_A -> Utilité de l'altruiste

W -> Pondération de l'utilité de n

U -> Paiement de n dans la matrice de jeu

n -> individu concerné (n = 0 -> soi-même).

Pour plus de simplicité nous ne considérerons qu'un jeu à deux participants, A et B alors :

$U_A = W_{A0}\, U_0 + W_{A1} U_1$ et $U_B = W_{B0}\, U_0 + W_{B1} U_1$

Bien entendu, une telle modification de l'utilité pourrait entraîner une transformation qualitative du jeu. Il devient alors intéressant de tester si une fonction altruiste d'utilité peut transformer le DP en un jeu où la coopération devient une stratégie dominante, solutionnant ainsi le problème.

	C	D
C	c,c	a,d
D	d,a	b,b

Fig. 19

	C	D
C	$(W_{A0}+W_{A1})$ c, $(W_{B0}+W_{B1})$ c	W_{A0} d+W_{A1} d, W_{B0} a+W_{B1} a
D	W_{A0} d+W_{A1} a, W_{B0} a+W_{B1} d	$(W_{A0}+W_{A1})$ b, $(W_{B0}+W_{B1})$ b

Fig. 20

D'après le DP de la fig. 19 (où $a < b < c < d$ et $(a+d)/2 < c$), nous obtenons en ajoutant les fonctions altruistes des deux joueurs A et B le jeu de la fig. 20. Afin que le résultat (C, C) soit à la fois Pareto-optimal et en équilibre de Nash, il s'agit que (pour le joueur A par exemple) :

a) (C, C) > (D, C), soit $(W_{A0} + W_{A1})\, c > W_{A0}\, d + W_{A1}\, a$

[19] En fait, Mueller ne pondère pas l'utilité propre de l'individu (Uo) mais ce détail ne faussera pas les conclusions.

b) (C, D) > (D, D), soit $W_{A0}\, a + W_{A1}\, d > (W_{A0} + W_{A1})\, b$

c) (C, C) > (D, D), soit $(W_{A0} + W_{A1})\, c > (W_{A0} + W_{A1})\, b$

Cette dernière condition se simplifiant par $c > b$, elle est automatiquement remplie. Pour ce qui est des deux autres, nous obtenons en isolant W_{A0} (Taylor 1987 : 113) :

a) $W_{A0} < (c-a)/(d-c)\, W_{A1}$

b) $W_{A0} < (d-b)/(b-a)\, W_{A1}$

Puisque $c > (a+d)/2$, il s'en suit que $c-a > d-c$ et $d-b > b-a$, donc $(c-a)/(d-c) > 1$ et $(d-b)/(b-a) > 1$. Dans le cas limite où ces deux rapports tendent vers 1, si le « facteur d'égoïsme » W_{A0} est plus petit ou égal au « facteur d'altruisme » W_{A1} alors le DP se transforme en jeu trivial où la stratégie de coopération devient dominante pour le joueur A. Lorsque la fonction d'utilité des deux joueurs affichent cette même caractéristique, la solution (C, C) sera toujours rationnellement atteinte. Par conséquent, sous certaines conditions, même le résultat (C^{∞}, C^{∞}) peut devenir un équilibre dans le DP répété plusieurs fois (Taylor 1987 : 120-21).

Le comportement altruiste permet donc de solutionner le DP mais malheureusement pour nous, cela ne constitue pas une norme sociale. Alors que les économistes voient l'altruisme comme une externalité, d'autres le voient comme une disposition psychologique (CS : 47) ou une attitude morale (Sen 1982 : 92, Mueller 1986 : 8). Dans tous les cas, l'altruisme n'est généralement pas une valeur partagée avec les autres et sujette à approbation ou désapprobation. Pour Elster, le don anonyme et inconditionnel à un organisme de charité constitue l'exemple parfait d'un comportement altruiste qui ne peut faire l'objet d'aucune sanction sociale (CS : 47). Il faut toutefois prendre garde à la confusion qui peut se glisser lorsque le résultat d'un comportement normatif ressemble à de l'altruisme, comme par exemple le kantien qui décide de produire seul un bien public (NB : 56-57). Il demeure possible que l'altruisme prenne la forme d'une norme sociale, notamment dans un petit groupe où les faits et gestes des individus sont observables par tous, mais la généralisation de ce type de comportement dans des groupes plus larges où l'anonymat devient la règle s'avère irréaliste (Taylor 1987 : 110, Williams 1988 : 12).

Une autre disposition psychologique déterminante en théorie des jeux est la confiance. Sa contribution à la solution coopérative du DP est simple : si les joueurs se font mutuellement confiance, ils peuvent s'entendre sur la stratégie C pour tous sans trop de problèmes. Encore ici, cette disposition demeure une affaire de petits groupes, car par définition on ne peut faire confiance qu'à quelqu'un que l'on connaît et avec qui on aura d'autres relations dans l'avenir impliquant donc un DP répété (Riker 1980 : 11 ; Kee, Knox 1970 : 363 ; Williams 1988 : 12). De toute façon, la confiance peut être modélisée en choix rationnel comme une décision en situation de risque, le degré de confiance que l'on attribue à son partenaire devient la probabilité subjective nécessaire à ce type de fonction d'utilité (Kee, Knox 1970 : 359 ; Riker 1980 : 3).

Il a été démontré précédemment qu'une forme subtile de confiance peut s'imposer dans un DP répété indéfiniment lorsque la stratégie TFT adoptée par un certain nombre constitue un équilibre stable. Cette stratégie possède la caractéristique de ressembler à une norme car elle implique des sanctions sociales aux resquilleurs et parallèlement, une récompense pour tous lorsque personne ne fait défection. Mais ce système de sanctions ne répond pas aux critères d'une norme sociale car il est avant tout *rationnel*, la menace de sanctions constituant la raison d'être de l'état d'équilibre. Lorsque le joueur A fait défection en réponse à la défection de B au tour précédent, ce n'est pas parce qu'il croit que la défection est un « mal » et donc sujette à une réprimande, mais bien pour ne pas se faire exploiter par B. Donc, A réagit uniquement en fonction de la maximisation de son utilité dans le jeu. En fait, c'est ce qui confère au TFT toute sa force de persuasion : la menace de sanctions étant fondée sur un calcul rationnel, il devient hautement prévisible qu'elle sera appliquée en cas de besoin.

La question du resquillage ne se pose plus une fois l'équilibre TFT atteint, le jeu devenant fonctionnellement équivalent à la coopération universelle. Toutefois, cet équilibre peut être déjoué par un individu ayant une certaine connaissance de la fin probable du jeu, du moins en ce qui le concerne. Dans ce cas il adoptera une stratégie du type TFT-n récoltant ainsi les bénéfices du resquilleur vers la fin du jeu. L'efficacité de TFT-n face à TFT accentue le caractère rationnel du système de sanctions en démontrant qu'un joueur n'accepte de les subir

que s'il s'attend à en sortir toujours gagnant à la conclusion de la partie. Ce faisant, la défection universelle suivant une défection unilatérale perd de sa force punitive tout en désavantageant les non-resquilleurs. L'équilibre TFT s'écroule alors, tout comme nous avons observé dans le chapitre précédent. Cette crainte de la présence d'un joueur TFT-n parmi un groupe de TFT peut amener certains individus à reconsidérer leur position sur le type de sanctions à imposer. Puisqu'une défection par un seul joueur s'effectue au détriment de tous les autres joueurs TFT, ceux-ci peuvent en venir à appliquer des punitions externes au jeu, comme l'ostracisme par exemple. Celui qui tente de profiter d'un groupe en équilibre TFT n'aura plus l'occasion de participer aux activités du groupe à l'avenir, *même si la prochaine interaction ne se qualifie pas comme un DP*. C'est alors que les membres du groupe internalisent la sanction du resquillage (Pettit 1990 : 736-37) ; leur stratégie devient une norme sociale en bonne et due forme, la « norme TFT », qui opère comme suit : appliquer la stratégie TFT mais ne jamais être le premier à faire défection. Le respect de cette norme par tous les joueurs assure un équilibre coopératif infaillible et donc une solution normative au DP répété, tandis qu'un respect partiel de la norme diminue d'autant les chances de dérivation de l'équilibre que le nombre de participants TFT « normatifs » est grand.

Selon cette conception de la norme TFT, les joueurs doivent d'abord atteindre l'équilibre TFT avant d'adopter une disposition à ne pas transgresser cet équilibre. Or, les calculs de Taylor ont démontré que plus le groupe est large, plus les conditions d'équilibre deviennent difficiles à respecter. Bicchieri (1993) propose une solution originale : les joueurs développent une habitude TFT lors de situations de DP impliquant de petits groupes et transposent cette habitude dans les interactions à participation plus élevée. Tout comme une norme et même au-delà, la stratégie TFT se prête très bien à l'apprentissage de type béhavioriste car elle punit les resquilleurs et récompense les coopérants de manière directe et perceptible (Mueller 1986 : 17). Selon la procédure de Bicchieri, un petit groupe (deux joueurs par exemple) participent à un DP répété indéfiniment et tentent d'en arriver à un équilibre autre que la défection universelle. Ce type d'interaction possède certains avantages sur une négociation à plus grande échelle. D'abord, du point de vue de la théorie formelle des jeux, la possibilité d'atteindre un équilibre coopératif tel que TFT est généralement inversement

proportionnelle au nombre de participants. Aussi à un niveau moins formel, la communication postérieure au jeu y est plus aisée, ainsi que la possibilité de confiance mutuelle. Bicchieri précise que les membres d'un petit groupe ont tendance à réfléchir stratégiquement à la situation, tenant compte des utilités des autres tandis que la tendance s'apparente plutôt à la décision paramétrique dans les groupes plus importants où la condition d'insignifiance est en vigueur (Bicchieri 1993 : 243). Ce faisant, les joueurs en arriveront probablement à préférer la stratégie TFT, principalement à cause de son paiement élevé en utilités, son mécanisme efficace de protection contre les resquilleurs et sa facilité d'apprentissage (Bicchieri 1993 : 245). Un succès répété de l'équilibre TFT au niveau local peut ainsi amener les individus à adopter instinctivement la même stratégie dans les situations où les joueurs sont plus nombreux[20]. Ici, la transformation de la stratégie en norme suit le même principe que précédemment.

III - Une typologie des motivations

III.1 - Normes sociales, personnelles et émotions -

Malgré le titre de ce chapitre, les motivations extra-rationnelles qui y furent étudiées ne font pas toutes partie de la catégorie des normes sociales. Ceci reflète en fait la tendance même de Elster d'inclure des motivations de nature controversée dans ses analyses, particulièrement dans CS. Cette confusion peut provenir de deux sources : une interprétation trop large de la définition opérationnelle des normes sociales ou une volonté d'examiner un ensemble de comportements le plus complet possible, quitte à diverger des normes sociales. Je crois qu'il nous faut préférer cette dernière explication, car la définition des normes sociales chez Elster demeure tout de même assez précise. Nous allons résumer ici toutes les motivations rencontrées au long de ce chapitre mais il sera avant tout nécessaire de postuler une formulation définitive des normes sociales.

---

[20] Cette dernière proposition ne fait pas partie du modèle de Bicchieri. A la place, elle élabore un système évolutif à la Axelrod où les petits groupes envahissent les plus gros et finissent par imposer TFT à tous. Pour elle, la *norme* TFT semble correspondre exactement à la *stratégie* TFT car elle affirme que la norme ne peut être respectée que lorsqu'elle implique un équilibre de Nash correspondant (Bicchieri 1993: 253). Pourtant, la force des normes est justement de permettre à un règlement hors-Nash de se réaliser.

Lorsqu'on met de côté le modèle rationnel de l'individu, on retrouve chez Elster deux types de motivations, les normes sociales et les dispositions psychologiques. La définition imprécise de ce dernier nous force à le diviser en deux sous-catégories. Si la disposition fait appel aux valeurs morales ou à un idéal de société, nous parlerons alors de normes personnelles. Si elle se fonde sur un penchant psychologique, une caractéristique de la personnalité de l'individu, il sera question de véritables émotions. Remarquez que la distinction entre les deux demeure extrêmement floue et arbitraire, mais elle s'impose car les normes personnelles ont la possibilité de se transformer en normes sociales lorsque l'individu *croit* qu'il est sujet aux sanctions d'autrui. Ces types se distinguent par le degré de respect des deux postulats fondamentaux propres aux normes sociales selon Elster, élaborés au tout début de ce chapitre : la non-maximisation de l'utilité et la crainte de sanctions. Examinons en détail ces postulats chez l'acteur normatif social (ANS).

1) *L'ANS ne maximise pas consciemment son utilité.*

Ce postulat, l'essence même de la dichotomie rationalité-normes, peut signifier deux choses : l'absence de choix lors de la décision ou l'absence d'évaluation des retombées de l'action.

1a) *L'ANS ne prend pas de décision.*

C'est l'injonction de Elster « Do it ! » opposé au mécanisme de décision conditionnelle du choix rationnel. L'individu est en quelque sorte forcé d'agir selon les diktats des normes sociales. Bien que ceci soit une vision assez juste du phénomène, il est important de remarquer que ce postulat ne tient pas lors de certaines situation extrêmes. Un concept minimal de choix demeure toujours à la disposition de l'ANS.

1b) *L'ANS n'évalue pas les conséquences de l'action selon son utilité personnelle.*

La distinction « utilité personnelle » doit être clairement établie car certaines normes sociales dépendent de l'utilité collective. Dans une situation de DP, ce postulat correspond à une norme plus générale interdisant le resquillage. En autant que la norme prescrive la collaboration, ce postulat sera toujours respecté car la collaboration dans le DP s'effectue toujours au détriment de l'utilité personnelle immédiate.

2) *L'ANS est motivé par une internalisation des sanctions sociales.*

Ce postulat permet de distinguer clairement entre normes sociales et normes personnelles. Seule la norme sociale implique la désapprobation (ou l'approbation) d'autrui ; la norme personnelle faisant plutôt appel aux convictions morales de l'individu. La norme kantienne constitue un excellent exemple de norme personnelle : un kantien transgressant sa propre norme n'éveillera pas de soupçons de la part de ses pairs. Le concept d'internalisation permet une autre distinction, cette fois-ci entre normes sociales et rationalisme. Nous avons discuté de cette dichotomie dans l'analyse de la discipline et de l'honneur chez Ullman-Margalit. En résumé, pour qu'une sanction soit internalisée il s'agit que a) les punisseurs respectent eux-même la norme sociale et que b) les émotions de l'individu joue un grand rôle dans le pouvoir persuasif de la norme.

Une dernière précision concerne une tendance marquante de la littérature anti-rationaliste au sujet du statut réel du choix rationnel. En effet, certains considèrent le processus rationnel de décision comme une norme sociale typique au monde occidental. Je veux éviter les débats concernant ce point de vue particulier. Il faut toutefois prendre conscience que si l'on accepte cette vision, alors la dichotomie rationalité - normes analysée jusqu'ici devient caduque. Pour des raisons de concision, nous nous en tiendrons à considérer la rationalité comme une entité distincte des normes.

Passons maintenant à la classification des diverses motivations extra-rationnelles étudiées dans ce chapitre. D'abord, certaines motivations respectent les conditions de sanctions sociales et de non-maximisation tout juste mentionnées ; elles se qualifient donc comme authentiques *normes sociales*. Ce sont les normes identitaire, TFT et d'honneur. La norme TFT a déjà fait l'objet d'une longue analyse. La norme identitaire est sociale en autant que l'individu ainsi motivé agit de crainte d'être pointé du doigt comme un profiteur alors qu'une majorité de personnes contribuent au bien public. La norme d'honneur demande une définition plus précise, la distinguant de phénomènes émotifs tels la fierté ou l'estime de soi. Selon Elster, la publicité des actes honorables - ou des atteintes à l'honneur - est essentielle à la norme. L'honneur est ici une question de statut social ; celui qui refuse de défendre son honneur se verra souvent ostracisé par ses pairs (NR : 884).

La norme kantienne est sans contredit une *norme personnelle*, son respect n'impliquan que l'individu concerné et personne d'autre. Pour cette même raison, la norme utilitariste s qualifie aussi comme norme personnelle. Pour ces deux normes, la question de sanction sociales ne se pose pas. Toutefois, celles-ci respectent le postulat de non-maximisation.

Les émotions jouissent d'une caractéristique supplémentaire par rapport aux norme personnelles : elles peuvent entrer dans la théorie (étendue) du choix rationnel par une simpl transformation du modèle individuel. Elster lui-même a élaboré cette révision en situant le émotions comme déterminants directs des désirs, tout comme l'information détermine le croyances (RESN : 35). Nous avons vu que la confiance pouvait être modelisée comme un décision risquée, l'élément de risque provenant d'un indice de fiabilité de ses adversaires dans l jeu. La norme de discipline correspond à une maximisation sous contrainte de sanction pouvant être reformulée en théorie des jeux. Les activistes considèrent une utilité alternative, l bénéfice de participation. L'altruisme correspond à une maximisation de l'utilité comprenant le utilités des autres dans le calcul. Il est aussi possible de considérer l'altruisme « débridé » un tel individu ne calcule pas d'utilités mais vient tout simplement en aide à ses pairs autan qu'il le peut. Un tel altruisme deviendra alors une émotion « authentique », inadaptable au choix rationnel.

III.2 - Solutions normatives au DP -

En guise de conclusion à ce chapitre, nous allons procéder à l'application des diverses motivations extra-rationnelles dans quelques variantes du problème d'action collective symbolisée par le DP. Nous exposerons d'abord l'exemple particulier d'Elster tiré de CS pour ensuite nous pencher sur des schémas de DP plus communs.

Elster se sert des types d'individus « normatifs » qu'il a lui-même élaboré (cf. section II.1) pour élaborer un scénario de production d'un bien public selon la courbe d'utilité de la fig. 21. Ce sont les kantiens, les activistes (d'élite et de masse), les utilitaristes et les gens motivés par la norme identitaire. L'auteur avoue lui-même que son scénario n'est qu'une vague

supposition, mais il nous aidera certainement à visualiser l'action collective d'un angle autre que celui de la théorie formelle des jeux. Il est à noter que ce graphique n'est pas une courbe de Schelling,
il représente seulement l'influence des différents types d'individus sur l'utilité collective.

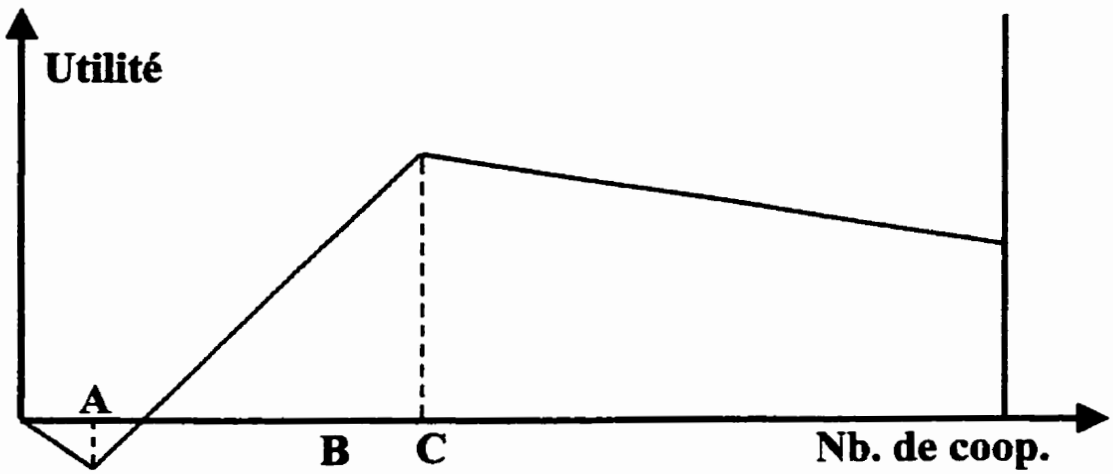


Fig. 21

Les premiers à démarrer l'action collective sont les kantiens, car la coopération universelle (extrême droite du graphique) est préférable à la défection universelle. Ceux-ci seront immédiatement accompagnés des activistes d'élite. Lorsque la population de ces deux groupes atteint A, les utilitaristes se joignent au mouvement. Au point B, les activistes de masse arrivent et les activistes d'élite commencent à faire défection. A ce niveau, les gens motivés par la norme identitaire se joignent graduellement. Lorsque tout ce monde atteint le point C, les utilitaristes se désistent. A ce moment, deux résultats deviennent possibles, dépendamment de l'importance des utilitaristes. S'ils ne sont pas très nombreux, la coopération augmentera vers l'universalité[21] sans trop de problème ; mais s'ils forment un groupe important, ils pourraient par leur défection faire changer d'idée les « identitaristes ». Dans cette éventualité, le mouvement peut très bien échouer complètement (CS : 204).

[21] L'universalité ne sera jamais atteint car il manquera toujours les participationnistes d'élite et les utilitaristes.

Bien que son scénario soit plutôt *ad hoc* et sommaire, Elster en tire deux généralités pouvant servir dans toutes les situations. La première est celle-ci : « *everyday Kantianism and the norm of fairness interact to produce much more cooperation than either could do by itself* » (CS : 205). En d'autres mots, ces deux groupes qui ne prennent jamais en considération les coûts et les bénéfices de la coopération peuvent agir comme catalyseurs pouvant attirer les gens aux motivations plus égoïstes. La seconde stipule que « *the strength of utilitarianism and that of fairness vary inversely with each other* » (CS : 206). Les utilitaristes ne prennent part au mouvement que lorsque c'est nécessaire et abandonnent lorsque leur tâche (activer le mouvement) est terminée. Les gens motivés par le devoir ne joignent le mouvement que si un nombre suffisant de participants s'y sont déjà engagés.

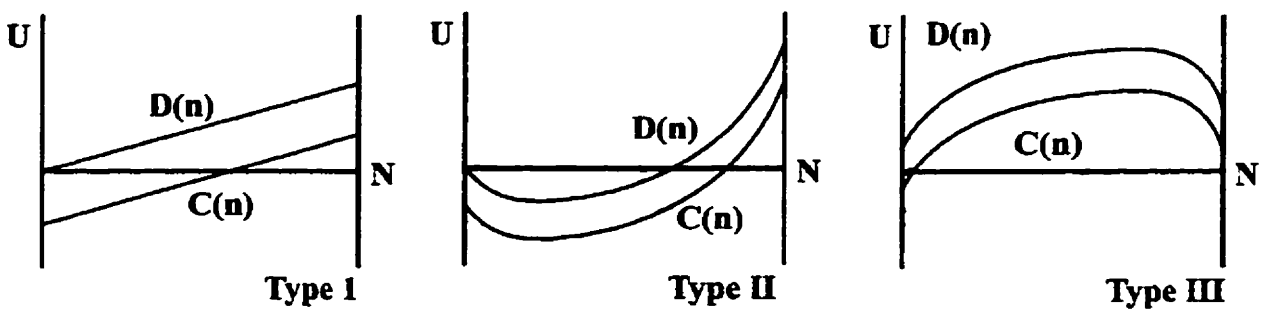


Fig. 22

Ces trois courbes de Schelling représentent le DP : la forme normale (I) et deux cas spéciaux où une coopération insuffisante peut s'avérer pire que la non-coopération universelle (II) et où la coopération universelle - bien que préférable à la non-coopération - ne constitue pas la meilleure solution (III). Nous avons appris dans le second chapitre que, sauf dans certains cas bien particuliers, la théorie classique des jeux prédit pour ces DP à n-personnes la défection universelle et l'échec de l'action collective. L'introduction de comportements normatifs va changer cet état de fait. Chacune des motivations étudiées jusqu'ici vont contribuer d'une façon ou d'une autre à expliquer les phénomènes auparavant quasi-insaisissables de la coopération universelle. L'influence des motivations extra-rationnelles va comme suit :

a) Certaines motivations peuvent résoudre par elles-mêmes le DP. Il s'agit des normes poussant l'individu à collaborer, peu importe les conséquences en termes d'utilité individuelle ou d'optimum social. La norme d'honneur répond à ce critère, ainsi que la norme kantienne. La norme TFT fait aussi partie de cette catégorie car elle enjoint ses adeptes à ne jamais faire défection en premier. D'autres motivations comme l'altruisme, la confiance et la discipline pourrait aussi figurer ici, mais elles dépendent toutes de la dimension du groupe : plus le groupe est important, moins ces motivations ont d'emprise sur l'individu car elles se révèlent inapplicables sur des gens ne se connaissant pas.

b) La norme identitaire dépend d'une coopération pré-établie. C'est d'ailleurs la seule condition d'opération de cette norme. Ce facteur peut s'avérer très utile dans des DP de type I et surtout II, où une participation totale est recherchée. Ses effets se font moins apprécier dans un type III, toutefois.

c) La norme utilitariste ne propose la coopération que si elle est utile, c'est à dire que le niveau de coopération se situe sur une pente ascendante. Donc, les utilitaristes collaborent toujours dans un type I, vont contribuer à sortir le groupe du pétrin dans un type II et va permettre au groupe de se stabiliser au niveau optimal dans un type III. Bref, ces individus qui placent les bénéfices immédiats du groupe devant leur propre utilité contribuent grandement au succès de l'entreprise dans de nombreux cas de DP.

d) Les activistes peuvent agir comme catalyseurs, dépendamment du type. Cette conclusion rejoint celle de Elster plus haut. Les activistes d'élite vont démarrer un mouvement alors que les activistes de masse vont joindre le mouvement plus tard. En fait, ceux-ci jouent le même rôle que les individus identitaires ; la différence réside dans le fait que la norme identitaire est sociale - donc plus intéressant pour nous - alors que l'activisme fait partie de la rationalité alternative.

e) La présence d'activistes pose un risque à l'entreprise collective. Les activistes sont les seuls à ne pas considérer les utilités de la matrice de jeu dans leur calculs. Ils vont se servir d'une autre matrice dont les utilités correspondent à la jouissance participative. De ce fait, ils peuvent très bien se situer dans un jeu autre que le DP, avec des conséquences potentiellement très graves pour l'équilibre coopératif. Prenons l'exemple d'un groupe composé de TFT et d'activistes. Tant que les activistes collaborent, la coopération universelle s'en suit. Au moment

où ceux-ci font défection dans leur propre jeu, ils entraîneront les TFT vers la défection et l'échec complet de l'entreprise. Si les participants au DP ne connaissent pas avec précision les motivations des activistes, leur présence aura l'effet d'une bombe à retardement créant bien entendu une grande incertitude au sein de la population.

CHAPITRE 5

CONCLUSION

Au début de ce texte nous avions soulevé deux questions devant servir de fondements à l'analyse de la rationalité et des normes en situation de Dilemme du Prisonnier. La première concernait la *nature* des solutions coopératives rationnelles. Nous avons découvert que les voies rationnelles vers le résultat Pareto-optimal sont particulièrement restreintes. La théorie des jeux n'admet pas de coopération dans le jeux à une seule instance. Dans les jeux répétés, les individus ne doivent pas connaître la fin du tournoi et doivent exhiber une préférence temporelle « raisonnable ». Plus il y a de joueurs, moins la coopération se révèle possible. Dans la quasi-totalité des autres situations, la théorie prédit la défection universelle. Ceci répondait à la seconde question, à savoir si la théorie des jeux est adéquate dans l'explication de phénomènes collectifs arborant les caractéristiques du DP. Jon Elster, comme plusieurs autres théoriciens, va répondre « non » et va tenter d'élaborer des alternatives viables. Il proposera une théorie des normes sociales permettant de rendre compte de comportements individuels insaisissables par le choix rationnel et pourtant déterminants dans le choix de stratégie dans le DP. Bien que sa théorie soit suffisamment claire pour être opérationnelle et réussir à certains moments où la théorie des jeux avait échoué, ses exemples de normes se révélèrent moins précis et il nous a fallu ajuster la définition pour y accommoder les normes personnelles et les dispositions psychologiques. Globalement nous pouvons affirmer que la tentative de Elster fut un succès, à la lumière de ses intentions même de vouloir provoquer un changement de mentalités à l'intérieur de la théorie du choix rationnel sans toutefois l'abandonner.

Bien que Elster ait prêché pour une théorie des normes *sociales*, la conclusion du troisième chapitre démontre clairement que les normes personnelles et autres dispositions psychologiques jouent aussi un rôle important dans l'explication de l'action collective. Nous allons donc proposer une théorie des normes (sociales et autres) comme complément idéal à la théorie du choix rationnel étendue, pourvu que cette dernière conserve une priorité

méthodologique : les normes ne seront utilisées que lorsque l'analyse rationaliste échoue. Deux bonnes raisons pour cela ; d'abord il faut rappeler que la théorie des jeux demeure *adéquate* lorsque l'on observe une faillite de l'action collective dans la réalité, phénomène quand même assez fréquent (pensons notamment au désarmement nucléaire). Ensuite, la théorie permet d'expliquer l'action collective sous certaines conditions contraignantes certes, mais cela se qualifie toujours comme une explication *valable*. L'avantage de la théorie du choix rationnel réside dans la simplicité de ses prémisses et dans ses pouvoirs explicatif et prédictif tout de même assez important, du moins comparativement aux autres théories en sciences sociales.

Pour que l'union des deux théories porte fruit, certaines précisions s'imposent. La théorie des normes doit offrir des définitions claires et opérationnelles des normes que l'on imposera aux sujets. Les normes kantienne, TFT et utilitaristes entre autres répondent à ces critères : leur application dans le DP s'est effectué sans heurts. Il faut aussi bien comprendre que la nature des normes dépend énormément du contexte. En milieu de travail par exemple, les normes régissant le comportement des travailleurs ne sont pas les mêmes que celles des patrons, et ces mêmes normes varient d'une entreprise à l'autre. Une analyse sociologique de la sphère socioculturelle à l'étude constitue une étape préliminaire essentielle à l'élaboration des normes.

Avant d'être vraiment utile, la théorie des normes demande à être raffinée sur certains points. Il faut bien faire la distinction entre motivations extra-rationnelles et extensions de la rationalité comme l'altruisme ou la confiance, pleinement compatibles avec la théorie de l'utilité. Elster a aussi soulevé une indétermination concernant le lien entre norme et action individuelle : bien que l'influence soit causale, l'individu conserve un pouvoir intentionnel sur la norme en ce qu'il peut décider de ne pas la respecter. Cette particularité des normes rend la tâche d'évaluer l'influence précise de celles-ci extrêmement ardue à déterminer. Des développements théoriques sont à espérer dans ce domaine. Enfin, les normes font appel au concept de socialisation, un terrain miné pour la théorie du choix rationnel. Pour qu'une définition des normes soit adéquate, le choix rationnel devra compenser avec un mécanisme de socialisation des comportements individuels. Entre le structuralisme à la Bourdieu ou Foucault

et le rationalisme plaidant l'indépendance complète de l'agent, il y a une place pour une théori
où le comportement individuel représente un amalgame d'intentionnalité et de causalité externe
Tout ceci pour le plus grand bien de la recherche en sciences sociales.

BIBLIOGRAPHIE

Ouvrages de Jon Elster fréquemment cités et leurs abbréviations:

-LFEC Leibniz et la formation de l'esprit capitaliste,
Paris, Aubier Montaigne, 1975.

-CPPT "Some Conceptual Problems in Political Theory",
in BARRY, Brian (ed), Power and Political Theory
London, Wiley, 1976.

-LS Logic and Society: Contradictions and Possible Worlds,
New York, Wiley, 1978.

-US Ulysses and the Sirens: Studies in Rationality and Irrationality,
New York, Cambridge Univ. Press, 1979.

-RAT "Rationality",
in FLOISTAD, Gunnar (ed), Contemporary Philosophy , vol. 2.
Boston, Nijhoff, 1982.

-ETC Explaining Technical Change: A Case Study in the Philosophy of Science,
New York, Cambridge Univ. Press, 1983.

-SG Sour Grapes: Studies in the Subversion of Rationality,
New York, Cambridge Univ. Press, 1983.

-RMCA "Rationality, Morality, and Collective Action",
Ethics, vol. 96, 1985, pp. 136-155.

-SBW "Sadder But Wiser? Rationality and the Emotions",
Social Science Information, vol. 24, 1985, pp. 375-406.

-MF "The Market and the Forum",
in ELSTER, Jon, HYLLAND, Aanund (eds), Foundations of Social Choice Theory
New York, Cambridge Univ. Press, 1986.

-IRC "Introduction",
in ELSTER, Jon (ed), Rational Choice,
Oxford, Blackwell, 1986.

-MSM Making Sense of Marx,
New York, Cambridge Univ. Press, 1985.

-NB Nuts and Bolts for the Social Sciences,
New York, Cambridge Univ. Press, 1989.

-CS The Cement of Society: A Study of Social Order,
New York, Cambridge Univ. Press, 1989.

-SJ Solomonic Judgements: Studies in the Limitation of Rationality,
New York, Cambridge Univ. Press, 1989.

-NR "Norms of Revenge",
Ethics, vol. 100, 1990, pp. 862-885.

-PP Political Psychology,
New York, Cambridge Univ. Press, 1993.

-RESN "Rationality, Emotions and Social Norms",
Synthese, vol. 98, 1994, pp. 21-49.

Bibliographie générale:

- ALMOND, Gabriel A. (1991), "Rational Choice Theory and the Social Sciences",
in MONROE, Kristen R. (ed), The Economic Approach to Politics,
New York, Harper Collins.

- AXELROD, Robert (1990), "The Emergence of Cooperation Among Egoists",
in MOSER, Paul K. (ed), Rationality in Action,
New York, Cambridge Univ. Press.

- BARRY, Brian (1985), "Comment on Elster's *Rationality, Morality and Collective Action*",
Ethics, vol. 96, pp. 156-158.

- BECKER, Gary (1976), The Economic Approach to Human Behavior,
Chicago, Univ. of Chicago Press.

- *Ibid.* (1986), "The Economic Approach to Human Behavior",
in ELSTER, Jon (ed), Rational Choice,
New York, New York Univ. Press.

- BERGSTEIN, Gordon S. (1985), "On the Role of Social Norms in a Market Economy",
Public Choice, vol. 45, pp. 113-137.

- BICCHIERI, Cristina (1993), Rationality and Cooperation,
New York, Cambridge Univ. Press.

- BINMORE, Ken (1994), Game Theory and the Social Contract,
vol. 1, Playing Fair,
Cambridge (MA), MIT Press.

- BOURDIEU, Pierre (1994), Raisons pratiques sur la théorie de l'action,
Paris, Seuil.

- BRENNAN, Timothy J. (1991), "The Trouble with Norms",
in KOFORD, Kenneth J; MILLER, Jeffrey B. (eds), Social Norms and Economic Institutions,
Ann Arbor, Univ. of Michigan Press.

- BRUNNER, Karl (1987), "The Perception of Man and the Conception of Society: Two Approaches to Understanding Society",
Economic Inquiry, vol. 25, pp. 367-389.

- BUNGE, Mario (1990), Game Theory is not a Useful Tool for the Political Scientist,
Montréal, UQAM, Groupe de Recherche en Epistémologie Comparée, cahier no. 9015.

- CUDD, Ann E. (1993), "Game Theory and the History of Ideas About Rationality", Economics and Philosophy, vol. 9, pp. 101-133.

- DANIELSON, Peter (1991), "Closing the Compliance Dilemma: How it's Rational to be Moral in a Lamarckian World", in VALLENTYNE, Peter (ed), Contractarianism and Rational Choice", New York, Cambridge Univ. Press.

- DAVIDSON, Donald (1980), Actions and Events, New York, Oxford Univ. Press.

- *Ibid.* (1986), "Judging Interpersonal Interests", in ELSTER, John; HYLLAND, Aanund (eds), Foundations of Social Choice Theory, New York, Cambridge Univ. Press.

- DAVIS, Lawrence H. (1977), "Prisoners, Paradox, and Rationality" American Philosophical Quarterly, vol. 14, pp. 319-327.

- DOWNS, Anthony (1957), An Economic Theory of Democracy, New York, Harper.

- FOLLESDAL, Dagfinn (1979), "Hermeneutics and the Hypothetico-Deductive Method", Dialectica, vol. 33, pp. 319-336.

- *Ibid.* (1982), "The Status of Rationality Assumptions in Interpretation and in the Explanation of Action", Dialectica, vol. 36, pp. 301-316.

- FRANSSEN, Maarten (1994), "Constrained Maximization Reconsidered - An Elaboration and Critique of Gauthier's Modelling of Rational Cooperation in a Single Prisoner's Dilemma", Synthese, vol. 101, pp. 249-272.

- GAUTHIER, David (1987), Morals By Agreement, New York, Oxford Univ. Press.

- *Ibid.* (1990), "Maximization Constrained: The Rationality of Cooperation", in MOSER, Paul K. (ed), Rationality in Action, New York, Cambridge Univ. Press.

- GIBBARD, Allan (1985), "Moral Judgment and the Acceptance of Norms", Ethics, vol. 96, pp. 5-21.

- *Ibid.* (1986), "Interpersonal Comparisons: Prefence, Good, and the Intrinsic Reward of a Life", in ELSTER, John; HYLLAND, Aanund (eds), Foundations of Social Choice Theory, New York, Cambridge Univ. Press.

- GREEN, Donald P.; SHAPIRO, Ian (1995), "Choix rationnels et politique: pourquoi en savons-nous toujours aussi peu?", Revue française de science politique, vol. 45, pp. 96-130.

- GRIFFIN, James (1985), "Some Problems of Fairness", Ethics, vol. 96, pp. 100-118.

- GRIFFITH, William B.; GOLDFARB, Robert S. (1991), "Amending the Economist's 'Rational Egoist' Model to Include Moral Values and Norms" (Parts 1 and 2), in KOFORD, Kenneth J; MILLER, Jeffrey B. (eds), Social Norms and Economic Institutions, Ann Arbor, Univ. of Michigan Press.

- HARDIN, Russell (1971), "Collective Action as an Agreeable n-Prisoners' Dilemma", Behavioral Science, vol. 16, pp. 472-481.

- *Ibid.* (1982), Collective Action, Baltimore (MD), John Hopkins Univ. Press.

- HARSANYI, John C. (1987), "Review of D. Gauthier, *Morals by Agreement*", Economics and Philosophy, vol. 3, pp. 339-351.

- *Ibid.* (1990), "Advances in Understanding Rational Behavior", in MOSER, Paul K. (ed), Rationality in Action, New York, Cambridge Univ. Press.

- HINDESS, Barry (1988), Choice, Rationality and Social Theory, Londres, Unwin Hyman.

- HIRSCHMAN, Albert O. (1970), Exit, Voice and Loyalty, Cambridge (MA), Harvard Univ. Press.

- *Ibid.* (1985), "Against Parsimony", Economics and Philosophy, vol. 1, pp. 7-21.

- HOVI, Jon (1986), "Binary Games as Models of Public Goods Provision", Scandinavian Political Studies, vol. 9, pp. 337-360.

- JANKOWSKI, Richard (1990), "Punishment in Iterated Chicken and Prisoner's Dilemma Games", Rationality and Society, vol. 2, pp. 449-470.

- JOHNSON, James D. (1991), "Rational Choice as a Reconstructive Theory", in MONROE, Kristen R. (ed), The Economic Approach to Politics, New York, Harper Collins.

- KEE, Herbert W.; KNOX, Robert E. (1970), "Conceptual and Methodological Considerations in the Study of Trust and Suspicion", Journal of Conflict Resolution, vol. 14, pp. 357-366.

- KOLM, Serge-Christophe (1986), Philosophie de l'économie, Paris, Seuil.

- KYMLICKA, Will (1990), Contemporary Political Philosophy, New York, Oxford Univ. Press.

- LANDRY, Réjean (1993), La théorie institutionnelle du Public Choice, Ste-Foy, Groupe de recherche sur les interventions gouvernementales, document polycopié.

- LECA, Jean (1985), "La théorie politique", in GRAWITZ, Madeleine; LECA, Jean, Traité de science politique, vol. 1, Paris, P.U.F.

- LINDBECK, Assar (1996), "Welfare State Disincentives with Endogenous Habits and Norms", in ANDERSEN, T.M.; MOENE, K.O.; SANDMO, A. (eds), The Future of the Welfare State, Oxford, Blackwell.

- LITTLE, Daniel (1991), Varieties of Social Explanation: An Introduction to the Philosophy of Social Science, Boulder (CO), Westview Press.

- LUCE, R. Duncan; RAIFFA, Howard (1957), Games and Decisions, New York, Wiley.

- MARGOLIS, Howard (1990a), "Equilibrium Norms", Ethics, vol. 100, pp. 821-837.

- *Ibid* (1990b), "Dual Utilities and Rational Choice", in MANSBRIDGE, Jane J. (ed), Beyond Self-Interest, Chicago, Univ. of Chicago Press.

- MILL, John Stuart (1979), "Utilitarianism", in WARNOCK, Mary (ed), John Stuart Mill: Utilitarianism, Glasgow, William Collins Sons.

- MONROE, Kristen R. (1991), "The Theory of Rational Action: Its Origins and Usefulness for Political Science", in MONROE, Kristen R. (ed), The Economic Approach to Politics, New York, Harper Collins.

- MUELLER, Dennis C. (1986), "Rational Egoism Versus Adaptive Egoism as Fundamental Postulate for a Descriptive Theory of Human Behavior", Public Choice, vol. 51, pp. 3-23.

- *Ibid.* (1991), "Comments" in KOFORD, Kenneth J; MILLER, Jeffrey B. (eds), Social Norms and Economic Institutions, Ann Arbor, Univ. of Michigan Press.

- NELSON, Alan (1988), "Economic Rationality and Morality", Philosophy and Public Affairs, vol. 17, pp. 149-166.

- NIDA-RUMELIN, Julian (1994), "Rational Choice: Extensions and Revisions", Ratio, vol. 7, pp. 122-144.

- OLSON, Mancur (1971), The Logic of Collective Action, Cambridge (MA), Harvard Univ. Press.

- ORDESHOOK, Peter C. (1986), Game Theory and Political Theory, New York, Cambridge Univ. Press.

- PETTIT, Philip (1986), "Preserving the Prisoner's Dilemma", Synthese, vol. 86, pp. 181-184.

- *Ibid.* (1990), "*Virtus Normativa*: Rational Choice Perspectives", Ethics, vol. 100, pp. 725-755.

- RAINBOLT, George W. (1989), "Gauthier on Cooperating in Prisoners' Dilemmas", Analysis, vol. 49, pp. 216-220.

- RAPOPORT, Anatol (1966), Two-Person Game Theory, Ann Arbor (MI), Univ. of Michigan Press.

- RIKER, William H. (1980), "Political Trust as Rational Choice", in LEWIN, Leif; VEDUNG, Evert (eds), Politics as Rational Action, Dordrecht (Holland), D. Reidel.

- RIKER, William H.; ORDESHOOK, Peter C. (1973), An Introduction to Positive Political Philosophy, Englewood Cliffs (NJ), Prentice-Hall.

- RORTY, Amelie Oksenberg (1980), "'Akrasia' and Conflict", Inquiry, vol. 23, pp. 193-212.

- ROSENBERG, Alexander (1988), Philosophy of Social Science, Boulder (CO), Westview Press.

- *Ibid.* (1992), Economics - Mathematical Politics or Science of Diminishing Returns?, Chicago, Univ. of Chicago Press.

- SANDVEN, Tore (1995), "Intentional Action and Pure Causality", Philosophy of the Social Sciences, vol. 25, pp. 286-317.

- SCHELLING, Thomas C. (1978), Micromotives and Macrobehavior, New York, W.W. Norton.

- SEN, Amartya (1982), "Rational Fools", in Choice, Welfare and Measurement, Cambridge (Mass), MIT Press.

- SHEPSLE, Kenneth A. (1989), "Studying Institutions: Some Lessons from the Rational Choice Approach", Journal of Theoritical Politics, vol. 1, pp.131-147.

- SIMON, Herbert A. *et al.* (1992), "Decision Making and Problem Solving", in ZEY, Mary (ed), Decision Making, Newbury Park (CA), Sage Publications.

- SKYRMS, Brian (1990), The Dynamics of Rational Deliberation, Cambridge (MA), Harvard Univ. Press.

- SMITH, Holly (1991), "Deriving Morality from Rationality", in VALLENTYNE, Peter (ed), Contractarianism and Rational Choice", New York, Cambridge Univ. Press.

- SOBEL, Jordan Howard (1991), "Constrained Maximization", Canadian Journal of Philosophy, vol. 21, pp. 25-52.

- SOBER, Elliott (1992), "Stable Cooperation in Iterated Prisoners' Dilemmas", Economics and Philosophy, vol. 8, pp. 127-139.

- SOWDEN, Lanning (1983), "That There Is a Dilemma in the Prisoners' Dilemma" Synthese, vol. 55, pp. 347-352.

- TAYLOR, Micheal (1987), The Possibility of Cooperation, New York, Cambridge Univ. Press.

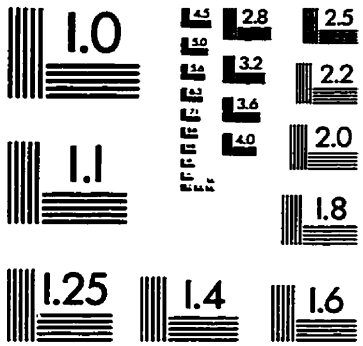
- ULLMANN-MARGALIT, Edna (1977), The Emergence of Norms, Oxford, Clarendon Press.

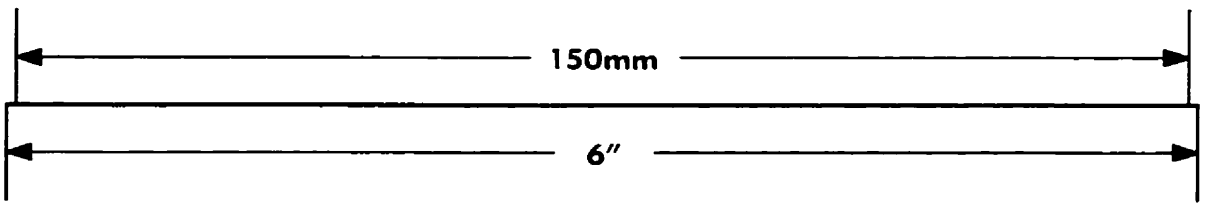
- VON NEUMANN, John; MORGENSTERN, Oskar (1953), Theory of Games and Economic Behavior, Princeton (NJ), Princeton Univ. Press.

- WILLIAMS, Bernard (1988), "Formal Structures and Social Reality", in GAMBETTA, Diego (ed), Trust: Making and Breaking Cooperative Behavior, New York, Basil Blackwell.

- WILLIAMS, E.E.; FINDLAY, M.C. (1981), "A Reconsideration of the Rationality Postulate", American Journal of Economics and Sociology, vol. 40, pp. 17-36.

TEST TARGET (QA-3)





APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved