Approaching preference change by partial deliberation

Niels Boissonnet

UNIVERSITÉ PARIS I - PANTHÉON-SORBONNE

SCIENCES ÉCONOMIQUES - SCIENCES HUMAINES - SCIENCES JURIDIQUES ET POLITIQUES

UNE APPROCHE DES CHANGEMENTS DE PRÉFÉRENCES PAR LA DÉLIBÉRATION PARTIELLE

Thèse soutenue pour le doctorat en sciences économiques
(arrêté du 30 mars 1992)
présentée et soutenue publiquement le 5 avril 2019 par

Niels BOISSONNET

Directeurs de Recherche :
Mme. Camille CHASERANT, Maitre de Conférence, Université Paris I Panthéon-Sorbonne
M. Jérôme LALLEMENT, Professeur Émérite, Université Paris Descartes

Membres du Jury :
Mme. Antoinette BAUJARD, (Rapporteure) Professeure, Université Jean Monnet Saint-Etienne
M. Franz DIETRICH, Directeur de Recherche CNRS, Université Paris 1 Panthéon-Sorbonne, PSE
M. Brian HILL, (Rapporteur) Professeur, École des Hautes Études Commerciales Paris
M. Jean-Marc TALLON, Directeur de Recherche CNRS, PSE, Université Paris 1 Panthéon-Sorbonne

L'université Paris 1 Panthéon-Sorbonne n'entend donner aucun approbation ni improbition aux opinions émises dans cette thèse ; elle doivent être considérées comme propres á leur auteur.

REMERCIEMENTS

Mes premières pensées vont à mes directeur.e.s de thèse : Jérôme Lallement et Camille Chaserant. Ils ont incontestablement su m'apporter attention, rigueur intellectuelle et soutien moral lorsque cela était nécessaire. Ils ont également forcé mon esprit à s'accrocher à celui des autres lorsque mes productions étaient obscures. Je ne peux me représenter leurs souffrances accumulées face à mon incorrigible créativité orthographique. Encore moins celle de leur voisin lorsque, dans l'avion, ils devaient me relire. Pour ces raisons, et de nombreuses autres, je leur dois une très grande reconnaissance. Ecrire cette thèse abrité par leurs regards bienveillants a été très agréable.

Je me dois ensuite de faire état de ma gratitude envers les membres du jury de cette soutenance : Antoinette Baujard, Franz Dietrich, Brian Hill et Jean-Marc Tallon. C'est un honneur de leur avoir fait lire ce qui m'a tenu en éveil tout au long de ces quatre années. Ils m'ont apporté, chacun de sa perspective, un regard affuté et qui anime encore mon esprit de multiples questionnements. Cette thèse constitue une réponse incomplète à l'ensemble de leurs remarques. A n'en pas douter, leurs réflexions rayonneront encore longtemps sur mon travail. Plus spécifiquement, et dans le respect de la chronologie, Jean-Marc Tallon a fait preuve d'une très grande bienveillance à mon égard. Ses conseils tant sur le fond, que sur la manière de positionner ce travail dans la communauté des sciences sociales, sont précieux. Sa porte m'a toujours été ouverte lorsque je le sollicitais. Il en va de même pour Franz Dietrich dont les connaissances et les contributions au domaine dans lequel se situe mon travail ont dès le départ inspirés mes réflexions. Ma rencontre avec Brian Hill a joué un rôle également déterminant sur les derniers mois de cette thèse. Ses remarques et conseils de lecture m'ont permis d'interroger mon travail sous des perspectives nouvelles. Enfin, je dois une grande reconnaissance à Antoinette Baujard qui, en dépit de ses contraintes, a accepté avec bienveillance de participer à la lecture de ce travail. J'attends avec impatience les échanges que nous pourrons avoir.

Table des matières

« Soulever la question : qu'est-ce que la liberté, semble une entreprise désespérée. Tout se passe comme si des contradictions et des antinomies sans âge attendaient ici l'esprit pour le jeter dans des dilemmes logiquement insolubles, de sorte que, selon le parti adopté, il devient impossible de concevoir la liberté ou son contraire, que de former la notion d'un cercle carré. Sous sa forme la plus simple, la difficulté peut être résumée comme la contradiction entre notre conscience qui nous dit que nous sommes libres et par conséquent responsables, et notre expérience quotidienne dans le monde extérieur où nous nous orientons d'après le principe de causalité. Dans toutes les choses pratiques et spécialement dans les choses politiques, nous tenons la liberté humaine pour une vérité qui va de soi, et c'est sur cet axiome que les lois reposent dans les communautés humaines, que les décisions sont prises, que les jugements sont rendus. Dans tous les champs de travail scientifique et théorique, au contraire, nous procédons d'après la non moins évidente vérité du *nihil ex nihilo* [rien ne vient de rien], du *nihil sine causa* [rien n'est sans cause], c'est-à-dire en supposant que «même nos propres vies sont, en dernière analyse, soumises à des causes » et que, s'il doit y avoir en nous-mêmes un moi ultimement libre, celui-ci, en tout cas, n'apparaît jamais sans équivoque dans le monde phénoménal, et ne peut donc jamais devenir l'objet d'une assertion théorique. D'où le fait que la liberté tourne au mirage au moment où la psychologie pénètre dans son domaine réputé le plus intime, car « le rôle que la force joue dans la nature comme cause du mouvement, a pour contrepartie dans la sphère mentale le motif comme cause de la conduite ». Il est vrai que l'épreuve de causalité – la prévisibilité de l'effet, si toutes les causes sont connues – ne peut être appliquée au domaine des affaires humaines ; mais cette prévisibilité pratique n'est pas une preuve de la liberté, elle signifie simplement que nous ne sommes jamais en mesure de connaître toutes les causes qui entrent en jeu ; cela, simplement à cause du grand nombre de facteurs, mais aussi parce que les motifs humains, la différence des forces naturelles, demeurent cachés à tous les regards, à l'observation des autres hommes comme à l'introspection. »

Hannah Arendt - *La crise de la culture*.- p. 186-187

INTRODUCTION GENERALE

Lors de ses premiers échanges avec Odette, Swann ne nourrit aucune affection à son égard. Son profil est « trop accusé, la peau trop fragile, les pommettes trop saillantes, les traits trop tirés », et bien qu'il le regrette parfois, sa beauté n'est pas « du genre de celles qu'il aurait spontanément *préférées*[1]. » A cette époque de la vie, la beauté qu'il accorde aux femmes est irréfléchie et ne dépend pas de sa passion pour les tableaux de grands maitres.

> Swann, lui, ne cherchait pas à trouver jolies les femmes avec qui il passait son temps, mais à passer son temps avec les femmes qu'il avait d'abord trouvées jolies. Et c'était souvent des femmes de beauté assez vulgaire, car les qualités physiques qu'il recherchait *sans s'en rendre compte* étaient en complète opposition avec celles qui lui rendaient admirables les femmes sculptées ou peintes par les maîtres qu'il préférait.
>
> Marcel Proust - *Du côté de chez Swann*, p.195[2]

Pourtant, dans la suite du roman, ses préférences changent. Au point que ses actions s'en trouvent tout entièrement rivées sur les agissements d'Odette, au point de la maintenir enfermée dans une prison imaginaire, souffrant de ses sorties, aveugle à ses trahisons.

Or, ce renversement de préférences n'est pas le fruit du hasard : il découle d'un raisonnement. C'est après avoir pris conscience qu'Odette ressemble à Zephora, figure d'une fresque de Botticelli ornant les murs de la chapelle Sixtine, qu'il s'éprend d'une inconditionnelle passion à son égard. Odette devient subitement « son genre » car sa façon d'attribuer la beauté à une femme change, et la ressemblance avec les figures peintes de ses grands maîtres constitue désormais un critère pour la reconnaitre.

Ces pages célèbres d'*A la recherche du temps perdu* montrent qu'une prise de conscience est susceptible de changer les préférences d'un individu. Aussi nous suggèrent-elles au moins trois choses.

1. D'abord, qu'il existerait un mécanisme mental par lequel une prise de conscience serait susceptible de causer un changement de préférences.
2. Ensuite, que ce mécanisme repose sur des connexions entre différentes préférences éprouvées par l'agent : la préférence qui veut que les figures peintes des grands

1. Je souligne.
2. Je souligne.

maîtres soient valorisées et la préférence qui veut qu'Odette, qui leur ressemble, le soit également.

3. Enfin, qu'un changement de préférences peut être non arbitraire et issu d'une forme de rationalité dont il convient de préciser la nature.

Il ne faut pas voir dans ces suggestions une simple fantaisie littéraire. Nombreux sont les phénomènes sociaux qui répondent à cette logique. Pensons par exemple à la récente « prise de conscience » célébrée par les médias après l'affaire Weinstein, à « l'éveil des consciences » appelée de leur voeux par les écologistes ou encore au rôle fondamental que joue, pour certains penseurs marxistes, le concept de « fausse conscience » ou de « conscience de classe » dans le maintien d'un ordre supposé établi. Un tel mode de transformation des préférences remplit, pour le décideur politique, une fonction normative considérable. Changer les préférences des individus, en s'appuyant sur leur capacité à raisonner, permet de satisfaire un *desiderata* essentiel de toute société démocratique [3]. Cela peut être considéré comme préférable à la manipulation subconsciente, qui ne reconnait pas le citoyen comme un être autonome, capable d'exercer son jugement propre.

Cette thèse se propose de prendre ces trois suggestions au sérieux et d'en donner un traitement formel. Ce faisant, elle apporte trois contributions :

1. Elle conceptualise le mécanisme articulant la formation des préférences au concept de prise de conscience.
2. Elle établit les conséquences analytiques de ce mécanisme pour la théorie du choix rationnel.
3. Elle en propose une application à un contexte au sein duquel un agent tente de manipuler les préférences d'un autre agent.

Il convient dans un premier temps de justifier l'étude de ce mécanisme et de situer son apport au regard de la littérature existante. Tel est l'objectif poursuivi dans cette introduction.

— La première tâche consiste à justifier l'intégration des changements de préférences dans la théorie économique. Car la tentation de modéliser les changements de préférences n'a pas toujours été au coeur de la discipline ; et de nombreuses raisons ont été invoquées pour en légitimer la négligence. La première section de cette introduction analyse quatre de ces raisons. Elle débouche sur l'idée que l'incorporation de changements de préférences dans les modèles d'économie est légitime, à condition de se fonder sur une théorie qui explicite le type de raisonnements à l'oeuvre dans les changements de préférences.

3. A partir de maitenant, sauf quelques exceptions que je justifierai, le terme de « préférences »sera employé au pluriel en français et au singulier en anglais.

— La seconde tâche consiste à articuler les changements de préférences à la théorie du choix rationnel, et le cortège d'outils conceptuels qu'elle met à disposition de l'économiste. Son objectif est donc de clarifier la nature de la rationalité à l'oeuvre dans les changements de préférence ; et d'établir la relation que cette forme de rationalité entretient avec les conceptions usuelles de la rationalité en théorie du choix (notamment la rationalité des préférences et la rationalité des croyances). La seconde section débouche sur la double affirmation que la *rationalisation* des changements de préférences 1) repose sur un concept axiologique de la rationalité et 2) nous contraint à l'adoption d'une position mentaliste.

— La troisième tâche consiste à justifier l'apport de ce mécanisme, et de le comparer avec des mécanismes employés par d'autres modèles avec changements de préférences. Dans ces modèles, soit le rôle de la rationalité n'est explicité que de façon confuse, soit la rationalité n'est pas comprise comme le moteur de ce changement. Le mécanisme étudié dans cette thèse a moins vocation à se substituer à ces modèles qu'à pallier ce manque. Il cherche à expliciter ce qui est de l'ordre de la rationalité dans un changement de préférences.

L'introduction de cette thèse s'achève par un résumé détaillé des trois chapitres qui la compose.

1 Pourquoi négligerait-on la formation des préférences ?

Eléments théoriques primitifs, susceptibles de fonder l'analyse économique, les préférences ont été traitées comme des données immuables, qui se révèlent en dernière instance dans le choix. En usant de cette stratégie, la microéconomie d'après-guerre a pu fournir d'importantes théories, qui permettent de mieux comprendre les mécanismes de marché. Sur la base de ces succès, une importante tradition d'économistes considère qu'il n'est pas nécessaire de s'intéresser à la formation des préférences. L'observateur doit prendre les préférences comme des causes du comportement sans s'intéresser à leur origine.

A ma connaissance, quatre raisons ont été mobilisées pour justifier une telle négligence. Dans cette section, je discute chacune de ces raisons. Au cours de cette discussion, je montre que, loin de proscrire les changements de préférences, elles nous invitent au contraire à en fournir une théorie structurée, qui rend compte des mécanismes mentaux à l'œuvre dans la formation des préférences. Cette thèse prend donc ces raisons au sérieux, elle cherche d'ailleurs à leur répondre.

1.1 La raison ontologique de cette négligence

La première raison de négliger la formation des préférences est ontologique. Elle affirme que le concept de préférences renvoie à une réalité immuable. A l'appui de cette interprétation, plusieurs enquêtes microéconométriques suggèrent que les comportements des

consommateurs respectent l'axiome faible des préférences révélées sur plusieurs périodes [Chalfant and Alston, 1988, Landsburg, 1981]. Cet axiome suppose qu'ils ne réagissent qu'à l'évolution des prix sans que leurs préférences ne se modifient au gré des années [4]. Toutefois, ce postulat ontologique est difficilement soutenable.

D'abord, les données de ces enquêtes portent sur un marché spécifique, à une époque spécifique : la consommation de viande entre 1900 à 1955 [5]. Ensuite, cette hypothèse va à l'encontre d'innombrables études empiriques faisant état d'une variation des préférences de consommation (voir [Kapteyn and Wansbeek, 1982]). Le travail d'Inglehart [1971] met en évidence, par exemple, l'existence d'une transition, dans les années 1960, d'une société matérialiste, où les valeurs des individus étaient centrées sur la possession matérielle, à une ère postmatérialiste où les valeurs d'expression personnelle auraient pris une place de plus en plus importante dans le spectre des valeurs qui orientent la consommation des agents [6].

Toutefois, refuser cette raison ontologique n'est pas suffisant pour confirmer l'intérêt potentiel de la formation des préférences. En définitive, ce n'est pas parce que les économistes font usage de préférences fixes qu'ils souscrivent à l'interprétation ontologique. « En réalité les articles défendant explicitement l'invariance des préférences sont plutôt rares [...]. La plupart d'entre eux considèrent la fixité des préférences comme une approximation ou une hypothèse satisfaisante » [Van de Stadt, Kapteyn and Van de Geer, 1985, p. 179]. Est-ce à dire que la fixité des préférences doit être vue comme une hypothèse, non pas simplement temporaire, mais constitutive de la discipline ?

1.2 La raison disciplinaire de cette négligence

Ces dernières remarques nous mènent à une seconde raison de négliger les changements de préférences. Selon cette raison, l'étude des changements de préférences ne relèverait tout simplement pas du domaine d'investigation de l'économiste. Je qualifie cette justification de « disciplinaire », car elle insiste sur l'idée que comprendre les changements de préférences n'est pas la *mission* que cherche à remplir la discipline économique. Que les préférences changent effectivement n'a d'intérêt que pour l'anthropologue, le psychologue ou le sociologue.

Cette idée prend sa source dans les travaux d'économistes comme Robbins [1932] ou de sociologues comme Parsons [1934]. Pour ces partisants du principe de division du travail en sciences sociales, la science économique se limite à l'analyse des moyens adéquats pour

4. L'axiome faible des préférences révélées dit que lorsque l'agent choisit un panier de consommation x lorsque y est disponible à son budget, il ne peut choisir y que si x n'est pas disponible à son budget.

5. Il est d'ailleurs douteux que ces résultats se retrouvent avec des données plus récentes. A l'heure où se développent les mouvements végétariens, on peut en effet supposer que l'évolution de la consommation de viande dans les pays occidentaux ne résulte pas exclusivement des prix, mais également de la prise en compte de considérations environnementales ou liées à l'éthique animale.

6. Une telle évolution se lit dans les stratégies mercatiques des grandes firmes de plus en plus enclines à se prétendre responsables sur le plan éthique, sur le plan écologique ou en matière de lutte contre les discriminations.

parvenir à des fins déterminées : elle est la science du choix, là où c'est à la sociologie qu'il revient d'étudier la manière dont les individus se sont mis à adopter les fins vers lesquelles ils conduisent leurs actions.

A charge pour les autres chercheurs en sciences sociales de déterminer si les préférences changent et la manière dont elles changent.

> La théorie économique se donne des préférences fixes. Cela illustre parfaitement le principe de division du travail. L'économiste n'a pas grand-chose à dire de la formation des préférences, problème qui est du ressort du psychologue. Sa tâche consiste bien davantage à identifier les conséquences de n'importe quel ensemble de préférences. Cette abstraction, au même titre que de nombreuses autres, tire légitimité et justification tant de sa clarté que du pouvoir de prédiction qu'elle porte avec elle [7].
>
> [Friedman, 1962, p. 13]

Toutefois, cette justification disciplinaire résiste difficilement à l'examen de l'évolution de la discipline :

— Paradoxalement, en faisant de la science économique une discipline sans autre objet que le comportement rationnel, Robbins ouvre la voie à l'extension de son domaine d'étude. Becker lui-même s'est d'ailleurs largement illustré en la matière, proposant que la discipline ouvre son spectre d'analyse à des « ensembles de valeurs et de préférences bien plus riches » que celles reposant sur le comportement égoïste de l'agent économique [Becker, 1996, p. 139]. De l'étude des mariages, à celle de la criminalité en passant par les phénomènes de discrimination, l'économie doit pouvoir s'intéresser à des phénomènes habituellement réservés à d'autres disciplines.

— Si la rhétorique de Becker est parfois perçue comme celle de l'impérialiste qui applique sans ménagement ses concepts à d'autres disciplines, la science économique s'est également employée à importer avec succès les concepts issus d'autres disciplines comme la psychologie. Une telle affirmation concerne tout aussi bien la psychologie sociale, qui fait état de changements d'attitude des consommateurs [Schwarz and Strack, 1991, Tourangeau, 1992], que les sciences cognitives. Non seulement ces travaux remettent en cause la stabilité des attitudes qui mènent à la consommation, mais ils permettent également de contester la généralité du concept de préférences professée par Friedman. De plus, les travaux de Tversky, Sattath and Slovic [1988] sur l'aversion au risque suggèrent que les préférences changent en fonction du contexte. L'interdisciplinarité se retrouve également à l'intersection entre la science économique d'une part, la mercatique et les sciences de gestion d'autre part [Holbrook and O'shaughnessy, 1988] ; ou encore dans la référence que certains économistes font à l'anthropologie [Bowles, 1998].

7. Je traduis ici *wants* par préférence par souci de clarté. Le terme de désir n'apportant, à mon sens, rien dans ce contexte.

L'interprétation selon laquelle, dans la division du travail, l'économiste ne porte que peu d'attention aux changements de préférences repose donc sur un constat erroné. Il est aujourd'hui difficile de soutenir que l'ensemble de la communauté des économistes constitue un bloc homogène qui ne collabore pas avec les autres disciplines. Toutefois, constater que certains économistes collaborent avec d'autres disciplines n'implique pas que cette collaboration soit utile pour aborder des problématiques dont l'appartenance à la discipline économique est incontestable.

1.3 La raison substantielle de cette négligence

C'est au nom de la défense d'une science économique portant sur un domaine substantiel d'objets que se formule une troisième raison du rejet de l'analyse de la transformation des préférences [8]. Elle consiste à dire que les changements de préférences ne sont tout simplement pas pertinents pour traiter les grands sujets traditionnellement abordés par les économistes. Certains économistes [Robinson and Acemoglu, 2012] soutiennent par exemple que les institutions expliquent à elles seules des phénomènes comme la croissance et c'est par un jeu subtile sur le système d'incitations (droit de propriété, système fiscal, politique de la concurrence) que se créent les conditions d'une croissance stable et équilibrée [9]. Une telle affirmation suppose cependant que les systèmes d'incitations dont il est ici question sont exogènes, indépendants du niveau des prix, des politiques publiques mises en œuvre ou du système institutionnel en vigueur [10]. De Schumpeter and Perroux [1935] à Bowles [2001], en passant par Polanyi and MacIver [1944] ou Hirschman [1983], nombreux sont les auteurs de renom à avoir insisté sur la manière dont « institutions et systèmes de valeurs coévoluent » [Tabellini, 2010]. Autrement dit, les préférences sont endogènes : elles dépendent des institutions, et les institutions dépendent d'elles. Dès lors, si les préférences sont endogènes, ne faut-il pas considérer qu'il en est de même des incitations sur lesquelles sont supposées reposer les institutions ? Comprendre l'évolution des préférences pourrait jeter la lumière sur la persistance dans le temps de certaines institutions, comme celles des cités médiévales italiennes [Guiso, Sapienza and Zingales, 2006], ou au contraire sur leur évolution rapide comme celles de l'Etat providence [Alesina and La Ferrara, 2005, Alesina and Ardagna, 2010].

De plus, la prise en considération de l'évolution des préférences conduit, depuis les années 1970, à réinterroger des domaines d'études fondamentaux pour les économistes. On

8. Pour une discussion méthodologique de la question des définitions substantielles et formelles de l'économie, on pourra consulter [Lallement, 2002].

9. Dans leur best-seller, *Why nations fail ?*, Acemoglu et Robinson considèrent que la croissance résulte avant tout de la structure institutionnelle des Etats. Ils prennent l'exemple de Nogales, ville à cheval entre les Etats-Unis et le Mexique et dont les criantes différences de développement s'expliquent, selon eux, uniquement par les différences institutionnelles.

10. Autrement dit, affirmer que les préférences changent n'implique pas qu'elles soient endogènes. Toutefois, l'exogénéité des préférences est une condition qui suffit à rendre ces changements non pertinents et donc à considérer les préférences comme fixes. C'est pourquoi endogénéité et changements de préférences sont généralement mentionnés sans distinction dans la littérature sur ces questions.

souligne l'intérêt des préférences endogènes, tant sur le plan positif de la théorie de la demande que sur le plan normatif de l'économie du bien-être [Pollak, 1978]. Des phénomènes comme la mode [Karni and Schmeidler, 1990] ou la publicité [Ashley, Granger and Schmalensee, 1980] deviennent objet d'attention et laissent penser qu'une étude exhaustive des mécanismes caractérisant la demande ne peut se réduire à l'étude des prix : les variations de la consommation sont plus subtiles et méritent d'être complexifiées [Pessemier, 1978] [11]

Cette nécessaire complexification se retrouve en économie normative, qui fonde son évaluation du bien-être sur l'hypothèse de fixité des préférences [Galbraith, 1958, Weckstein, 1962].

> En tant que conseiller politique, l'économiste est censé chercher l'efficacité. Toutefois, qu'une politique donnée soit efficace ou non, dépend des préférences de ceux qu'elle affecte, préférences qui dépendent en retour de la politique menée.
>
> [Marschak, 1978, p. 387]

Si elle prenait au sérieux les phénomènes de transformation des préférences, la théorie devrait alors se mettre en quête de nouveaux critères d'évaluation du bien-être [Gintis, 1974].

Cet intérêt des économistes pour les préférences endogènes découle également du développement de nouvelles branches en économie appliquée, comme l'économie politique ou l'économie du développement. On s'intéresse désormais à des formes d'incitation qui ne dépendent pas seulement du revenu, en distinguant la motivation intrinsèque de la motivation extrinsèque [Benabou and Tirole, 2003], où en étudiant la relation entre normes sociales et incitations [Bénabou and Tirole, 2006]. On théorise la manière dont les agents acquièrent des attitudes, parfois sociales (*other regarding preferences*) comme l'altruisme [Bergstrom and Stark, 1993] ou parfois individuelles comme l'aversion au risque [Levy, 2015]. On étudie comment certains « systèmes de valeurs » [Corneo and Grüner, 2002] peuvent être encouragés, comme ceux offrant la part belle à la tolérance [Corneo and Jeanne, 2009] ; comment s'apprend la démocratie [Ticchi, Verdier and Vindigni, 2013], ou bien comment se combat la corruption [Hauk and Saez-Marti, 2002]. On s'intéresse également à l'évolution des préférences politiques comme les préférences pour la redistribution [Alesina and Fuchs-Schündeln, 2007, Bisin and Verdier, 2004], qui permettent de comprendre l'évolution de l'Etat providence et le rôle des politiques mises en œuvre dans cette évolution [Bowles, 1998]. Enfin, de nombreux phénomènes liés à la mondialisation des échanges impliquent l'interaction d'agents dont les préférences sont hétérogènes. Comprendre les effets de cette hétérogénéité devient un enjeu fondamental pour la science

11. Pessemier écrit par exemple : « Individual preference behavior is highly varied and choice behavior cannot be easily predicted on any given purchase occasion. These observations do not flow from irrational behavior, faulty measurmement or imperfect model building. A varied choice repertoire can be attributed largely to the adaptive characteristics and social needs of the species. Instead of ignoring or demeaning these inconvenient facts of life, their economic effects should be examined with care » [Pessemier, 1978, p. 7].

économique. On peut penser à l'étude de l'immigration [Algan and Cahuc, 2013] ou de l'expansion du commerce international [Guiso, Sapienza and Zingales, 2004].

Il semble donc fondamental de comprendre comment les préférences évoluent pour traiter de problèmes qui relèvent traditionnellement de l'objet économique, comme la croissance, le commerce ou le développement. Toutefois, comme nous le verrons dans la troisième section de cette introduction, si la plupart des travaux cités permettent d'étayer l'impact des changements de préférences, les modèles sur lesquels ils s'appuient articulent de façon confuse les concepts de rationalité et de changement de préférences. Soit ils forcent les changements de préférences à entrer dans le cadre d'une conception restrictive de la rationalité, soit ils refusent de faire de la rationalité un moteur des changements de préférences. Dans ces conditions, la rationalité ne rend plus le comportement intelligible, elle est une simple métaphore permettant d'aboutir à un « concept d'équilibre » préétabli, et pensé pour s'appliquer à des situations qui n'ont rien à voir avec les changements de préférences. Aussi conduisent-ils à des explications *ad hoc*. Ce qui nous mène à une quatrième raison de négliger la formation des préférences.

1.4 La raison méthodologique de cette négligence.

Une quatrième raison de ne pas prêter attention aux changements de préférences tient à une difficulté méthodologique que Grüne-Yanoff and Hansson [2009] déclinent en deux problèmes. Tout d'abord, en faisant varier les préférences, le modélisateur serait capable d'expliquer n'importe quel phénomène. Dans ces conditions, les changements de préférences conduiraient nécessairement à des explications *ad hoc* et il serait impossible de discriminer plusieurs théories sur la base de données empiriques [12]. Il en découlerait un faible pouvoir explicatif pour la théorie. Cet argument méthodologique est important, mais il ne conduit pas nécessairement à la conclusion que les changements de préférences doivent être ignorés. Il nous invite plutôt à sélectionner les changements de préférences susceptibles de reposer sur des fondements théoriques. Comme l'écrivent [Grüne-Yanoff and Hansson, 2009, p. 7], « dans un contexte où les préférences changent, notre pouvoir de discrimination empirique des théories est probablement faible, à moins que les changements de préférences soient suffisamment structurés ». Cela suppose de construire une théorie qui conceptualise de façon rigoureuse les états mentaux qui guident l'évolution de l'état psychologique de l'agent. C'est pourquoi nous avons besoin de modèles procurant une structure (psychologique) à ces changements. Un tel mécanisme permettrait de sélectionner les changements qui sont susceptibles d'être rationalisés, à la différence de ceux

12. Becker formule la même critique en ces termes :

« Our assumption that extended preferences are stable was intended not as a philosophical or methodological "law," but as a productive way to analyze and explain behavior. We were impressed by how little has been achieved by the many discussions in economics, sociology, history, and other fields that postulate almost arbitrary variations in preferences and values when confronted by puzzling behavior. We hoped that making these puzzles explicit would hasten the development of more rewarding approaches » [Becker, 1996, p. 6]

qui ne le sont pas. La critique méthodologique de la formation des préférences nous invite donc à ouvrir la boite noire des changements de préférences.

Le deuxième problème a trait au type de données à partir desquelles nous pourrions tester les changements de préférences. Selon Grüne-Yanoff and Hansson [2009], supposer les préférences susceptibles de changer impliquerait de ne plus être en mesure d'« expliquer » les préférences en matière de choix. S'imposerait alors le recours à des données introspectives. Pourtant, on considère traditionnellement en théorie du choix que seuls les comportements « révèlent » les préférences des agents. La plupart des états mentaux ne sont pas observables. Avoir recours à des états mentaux pour expliquer les changements de préférences nous conduirait donc à employer des concepts inobservables et donc impossibles à tester empiriquement. Négliger le contenu mental des préférences relève donc d'un impératif de falsification. Si c'est le comportement des agents qui nous intéresse, nous pouvons nous limiter à l'observation du comportement et donc à une conception opaque et immuable des préférences. Toutefois, il n'est pas certain que la discipline économique doive concentrer son analyse sur la seule étude du comportement.

> En tant qu'économistes, nous ne nous intéressons pas qu'au comportement des individus, mais également à la satisfaction d'états psychologiques comme le bien-être par exemple. Lorsque nous analysons les mécanismes sociaux et que nous les évaluons en termes de bien-être, ce sont les préférences des individus que nous avons en tête, en tant qu'elles reflètent leur bien-être.
>
> [Rubinstein and Salant, 2011, p. 118]

C'est pourquoi nous avons besoin de théoriser ces états mentaux. D'un point de vue empirique, cela suppose une approche pluraliste, avec le recours à des données relativement moins répandues en économie, comme les données introspectives ou les données déclaratives. De plus, si les données introspectives ne sont pas à exclure, les données de choix ne perdent pas toute leur pertinence dans le contexte de préférences mouvantes. Il reste envisageable de comprendre les préférences comme révélées par le choix à un moment donné du temps, et d'étudier leur évolution à l'aide de ce que le consommateur révèle par ces choix à chaque période. Pour cela, une conceptualisation théorique dans les termes de la théorie de choix rationnel est requise.

> Les économistes [...] ne se sont pas attaqués au problème de la complexité et de l'endogénéité des motivations humaines, non parce qu'ils voient en l'*homo economicus* et sa simplicité une représentation adéquate, mais par manque d'outils conceptuels et d'informations empiriques sur le processus de formation des préférences.
>
> [Bowles, 2001, p. 2]

Nous avons bel et bien besoin de théories expliquant la façon dont les préférences se transforment. C'est vers cette conclusion que nous oriente l'analyse des quatre raisons

évoquées de négliger les changements de préférences. Nous devons développer des modèles permettant de comprendre non seulement pourquoi les agents font tel ou tel choix, mais également pourquoi ils adoptent telle ou telle préférence.

Loin de mettre de côté le concept de rationalité, une telle théorie doit pouvoir rendre compte de la rationalité des changements de préférences. De même que le changement de Swann découle d'un raisonnement, de même la rationalité des agents pourrait expliquer certains changements de préférences. Encore faut-il préciser ce que l'on entend par rationalité dans le cadre de la théorie des changements de préférences. La section suivante s'attaque à définir ce que j'appellerai un changement de préférences rationnel.

2 Qu'est-ce qu'un changement de préférences rationnel ?

Dans le contexte des sciences sociales, attribuer des états mentaux à celui qui agit pour en restituer la logique constitue une condition d'intelligibilité de son comportement. *Comprendre* un individu c'est associer une forme de rationalité aussi bien à sa conduite qu'à la façon dont elle peut évoluer [13]. En adoptant ce principe dit de « charité » [14], l'observateur rejette « les hypothèses diverses du hasard absolu, du conditionnement rigide, de la décision capricieuse et du désordre des idées » [Mongin, 2002, p. 302]. Pour donner sens aux changements de préférences il convient donc d'en éclaicir la rationalité.

Cette tâche suppose d'établir la règle de transition d'une relation de préférences à une autre, et de montrer le mécanisme psychologique sur lequel elle repose. Une telle règle n'a rien d'évident. Rien n'indique *a priori* qu'une préférence doit évoluer de telle ou telle manière. Il n'y a rien d'irrationnel à conserver un désir qui ne peut être assouvi. En réalité, je voudrais défendre la thèse que la rationalité des changements de préférences passe nécessairement par la rationalité épistémique de l'agent. Cela ne signifie pas, pour autant, que la rationalité bayésienne suffise à rendre compte de la rationalité à l'oeuvre dans les changements de préférences. C'est pourquoi une analyse de la rationalité spécifique à l'oeuvre dans les changements de préférences est requise.

Après un retour sur les principales façons dont la théorie du choix rationnel aborde ces questions, je montre qu'il est possible de penser des changements de préférences rationnels

13. « L'ethnologue étudiant une culture étrangère vivante doit s'être souvent posé la question : 'Que pensent les gens de cette culture ? Comment pensent-ils ? Leurs processus rationnels et intellectuels sont-ils semblables aux nôtres ?'» [Whorf, 1956, p. 65]

14. Formulé en premier lieu par Quine [1960], le principe de charité tient originellement à l'idée qu'en situation de « traduction radicale », il est toujours méthodologiquement préférable de considérer que la traduction de l'observateur est mauvaise, si elle rend les propos du locuteur irrationnels. « La maxime de traduction qui est à la base de tout ceci, c'est qu'il est probable que les assertions manifestement fausses à simple vue [comme p et non-p] fassent jouer des différences cachées de langage... La vérité de bon sens qu'il y a derrière cette maxime, c'est que la stupidité de notre interlocuteur, au-delà d'un certain point, est moins probable qu'une mauvaise traduction » [Quine, 1960, p. 101]. Ce principe est ensuite développé par Davidson, qui lui donne notamment une dimension pratique : il est préférable de penser que la théorie qui conduit à interpréter le comportement de l'agent est mauvaise, plutôt que de dire que l'agent est irrationnel.

en s'inspirant des outils qu'elle propose [15]. Je défends enfin l'idée que, pour être rationnel, ce changement ne doit pas être conçu comme procédant des préférences à elles-seules : il découle d'une révision des états épistémiques de l'agent : de ses croyances et également d'une prise de conscience.

2.1 Rationalité des préférences

Dans sa version « formelle », la théorie du choix propose une interprétation minimale de la rationalité des préférences. Inspirée de la théorie humienne de l'action, elle fait sienne la célèbre formule du *Traité de la nature humaine* : « il n'est pas contraire à la raison de préférer la destruction du monde entier à une égratignure de mon doigt » [p.273] [16]. La rationalité est une notion formelle qui présuppose simplement que l'agent adopte les moyens qui sont à sa disposition pour parvenir à ses fins. Ce faisant, elle laisse ouverte la question des fins qui méritent d'être valorisées.

> Dans la tradition de la science économique, du moins telle qu'elle est exprimée par Pareto (*Traité de sociologie générale*), une action est dite rationnelle lorsqu'elle est objectivement bien adaptée au but poursuivi par le sujet. Rationalité signifie dans ce cas : adaptation des moyens aux fins. L'économiste moderne, pour sa part, définit le comportement rationnel comme le choix par l'individu de l'action qu'il préfère parmi toutes celles qu'il a la possibilité d'accomplir, en bref comme un choix conforme à ses préférences. Cette définition tend - notons-le incidemment - à introduire une hypothèse irréfutable à partir du moment où les préférences sont, comme c'est généralement le cas, induites à partir des comportements observés. En général, l'économiste s'abstient d'appliquer le concept de rationalité aux fins elles-mêmes.
>
> [Boudon and Bourricaud, 1983, p. 479]

Les préférences sont rationnelles simplement parce qu'elles sont structurée par des impératifs de « cohérence interne ». C'est-à-dire qu'elles fournissent une hiérarchie permettant de classer des actions, des biens ou des états du monde de façon non contradictoire [17]. Elles ne rationalisent le comportement que dans la mesure où elles impliquent que l'agent

15. Dans cette thèse je me concentre sur les préférences individuelles. Je laisserai donc de côté les interrogations sur la rationalité collective qui portent sur les concepts de *common prior*, de *common knowledge* ou bien de *common belief in rationality*.

16. On peut noter que selon de nombreux commentateurs, la théorie humienne de l'action telle qu'elle se donne dans l'imaginaire collectif est différente de celle défendue par l'auteur du *Traité de la nature humaine*. Selon Sugden [2005], la théorie humienne de la rationalité est plus minimale encore que celle défendue par la théorie du choix rationnel. De même, Broome [1999] propose de distinguer une interprétation « modérée » de la théorie humienne, qui s'accorderait avec la théorie du choix rationnel, à une interprétation plus minimale encore de la rationalité, qui n'imposerait quasiment aucune contrainte sur le choix. Voire également [Diaye and Lapidus, 2012].

17. Dans le cas le plus simple, où l'ensemble des alternatives est fini, on attribue aux préférences généralement la transitivité et la réflexivité. On considère également que les préférences sont complètes, même si, comme nous l'évoquons plus bas, cette hypothèse ne semble pas nécessaire pour attribuer la rationalité à l'agent.

choisit toujours ce qu'il préfère dans l'ensemble des choses que ses moyens mettent à sa portée. C'est exactement l'idée de l'axiome faible des préférences révélées, qui veut que si l'agent choisit a lorsque b est disponible, alors il faut que, dans chaque situation où a et b sont disponibles, le choix de b implique celui de a également [18]. La cohérence est *interne* au sens où il n'est pas besoin de données extérieures au choix pour en rendre compte [Dowding, 2002, Sen, 1993]. Que ma préférence de a sur b repose subjectivement sur une entité psychologique qui lui donne substance (une motivation, un but, un principe normatif), n'implique pas que l'observateur ait à lui assigner un contenu externe au choix. La rationalité des préférences s'exprime en termes de cohérence des choix et non en termes de cohérence mentale.

Ce refus de donner une substance au concept de préférences qui soit *externe* au choix la rend incapable d'expliquer pourquoi les agents préfèrent ce qu'ils préfèrent. La relation causale qu'elle met en exergue entre préférences et choix n'est pas à proprement parlé explicative. Le choix est causé parce qu'il est causé, l'agent fait ce qu'il préfère parce qu'il le préfère [19], c'est-à-dire parce qu'il se comporte ainsi. Une telle cécité n'a pas pour seule conséquence d'affaiblir le lien supposé entre psychologie du sens commun et théorie du choix rationnel [20]. Elle rend impossible la description de ce qui pourrait faire varier les préférences des individus. Il faut bien que nous soyons en capacité d'assigner un contenu externe aux préférences, comme une motivation, une raison, pour comprendre ce qui, au cours du passage d'une relation de préférences à une autre, change dans l'esprit de l'agent.

Par conséquent, dans sa conception formelle, la rationalité des préférences est trop restrictive d'un point de vue dynamique précisément parce que, en statique, elle se veut agnostique. En un sens, c'est bien ce souci de rester agnostique qui conduit à se donner la stabilité des préférences pour hypothèse : soucieuse de ne pas caractériser les préférences, la conception formelle se condamne au mutisme quant à leur évolution. Dans ces conditions, il est tout naturel de voir les changements de préférences comme *ad hoc*. Construire une théorie des changements de préférences suppose donc de leur assigner un contenu externe au choix.

C'est à ce titre que le concept de rationalité axiologique mérite l'attention de celui qui

18. Il s'agit donc bien de dire que les choix des agents ne sont contraints que par les moyens qu'ils ont à leur disposition. Pour le voir on peut, à l'instar de Sen, décortiquer cette condition entre le cas où les possibilités (les moyens) de l'agent diminuent et celles où elles s'étendent. *La propriété* α veut que lorsque les possibilités de l'agent diminuent, mais que le choix initial de l'agent reste disponible alors son choix reste le même. $\forall S, T$ si $a \in S \cap C(T)$ et $S \subseteq T$ alors $x \in C(S)$. *La propriété* β veut qu'à l'inverse, lorsque ses possibiltés s'étendent, l'indifférence de l'agent reste stable. $\forall S, T$ tels que $S \subseteq T$ si $a, b \in C(S)$ alors $a \in C(T)$ si et seulement si $b \in C(T)$.

19. C'est en ce sens que, comme l'écrivent Boudon and Bourricaud [1983], la théorie du choix rationnel introduit une « hypothèse irréfutable »

20. Davis Lewis fait de la théorie de la décision l'expression même de la psychologie de sens commun. « Bien que de prime abord peu intuitive aux non-initiés, la théorie de la décision (si, du moins, on en exclu la fioriture) n'a rien d'une science ésotérique. Au contraire, elle consiste en une exposition systématique des conséquences de certaines banalités intelligemment choisies sur les croyances, les désirs, les préférences et le choix. Elle forme le cœur de notre psychologie de sens commun des personnes, savamment disséqué et élégamment systématisé. » [Lewis, 1983, p. 337].

cherche à comprendre comment les préférences se forment.

2.2 La rationalité axiologique

Dans son interprétation substantielle, la rationalité des préférences ne se limite pas à une analyse interne, i.e., qu'elle s'appuie sur des données extèrieures au choix pour expliquer les préférences de l'agent. Dire qu'un agent est substantiellement, rationnel c'est dire qu'il préfère a à b parce qu'il a des états mentaux, comme des motivations, des désirs qui le conduisent à préférer a à b. Un concept substantiel de la rationalité des préférences suppose donc d'assigner un contenu mental aux préférences, contenu mental qui fournit la « base causale » des préférences. Dès lors, expliquer comment les préférences varient c'est expliquer par quel mécanisme cette base causale change. Dans le premier chapitre de cette thèse, je définis ces états mentaux comme des valeurs, la cohérence entre ces valeurs donnant sens à la rationalité axiologique de l'agent.

Contrairement à sa version instrumentale, la rationalité axiologique ne porte pas sur les moyens de parvenir à une fin, mais sur ces fins elles-mêmes. La distinction kantienne entre « impératif hypothétique » et « impératif catégorique » illustre une telle conception sur le plan de la morale Dietrich and List [2013*b*] [21]. Alors que l'impératif hypothétique offre un critère de sélection des moyens susceptibles d'aboutir à la meilleure des fins selon un ordre fixé au préalable, l'impératif catégorique suppose de délibérer sur les fins qui valent la peine d'être poursuivies. D'après Pettit [1991], puisqu'elle se donne l'étude de la psychologie de sens commun pour objet, la théorie de la décision est incomplète : elle est incapable de conceptualiser la manière dont le agents délibèrent sur les fins qu'ils devraient adopter.

Le modèle que je propose dans cette thèse vise à combler ce manque en proposant une théorie où les individus délibèrent sur leurs valeurs : elle suppose donc bien d'adopter une conception axiologique de la rationalité. Deux ecueils méritent toutefois d'être explicités.

— D'abord, le concept de rationalité axiologique suppose que le relativisme axiologique est erroné. Certains systèmes de valeurs, certaines motivations sont plus rationnels axiologiquement que d'autres, et donc certains changements de préférences méritent d'être exclus des sphères de la rationalité. Ce sera d'ailleurs la contribution principale du second chapitre que de caractériser ces changements de préférences. Autrement dit, la conception formelle des préférences permet de rationaliser un grand nombre de choix, mais elle sous-détermine, en dynamique, l'ensemble des changements de préférences possibles. Pour le modélisateur l'arbitrage est alors le suivant : plus il assigne de contenu aux préférences, plus il est capable de déterminer leur évolution,

21. On peut également penser à la fameuse distinction de Weber entre « rationalité en finalité » et « rationalité en valeur ». Toutefois, dans l'idée de Weber, il s'agit bien de distinguer deux modes de choix dont les motifs sont différents. Alors que dans le premier type les agents sont mus par le désir de parvenir à leur fin, ils prennent, dans le deuxième type, leur décision indépendamment des conséquences possibles de leur choix. Voir [Weber et al., 1995].

mais plus il restreint l'ensemble des préférences admissibles. Le curseur entre les différentes positions de cet arbitrage est donné par le type de restriction logique imposé à la rationalité axiologique de l'agent. Le modèle proposé dans cette thèse offre les premiers rudiments pour placer ce curseur.

— Ensuite, la rationalité axiologique ne permet pas d'expliquer l'hétérogénéité des préférences lorsque les agents ont des ressources cognitives illimitées. Si l'agent délibère sur les fins à poursuivre, soit il n'y pas de raison qu'il choisisse des valeurs différentes de ses congénères, soit toutes les préférences ont la même valeur axiologique et la rationalité axiologique ne permet pas d'expliquer le passage d'une relation de préférences à une autre. Il est donc nécessaire de supposer que les agents disposent de capacités limitées à faire usage de leur rationalité axiologique. Cette solution est également proposée par Boudon [1999] [22].

Ce dernier point suggère que les failles cognitives des agents sont susceptibles d'être moteur des changements de préférences. Dans l'armature conceptelle de la théorie du choix, la cognition des agents est données par leurs états épistémiques, i.e., par un cortège de notions comme celles de croyances ou de conscience (*awareness*). Il est donc nécessaire de s'interroger sur les connexions que les états epistémiques entretiennent avec la transfomation des préférences.

2.3 La non-séparation des préférences et des croyances

Admettre que la rationalité des changements de préférences passe par la prise en considération de la capacité des agents à délibérer sur les fins qu'ils poursuivent, suppose de comprendre à partir de quelle règle de révision l'agent est susceptible delibérer sur les fins. Or, il n'existe pas de règle de révision évidente sur la transition d'une relation de préférences à une autre. Pour comprendre cette affirmation, reprenons à notre compte les remarques de Searle [1995] à propos du désir pour les appliquer au concept de préférences.

Selon Searle, croyance et désir constituent des états intentionnels, c'est-à-dire des attitudes propositionnelles qui renvoient à des conditions de satisfaction. Dire « je crois que p », c'est suggérer implicitement qu'il existe une condition pour que ma croyance soit satisfaite. Pour la croyance, une telle condition est donnée par le concept de vérité : ma croyance est satisfaite si elle est vraie. Parallèlement, dire, « je désire que p », c'est renvoyer à une condition de satisfaction différente : mon désir que p est satisfait si j'obtiens p. Les conditions de satisfaction de la croyance et du désir diffèrent par leur direction

22. Une telle approche se retrouve chez de nombreux sociologues, comme Max Weber ou Durkheim pour qui les différences de cultures s'expliquent par une distribution des ressources cognitives différenciée."A belief is held by a given individual as true in a first stage because it is collective ; in a second stage, it becomes collective because it is true.' To Durkheim, while values vary from one society to another, they should in other words not be analyzed as illusions. They vary from one society to another because the context, the cognitive re- sources, the level of information of people on such and such subjects and many other factors vary." [Boudon, 1999, 479]

d'ajustement. La direction d'ajustement établit celui qui, du monde ou de l'agent, est responsable si les conditions de satisfaction ne sont pas remplies. La direction d'ajustement du désir va du monde vers l'agent. Si le désir de l'agent n'est pas satisfait, l'agent n'est pas responsable, c'est le monde qui est à blâmer. Il n'y a donc pas de raison évidente pour que l'agent change son désir. Tout au plus peut-il tenter d'acquérir de nouveaux moyens pour que son désir soit satisfait. A l'inverse, la direction d'ajustement de la croyance va de l'agent vers le monde : si la croyance de l'agent n'est pas vraie, ce n'est pas le monde qui est responsable, mais l'agent, qui a maintenant intérêt à réviser sa croyance. Parler d'un changement de désir rationnel n'a rien d'évident. Dès lors, qu'en est-il des préférences ? S'il doit y avoir changement rationnel des fins qui induisent les préférences de l'agent, il semble qu'il doive être la conséquence d'une révision des croyances de l'agent, et plus généralement, de ses états épistémiques. Avant de poursuivre dans cette voie, il est cependant nécessaire de justifier la possibilité d'une connexion entre croyances et préférences.

Héritière d'une conception humienne de l'action, la théorie de la décision dans l'incertain postule la séparation des préférences et des croyances [23]. Toutefois, notre vocabulaire commun est plein de concepts mixtes, qui expriment quelque chose entre le concept de croyance et de préférence. Dire « je crois qu'il faut être communautarien » ne revient pas à formuler une croyance portant sur un état de fait empiriquement vérifiable. Cela ne signifie pas cependant qu'une telle affirmation ne soit pas constituée d'éléments cognitifs avec lesquels mes croyances sont susceptibles d'entrer en interaction. Dire « je crois qu'il existe un Dieu qui nous regarde et qui attend qu'on lui exprime son amour », c'est formuler une vision du monde bien plus qu'une croyance telle qu'elle se donne dans la théorie bayésienne. Qu'il s'agisse de vision du monde, d'idéologies, de valeurs, de conceptions, les modèles d'économie appliquées enrichissent volontiers les préférences de ces concepts « mixtes ». Comme le souligne Buchanan [1991], l'acceptation courante du concept de préférences est plus restreinte que l'usage. Le concept de préférences ne renvoie pas qu'à des évaluations subjectives, il peut revêtir une dimension cognitive, et être assimilé à des visions du monde, des attitudes ou des valeurs.

> Dans sa compréhension commune, le concept de préférences renvoie à un ensemble de valeurs purement subjectives. Il réfère aux évaluations de l'agent de ses choix potentiels. Dans son usage quotidien, cependant, il ne saurait être réduit à cette dimension évaluative. Son usage mêle des composantes évaluatives et cognitives, ou, pour le dire autrement, il mêle l'évaluation d'un individu [...] des conséquences possibles de son choix et sa vision du monde.
>
> [Buchanan, 1991, p.168]

Bien qu'elles renvoient à des entités conceptuellement différentes, les préférences et les états épistémiques de l'agent méritent d'être pensées dans leur interaction. S'il est

23. Cette séparation se manifeste notamment chez Savage [1954] au travers du fameux axiome de la chose sûre.

possible de parler de rationalité des changements de préférences, c'est parce que certains états épistémiques comme les croyances ont un effet sur les préférences. La question est alors de savoir si nous pouvons nous contenter d'utiliser la théorie classique de révision des croyances fournie par les modèles bayésiens pour comprendre l'évolution des préférences.

2.4 La révision des croyances

En conformité avec une interprétation formelle des préférences, on pourrait soutenir que les changements de comportement découlent exclusivement de l'acquisition par l'agent de nouveaux moyens pour parvenir à ses fins. L'un des moyens dont dispose l'agent rationnel pour parvenir à ses fins est l'information [24]. C'est par ce truchement qu'intervient le concept de croyance. Dans le langage de la théorie bayésienne, les croyances d'un individu se formulent par l'assignation d'une probabilité à des évènements, c'est-à-dire à des sous-ensembles de l'ensemble des états du monde, Ω. On dit alors que des croyances sont rationnelles si elles répondent à des impératifs statiques et des impératifs dynamiques.

— Les impératifs statiques supposent que les croyances des agents obéissent à des principes probabilistes rudimentaires. Si par exemple l'évènement A implique l'évènement B, i.e. $A \subset B$, alors l'agent croit que l'évènement B est plus probable que l'évènement A, $P(A) < P(B)$ [25].

— Les impératifs dynamiques supposent qu'à l'acquisition de nouvelles informations, l'agent révise ses croyances selon la règle du conditionnement bayésien [26].

Clarifions cela par un exemple.

Tout saltimbanque qu'il est, Pierrick fréquente régulièrement ces fêtes de village où des tombolas sont organisées. En première instance, Pierrick est tenté d'acheter un ticket pour 10 euros, notamment parce que le prix gagnant est le synthétiseur de ses rêves. Toutefois, il apprend en seconde instance que le nombre de tickets distribués est si important que ses chances de succès sont quasi nulles. En changeant la croyance de Pierrick, cette information a pour effet de changer son choix ; il juge maintenant plus profitable d'utiliser ses 10 euros pour acheter un poireau à sa belle. Une telle révision est purement informationnelle [Dietrich and List, 2011]. Avec l'information, ce n'est pas l'utilité de son action qui change, mais celle qu'il peut en espérer, étant donné les moyens dont il dispose [27]. Or, ce type de changement est trop restrictif. Supposons par exemple que Pierrick n'apprenne rien sur l'éventualité de remporter la mise, mais qu'il vienne à son esprit que, faute de poireau,

24. On peut également penser à la productivité ou aux dotations de l'agent comme constituant des moyens de l'action. Mais je laisse de côté ces questions dans cette thèse.

25. A noté que Tversky and Kahneman [1983] démontrent l'invalidité de ce principe à travers le biais de représentativité.

26. L'information se donne comme une partition sur les états du monde qui permet à l'agent de distinguer l'ensemble des états du monde compatibles avec le monde actuel de ceux qui ne le sont pas.

27. Après avoir acquis l'information $\mathcal{I}$, l'espérance d'utilité de l'acte f de l'agent passe de $\mathbb{E}u(f)$ à $\mathbb{E}u(f|\mathcal{I})$, où $\mathbb{E}u(f|\mathcal{I})$ représente son espérance d'utilité de conditionnellement à l'information contenue dans $\mathcal{I}$.

sa belle pourrait en venir à le congédier. Pierrick décide de changer son choix. Cette dernière situation ne résulte pas d'un changement d'information tel qu'il est compris par le paradigme bayésien. Elle suppose que Pierrick assigne maintenant une utilité à une conséquence qu'il ignorait au préalable. Or, ce n'est pas compatible avec le paradigme bayésien où l'agent est capable de tout imaginer et donc de tout évaluer en matière d'utilité [28].

Dans un cadre bayésien, l'acquisition d'information suppose que l'agent restreigne l'ensemble des possibles, excluant les évènements incompatibles avec l'information recueillie. Une nouvelle information n'implique pas la prise en compte de nouvelles possibilités.

> Dans le paradigme bayésien, à mesure que de nouvelles découvertes sont établies et que l'information devient disponible, l'ensemble des possibles rétrécit [...]. Ce processus de « destruction » reflète l'impossibilité, au sein du cadre bayésien, d'étendre l'espace des mondes possibles [...]. Néanmoins, autant l'expérience que l'intuition contredisent cette vision du monde. S'accoutumer à des possibilités qui étaient précédemment inconcevables est au cœur de l'Histoire et de notre expérience de la vie. Il y a donc un sens à dire que l'univers s'étend à mesure que nous prenons conscience de ses nouvelles opportunités.
>
> [Karni and Vierø, 2013, p. 2790]

Ce défaut du paradigme bayésien découle d'un axiome important de la rationalité épistémique sur laquelle il repose : l'introspection négative. Cet axiome affirme que lorsque l'agent ne connait pas une chose, alors il sait qu'il ne la connait pas. L'agent est capable d'attribuer une probabilité (même subjective) à n'importe quel évènement. Or, il est parfaitement envisageable que l'agent ne dispose pas de certains concepts pour envisager la réalité, que son attention flanche ou qu'il échoue à être logiquement omniscient, sans pour autant qu'il soit nécessaire de le qualifier d'irrationnel. Je ne suis pas perpétuellement en train de me demander si les chats poussent dans les arbres, cela ne fait pas de moi quelqu'un d'irrationnel. Etre incapable d'imaginer, de concevoir ou de se questionner sur certains évènements n'a rien d'irrationnel. Cela ne conduit pas l'agent à adopter un système de croyances contradictoires. Prétendre le contraire reviendrait à voir dans la contradiction entre des grandes théories scientifiques, comme la théorie de la relativité générale et la mécanique quantique, un symptôme de l'irrationalité des physiciens. De même que la complétude des préférences n'est pas vue comme une propriété nécessaire de la rationalité, de même l'incomplétude des croyances n'a rien d'irrationnel [29].

Dès lors, une autre façon de comprendre le renversement de préférences de Pierrick consiste à souligner que dans cet exemple, il prend subitement conscience d'une nouvelle conséquence de son choix : son choix risque de lui faire perdre sa belle. Nul besoin de parler de changement de préférences. Il suffit de prendre ses préférences pour cette conséquence

28. Malgré l'importance de cette littérature, je laisse de côté ce qui a trait à la question de l'ambigüité. Pour une revue en langue française de ces travaux, on pourra consulter [Cohen and Tallon, 2000].

29. "In fact, it appears that the completeness axiom does not really correspond to an unexceptionable trait of rationality, but it is rather useful on the grounds of analytical tractability." [Ok, 2002, p. 2]

comme déjà données. Toutefois, cette solution n'est pas susceptible d'être appliquée dans l'exemple inaugural que nous avons pris. Dans cet exemple, rappelons-le, ce sont bien les préférences de Swann qui changent. Swann passe d'un état psychologique où il se dit qu'Odette n'est pas son genre à un état où il la trouve appréciable. S'il est bien question de la prise de conscience par Swann de la ressemblance d'Odette avec la figure peinte de Zephora, celle-ci porte bien davantage sur l'objet de son affection. Sa façon d'attribuer la beauté change et ses préférences avec elle.

Cela nous suggère, que pour comprendre la manière dont les préférences changent, il faut rendre compte de la connexion entre la conscience des individus et le contenu axiologique qui constitue les préférences de l'agent. La construction conceptuelle de ce lien fait l'objet du chapitre premier de cette thèse. Cette connexion n'a rien d'immédiat et suppose d'adopter une position mentaliste. C'est à partir de cette position que nous pourrons définir le concept de rationalité.

2.5 Mentalisme et rationalité des changements de préférence

Dans cette thèse, le concept de préférences est compris sous un angle mentaliste. Ontologiquement, cela suppose qu'il existe une entité mentale à laquelle renvoie le concept de préférences, que cette entité entretient des relations avec d'autres états mentaux. Ce choix se justifie par l'objectif qui est le mien, prendre les préférences non seulement comme des causes, mais également comme des effets.

Sur le plan méthodologique, adopter une position mentaliste ne signifie en rien que nous devons laisser les données comportementales de côté. Car notre source d'accès la plus fiable aux états mentaux des agents demeurent leurs choix [30]. Le mentalisme adopté dans cette thèse trouve sa justification dans le fait qu'il est nécessaire de déterminer la base causale des préférences pour comprendre la façon dont elles changent. Il n'implique pas qu'en statique, nous ne pourrions pas nous limiter aux choix des agents pour avoir accès à leurs préférences. Il affirme simplement que si nous voulons comprendre leur évolution, nous devons leur assigner un contenu. Nous y reviendrons au premier chapitre de cette thèse à travers le principe d'efficacité dynamique des concepts de valeur et de conscience.

La conception générale de la rationalité sur laquelle elle repose est donc interne. Elle s'intéresse aux connexions à l'œuvre entre les états mentaux et s'interroge sur la cohérence de ces connexions. La cohérence n'est donc pas essentiellement celle des choix, mais bien des états mentaux de l'agent. L'étude de la rationalité axiologique ne peut se réduire aux choix. Elle s'interroge sur les liens entre états épistémiques et préférences de l'agent.

Dans ce cadre, pour comprendre ce qui est rationnel, il peut être utile de définir ce que j'entends par irrationalité. Est irrationnel un agent qui préfère a à b alors qu'*il a conscience* que tout le contenu mental dont il dispose pourrait lui indiquer que b doit être préféré à a

30. On peut aussi penser aux données déclaratives des agents. Toutefois, il est toujours délicat de conclure à partir de telles données. Comme le suggèrent de nombreuses expériences les agents ont tendance à rationaliser leur choix de façon impropre.

et qu'il *croit* que ce contenu lui indique que b doit être préféré à a. L'agent irrationnel est non seulement contradictoire, mais il a conscience de l'être. L'irrationalité, n'est pas une erreur ou l'effet d'un cause externe qui s'appliquerait sur l'agent :

> L'irrationalité est un échec au domicile de la raison elle-même. Quand Hobbes dit que seul l'homme a « le privilège de l'absurdité », il suggère que seul un être rationnel peut être irrationnel.
>
> [Davidson, 1980, p. 22]

Les erreurs de raisonnement et les biais cognitif ne sont pas à proprement parlé irrationnels. Mais ils ne sont pas rationnels non-plus. Ils reposent sur des causes non rationnelles, qui s'exercent en dehors de la conscience de l'agent. La rationalité, cependant, ne s'exerce jamais en dehors de ces causes non rationnelles. Elle surgit de causes externes. Lorsqu'il raisonne, l'agent utilise un langage, des données sensorielles, des conventions, des croyances tacites. Autant d'ingrédients qu'il n'est pas capable de remettre en question [31]. Bien que le contenu qui est associé à ses préférences, comme ses raisons, ses valeurs, ses motivations reposent parfois à l'arrière-plan de sa conscience, ils ont un effet sur son choix. Cela n'implique pas que son raisonnement soit incohérent. Est donc rationnel un agent qui fait au mieux compte tenu des ressources *de sa conscience*. Toutefois, cette définition est très générale. Nous parlerons donc désormais de *rationalité générale* pour nous y référer. Le chercheur en sciences sociales est à la recherche de rationalités spécifiques, assignant un contenu spécifique aux états mentaux de l'agent.

En résumé, est rationnel au sens général du terme un agent qui fait au mieux étant donné ce qu'il est capable de prendre en compte. Cette conception générale de la rationalité s'appuie sur une méthodologie mentaliste au sens où elle s'interroge sur la cohérence des états mentaux de l'agent et n'attribue pas de privilège à la cohérence de ces choix. Elle suppose qu'un changement de préférences s'opère lorsque l'agent révise ses états de conscience, et donc qu'il existe une connexion entre les états épistémiques de l'agent et ses préférences. Enfin, elle suggère que sa rationalité repose sur l'articulation de deux types de causes :

1. les causes externes, sur lesquelles l'agent ne peut, à un moment donné, rien changer,
2. les causes internes sur lesquelles il peut réfléchir et qu'il peut, par conséquent, modifier.

L'intervention de causes externes est compatible avec la rationalité générale de l'agent et c'est dans la complémentarité des causes internes et causes externes que se formule la rationalité spécifique d'un agent. Comprendre l'articulation entre ces deux types de causes est donc essentiel pour comprendre la rationalité des changements de préférences. Cela permettra de construire un mécanisme offrant une structure à l'ensemble des rationalités spécifiques possibles pouvant expliquer un changement de préférences. Pour cette raison

31. Voir les écrits sur la modularité de l'esprit chez Fodor [1983].

je parlerai de ces causes comme *les causes internes et les causes externes de la rationalité*, renvoyant dans ce cas à la rationalité spécifique que supposent les changements de préférences. Dans la section suivante, j'expose quelques-uns des mécanismes utilisés dans la littérature pour rendre compte des changements de préférences et montre leurs limites quant à l'articulation de ces deux types de causes.

3 Les mécanismes de transformation des préférences

Dans cette section, je m'intéresse à différents mécanismes employés par les économistes pour rendre compte des changements de préférences. Bien qu'ils s'inspirent de nombreuses disciplines, comme la biologie, la sociologie ou la psychologie, ces différents mécanismes peuvent s'analyser à la lueur de la difficulté conceptuelle qu'ils essaient de surmonter : articuler causes externes et causes internes de la rationalité. J'explicite donc la teneur qu'ils donnent à cette articulation, pour montrer en quoi elle se distingue d'un changement de préférences qui s'opère rationnellement comme celui que j'entends décrire dans cette thèse.

3.1 Evolution par adaptation : l'approche évolutionnaire

John Maynard Smith remarquait que, « paradoxalement, il s'avère que la théorie des jeux s'applique plus aisément à la biologie qu'au domaine du comportement économique, pour lequel elle avait pourtant été élaborée » [Smith, 1982, p. vii]. Ce faisant, il indiquait que le type de rationalité assignée aux agents économiques n'était pas nécessaire pour décrire la convergence d'une population vers un équilibre. La théorie des jeux s'applique parfaitement à la biologie puisqu'elle permet de restituer certaines intuitions de la sélection darwinienne. Le comportement des individus étant déterminé par leurs phénotypes, leur survie ne dépend que de leur adaptation à l'ensemble de la population.

Formellement, considérons une population avec deux phénotypes, a et b. On attribue à chaque individu un taux de reproduction, son *fitness* f, qui dépend de $x = (x_a, x_b)$, avec x_a qui représente la proportion d'individus de phénotype a et x_b, la proportion d'individus de phénotype b. L'évolution de la population est donnée par l'équation de réplication, qui consiste en l'équation différentielle suivante :

$$\dot{x}_i = x_i\Big(f_i(x) - (x_a f_a(x) + x_b f_b(x))\Big) \text{ avec } i \in \{a, b\} \tag{1}$$

Cette équation traduit l'idée que si le taux de reproduction d'un type de population, i, est plus grand que le taux de reproduction moyen de la population entière, alors la proportion d'individus caractérisés par ce phénotype devrait augmenter ($\dot{x}_i > 0$).

Ce qui vaut pour les phénotypes vaut également pour les stratégies des agents. Nul besoin de les voir comme résultant d'une rationalité démesurée. Il suffit de considérer la dynamique qui procède de la sélection naturelle pour comprendre comment leurs stratégies

(et donc leurs comportements) évoluent. Cette théorie permet d'expliquer comment des caractéristiques individuelles, comme l'altruisme [Fletcher and Zwick, 2007, Gintis et al., 2003] ou les comportements moraux [Alexander, 2009, Boehm, 1982], constituent des situations *évolutionnairement stables* Smith [1982] [32]. Le comportement ne résulte pas de préférences au sens où l'entend la théorie du choix rationnel. Ni la stratégie des individus, ni leur succès ne procèdent d'une délibération des agents sur leurs propres préférences. Ces derniers sont comme les particules du physicien et leur évolution est indépendante de ce dont ils ont conscience. Ils sont tout entièrement soumis à des causes externes et non rationnelles.

Bien qu'elle fournisse d'importantes intuitions sur l'évolution des sociétés humaines, la théorie des jeux évolutionnaires laisse de côté la question de la rationalité des changements de préférences. La question de la complémentarité entre causes externes et causes internes de la rationalité est laissée de côté au profit d'une hypertrophie de ses causes externes. Il n'est donc pas question d'un changement de préférences issu de la rationalité.

3.2 L'approche évolutionnaire indirecte

Les modèles dits d'évolution indirecte (*indirect evolutionary approach*) s'inspirent de l'approche évolutionnaire tout en tenant compte de cette difficulté : ils tentent de « combler le fossé entre rationalité et adaptation » [33]. Usant des concepts d'équilibre de la théorie des jeux évolutionnaire, le concept de préférences individuelles se substitue aux gènes : les préférences déterminent le comportement qui, plus ou moins adapté à l'environnement de l'agent, se reproduit selon certaines fréquences : seules les préférences les plus efficaces relativement à la population présente survivent à long terme. Les préférences sont donc sélectionnées, mais les agents, à chaque instant, prennent des décisions rationnelles au sens où ils anticipent les choix de leurs semblables pour adopter un comportement d'équilibre. Cela permet à l'approche évolutionnaire indirecte de distinguer « les dispositions qui poussent l'agent de derrière » et « ses projections dans le futur ».

> Dans une large mesure, expériences et évènements du passé affectent les choix des individus. Bien qu'elle attribue des anticipations aux agents, la théorie du choix rationnel doit incorporer le fait que le comportement ne résulte pas exclusivement de ce qui tire l'agent vers son futur, et est également poussé par derrière.
>
> [Güth and Kliemt, 1998, p. 378]

L'objectif même de l'approche évolutionnaire indirecte consiste donc à combiner causes externes et causes internes de la rationalité. Toutefois, elle ne permet pas de décrire la façon dont un agent délibère intérieurement pour changer ses préférences. Le type de

32. Une stratégie est évolutionnairement stable si, lorsqu'elle est adoptée par l'ensemble de la population, cette dernière n'est pas susceptible d'être envahie par une (petite) population de mutants, adoptant une stratégie différente.

33. Cette formule est d'ailleurs le sous-titre de l'article de Güth and Kliemt [1998].

changement de préférences n'est pas causé par des causes internes à l'individu, mais par un processus de sélection qui s'apparente à celui de la théorie des jeux évolutionnaires. Il s'agit donc d'une approche intéressante, mais qui ne permet pas de remplir la mission que nous nous sommes fixés : décrire un changement de préférences qui procède de la *délibération*. Elle ne permet pas non plus de comprendre par quels processus un individu peut chercher à s'appuyer sur la rationalité de ses semblables pour changer ses préférences.

3.3 Les transmissions culturelles

S'inspirant du modèle des anthropologues Cavalli-Sforza and Feldman [n.d.] et Bisin and Verdier [2001*b*, 2004] expliquent la transformation des préférences à l'aide d'un mécanisme de transmissions culturelles. Dans ce modèle, les préférences sont reflétées par une caractéristique culturelle (*cultural trait*) [34], susceptible de renvoyer tout aussi bien à une certaine « éthique du travail », à une « attitude politique » ou à un goût pour les études qui se transmettent de génération en génération [35]. La transmission se fait par le canal de l'éducation ; les parents de chaque génération investissent sur l'éducation de leur enfant afin d'augmenter la probabilité qu'ils adoptent leur propre trait culturel (socialisation verticale). S'ils ne parviennent pas à l'influencer, la société s'en charge (socialisation oblique) et, à l'enfant, est associée la caractéristique culturelle d'un individu pris au hasard. Le modèle repose sur l'arbitrage que les parents doivent effectuer entre ces deux formes de socialisation. Si elles sont substituables, les minorités ont intérêt à concéder un coût à l'éducation de leur progéniture et le modèle abouti à une société hétérogène sur le plan culturel ; sinon elles n'y ont pas intérêt et la majorité l'emporte. Le modèle des transmissions culturelles repose donc sur l'idée que les parents, soucieux de transmettre leur caractéristique culturelle à leur enfant, peuvent espérer que la société, si elle se compose d'une proportion importante d'individus semblables, fasse le travail d'éducation à leur place.

Bien qu'ils ne fassent pas preuve d'égoïsme [36], les parents sont conçus comme parfaitement rationnels. A l'inverse, les enfants sont parfaitement passifs dans le processus éducatif. Ils sont pris en tenaille entre deux formes de socialisation qui les dépassent et sur lesquelles ils n'ont aucun contrôle. Le modèle repose sur la relation unidirectionnelle de ces deux types d'agent, l'un parfaitement rationnel, l'autre passif. L'articulation entre causes externes et causes internes de la rationalité correspond à ce partage. Or certains anthropologues, spécialisés dans l'étude des transmissions culturelles, suggèrent que les

34. Les analogies entre biologie et culture sont extrêmement répandues dans ces travaux, comme dans ceux des anthropologues et sociologues qui leur servent d'inspiration. Rappelons que le concept de « même » s'inspire de celui de gêne mais pour s'appliquer à la culture.

35. Le concept de trait culturel renvoie donc tout aussi bien à des attitudes, des valeurs qu'à des préférences.

36. Verdier et Bisin justifient l'intérêt que les parents portent à leurs enfants par un mécanisme d'« empathie imparfaite ». Les parents veulent le meilleur pour leurs enfants, mais ils envisagent le meilleur à travers leurs propres préférences, sans préjuger du bonheur réel que procurerait à leurs enfants l'adoption d'un trait différent du leur.

transmissions se font selon un schéma davantage bidirectionnel.

> Parents et enfant sont susceptibles d'avoir des besoins et des objectifs séparés, ce qui conduit à des tensions [...]. Ces tensions peuvent être porteuses de nouveaux sens qui, en des termes dialectiques, génèrent des synthèses, susceptibles de résoudre temporairement ces contradictions.
>
> [De Mol et al., 2013, p. 9]

De l'aveu des anthropologues, la construction par l'enfant de ses préférences résulte bien de l'effet d'une forme de rationalité. Cette rationalité suppose l'intervention de causes externes que l'enfant ne maîtrise pas (apprentissage de langage et de mode de vie particuliers) mais également de causes internes. Se projetant sur son existence, l'enfant essaie de lui donner son propre sens. C'est pourquoi un modèle faisant cohabiter les causes internes et les causes externes de la rationalité serait à même d'offrir un fondement théorique au modèle des transmissions culturelles [37].

3.4 La formation des habitudes

Comme l'écrit Gorman [1967], il est trivial de faire remarquer que les « choix dépendent des goûts, tout autant que les goûts dépendent des choix ». Autrement dit, l'habitude joue un rôle considérable dans la formation des préférences. Dans l'*Ethique à Nicomaque*, Aristote faisait d'ailleurs de l'habitude la condition par laquelle l'homme est en mesure de faire preuve de vertu pratique. Généralement, l'effet de l'habitude sur les préférences est compris comme l'intervention d'une cause externe à l'individu. Elle s'explique par les théories behavioristes de l'apprentissage, par renforcement, par des réflexes pavloviens. Elles peuvent se donner comme une modification physiologique qui précède la conscience : à mesure qu'un agent fume, son cerveau lui envoie des signaux qui manifestent son désir de nicotine.

von Weizsäcker [1971] et Pollak [1978], deux contributeurs pionniers dans l'analyse des préférences endogènes, font de la formation des habitudes un des concepts centraux de leur analyse. Le problème qu'ils essaient de résoudre consiste à rationaliser une fonction de demande de long terme, notamment pour être en mesure d'effectuer des évaluations en termes de bien-être dans un contexte où les préférences changent. Pour formaliser cette fonction, von Weizsäcker [1971] propose un modèle à deux biens qui, comme cela est montré par El-Safty [1976], est susceptible de s'étendre à n biens si et seulement si les agents sont myopes, c'est-à-dire totalement incapables d'anticiper les effets de la formation de ses habitudes. Dans ce contexte, non seulement les changements de préférences sont traités comme purement non rationnels mais, de plus, les agents n'ont pas la capacité d'anticiper ces changements [38].

37. Un tel modèle pourrait par exemple permettre d'axiomatiser la probabilité de transmission par l'adulte de son trait culturel à l'enfant.

38. Pollak [1978] fait remarquer que, dans ces conditions, l'analyse en termes de bien être est quasiment impossible.

Cette capacité est octroyée à l'agent dans le modèle de Becker [1996] qui intègre le concept d'habitudes et propose qu'additionnellement à la consommation de l'agent, x_t, soit ajouté à sa fonction d'utilité un paramètre représentant son capital personnel, P_t :

$$U = U(x_t, P_t)$$

Le capital personnel de l'agent est déterminé par ses choix passés. Il reflète l'impact des habitudes de consommation de l'agent sur ses préférences. L'agent, à chaque période, choisit d'investir dans ce type de capital, sachant qu'un tel investissement aura un effet sur ses préférences futures. Son choix dépend aussi bien de son utilité présente que de celle à venir. Becker fait l'hypothèse que « les agents anticipateurs sont conscients que leur choix présent et leur expérience affectera leur capital personnel futur, et que ce capital affectera directement leur utilité. Par conséquent, le choix actuel dépend non seulement de la façon dont il affecte l'utilité, mais également de la façon dont il affectera l'utilité future de l'agent » [Becker, 1996, p. 9]. Mais cet investissement se déprécie d'une période à l'autre. Dès lors, l'évolution du capital personnel est donnée par l'équation suivante :

$$P_{t+1} = x_{t+1} + (1 - d_p)P_t$$

où d_p est un taux constant de dépréciation du capital personnel. L'investissement en capital permet non seulement son accumulation, mais il change également la consommation des biens complémentaires à son accumulation. Le concept de capital personnel permet de penser un agent qui, par sa consommation, travaille son regard et valorise les biens qu'il consomme à l'accoutumée. La notion de complémentarité entre bien de consommation et investissement est fondamentale. Elle permet, selon Becker, de rendre compte d'une panoplie conséquente de comportements.

> Les phénomènes de complémentarité et de renforcement, au sein du comportement fondé sur des habitudes, expliquent, par exemple, pourquoi l'envie de fumer est plus importante pour une personne fumant depuis des années, pourquoi manger des *corn flakes* régulièrement pour le petit déjeuner augmente la demande future de cette céréale, pourquoi dire des mensonges et agir avec violence augmentent la tendance des agents à dire des mensonges et faire preuve de violence, pourquoi il devient habituel d'épargner, même pour des individus dont les jours sont comptés, pourquoi grandir dans une famille religieuse affecte largement la probabilité de devenir une personne religieuse une fois adulte, ou pourquoi vivre avec une femme des années durant induit une telle dépendance chez son mari que ses capacités physiques s'éteignent à la mort de cette dernière.
>
> [Becker, 1996, p. 8]

Le modèle de Becker est très séduisant. Le problème est que son *behaviorisme* notable

le pousse à confondre ce qui pousse l'agent à reproduire le même comportement, et sa propension à changer lui-même ses caractéristiques. Il en découle qu'il est difficile de distinguer ce qui dans le processus de détermination des préférences de l'agent relève, d'une part, de causes externes à sa conscience, de ses croyances tacites ou de ses déterminations biologiques et, d'autre part, de causes internes à sa conscience, comme son désir de changer de préférences. Becker confond ces deux mécanismes.

On pourrait cependant rétorquer à cette critique que les deux déterminants de l'évolution des préférences de l'agent sont constitués par son stock de capital personnel initial et son taux d'escompte [39], le premier constituant l'inertie du passé, quand le second est une incapacité à se projeter dans l'avenir (un défaut d' « imagination » [40]). Si cette interprétation est tout à fait recevable, il n'en demeure pas moins que l'on considère habituellement le taux d'escompte comme relevant des préférences de l'agent. Conçu comme stable dans le modèle, il ne serait pas capable d'être modifié par l'investissement en capital personnel de l'agent. Et, bien qu'il nous suggère le contraire, Becker ne propose pas de modéliser comment ce taux d'escompte pourrait relever de l'imagination ou de l'*awareness* [41]. Son modèle ne s'appuie donc pas sur une conceptualisation rigoureuse des états épistémiques de l'agent. Dans ces conditions, il peut paraître héroique de dire que le modèle de l'*homo economicus* à la Becker permet de clarifier la formation des préférences [42], alors qu'à vouloir imposer une interprétation préconstruite de la rationalité, il crée beaucoup de confusions. Mon approche est différente en ce qu'elle tente de reconstruire un concept de rationalité qui permette de clarifier les changements de préférences.

De plus, son modèle ne permet pas de dire en quoi la capacité d'un individu à s'affranchir de ses conditionnements est limitée. Pourtant, il est possible que les changements de préférences soient plus erratiques, lorsqu'affectés par un prise de conscience. Becker le souligne lui-même, la consommation de tabac a fortement diminué lorsque se sont accumulées les preuves de sa nocivité. Mais comment son modèle rend-il compte d'une telle prise de conscience si les agents anticipent parfaitement ce qui est à l'œuvre ? En définitive, le modèle de Becker formule une métaphore habile et très parcimonieuse des changements de préférences. Néanmoins, pour ce qui est du capital personnel, son modèle ne permet

39. A ce stade, ce sont les seuls manières susceptibles de distinguer deux agents dans le modèle de Becker.

40. Becker parle bien d'imagination à plusieurs reprises : "Imagination capital not only affects the discount on future utility, but it also alters preference over goods by affecting present and future choices."

41. "They may choose greater education in part because it tends to improve the appreciation of the future, and thereby reduces the discount on the future. Parents teach their children to be more aware of the future consequences of their choices (Akabayashi, 1995, studies the conflict between parents and children over the weight attached to the future). Addictions to drugs and alcohol reduce utility partly through decreasing the capacity to anticipate future consequences. Religion often increases the weight attached to future utilities, especially when it promises an attractive afterlife. Imagination capital not only affects the discount on future utility, but it also alters preference over goods by affecting present and future choices. Someone who places greater weight on the future consequences of current choices is more likely to engage in activities that raise future utilities, perhaps partly at the expense of current utility".p. 11

42. Dans *Accounting for taste*, il soutient que le défaut des théories anthropologiques et sociologiques tient au fait qu'elles ne munissent pas leur analyse d'un « cadre analytique puissant » [Becker, 1996, p. 1996]

pas de comprendre le type de rationalité à l'œuvre dans un changement de préférences.

D'autres auteurs ont alors insisté sur la capacité de l'agent à contrôler davantage l'effet de l'habitude. Gul and Pesendorfer [2001, 2005] proposent un modèle à la Kreps [1979], et construisent la représentation d'une utilité où l'agent choisit non seulement un sentier de consommation mais également les choix qu'il aura à sa disposition. De même, Rozen [2010] fournit un théorème de représentation de l'utilité d'un agent pour qui l'historique des choix devient un point de référence (*reference point*) à partir duquel penser ses choix futurs. Dans chacun de ces cas, si la rationalité joue un rôle dans le processus de formation des habitudes, c'est en tant que l'agent sait (ou non) que ses préférences changent avec les habitudes qu'il prend. Celui qui refuse de prendre de l'héroïne une fois dans sa vie, par exemple, anticipe que ses préférences pourraient changer et qu'il pourrait préférer, suite à une première prise, désirer en prendre davantage ultérieurement. Autrement dit, si la rationalité est à l'œuvre dans les changements de préférences qui découlent du processus d'habituation, c'est en tant que l'agent tente de maitriser ce changement sans en être le moteur. En tant que telle, l'habituation induit un changement de préférences non rationnel. La rationalité n'y est pas le moteur du changement de préférences, elle consiste à contrôler ce changement qui lui est exogène.

3.5 Résumé des chapitres

La brève revue des mécanismes de transformation des préférences dans la théorie économique suggère qu'il est absolument essentiel de construire un mécanisme qui permette de prendre au sérieux une rationalité spécifique des changements de préférences et qui clarifie le type de complémentarité qui existe entre causes externes et causes internes de la rationalité dans les changements de préférences. Cette thèse formalise un tel mécanisme de transformation des préférences : *la délibération partielle*.

C'est un mécanisme qui donne un rôle primodial à la conscience des agents. Autrement dit, la transition que l'agent y effectue d'une relation de préférences à une autre dépend de sa conscience et de sa capacité à délibérer sur ce qui en détermine le contenu. La délibération partielle repose donc sur des intuitions que l'on retrouvent dans la psychologie de sens commun : les individus peuvent changer rationnellement ce qui détermine leurs préférences, dès lors qu'ils délibèrent consciemmment sur ces dernières. Le premier chapitre est conceptuel : il expose ces intuitions en s'appuyant sur un ensemble de cinq hypothèses dont les ressorts philosophiques sont décortiqués. Le second est axiomatique : il définit formellement ces intuitions et les caractérise à l'aide de six axiomes portant sur la règle de transition d'une relation de préférences à une autre. Ce faisant, il permet d'exposer les types de changements de préférences que la délibération partielle rend rationnels et, surtout, ceux qu'elle rend irrationnels. Or, si certains changements de préférences peuvent être exclus par la théorie du fait de leur irrationalité, expliquer n'importe quelle situation par un changement de préférences devient impossible. Par ce résultat, cette thèse offre

donc une réponse au problème de l'*adhocité* des changements de préférences. Enfin, le troisième chapitre propose une application du modèle qui jette les bases d'une théorie de la manipulation des préférences.

Cette thèse contribue ainsi à clarifier la rationalité à l'œuvre dans les changements de préférences.

Chapitre 1 :

Le premier chapitre consiste en une étude conceptuelle de la délibération partielle. Après une présentation succincte des rudiments formels du modèle, sont développées et justifiées les cinq hypothèses conceptuelles sur lesquelles repose ce mécanisme.

— Selon la première hypothèse, les préférences sont induites par des valeurs. Les valeurs causent les préférences, c'est-à-dire que les préférences d'un individu sont entièrement déterminées par ses valeurs, et qu'un changement de préférences doit donc passer par un changement dans l'ensemble de ses valeurs de l'individu. Pour formaliser cette hypothèse, je m'inspire du modèle fondé sur les raisons, proposé par Dietrich and List [2013*b*]. Cette définition permet de résoudre le problème évoqué à la section 2.1 de cette introduction, et donne aux préférences le contenu substantiel nécessaire à leur transformation (cf section 2.2).

— Selon la deuxième hypothèse, les valeurs forment des systèmes susceptibles d'être hiérarchisés par leur degré de cohérence axiologique. L'existence d'une telle hiérarchisation est justifiée par les contradictions ou bien, au contraire, la relation de complémentarité que peuvent entretenir certaines valeurs. Elle implique l'existence d'une relation $\trianglelefteq$ permettant d'ordonner les ensembles de valeurs à l'aune d'un critère axiologique[43]. Les conséquences de certaines des propriétés de cette relation sont précisées dans le chapitre 2 de la thèse.

— Selon la troisième hypothèse, certaines valeurs ont un statut particulier : elles peuvent être conscientisées par l'agent. Les valeurs dont l'agent a conscience sont nommées « raisons ». Je défends l'idée que ces raisons sont des valeurs sur lesquelles l'agent peut faire un retour introspectif. C'est-à-dire que l'agent est capable de se demander si ces valeurs valent la peine qu'on y adhère. Je défends ensuite l'idée que ce concept de conscience est nécessaire dans un cadre dynamique.

— Selon la quatrième hypothèse, l'agent est capable de modifier l'importance qu'il assigne à une valeur si et seulement si il en a conscience. C'est donc bien par une prise de conscience que l'agent change ses préférences.

— Selon la cinquième hypothèse, l'agent choisit effectivement parmi ses raisons celles qui maximisent la cohérence de ses préférences, étant donné l'ensemble des valeurs

43. L'usage d'une relation d'ordre sur les ensembles et la façon dont elle peut être induite par une relation sur les élements qui composent ces ensembles a été étudié dans le cadre de la littérature visant à caractériser la liberté. Pour une revue de ces travaux on pourra consulter [Barbera, Bossert and Pattanaik, 2004] ou [Baujard, 2007].

dont il n'a pas conscience et dont il ne peut changer le statut motivationnel. On suggère qu'une analyse des changements de préférences rationnels repose sur une extension du concept de rationalité au concept d'autonomie : l'individu dont les changements de préférences sont rationnels est autonome au sens où il est le moteur des modifications de ses valeurs. Il l'est de façon locale, cependant, car la maîtrise de ses propres valeurs, se limite à l'ensemble des valeurs dont il a conscience.

De ces cinq hypothèses, on déduit que la délibération partielle donne lieu à un changement de préférences rationnel, puisqu'il résulte de la prise de conscience, par l'individu, de certaines des valeurs qui causent son choix. Toutefois, la délibération partielle concède à la thèse déterministe que le pouvoir d'autodétermination des agents est contraint par des valeurs implicites, dont ils n'ont pas la maitrise. Ainsi, se disposent les causes internes et les causes externes de la rationalité au sein du mécanisme de délibération partielle.

Chapitre 2 :

Le second chapitre déploie les ingrédients formels du modèle afin de s'interroger sur la structure de la délibération partielle et montrer le type de changement de préférences qu'une telle théorie autorise. A l'instar de Dietrich et List (2013), je définis une relation de préférences sur des alternatives $\prec_V$ induite par un ensemble V^i de valeurs auxquelles l'agent adhère. Aussi la délibération partielle porte-t-elle sur la règle permettant le passage de l'ensemble de V à V'. Notée $\rightarrow_A$, cette règle dépend d'un ensemble de valeurs, noté A, dont l'agent a conscience en période 1. Dès lors, l'étude de cette règle se divise en quatre parties :

1. La première partie caractérise à l'aide de six axiomes la règle de passage, $\rightarrow_A$, afin qu'elle puisse être représentée par une unique relation de cohérence axiologique $\trianglelefteq$, qui vérifie les hypothèses conceptuelles de la délibération partielle évoquées au chapitre 1, i.e.

$$V \rightarrow_A V' \iff \begin{cases} \exists B \subseteq A, V' = B \cup (V \backslash A) \\ \forall B' \subseteq A, B' \cup (V \backslash A) \trianglelefteq V' \end{cases} \tag{2}$$

   Ce résultat permet d'établir les conditions à partir desquelles la règle de passage, $\rightarrow_A$, est rationnelle au sens général que j'ai donné à ce concept dans la section 2 de cette introduction.

2. La seconde partie relie cette dynamique des systèmes de valeurs avec les changements de comportement. Elle suggère deux choses. D'abord, des restrictions supplémentaires sur le lien entre options de choix et valeurs doivent être mobilisées pour pouvoir obtenir une révélation des changements de préférences à partir des changements de choix. Ensuite, elle suggère que l'esprit de la délibération partielle contraste avec l'analyse des changements de préférences en termes de dissonnance cognitive.

La primauté est donné au choix des valeurs dans la délibération partielle, tandis qu'avec l'analyse fondée sur la dissonance cognitive le choix des options est premier.

3. La troisième partie porte sur la structure de certaines rationalités spécifiques, et s'intéresse à l'arbitrage évoqué dans cette introduction (et au chapitre 1) entre d'une part la nécessité d'avoir une règle de passage qui détermine un changement de préférences unique au sens où le couple (A, V) donne un unique V', et d'autre part de ne pas contraindre démesurément l'ensemble des préférences dont l'agent peut être porteur au gré du processus de délibération partielle. Pour traiter ce problème, je spécifie deux différentes formes de relation, chacune renvoyant à une rationalité spécifique.

   La première est monotone et stipule que les valeurs sont intrinsèques, i.e. que leur adoption par un agent est indépendante du système de valeur initial de cet agent. Il en résulte que la règle de passage détermine un ensemble unique. Mais elle rend impossible, à valeur initiale égale et niveaux de conscience différents, que deux agents puissent avoir des préférences similaires. Surtout, elle impose que l'agent soit contraint d'adhérer aux valeurs dont il a conscience. La seconde structure l'ensemble des valeurs de sorte que les oppositions ou complémentarité forme une partition. Les valeurs complémentaires sont regroupées entre-elles dans les éléments de cette partition et elles s'opposent systématiquement aux valeurs qui appartiennent à d'autres éléments de la partition. Sous certaines hypothèses, on montre que la règle de passage est bien déterminée, mais qu'elle suppose qu'un agent ne puisse adhérer à des valeurs qui s'opposent, s'il a conscience de cette opposition.

4. Enfin, je propose deux applications de la délibération partielle en m'appuyant sur les structures développées précédemment. Je formalise l'*endowment effect* et le renforcement des comportements addictifs.

Chapitre 3 :

Dans le dernier chapitre de cette thèse, je formalise une situation de manipulation. Un envoyeur doit choisir simultanément le projet qu'il souhaite conduire et le type de valeur qu'il doit révéler afin de changer les préférences d'un receveur. L'exemple canonique sur lequel je m'appuie est celui d'un conseiller en fiscalité qui propose un projet et tente de changer les préférences d'un décideur politique. Dans l'analyse de cette situation, j'aborde trois questions :

— En premier lieu, je développe un cadre alternatif de la délibération partielle en exhibant le lien qu'entretient ce cadre avec le modèle développé dans le chapitre 2. Dans ce cadre, les agents assignent maintenant une intensité axiologique aux valeurs. C'est-à-dire qu'ils adhèrent à une valeur avec une certaine intensité. Exprimé en ces termes, le modèle permet d'utiliser un concept de distance entre les préférences des agents et le projet proposé.

— En second lieu, et à l'appui de ce nouveau cadre, je modélise le concept d'empathie imparfaite, i.e., le fait que l'envoyeur puisse être incertain de la façon dont le receveur est suseptible de réagir à sa stratégie de divulgation de nouvelles valeurs. J'étudie séparément cette situation lorsque les valeurs sont axiologiquement dépendentes et lorsqu'elles indépendentes. Je montre que l'indépendence implique que l'envoyeur peut simplement de révélés les valeurs en les considérants une à une. A l'inverse, la dépendence implique que l'envoyeur peut avoir intérêt *en soi* à révéler l'une des deux valeurs .

— En troisième lieu, je m'intéresse à l'effet d'une propriété intéressante de la délibération partielle dans ce cadre : le fait qu'une divulgation séquentielle des valeurs n'induit pas nécessairement le même changement de préférences qu'une divulgation simultanée de ces valeurs. Je caractérise les situations où l'envoyeur peut tenter de faire usage d'une telle politique de divulgation séquentielle.

Chapitre 1

LES CINQ HYPOTHÈSES DE LA DÉLIBÉRATION PARTIELLE

Les premières réflexions conduites dans l'introduction, suggèrent que s'interroger sur les causes de l'évolution des préférences suppose de se positionner par rapport à deux grilles de lecture distinctes. La première consiste à attribuer la responsabilité du changement à l'exercice de causes externes, sur laquelle l'agent n'a pas de prise. Avec cette *conception externe* des changements de préférences, les individus sont hétérodéterminés. La second, à l'inverse, consiste à placer la cause du changement dans la conscience de l'agent. Avec cette *conception interne* des changements de préférences, les agents sont capables de s'autodéterminer. Traditionnellement ces deux approches ont donné naissance à des écoles qui s'opposent.

Chacune a ses avantages, chacune ses inconvénients. La conception externe permet aisément de décrire un changement de comportement. Mais elle ne restitue pas la psychologie qui est à l'oeuvre dans ce changement. Le changement a lieu, mais à l'intérieur d'une « boite noire ». La conception interne présuppose de l'agent qu'il choisisse lui-même ses déterminations du mieux qu'il le peut, en utilisant les capacités réflexives qui lui sont offertes par sa conscience. Elle traduit la « platitude de la psychologie de sens commun » qui veut que la conduite d'un agent soit déterminée par ses capacités délibératives [Pettit, 1991]. Sa difficulté est qu'elle parvient difficilement à expliquer pourquoi l'agent devrait changer ses préférences dans ces conditions. Car s'il est capable changer ses préférences comme il le souhaite, pourquoi l'agent ne choisirait-il pas à tout instant les meilleures préférences qui soient.

Dans une certaine mesure, ces approches ne sont pas incompatibles et, comme nous l'avons déjà vu à la fin de l'introduction générale de cette thèse, certaines tentatives de conciliation ont déjà été conduites dans divers champs des sciences sociales. J'ai montré en introduction que, cependant, ces conciliations ne permettent pas de penser comment la

conjonction de causes externes et de causes internes est susceptible de fournir en elle-même une cause aux changements de préférences. Dans ce chapitre, j'entame la fondation d'un mécanisme de changement de préférences qui permet de combler ce manque.

L'idée selon laquelle changer les préférences de nos semblables nécessite parfois d'attirer leur attention sur des éléments neufs, qu'ils avaient jusqu'alors sincèrement ignorés, est une platitude de la psychologie de sens commun. C'est le vendeur qui tente d'amadouer son client en soulignant les propriétés avantageuses de son produit, le politique qui discute la vision portée par son programme [1], les amis qui échangent leurs vues sur la situation du monde, les parents qui tentent d'inculquer une éthique à leur progéniture ou les amateurs de théâtre qui s'écharpent sur les règles incontournables de la représentation dramatique.

Prenons la bataille d'*Hernani*. Soucieux de défendre une vision rénovée de la représentation théâtrale, les acteurs de cette querelle ne se contentent pas d'invectives, ils fournissent des textes théoriques pour défendre la pièce de Victor Hugo et justifier leur vision de l'art. Fussent-ils dans l'erreur, ils se comportent comme si le fameux *de gustibus non est disputandum* devait se suspendre [2], comme si un changement de préférences pouvait être *engendré par une prise de conscience* (*growing awareness*). Qu'est-ce à dire ? Pour préciser l'idée que j'ai en tête, deux remarques s'imposent.

D'abord, par prise de conscience, j'entends le fait que l'agent se trouve *subitement* capable de concevoir de nouvelles données dont il était, jusqu'alors, incapable d'imaginer l'existence même. Il n'y a donc aucune raison *a priori* qu'un agent soit susceptible d'anticiper un changement de préférences dès lors qu'il est engendré par une prise de conscience. Il ne s'agit pas, à l'instar de Strotz [1955], Kreps [1979] ou encore Gul and Pesendorfer [2005], d'interroger la capacité qu'il aurait à s'engager sur un sentier de consommation, exerçant un contrôle de soi, par la contrainte imposée sur ses choix futurs [3]. Bien davantage, mon approche présuppose que l'agent ne s'imagine pas changer de préférences et que, partant, il n'ait pas à se soucier de leurs éventuelles altérations. Il ne s'agit donc pas non plus de dire que l'agent *découvre* ses préférences par la consommation comme cela a

1. Certains mouvements politiques sont même fondés sur cette idée, comme par exemple les mouvements écologiques qui mettent en oeuvre des campagnes d'éveil des consciences (*awareness campaign*) ou les mouvements féministes comme *raising awareness* ou la *Vegan awareness fundation*.

2. Dans *Ainsi parlait Zarathoustra* Nietzsche écrit que « des goûts et des couleurs on ne discute pas, et pourtant on ne fait que ça ». Pourtant, rien ne nous certifie que ce type de conversation mène à un changement de préférences. Peut être que l'enjeu pour les protagonistes de la bataille d'*Hernani* consiste moins à tenter d'influencer les préférences de l'autre qu'à faire état de son identité. Sans nier l'intérêt de cette dernière possibilité, je postule dans cette thèse qu'il est possible que ce genre de conversation conduise à un changement de préférences.

3. Etant donné que mon concept de conscience se rapproche de l'*unawareness*, il fournit un concept satisfaisant de l'incapacité à anticiper. Il en résulte que l'évolution de cette variable doit donner lieu à des changements de préférences non anticipés, type de changement que je crois être extrêmement répandu. On peut par exemple mentionner ceux qui interviennent chez ces migrants qui, simplement à la recherche d'un travail au moment de leur immigration, s'accoutument aux pratiques culturelles du pays d'accueil et y prennent goût sans l'avoir prévu. Sur cette question de l'anticipation des changements de préférences, on pourra se référer à [Kahneman and Snell, 1992]. *A contrario*, l'objectif de [Cyert and DeGroot, 1975], ainsi que de la littérature à laquelle ils ont donné naissance, est assez différent du mien. Dans leur modèle les agents sont incertains des préférences qu'ils auront et essaient de les anticiper.

été très récemment suggéré Delaney, Jacobson and Moenig [2017]. Mon objectif consiste plutôt à expliquer comment ses préférences deviennent ce qu'elles sont, comment elles se construisent [Slovic, 1995] [4].

Ensuite, l'idée qu'un changement de préférences puisse être engendré par une prise de conscience est susceptible de renvoyer à deux mécanismes distincts :

— Par le premier, l'agent prend conscience qu'une alternative sur laquelle pourrait porter ses préférences, *Hernani* par exemple, est caractérisée par des propriétés conformes à son système de valeurs. La préférence de l'agent change, localement, car seule la valeur qu'il accorde à *Hernani* est modifiée, les autres oeuvres gardant la même valeur à ses yeux [5]. Cette prise de conscience change sa perception de l'objet, qui revêt un intérêt nouveau, mais laisse inaltéré son système de valeurs.

— Par le second, *a contrario*, l'agent prend conscience de nouvelles sources de valorisation qui le poussent à modifier son système de valeurs. Le changement perceptif instauré par la prise de conscience ne porte plus sur un objet particulier, mais sur l'ensemble des valeurs auxquelles l'agent adhère, et doit se traduire par une reconfiguration générale de ses préférences. Ce n'est pas seulement la valeur qu'il accorde à *Hernani* qui change, mais c'est également la valeur qu'il accorde à d'autres objets de choix, c'est-à-dire à d'autres oeuvres.

Dans le premier cas, l'agent change de point de vue sur la conformité de l'objet de choix à son système de valeurs, V, qui reste inchangé ; dans le second cas, il change de point de vue sur son propre système de valeurs, qui doit alors se transformer, passant de V à V'.

Si le premier mécanisme est parfaitement illustré par le vendeur dépeignant les caractéristiques de son produit, c'est le second qui est à l'oeuvre dans la bataille d'*Hernani*. Car cette querelle ne saurait être cantonnée à la défense du texte de Victor Hugo, ni même à une vision particulière de l'art : elle porte sur la façon dont la beauté s'articule avec le monde, sur ce qui doit être valorisé en politique [6], la mécanique de l'histoire, la condition humaine. Ce faisant, elle invite son spectateur à réformer l'ensemble de son système de valeurs et non pas seulement sa préférence pour l'objet particulier que serait *Hernani* [7]. Il n'est pas inconcevable que ces deux mécanismes puissent être à l'oeuvre de concert, mais ils sont conceptuellement distincts et méritent un traitement séparé. Ce chapitre, ainsi que l'ensemble de cette thèse, consistent à analyser dans les termes de la théorie du choix rationnel le second de ces mécanismes.

4. Cette idée que les préférences se construisent est très répandue parmi les psychologues. [Tversky, Slovic and Kahneman, 1990, p. 489] écrivent par exemple « [V]alues or preferences are commonly constructed in the process of elicitation (they ain't nothing till I call them) ».

5. J'emploie le terme de « préférence » au singulier ici pour bien souligner le caractère local du changement de préférences.

6. Il est courant d'identifier l'oeuvre avec une défense du libéralisme. Aussi Victor Hugo nous parle-t-il dans la *Préface de Cromwell* (1827) de « cette élite de jeunes hommes, intelligente, logique, conséquente, vraiment libérale en littérature comme en politique, noble génération qui ne se refuse pas à ouvrir les deux yeux à la vérité et à recevoir la lumière des deux côtés »

7. Pour la même raison qu'à la note de bas de page 5, je mets « préférence » au singulier.

Une telle entreprise n'a rien d'immédiat. Elle suppose de connecter les concepts conatifs de préférences et de système de valeurs, au concept cognitif de conscience. Aussi nous invite-t-elle à clarifier l'interprétation de ces nouveaux concepts, jusqu'alors inemployés dans le corps standard de la théorie du choix rationnel ; à interroger la relation qu'ils entretiennent avec les préférences et le rôle qu'ils jouent dans leur transformation. Ce chapitre poursuit donc trois objectifs :

1. D'abord, il fournit la présentation conceptuelle de ce mécanisme que je qualifie de *transformation des préférences par délibération partielle sur les valeurs* (pour simplifier dans la suite du texte j'emploie le terme de *délibération partielle*).
2. Ensuite, il suggère que l'usage de ces concepts constitue une extension naturelle de la théorie du choix rationnel, et discute les antécédents de cet usage au sein de la discipline.
3. Enfin, il propose de justifier l'ajout de ces concepts à l'ontologie de la théorie du choix rationnel et s'appuie, pour ce faire, sur un principe d'efficacité dynamique selon lequel les concepts de valeur et de conscience revêtent une importance cruciale dans l'étude *dynamique* des préférences.

A ma connaissance, la délibération partielle n'a pas été étudiée jusqu'à présent dans la littérature sur les changements de préférences.

Elle porte sur les *valeurs* en ce que celles-ci induisent les préférences de l'agent.

Elle est *délibération* car elle procède d'un changement délibéré qui, contrairement aux changements qu'Hirschman [1984] qualifie d'impulsifs (*wanton*), suppose de l'agent qu'il fasse preuve de réflexivité sur ce qui motive ses choix.

Elle est *partielle* parce que le domaine d'exercice de ces capacités délibératives est restreint, susceptible de varier. Cela signifie que l'agent ne peut rationnellement abandonner que *certaines* de ses valeurs et donc cultiver des préférences différentes de ses semblables.

Il en découle, et c'est le principal résultat de ce chapitre, qu'avec la délibération partielle, l'agent change ses préférences *parce qu'il* est ni totalement actif, ni totalement passif le processus par lequel ses préférences sont déterminées. Autrement dit, le type de changement de préférences décrit par la délibération partielle n'est rendu possible que par la conjonction de causes externes et de causes internes.

Pour donner sens à ces ingrédients conceptuels, nous avons besoin de cinq hypothèses qui permettront de clarifier le processus de changement de préférences auquel donne lieu la délibération partielle. Alors que la section 1 résume le mécanisme de délibération partielle, en s'appuyant sur des définitions provisoires, les sections suivantes approfondissent ce travail de définition et exhibent chacunes des hypothèses. La section 2 avance l'idée que les préférences sont induites par des valeurs. La section 3 ajoute que ces valeurs forment des *systèmes* qui peuvent être hiérarchisés par un critère axiologique. La section 4 introduit la notion de conscience, qui offre à l'agent la capacité à faire (partiellement) usage de réflexivité et de faire retour sur les valeurs qui induisent ou non ses préférences. La section

5 postule que l'agent est capable de décider d'adhérer ou bien d'abandonner certaines valeurs à condition d'avoir conscience de ces valeurs. La section 6 définit l'agent rationnel comme celui qui utilise systématiquement cette capacité afin d'adopter le meilleur système de valeurs (celui qui maximise son critère axiologique) à sa portée. Dès lors, dans le cadre de la délibération partielle, la rationalité se comprend comme une forme d'autonomie. Ce faisant, et c'est le thème développé par la section 7, la délibération partielle offre une voie médiane entre une conception interne du comportement humain et une conception externe, sans pour autant prétendre à la non-validité de ces approches.

1 Exposition synthétique de la délibération partielle

Cette section vise à décrire le mécanisme à l'oeuvre dans la délibération partielle. Dans la mesure où l'apport de ce chapitre est principalement conceptuel, je me donne un appareil mathématique minimal, reléguant les questions formelles aux chapitres suivants de cette thèse.

Chercher à expliquer comment les préférences changent c'est, formellement, exprimer la règle de transition, $\rightarrow$, d'une relation de préférences *ex ante* $\preceq$ à une relation de préférences *ex post* $\preceq'$.

$$\preceq \quad \rightarrow \quad \preceq'$$

La stratégie que j'emploie pour établir cette règle consiste dans un premier temps à faire dépendre les préférences $\preceq$ et $\preceq'$ des ensembles V et V', qui forment *les systèmes des valeurs* auxquelles l'agent adhère [8]. Aussi la règle de transition peut-elle s'exprimer de la façon suivante :

$$\preceq_V \quad \rightarrow \quad \preceq_{V'}$$

Je pose $\hat{V}$ l'*ensemble total des valeurs*, constitué de toutes les valeurs imaginables qu'un agent est susceptible d'adopter. Ainsi, on a :

$$V \subseteq \hat{V} \text{ et } V' \subseteq \hat{V}$$

Je définis et discute la notion de système de valeurs dans les deuxième et troisième sections [9]. Provisoirement, on peut cependant indiquer que, dans le modèle de délibération

8. Cyert and DeGroot [1979], qui utilisent la même stratégie, parlent « d'état d'esprit » de l'agent. Toutefois, comme l'écrit Houlding [2008, p. 40], « en tant que tel, Cyert et De Groodt ne s'intéressent pas à la nature ontologique de V ». Les deuxième et troisième sections spécifient la nature ontologique de cet ensemble.

9. Par exemple les propositions v=« un système politique démocratique est bien » et v'=« un système politique qui respecte la liberté est bien » constituent des valeurs potentielles. Un système de valeurs potentielles peut être constitué par la conjonction de ces deux valeurs $\{v, v'\}$.

partielle, un système de valeurs remplit deux fonctions :

1. Il détermine la relation de préférencesd'un individu $\preceq_V$ à un instant donné,
2. Il conditionne, *en partie*, la manière dont l'individu change le système de valeurs auquel il adhère, le faisant passer de V à V'.

Le second point indique qu'il ne le conditionne qu'en partie, dans la mesure où les changements de préférences que je prends pour objet procèdent d'une *prise de conscience*. Aussi doivent-il également dépendre d'un ensemble $A \subset \hat{V}$ de valeurs dont l'agent a conscience *ex post*. Je qualifie de *raisons* les valeurs appartenant à cet ensemble.

Définition 1 : *Une raison est une valeur dont l'agent a conscience.*

Notons qu'une raison de l'agent n'est pas nécessairement une valeur à laquelle il adhère. Il est possible d'être conscient que la valeur « être raciste est légitime » existe et de ne pas y adhérer. Cet élément conceptuel est confirmé par la figure 1.1.

Dans la délibération partielle, un changement de préférences est intégralement déterminé par le couple (A, V) constitué des raisons qu'envisage l'agent et des valeurs auxquelles il adhère *ex ante*. Autrement dit, l'objectif devient d'exprimer la règle de transition $\rightarrow_A$ telle que :

$$\preceq_V \quad \rightarrow_A \quad \preceq_{V'}$$

A partir de ces deux ensembles de valeurs (celles dont l'agent a conscience et celles auxquelles il adhère), on peut définir quatre types de valeurs.

— Les *raisons motivantes* constituent l'ensemble des valeurs auxquelles l'agent adhère et dont il a conscience ($V \cap A$). C'est, en principe, les valeurs sur lesquelles l'agent est capable de faire un retour introspectif. Par exemple, une raison avancée par Victor Hugo durant la bataille d'*Hernani* est que les règles de l'art classique représentent les valeurs d'un temps ancien, non compatibles avec l'ordre nouveau que la révolution française et la période napoléonienne sont en train de faire advenir.

— Les *raisons irrecevables* sont les valeurs dont l'agent a conscience mais auxquelles il n'adhère pas ($A \backslash V$). Victor Hugo a conscience de la règle du théâtre classique qui veut que la « bienséance externe » soit respectée, mais il la juge irrecevable ; l'art n'a pas pour objet le respect des conventions sociales.

— Les *valeurs d'arrière-plan* sont les valeurs auxquelles l'agent adhère sans en avoir conscience ($V \backslash A$). Implicitement, Victor Hugo cautionne certaines thèses défendues par Hegel sur la dialectique de l'histoire et de l'art, mais il n'est pas certain qu'il en ait eu connaissance.

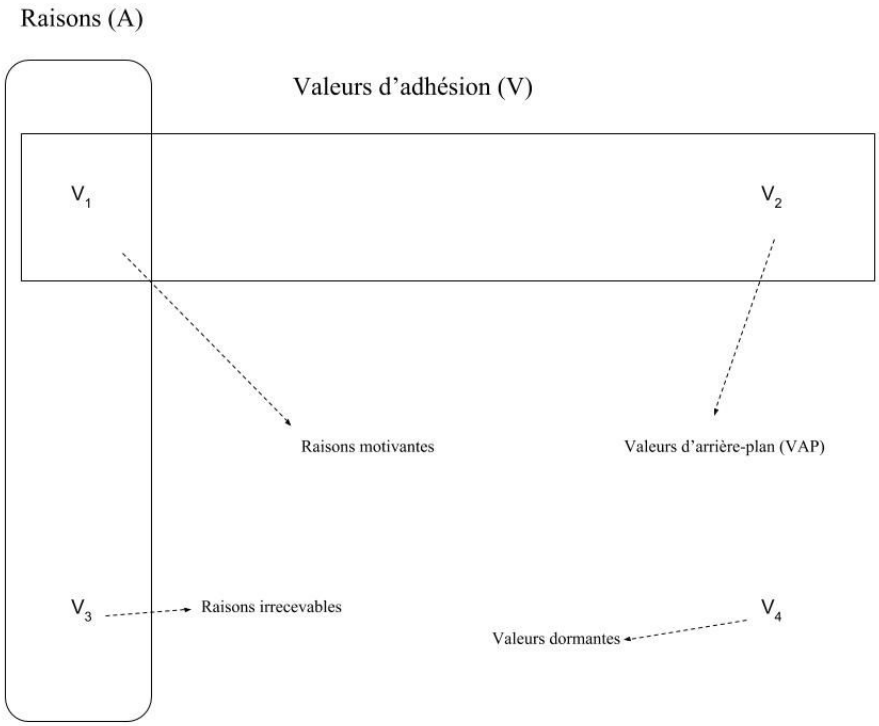

FIGURE 1.1 – Définition des différents types de valeurs

— Les *valeurs dormantes* sont celles dont l'agent n'a pas conscience et auxquelles il n'adhère pas ($\neg A \cap \neg V$). La bataille d'*Hernani* fait suite à une longue période durant laquelle se construit le concept d'esthétique. L'idée selon laquelle le beau est une valeur en soi, indépendante de l'utile, n'est pas alors un conception externeère d'appréciation [10] : le jugement esthétique n'existe pas à proprement dit. A n'en pas douter, l'émergence de cette discipline a joué un rôle important dans l'édification des valeurs de « l'art pour l'art » telle qu'elle fut professée par les romantiques durant la bataille d'*Hernani*. Elle met sur le marché des valeurs de nouvelles valeurs comme celle de « l'art pour l'art » qui était, jusqu'alors, inconcevable. La figure 1.1 résume l'ensemble de ces définitions.

Ce dernier exemple suggère qu'il est nécessaire, pour qu'une valeur soit adoptée (ou abandonnée), qu'elle surgisse à la conscience de l'agent, qu'elle soit cognitivement disponible. Si la valeur de « l'art pour l'art » n'avait pas été mise sur le marché des idées, nul individu n'aurait pu l'ajouter à son système de valeurs ; si les valeurs de la représentation théatrale classique n'avaient pas été questionnées, nul individu n'aurait pu les retirer de son système de valeurs. Si je définis et discute ce que j'entends par le fait d'avoir conscience de certaines valeurs dans la section 5, on peut d'ores et déjà préciser que cette condition

10. Dans la *Critique de la Faculté de Juger*, Kant distingue le beau de l'agréable précisément à partir de ce critère d'utilité.

impose que les valeurs qui différencient V et V' doivent appartenir à l'ensemble des valeurs dont l'agent a conscience, i.e. $V \triangle V' \subseteq A$ [11]. Nous pouvons dès lors définir la première caractéristique de la délibération partielle comme suit.

Définition 2 : *On dit qu'un changement de préférences est engendré par une prise de conscience si les valeurs qui changent de statut motivationnel sont celles dont l'agent prend conscience, i.e., si :*

$$V \triangle V' \subseteq A$$

Autrement dit, les seules valeurs que l'agent est susceptible d'adopter (ou d'abandonner) d'une période à l'autre sont les valeurs dont il a conscience *ex post*. Cette caractéristique est illustrée par la figure 1.2. Le schéma de gauche donne l'ensemble des valeurs auxquelles l'agent adhère (V) et celles dont il a conscience *ex ante* (A_1). Sur le schéma du centre, seules les valeurs dont l'agent a conscience changent : sont superposées les valeurs auxquelles il adhère *ex ante* et celles dont il aura conscience *ex post* (A_2). Conformément au schéma de droite, le changement d'état de conscience induit alors un changement dans l'ensemble des valeurs auxquelles il adhère *ex post* (V'), mais ce changement ne peut se produire que sur l'ensemble des valeurs dont l'agent a conscience *ex post* (A_2).

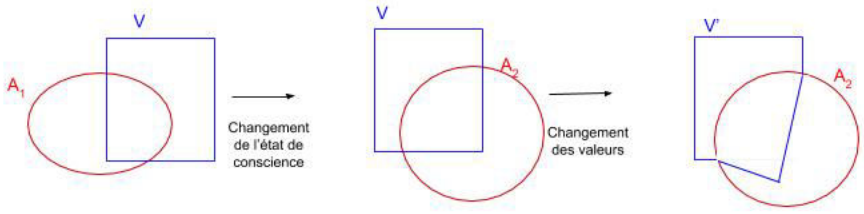

FIGURE 1.2 – Résumé schématique du mécanisme de formation des préférences par délibération partielle.

Pourquoi un tel changement s'opère-t-il ? La réponse apportée par la délibération partielle repose sur le fait que les valeurs forment des systèmes susceptibles d'être hiérarchisés. Cette hiérarchie est incarnée, dans le modèle de délibération partielle, par une relation d'ordre $\trianglelefteq$, nommée *critère axiologique* et qui se définit comme suit.

Définition 3 : *On dit que le système de valeur $V \subseteq \hat{V}$ est axiologiquement dominé par que le système de valeur $V' \subseteq \hat{V}$ si le critère axiologique le place plus haut dans sa hiérarchie, i.e., :*

$$V' \trianglelefteq V$$

11. $\triangle$ est l'opération ensembliste nommée différence symétrique, c'est-à-dire $A \triangle B = (A \backslash B) \cup (B \backslash A)$. $V \triangle V'$ est constitué des valeurs qui sont soit dans V, soit dans V', mais pas dans les deux.

Cette affirmation, ainsi que l'usage d'un critère axiologique susceptible de classer les systèmes de valeurs, est discutée dans la troisième section de ce chapitre. Elle signifie qu'il est, en principe, possible de classer des ensembles de valeurs à l'aide d'une relation d'ordre $\trianglelefteq$, de sorte que le passage de V à V' s'explique par le fait que le critère axiologique de l'agent le conduit à adopter (ou abandonner) des valeurs dont il a pris conscience. Cela suppose donc que l'agent décide *partiellement* des valeurs auxquelles il adhère. Le fait que 1) l'agent soit capable de changer les valeurs auxquelles il adhère, 2) qu'il effectue effectivement ce changement est discuté dans les sections 5 et 6. Formellement, cela se traduit par la définition suivante.

Définition 4 : *On dit qu'un changement de valeur procède de la délibération partielle s'il existe $\trianglelefteq$ tel que :*

$$V \rightarrow_A V' \iff \begin{cases} \exists B \subseteq A, V' = B \cup (V \backslash A) \\ \forall B' \subseteq A, B' \cup (V \backslash A) \trianglelefteq V' \end{cases}$$

Remarquons que le nouvel ensemble de valeurs auxquelles l'agent adhère, V', contient donc l'ensemble de ses valeurs d'arrière-plan $(V \backslash A)$ qui, joint à un ensemble de raisons, forme l'ensemble des valeurs maximal qu'il puisse adopter. Autrement dit, les valeurs que l'agent choisit d'adopter (ou d'abandonner) sont celles dont 1) il a conscience, 2) qui lui permettent d'améliorer (au sens de $\triangleleft$) son système de valeurs, 3) conditionnellement à ses valeurs d'arrière-plan $(V \backslash A)$.

Notons que, bien qu'il ne soit pas en mesure d'adopter (ou d'abandonner) les valeurs dont il n'a pas conscience *ex post*, le changement de valeurs effectué par l'agent est conditionné par l'ensemble de ses valeurs d'arrière-plan $(V \backslash A)$. En d'autres termes, ses valeurs d'arrière-plan jouent un rôle dans la mutation du système de valeur de l'agent, même s'il n'est pas en mesure d'adopter (ou d'abandonner) leur valeurs dont il se compose. Cette idée est illustrée par la figure 3, qui met en scène le changement de préférences de deux agents : l'agent A dont le système de valeurs *ex ante* est donné par l'ensemble V_A et l'agent B dont le système de valeurs *ex ante* est donné par l'ensemble V_B. Bien que les agents A et B prennent exactement conscience des mêmes valeurs (de C_1 on passe à C_2), le changement de préférences ne conduit pas à ce que A et B décident d'adopter (ou d'abandonner) les mêmes valeurs. Cette différence est due au fait que les valeurs d'arrière-plan des deux agents diffèrent et que le processus de formation des préférences dépend des valeurs d'arrière-plan de l'agent.

Le chapitre suivant de cette thèse donne les axiomes permettant de caractériser les conditions dans lesquelles la définition 4 est vérifiée. Les sections suivantes du présent chapitre justifient son cadre conceptuel à l'appui des cinq hypothèses mentionnées à la fin de l'introduction de ce chapitre.

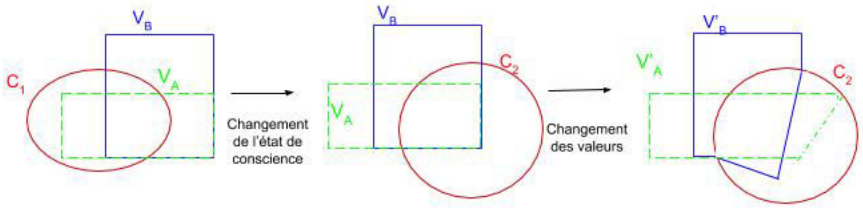

FIGURE 1.3 – Illustration du rôle des valeurs d'arrière-plan : comparaison entre deux agents.

2 Les préférences sont déterminées par des valeurs

L'expérience quotidienne nous suggère que nous sommes en capacité de rationaliser nos préférences. « En tant que créature rationnelle, écrit le biologiste Jones [1908], tout un chacun se sent en mesure de rendre compte de sa personne, de sa conduite ou de ses opinions, en s'appuyant sur un discours fait de connexions logiquement et continument agencées » [12]. Imaginez-vous demander à Théophile Gautier *pourquoi* il préfère les oeuvres où sont brisées des règles du théâtre classique. Ne seriez-vous pas insatisfait s'il vous répondait simplement, « parce que je les préfère » ? Conçu comme tautologique, le concept de préférences, tel qu'il est pensé dans la théorie des préférences révélées, ne correspond pas à l'intuition ordinaire que nous en avons. Nous nous imaginerions Théophile Gautier discourir sur les implications de la règle des trois unités, affirmer que les règles du théâtre classique ont été établies pour ajourner l'avènement du monde de demain, qu'elles sont arbitraires et ont pour unique objectif de servir les intérêts du pouvoir en place, etc. En d'autres termes, nous attendrions de Théophile Gautier qu'il donne ses *raisons* de préférer telle oeuvre à telle autre.

Le terme de « raison » nous suggère ici que les préférences sont susceptibles d'être *expliquées*, que ce soit par l'agent ou par un observateur extérieur. Dans le cadre de la délibération partielle, cette capacité de rationaliser les préférences d'un invividu est fournie par l'ensemble V de ses valeurs ; les raisons étant comprises comme rationalisant le comportement de l'agent à la fois pour l'observateur et pour l'agent (cf. section 4). La première hypothèse de la délibération partielle consiste donc à prendre cette intuition au sérieux et à postuler que les préférences sont induites par de telles valeurs :

Hypothèse 1 : *Les préférences de l'agent dépendent d'un ensemble limité de valeurs.*

La notion de valeur fait l'objet d'une importante littérature à l'intersection de la psychologie, la sociologie et la science politique. A ma connaissance, il n'existe pas de définition

12. "Everyone feels that as a rational creature he must be able to give a connected, logical and continuous account of himself, his conduct and opinions, and all his mental processes are unconsciously manipulated and revised to that end."

[Jones, 1908, p. 166]

nette et consensuelle de ce terme. Pour mon propre usage, j'en adopte une définition large mais précise qui tient en deux caractéristiques :

1. les valeurs sont des entités psychologiques susceptibles de déterminer les préférences d'un individu à un instant donné (les préférences d'un agent étant déterminées par l'ensemble des valeurs auxquelles il adhère),
2. les valeurs forment le contenu d'un état mental du type : l'agent « adhère à v » , état mental qui est pourvu d'une condition de satisfaction qui porte sur l'ensemble des valeurs auxquelles l'agent adhère. C'est ce que j'entends par l'idée *les valeurs forment des systèmes*.

Conformément à l'hypothèse 1, seule la première caractéristique est discutée dans cette section, la discussion de la seconde caractéristique étant reléguée à la section suivante.

2.1 Raisons, valeurs et préférences

En définissant les valeurs comme ce qui détermine les préférences, j'indique qu'une valeur peut être pensée comme une proattitude. Selon la formule de Davidson, dans la rubrique des pro-attitudes « il faut inclure des désirs, des volontés, des envies, des incitations, et une grande variété de conceptions morales, de principes esthétiques, de préjugés économiques, de conventions sociales, d'objectifs, et de valeurs, publiques ou privés, pour autant que l'on puisse interpréter ceux-ci comme les attitudes d'un agent dirigées vers des actions d'un certain type. » [Davidson, 1993].

Ainsi, à première vue, une valeur ne se distingue pas d'un ensemble d'entités mentales telles que le désir, la motivation, etc.. Il ne s'agit pas de le nier. En ce sens mon concept de valeur est large. Il convient cependant de souligner deux choses. D'abord, le terme de valeur tel que je l'emploie ne pointe pas vers ce qui est désiré, i.e., une option de comportement de l'agent, elle pointe vers ce qui est désirable dans un ensemble d'options de comportement. Par ailleurs, une valeur ne constitue pas l'opérateur modal d'une attitude propositionnelle, mais le *contenu* d'une attitude propositionnelle.

Que les valeurs déterminent les préférences impliquent quelles sont à même de jouer un rôle dans l'explication de l'action. A l'instar des raisons en philosophie de l'action, les valeurs permettent donc d'expliquer, tant du point de vue de l'agent que de l'observateur, le choix effectué par l'agent, et ce même si l'explication de chacun d'eux diffère. Toutefois, la définition d'une raison donnée par Davidson [1993] est constituée de deux clauses. Pour être une raison R de faire A, il faut que :

1. l'agent ait une inclination (une pro-attitude) envers une propriété P qui renvoie à R,
2. l'agent croit que A satisfait P.

Les préférences de l'agent peuvent changer soit parce que 1) l'agent modifie l'inclination qu'il porte à P, soit parce que 2) il modifie sa croyance que telle action A à la propriété

P, ou, plus généralement, il prend conscience que l'action dispose de P. On retrouve ici les deux formes de mécanismes évoquées dans l'introduction de ce chapitre. De fait le point 2. conduit à un changement de préférence local car il suppose de l'agent qu'il ait simplement changé sa façon de voir l'action A et qu'il la croit maintenant pourvue d'une valeur neuve [13]. Alors que 1) induit un changement de l'ensemble des valeurs de l'agent.

Duperie de soi écartée, adhérer à une valeur « l'égalité est bonne » ne me prédispose pas à juger les Etats-Unis inégalitaires davantage que ne pas y adhérer. En revanche, il est clair qu'on peut avoir des raisons (epistémiques) de croire que les Etats-Unis sont inégalitaires et que, toutes choses égales par ailleurs, ces raisons peuvent constituer une motivation de ne pas aimer le système politique américain lorsque l'on est dans le cadre du mécanisme 2). Toutefois, comme la délibération partielle porte sur le mécanisme 1), il est légitime de parler de valeur plutôt que de raison. En effet, en supposant que le mécanisme de transformation des préférences surgit de la clause 1), on ne s'intéresse qu'à la transformation des proattitudes de l'agent.

Je souscris donc à la « théorie causale de l'action », telle qu'elle est défendue par Davidson [1993]. Je comprends les valeurs comme des évènements mentaux qui, à l'instar des évènements physiques, causent les actions [14]. Cette caractéristique épuise la définition que j'attribue aux valeurs sur le plan statique. Ce qui différencie les valeurs des désirs tient à leur effet dynamique : les valeurs sont susceptibles d'*orienter* l'évolution des préférences. Avant de revenir sur cette caractéristique dans la prochaine section, il convient de remarquer qu'il existe de nombreuses tentatives, dans la théorie du choix rationnel, de fournir une explication psychologiques aux préférences comme je le fais en m'appuyant sur le concept de valeur.

2.2 L'explication psychologique des préférences

L'idée que les préférences individuelles peuvent être expliquées par des entités psychologiques comme des valeurs, des motivations ou des raisons n'est pas neuve en théorie du choix. L'approche de Lancaster [1966], qui fonde la valeur des biens sur leurs caractéristiques, constitue l'une de ses occurrences les plus célèbres. Cette approche porte sur les propriétés des objets de choix, et laisse de côté la manière dont elles affectent l'esprit des agents. Ce faisant, elle s'interdit tout discours sur l'évolution des préférences car elle ne permet pas de comprendre comment les agents attribuent une valeur aux caractéristiques de ces biens. Pour ma part, si je suppose que les valeurs portent sur les propriétés des biens, je m'intéresse plutôt à la manière dont un point de vue psychologique peut en éclairer l'évolution.

Les psychologues Shafir, Simonson and Tversky [1993] proposent une approche alternative à la théorie standard de l'utilité : une approche fondée sur les raisons (*reason*

13. *Idem*, je mets préférence au sigulier car le changement est local.
14. On pourra consulter également Proust [2005] pour un exposé détaillé de cette théorie.

based theory of choice). Selon ces auteurs, bien que réputée moins rigoureuse que la théorie standard de l'utilité, une telle approche permet de souligner les contradictions internes qui conduisent à la décision. Ce faisant, elle traduit mieux l'intuition qui veut que certaines de nos valeurs peuvent entrer en contradiction [15].

Ce type de contradiction se retrouve dans de nombreux modèles d'économie comportementale, comme ceux qui font l'hypothèse des « moi pluriel » [16]. On peut classer ces modèles en deux catégories. La première suggère que la prise de décision individuelle résulte d'une procédure d'agrégation des préférences de plusieurs « identités » [Ambrus and Rozen, 2014, Kavka, 1991, May, 1954], qui sont parfois opposées. La seconde porte sur les choix dynamiques et suggère que l'agent, d'une période à l'autre, résout le conflit intertemporel qui est à l'oeuvre entre différentes identités potentiellement contradictoires [Benabou and Tirole, 2004, Bernheim and Rangel, 2009, Fudenberg and Levine, 2012]. La question de la maîtrise de soi, du sang-froid, devient alors cruciale.

Ces deux approches ont cependant recours au concept métaphysiquement chargé de « moi ». L'approche par les valeurs ou les raisons me semble renvoyer à des entités psychologiques mieux définies et qui ne supposent pas les mêmes embarras métaphysiques [Scanlon, 1998].

En théorie du choix rationnel, plusieurs articles se sont penchés sur la notion de rationalisation des préférences. Manzini and Mariotti [2007] proposent ainsi de comprendre le choix comme l'application séquentielle de plusieurs relations d'ordre, chacune étant justifiée par des critères normatifs, ou des rationalisations (*rationals*). Autrement dit, lorsqu'il est confronté à un choix, l'agent sélectionne en premier lieu un sous-ensemble d'alternatives, maximales au sens d'un premier critère C_1, puis applique un second critère C_2 pour sélectionner l'alternative de son choix à l'intérieur de ce sous-ensemble [17]. Salant and Rubinstein [2008] reprennent une idée similaire, insistant sur le fait que le décideur prend ses décisions à la lueur d'un certain cadre d'analyse. De même, Cherepanov, Feddersen and Sandroni [2013] insistent sur le rôle des rationalisations. On peut également citer les travaux réalisés en théorie de la décision multicritères [18], ou encore l'armature conceptuelle fondée sur les raisons De Clippel and Eliaz [2012]. Toutefois, la plupart de ces modèles n'usent des concepts de raisons ou de valeurs que de façon implicite. Dietrich and List [2013*a*,*b*, 2016] renouvellent ce type d'approche en modélisant explicitement une relation de préférences fondé sur les raisons. Cette thèse prolonge cette approche, et lui donne une portée dynamique.

15. Jackson [1985] montre que tant en matière de désirs qu'en matière de croyances morales, il est désirable que notre théorie de l'esprit (et de l'action) puisse accorder une place à ce type de contradiction.

16. Je prends à mon compte la traduction de Ferey [2011]

17. Une fonction de choix est séquentiellement rationalisée par une paire ordonnée de $(<_1,<_2)$ si pour tout S, $\{C(S)\} = max(max(S;<_1),<_2)$. Manzini et Mariotti montrent que cette caractérisation est obtenue à partir de deux axiomes.

18. La théorie de le décision multicritères a un statut à part. Elle relève tout aussi bien de la recherche opérationnelle, de l'ingéniérie, que de la théorie du choix rationnel. Pour une revue des approches théoriques récentes, fondées sur l'intégrale de Choquet on pourra consulter Grabisch and Labreuche [2010].

Dietrich and List [2013*b*] proposent de modéliser une *structure de préférences fondée sur les raisons* qui s'applique sans encombre, dans notre cadre, aux préférences fondées sur des valeurs. Pour ce faire, ils définissent en premier lieu un ensemble d'objets X, qui peuvent être caractérisés par des valeurs, $v \in \mathcal{P}(X)$ [19]. Une valeur constitue ainsi l'ensemble des objets de choix dont les propriétés vont dans le sens de cette valeur. Autrement dit $x \in v$, si l'alternative x est dotée de propriétés satisfaisant la valeur v [20]. L'ensemble V se constitue de ces valeurs.

Ils définissent également une collection d'ensembles $\mathcal{V}$ dont les éléments sont les V ; une famille de relations de préférences induite par $(\preceq_V)_{V\in\mathcal{V}}$ et une relation de pondération (*weighing relation*) [21], notée $\leq$, qui permet de classer des ensembles de valeurs [22]. On dit alors qu'une famille de relations de préférences $(\preceq_V)_{V\subset\hat{V}}$ admet une structure de raison si pour tout $x, y \in X$:

$$x \preceq_V y \iff \{v \in V : x \in v\} \leq \{v \in V : y \in v\}$$

Autrement dit, lorsque l'ensemble des valeurs de l'agent est limité à V, l'agent préfère une alternative y à une alternative x si, et seulement si les propriétés de V que détient x sont meilleures au sens de $\leq$, que celles détenues par y.

Ils montrent alors que deux axiomes permettent d'induire la relation $\preceq_V$ par une relation de pondération $\leq$:

1. Le premier indique que seules les valeurs auxquelles l'agent adhère (qui sont dans V) et qui caractérisent effectivement les objets de choix $x \in v$, induisent la relation de préférence. Autrement dit,

$$\{v \in V : x \in v\} = \{v \in V : y \in v\} \Leftrightarrow x \backsim_V y$$

2. Le second axiome dit que si les ensembles de valeurs auxquelles l'agent adhère changent sur un ensemble de propriétés qui ne permettent pas de distinguer deux objets de choix, alors ses préférences pour ces deux objets ne s'inversent pas. Autrement dit, pour tout $V_0, V_1 \in \mathcal{P}(V)$, et tout $v \in V_0/V_1$ si ni x ni y n'appartiennent à v alors

$$x \preceq_{V_0} y \Leftrightarrow x \preceq_{V_1} y$$

Ce mode de représentation permet de conceptualiser des agents dont les préférences

19. Je note $\mathcal{P}(X)$ l'ensemble des parties de X.

20. Par exemple, si l'objet de choix est une distribution de revenu dont la propriété est « équitable », alors la propriété satisfait la valeur « être équitable est bien »et la distribution en question appartient à l'ensemble dénoté par la valeur « être égalitaire ».

21. A propos de ce type de *weighing relation* voir aussi [Nozick, 1993].

22. Bien qu'il puisse sembler plus commode d'utiliser la logique propositionnelle pour manifester ces relations, c'est par exemple le cadre adopté par [Liu, 2011], mon analyse s'appuie exclusivement sur des outils issus de la théorie de la décision ou de la théorie des jeux. Or la relation d'ordre permet de traduire cette logique notamment par les notions de complémentarité et de substituabilité.

différent. Ce faisant, il satisfait un *desiterata* non négligeable de la théorie des transformations des préférences. Les préférences étant déterminées par des ensembles *limités* de valeurs, chaque agent construit ses préférences sur la base d'un ensemble restreint de valeurs. Dans ce cadre, les partisans de Victor Hugo et les défenseurs du modèle classique de l'art dramatique ne fondent tout simplement pas leurs préférences sur le même ensemble de valeurs.

En revanche, Dietrich et List restent délibérément agnostiques sur la nature exacte de l'ensemble V. Il peut, selon eux, représenter diverses contraintes psychologiques comme « la saillance, l'attention, la compréhension qualitative, une réponse émotive ou alors tout simplement la conscience de l'individu (*awareness*) » [Dietrich and List, 2013*b*, p. 4]. Mon cadre donne plus de substance au concept de valeur. Cette substance permet de comprendre pourquoi et comment, avec la délibération partielle, les préférences se forment.

3 Les valeurs forment des systèmes susceptibles d'être axiologiquement ordonnés

A l'appui des travaux de Shafir, Simonson and Tversky [1993], nous avons vu que prendre les valeurs en considération, en statique, peut redonner sens à l'idée que les préférences sont constituées par des valeurs potentiellement conflictuelles. Cette idée est au centre de l'ouvrage de Levi (1990). Selon lui, en tant qu'individus rationnels, nous sommes susceptibles de sentir que certaines de nos valeurs entrent conflits. Un tel conflit ne relève pas, comme dans le cas de la faiblesse de volonté, d'un conflit entre ce que l'agent voudrait faire et ce qu'il juge préférable, « tout bien considéré ». Il relève du fait que certains groupes de valeurs ne peuvent susciter l'adhésion sans créer de tension qui, pour être résolues, sont susceptibles d'engendrer des changements de système de valeur de l'agent.

Les conflits potentiels entre les valeurs sont moins essentiels pour comprendre le choix en statique, que pour les conséquences qu'ils peuvent revêtir en dynamique. La théorie de la délibération partielle ne s'intéresse pas au fait que les agents sont contradictoires à un instant donné. Elle tend à soutenir qu'étant donnés les moyens cognitifs qui sont les leurs, ils choisissent l'ensemble des valeurs qui est le plus satisfaisant sur le plan *axiologique*.

Que signifie, cependant, être « plus satisfaisant sur le plan axiologique » ? Pour donner sens à cette expression, nous avons besoin d'une seconde hypothèse, qui suppose l'existence d'un principe de classement, à même d'établir une hiérarchie entre les ensembles de valeurs et qui renvoie à la rationalité axiologique de l'agent évoquée en introduction de cette thèse. La seconde hypothèse peut donc se formuler en ces termes.

Hypothèse 2 : *Les valeurs peuvent être articulées en des systèmes susceptibles d'être classés par un critère axiologique.*

Une telle hiérarchie découle d'une seconde caractéristique des valeurs que je vais main-

tenant développer : les valeurs fournissent un contenu non factuel à des états mentaux intentionnels. Or pour susciter l'adhésion, un tel contenu doit être conforme à la rationalité axiologique de l'agent.

Les valeurs ne servent pas simplement à évaluer les actions que nous avons à entreprendre, elles constituent des jugements de second ordre, susceptibles d'indiquer les valeurs qui doivent être préférées. Cette propriété des valeurs suggère que les ensembles de valeurs peuvent être classés par leur degré d'articulation axiologique.

Cette section vise à discuter la nature de cette articulation axiologique, ainsi que sa représentation formelle par ce que j'appelle un *critère axiologique* ($\trianglelefteq$).

3.1 Les valeurs fournissent un contenu non factuel à des états mentaux intentionnels

Dans le cadre de la délibération partielle, on s'intéresse à un état mental du type "l'agent adhère à la valeur v". Une valeur forme donc le contenu d'un état mental, qui conformément à Searle [1983], est pourvu d'une intentionalité, c'est-à-dire un état mental qui porte sur quelque chose, qui met l'agent en rapport avec le monde. Mais le monde dont il est question n'est pas empirique à proprement parler, car la condition de satisfaction d'un état mental du type, "l'agent adhère à v" est axiologique. Qu'est-ce à dire ?

Un exemple paradigmatique d'état intentionel est la croyance. Dire que l'agent croit la proposition p, c'est dire que l'agent attend de p qu'elle soit vrai. S'il était donc montré empiriquement que p n'est pas vrai, l'agent devrait modifier sa croyance. C'est que la condition de satisfaction donnée par le critère de la vérité impose qu'une croyance puisse être révisée dès lors qu'elle porte sur un énoncé qui est faux.

A l'instar des croyances, une valeur v renvoie à une condition de satisfaction. Contrairement aux croyances cependant, cette condition de satisfaction ne porte pas sur un critère de vérité, mais sur un *critère axiologique*. Autrement dit, l'état mental qui veut que l'agent adhère à une valeur v, suppose que v soit en adéquation avec le critère axiologique de l'agent.

L'hypothèse centrale que je fais ici est que, pour savoir si une valeur entre en adéquation avec le critère axiologique de l'agent, il ne faut pas se référer au contenu empirique de v, mais à la compatibilité de v avec les autres valeurs auxquelles l'agent peut adhérer (ou abandonner). Il est peu probable que les valeurs du romantisme satisfassent le critère axiologique de Corneille dans la mesure où les valeurs auxquelles Corneille adhère sont celles du clacissisme et qu'elles sont incompatibles avec les valeurs de romantiques.

Bien entendu, certaines valeurs portent sur des énoncés qui incorporent une dimension factuelle. Prenons le cas de la croyance que « le monde est juste ». Développée par le psychologue Lerner [1980], et reprise par Bénabou and Tirole [2006], cette valeur se définit comme le fait de croire que les individus, contre toute détermination, obtiennent ce qu'ils méritent. Pourquoi affirmer qu'une telle croyance n'est pas factuelle ? La raison est double.

D'abord, adopter cette croyance revient à combiner dans le même énoncé une dimension évaluative - « il est bon que les individus obtiennent ce qu'ils méritent » - et une dimension factuelle - « les individus disposent de la liberté suffisante pour faire ce qu'ils désirent ». Elle n'est donc pas purement factuelle. Ensuite, la dimension factuelle de cette croyance porte sur un énoncé dont ni l'affirmation, ni la contradiction ne sont expérimentalement vérifiables.

Les valeurs contiennent des jugement de second ordre [23]. La bataille d'*Hernani*, par exemple, porte sur le type de valeur esthétique qu'il faut adopter. Elle vise, entre autres, à défendre l'autonomie de l'art, arguant qu'il détient une valeur en soi, indépendamment de la morale ou de la vérité. L'enjeu de la querelle est de défendre un jugement portant non seulement sur les objets du choix, mais surtout sur les valeurs qui doivent susciter l'adhésion.

Dans la section précédente, j'ai donné une caractéristique statique aux valeurs arguant qu'elles déterminent les préférences de l'agent. Dans cette nouvelle section, j'attribue une seconde caractéristique, dynamique cette fois-ci, qui permet de comprendre ce qui motive ou non l'adhésion à une valeur : l'adhésion à une valeur dépend d'un critère axiologique, qui établit si des valeurs sont compatibles ou non d'un point de vue axiologique. Bien que j'entende rester agnostique sur ce critère, la suite de cette section cherche à éclaicir certains des enjeux ontologiques qui l'accompagnent.

3.2 Par quel biais les valeurs s'articulent-elles ?

Déterminer comment certaines valeurs s'articulent entre elles est un enjeu primordial des sciences sociales. Un exemple célèbre est le concept d'affinité élective, employé par Weber [2003] dans *L'éthique protestante et l'esprit du capitaliste*. Selon lui, la société capitaliste que nous connaissons tirerait ses origines d'une articulation entre les valeurs du protestantisme et les valeurs de la bourgeoisie naissante [24]. Si le concept de Weber a pour principal objectif d'éviter les écueils de la causalité ou de la corrélation, les psychologues des valeurs emploient plus volontiers ce dernier terme, et établissent des typologies qui suggèrent que l'adhésion à une valeur v_1 prédispose à l'adhésion de v_2. Mais cette corrélation statistique ne dit rien en elle-même de la connexion à l'oeuvre entre ces valeurs.

Dans le cadre de la délibération partielle, cette articulation peut, par exemple, tenir au fait que certaines valeurs sont incompatibles : la présence d'une de ces valeurs dans le système auquel l'agent adhère, le prédispose à ne pas adhérer à l'autre valeur. Pour reprendre un exemple de Dewey, les valeurs de patriotisme et de pacifisme sont potentiellement incompatibles.

23. « Les valeurs sont les *conceptions centrales* du *désirable* pour chaque individu et pour la société. Elles servent de standards ou de critères pour *guider non seulement l'action* mais aussi le jugement, le choix, l'attitude, l'évaluation, l'argumentation, l'exhortation, la rationalisation, et, on pourrait ajouter, l'attribution de causalité ». [Rokeach, 1973, p. 2]

24. Le type de relation supposé par le terme d'affinité élective est, à dessein, plutôt vague. Il provient de l'alchimie médiévale et évoque la propension de certains éléments à s'attirer.

> Prenons l'exemple [...] d'un citoyen dont l'Etat vient juste de déclarer la guerre à un autre pays. Il est si profondément attaché à son Etat, si habitué à lui être fidèle, à respecter ses lois, qu'il décrête en première instance qu'il doit soutenir l'effort de guerre. D'autant qu'il ressent gratitude et affection pour ce pays qui l'a logé et nourri. Toutefois, il croit également que cette guerre est injuste, et nourrit même la conviction que toute forme de guerre est inutilement meurtrière, et donc, dans l'erreur. Quand une part de lui-même, un ensemble d'habitudes et de convictions le conduit à consentir à la guerre ; une autre, non moins profonde de lui-même, le pousse à en contester les fondements. Il est tiraillé entre deux devoirs et fait l'expérience d'un conflit entre des valeurs incompatibles.
>
> [Dewey and Tufts, 1932, p. 174-175]

Certaines valeurs, à l'inverse, se complètent, de sorte qu'adhérer à l'une conduit à adhérer à l'autre. Par exemple, selon Hans Jonas, pour qui veut convaincre quelqu'un d'adopter les valeurs de l'écologie, il convient de remarquer que la valeur de liberté ne peut être satisfaite dans l'*hubris* et que la contrainte (écologique) en est la condition nécessaire :

> S'imposer des limites est la première obligation de toute liberté, la condition même de son existence, car c'est seulement ainsi qu'une société [...] est possible.
>
> [Jonas, 2017, p. 181-182]

Autrement dit, pour défendre une valeur, il est parfois commode de s'appuyer sur la complémentarité entre cette valeur et une autre valeur, à laquelle l'agent (ici le lecteur de Hans Jonas) est supposé adhérer. Quelle est l'origine de cette articulation ?

— En premier lieu, elle peut reposer sur des fondements logiques, dès lors que l'on admet que certaines valeurs sont susceptibles de se contredire dans les termes, i.e., qu'elles portent en elles une signification susceptible, *a priori*, d'entrer en contradiction. Par exemple la valeur selon laquelle il est « bon d'être raciste » et celle qui indique qu'il « faut faire preuve d'une égale tolérance à l'égard de tous les êtres humains » se contredisent dans les termes. L'une suppose l'intolérance à l'égard d'une partie des être humains quand l'autre professe l'énoncé contradictoire.

— En second lieu, elle peut reposer sur des fondements empiriques. Notamment parce que l'expérience de l'agent lui révèle que certaines valeurs promeuvent des actions qui conduisent à des conséquences incompatibles. C'est sur ce type de d'incompatibilité que semble se focaliser Dewey et Tuft cités précédemment. Reprenant l'exemple donné par Dewey et Tufts, l'analyse que Levi [1990] donne du conflit entre les valeurs de l'agent repose sur cette idée. C'est la thèse défendue par le psychologue Schwartz [2012] : « Le fondement de la structure reliant les valeurs tient au fait qu'elles conduisent à des actions dont les conséquences peuvent entrer en conflit avec

> les conséquences engendrées par d'autres valeurs, ou bien leur être congruentes » [25]. Par exemple, Schwartz suggère qu'adhérer à l'esprit de compétition implique d'essayer de se distinguer des autres, au détriment de l'esprit de coopération. Notons cependant que la contradiction est susceptible de se diluer en cas d'ajout d'une troisième valeur. Une littérature importante à la croisée de la gestion et de la théorie des jeux remet en cause cette contradiction et évoque la possibilité d'un coopétition [Nalebuff, Brandenburger and Maulana, 1996].

Dans cette thèse, je n'ai pas l'ambition de discuter la différence entre ces deux façons de voir le type de relations que sont susceptibles d'entretenir les valeurs [26]. Si un tel travail est d'une importance considérable, il relève de l'ensemble des sciences sociales et doit se fonder sur des analyses *a posteriori* qui dépassent la portée de l'analyse que je peux construire ici.

3.3 Hiérarchisation des systèmes de valeurs

Que les valeurs s'articulent entre elles nous invite à penser les ensembles de valeurs comme des systèmes susceptibles d'être hiérarchisés par un critère axiologique.

Formellement, ce critère est donné par la relation $\trianglelefteq$ qui fournit un ordre - réflexif, complet et transitif - susceptible de classer les ensembles de valeurs [27]. Dès lors, dire $B \trianglelefteq A$, c'est dire que le système de valeurs A constitue une amélioration axiologique du système B.

Il peut sembler contraignant d'affirmer que l'agent utilise un tel classement. En particulier, le critière axiologique n'implique-t-il pas qu'un individu suffisamment informé, rejète certaines valeurs ou effectue certains choix ? Postuler que certains systèmes de valeurs sont axiologiquement dominés par d'autres systèmes, ne suppose-t-il pas de faire entorse au principe de neutralité axiologique ?

Une première réponse est que rien n'impose que le critère axiologique soit universel, au sens où il serait le même pour tous les individus. Le problème ici est qu'appliquer la délibération partielle à une situation interactive, où un agent se projette dans les préférences d'un autre agent, afin de prédire ses changements de préférences, suppose que les agents pensent ce critère comme universel. C'est l'hypothèse qui sera posée au chapitre 3. Il peut donc être utile d'attribuer une universalité, même locale [28], au critère axiologique.

25. "One basis of the value structure is the fact that actions in pursuit of any value have consequences that conflict with some values but are congruent with others". [p. 8]

26. Ces deux options nous inviteraient certainement les ensembles de valeurs à partir des options existantes (pas nécessairement accessibles) pour le DM. En ce sens deux valeurs pourraient être jugées incompatibles dès lors qu'il n'existe pas d'alternative de choix qui satisfaisant ces deux valeurs de concert. Autrement dit, $\forall x \in X$, si $v \in x$ alors $v' \notin x$. A l'inverse, deux valeurs sont compatibles si les alternatives qui sont compatibles avec l'une de ces valeurs est également compatible avec l'autre. Je n'explorerai pas cette voie et ne l'utiliserais que dans des exemples.

27. Bien qu'il puisse sembler plus commode d'utiliser la logique propositionnelle pour manifester ces relations. C'est le cadre adopté par Liu [2011]. Mon analyse s'appuie exclusivement sur des outils issus de la théorie de la décision ou de la théorie des jeux Liu [2011]

28. Par "local", je veux dire dans un contexte donné (une même culture par exemple) ou un ensemble

Une seconde réponse consiste à souligner que l'universalité n'est pas si contraingante : même universel, le critère axiologique est susceptible de s'appliquer à un ensemble vaste de systèmes de valeurs d'une façon plus ou moins structurée. En particulier, ce n'est pas parce que, dans certaines circonstances, deux valeurs sont incompatibles qu'elles le sont universellement. Imaginons que deux valeurs v_1 et v_2 soient placées dans la hiérarchie donnée par $\triangleleft$ de telle sorte qu'adhérer à une seule d'entre elles soit axiologiquement plus satisfaisant qu'adhérer conjointement aux deux ($\{v_1, v_2\} \triangleleft \{v_i\}$ avec $i = 1, 2$). Cela semble indiquer que v_1 et v_2 sont incompatibles. Est-ce à dire que ces valeurs sont incompatibles en toute circonstance ? Pas nécessairement, si $\hat{V}$ et $\triangleleft$ sont structurés de telle sorte qu'il existe une valeur v_3 telle que $\{v_1, v_2, v_3\} \triangleright \{v_i\}$. Il peut exister, par l'ajout d'une troisième valeur v_3, une façon de dépasser l'opposition entre v_1 et v_2.

Cela implique qu'observer deux individus qui classent, à première vue, d'une façon différente $\{v_1, v_2\}$ et l'un des $\{v_i\}$ ne suppose pas nécessairement de voir $\triangleleft$ comme propre à chacun de ces individus. Peut-être ce manque d'universalité provient-il du fait qu'en tant qu'observateur nous avons omis une troisième valeur à laquelle l'un des agents adhère, tandis que l'autre n'y adhère pas et qui explique cette disparité. En revanche, si par exemple nous imposons que le critère axiologique soit structuré de telle sorte que pour tout j :

$$\{v_1, v_2\} \triangleleft \{v_i\} \implies \{v_1, v_2, v_j\} \triangleleft \{v_i, v_j\}$$

Il serait alors nécessaire de justifier la différence entre ces deux individus par le manque d'universalité de $\trianglelefteq$. Or, de telles restrictions peuvent s'imposer par le souci de prédire une évolution avec un nombre limité d'informations.

C'est pourquoi il est crucial de comprendre comment restreindre $\trianglelefteq$ pour rendre compte d'un grand nombre de préférences, tout en permettant de déterminer leur évolution. Cet arbitrage peut être étudié dans le cadre de la délibération partielle. *Comprendre les conséquences conceptuelles que peuvent impliquer différentes contraintes formelles sur la relation $\trianglelefteq$ constitue le principal objectif du chapitre 2 de cette thèse.*

Par ailleurs, l'universalité n'exclut pas *a priori* certaines préférences et laisse une place à l'hétérogénéité : elle en fournit même une explication. Pour le comprendre nous avons besoin d'une troisième hypothèse.

4 Prendre de conscience des valeurs

D'après l'hypothèse 1, une proportion limitée de valeurs détermine les préférences des agents. D'après l'hypothèse 2, ces valeurs forment des systèmes susceptibles d'être classés par un critère axiologique. Toutefois, à elles seules, ces deux hypothèses ne suffisent

de valeurs donné. Il peut être raisonnable de penser que deux agents partagent un critère axiologique relativement proche l'un de l'autre.

pas à rendre de compte notre intuition de départ : les changements de préférences procèdent d'une prise de conscience de nouvelles valeurs par l'agent. Aussi peut-on formuler la troisième hypothèse de la délibération partielle comme suit :

Hypothèse 3 : *L'agent a conscience d'un ensemble restreint de valeurs susceptibles de déterminer ses préférences.*

Il m'incombe de préciser ce que j'entends par conscience. Je donne d'abord une définition positive de ce que c'est que d'avoir conscience d'une valeur. Puis, je montre les usages du concept de conscience en théorie du choix. Toutefois, cet usage suppose que ce dont l'agent n'a pas conscience n'a pas d'impact sur le comportement. A rebours de cette interprétation de la conscience, je suggère qu'il est satisfaisant tant sur le plan conceptuel que sur le plan empirique, d'accorder un rôle aux valeurs dont l'agent n'a pas conscience dans sa décision. Je m'interroge ensuite sur les raisons épistémologiques qui peuvent conduire à rejeter ce rôle, pour défendre l'idée que son intérêt se situe dans un cadre dynamique, et qu'il est parfaitement possible de les négliger dans un cadre statique. Enfin, je montre que cette idée permet de rendre compte du fait que les agents sont suceptibles de déclarer être mus par des valeurs lorsque leur comportement révèle le contraire.

4.1 Définition de la conscience

Dans le cadre de la délibération partielle, il existe deux façons de caractériser la conscience. La première version suppose un faible engagement ontologique. Elle fait d'une valeur dont l'agent a conscience, c'est-à-dire d'une raison, une valeur dont il est en capacité de modifier le statut motivationnel, i.e., qu'il peut décider d'adhérer à (ou d'abandonner) cette valeur. L'ensemble des raisons est alors simplement compris comme un ensemble de valeurs « manipulables ». Sous cette version, la troisième hypothèse de la délibération partielle est contenue dans la quatrième hypothèse, que je donne à la section prochaine. Cela implique qu'un lecteur déflationiste, qui n'accepterait pas le coût conceptuel que représenterait l'introduction d'un concept de conscience plus ontologiquement chargé, peut toutefois se retrouver dans le mécanisme de la délibération partielle dès lors qu'il accepte les deux premières hypothèses, ainsi que les deux suivantes. Néanmoins l'économie réalisée par l'adhésion à cette caractérisation faible de la conscience n'est pas sans inconvénients.

— Elle suppose d'abord un inconvénient conceptuel, car sans information supplémentaire il est difficile de savoir pourquoi certaines valeurs seraient manipulables quand d'autres ne le seraient pas. Cela implique d'abord qu'un décideur soucieux d'aider les agents à changer leurs préférences pourrait difficilement comprendre le type de stratégie à adopter. Cela implique également qu'il devient difficile d'affirmer que la délibération partielle renvoie à un changement de préférences rationnel comme cela a été soutenu en introduction. En définitive, un tel nominalisme reviendrait à retomber dans une forme de mutisme conceptuel.

— Elle suppose ensuite un inconvénient empirique car si les raisons de l'agent sont simplement des valeurs manipulables, seul le comportement est à même de les révéler. Or il n'est pas certain qu'à elle seule cette source suffise. Avoir recours à ce que les agents déclarent de leurs préférences est une option utile et qui s'accomode difficilement de cette version faible du concept de conscience.

C'est pourquoi je voudrais soutenir une version plus forte sur le plan ontologique de la conscience des agents, et donc de leurs raisons. Les raisons sont des valeurs sur lesquelles l'agent est capable de faire un retour réflexif, au sens où non seulement il est conscient d'être (potentiellement) motivé par ces raisons, mais il est conscient d'être conscient d'être (potentiellement) motivé par ces raisons [29]. Les valeurs qui échappent à la conscience de l'agent ne satisfont pas ce critère. Par cette version plus forte du concept de conscience, il devient possible d'articuler la délibération partielle avec des concepts fondamentaux de la théorie de l'agir rationnel. Il est également davantage possible de voir les raisons comme ce que l'agent est capable de dire sur les causes de son comportement.

Comme je le suggère par la suite, cela signifie pas que les causes évoquées par les agents soient les bonnes. Car faire jouer un tel rôle à la conscience, c'est souligner l'impact des valeurs d'arrière-plan de l'agent, dont l'agent n'a pas conscience bien qu'elles déterminent son comportement. Les valeurs d'arrière-plan sont des valeurs auxquelles l'agent adhère *pour* l'observateur, mais l'agent n'est pas conscient qu'elles sont des valeurs auxquelles il adhère. Il n'envisage même pas l'existence de ces valeurs qui, pourtant, causent son comportement. Elles servent donc d'explication à l'observateur, et non à l'agent. A l'inverse, les raisons motivantes constituent des raisons *pour* l'agent et pour l'observateur.

Ces dernières remarques soulignent le fait qu'introduire le concept de conscience produit une assymétrie entre observateur et agent. C'est une difficulté que j'assume et dont je voudrais montrer non seulement qu'elle est utile, mais qu'elle s'inclut parfaitement dans les développements récents de la théorie du choix rationnel.

4.2 Antécédents du concept de conscience dans la théorie du choix rationnel

En théorie du choix rationnel, le concept de conscience a été mobilisé au sein de (au moins) deux programmes de recherche.

Le premier s'inscrit dans l'analyse du comportement de l'agent face à l'incertitude et tente d'interroger les conséquences sur le comportement de l'agent de l'occurrence d'évènements imprévisibles (*unforeseen contingencies*). L'imprévisibilité doit être distinguée de l'incertitude (même radicale), puisqu'elle implique de ne pas être en capacité d'assigner une probabilité non seulement à un évènement, mais également à son complémentaire. En d'autres termes, l'inconscience ou, pour reprendre le terme anglo-saxon, l'*unawareness*,

29. Dans mon analyse, être conscient d'être motivé par une raison motivante revient au même qu'être conscient d'être conscient d'être motivé par une raison.

renvoie à l'incapacité épistémique de second ordre qui porte sur les capacités introspectives de l'agent, sur sa capacité à faire retour sur son ignorance : l'agent est ignorant et il ne sait pas qu'il est ignorant. Cela conduit à remettre en cause le cinquième postulat de la logique épistémique standard [30], dit d'introspection négative [31] [32].

Selon la situation modélisée, l'inconscience est susceptible de renvoyer à au moins trois types de déficits cognitifs : un déficit attentionnel, une incapacité à concevoir ou encore un défaut d'omniscience logique Fritz and Lederman [2016]. Dans mon analyse, il n'y a pas de raison d'exclure *a priori* l'un de ces déficits cognitifs. Toutefois, les défauts d'attention et d'omniscience logique décrivent une inconscience passagère, sensible au contexte, tandis que l'inconcevabilité décrit un déficit plus massif, dont les conséquences perdurent et qui transcendent les contextes. Par conséquent, je suggère que les deux premiers types de déficits sont susceptibles d'exprimer comment les préférences varient d'un contexte à l'autre, tandis que l'inconcevabilité renvoie à des transformations de préférences qui s'apparentent à des mutations culturelles ou des changements de caractéristiques individuelles durables. Cette littérature a le mérite d'insister sur le rôle de la conscience dans la prise de décision. Ce faisant, elle suppose que la conscience informe le choix à un instant t. Je reprends à mon compte l'idée, portée par cette littérature, que le déficit de conscience correspond à un manque de capacités introspectives [33].

Le second programme de recherche tire le concept d'inconscience du côté du déficit attentionnel et met en avant que le choix de l'agent est restreint à un ensemble d'options, qu'il est capable de prendre en considération (*consideration set*). Entre marketing et théorie du choix (stochastique), l'idée est que certaines alternatives de choix ne sont pas disponibles à la conscience de l'agent. On peut penser aux travaux de Manzini and Mariotti [2014], Eliaz and Spiegler [2011] ou encore Masatlioglu, Nakajima and Ozbay [2016].

Ces approches ont le mérite d'illustrer le rôle de la prise de conscience dans l'exercice du choix, mais il convient de les distinguer de ma propre approche pour au moins deux raisons. D'abord, elles ne cherchent pas à rendre compte des changements de préférences. Dans leur perspective, les préférences sont fixes, mais le choix est contraint par les capacités cognitives de l'agent. Dans la perspective qui est la mienne, les préférences ne sont pas fixes, elles sont susceptibles de changer, mais ce changement procède des contraintes cognitives

30. La logique epistémique qui sert de fondement tout aussi bien à la théorie des jeux épistémiques qu'à la théorie de la décision.

31. Si $K(E)$ signifie l'agent sait que E, l'introspection négative postule que pour tout évènement E, $\neg K(E) \subset K\neg K(E)$. Autrement dit si l'agent est incertain vis-à-vis de E alors il sait qu'il est incertain vis-à-vis de E. Dès lors dire qu'un agent peut être inconscient c'est dire qu'il existe un évènement E tel que $\neg K(E) \cap \neg K\neg K(E) \neq \emptyset$, ou encore, selon Schipper [2014], qu'il existe des évènements auxquels l'agent assigne une probabilité 0, tout en assignant la probabilité 0 à leur complémentaire.

32. Savage [1954], lui-même, évoquait cette imperfection et tentait de le résoudre dans son analyse des *small world*.

33. Sans rendre justice aux raffinements formels auxquels les travaux sur l'*unawareness* ont donné lieu, mon objectif dans cette thèse étant de rester général. Depuis l'article de Dekel, Lipman and Rustichini [1998], on sait en effet qu'un tel concept nécessite, pour être modélisé d'une façon conceptuellement satisfaisante, d'employer un espace d'états non standard.

auxquelles l'agent est soumis. Ensuite, ces auteurs présupposent que le choix de l'agent repose exclusivement sur ce dont il a conscience. Il n'y a pas d'effet possible pour des valeurs à l'arrière-plan. Ils considèrent que la conscience a un effet statique. Comme cela paraitra dans la suite de ce chapitre, la délibération partielle présuppose que le rôle de la conscience n'est pas tant d'influencer les préférences en statique. Elle permet plutôt à l'agent de réformer ses préférences et donc son influence est davantage dynamique. Cela permet de contrer la thèse que j'appelle de l'*inutilité de la conscience*. Avant d'exposer ma position quant à cette thèse il convient de justifier l'idée qu'il puisse y avoir un effet des valeurs non-conscientes sur la décision.

4.3 L'effet des valeurs dont l'agent n'a pas conscience sur ses décisions

Le philosophe Bernard Williams [1981] suggère qu'une disposition de base comme la générosité doit pouvoir faire l'objet d'une représentation par l'agent qui agit généreusement, sans quoi elle ne pourrait pas rendre intelligible le processus délibération morale à l'oeuvre et notamment le poids de cette disposition dans la décision. Autrement dit, selon cette analyse, il n'y a pas d'influence sur la décision qui soit inconsciente.

Mais une telle façon d'articuler les concepts de conscience et de valeurs est trop forte. Reprenons un exemple donné par Pettit and Smith [1990] pour le voir[34]. Supposons que j'adhère à la valeurs « je crois qu'exprimer des idées nouvelles est bon » (v) et que, par conséquent, j'ai aussi la valeur « je crois que publier ces nouvelles idées dans un journal est une bonne chose » (v'). Il se peut qu'une année plus tard, au moment de la publication, la valeur v sorte de mon esprit. Est-ce à dire qu'il en est de même pour v' ? Certainement pas. Pourtant, la valeur qui me conduit à v, me conduit également à v'. Si donc mes valeurs devaient voir leurs effets reposer sur une délibération consciente, la disparition de la valeur v devrait engendrer celle de v'. Mais il n'est pas inenvisageable que je conserve la valeur v', tout n'ayant plus à l'esprit la valeur v. Aussi, la disparition de v de ma conscience n'implique pas nécessairement celle de v', contrairement à ce que laisse entendre Williams. Un tel problème se pose pour toute théorie selon laquelle une cause du comportement ne peut échapper à la conscience de l'agent. Il est donc essentiel d'accepter que certaines causes du comportement puissent echapper à la conscience de l'agent.

Des neurosciences à la psychanalyse, en passant par la sociologie ou la biologie, plusieurs théories du conditionnement humain sont susceptibles de clarifier le rôle que des entités psychologiques comme les désirs, les motivations ou les valeurs, peuvent avoir sur le choix, indépendamment de la conscience. Aussi pourraient-elles tout aussi bien faire référence à des désirs inconscients[35], des heuristiques, des influences subliminales ou

34. Smith et Petit prennent cet exemple à propos du désir et non des valeurs. Toutefois, ma conception des valeurs englobe celle de leur concept de désir.

35. Comme le note Elster cependant, il est problématique de traiter des états de conscience sur le même mode que des états mentaux conscients. En ce sens, le terme désir non conscient n'est peut-être pas tout à fait adéquat.

même des contraintes physiologiques sur la sensibilité [36]. D'une nature variée, les valeurs non-conscientes sont des entités psychologiques susceptibles de traduire de multiples phénomènes. La notion centrale d'arrière-plan, *background*, mérite d'être évoquée ici. Dans *L'intentionalité*, Searle [1983] insiste sur ce concept pour rendre compte du support sur lequel peut surgir un état intentionnel comme un désir ou une croyance. L'énoncé $p =$« je désire devenir président des Etats-Unis » présuppose d'entretenir des croyances, des désirs qui n'ont pas à être formulés pour constituer mon désir. De même que je dois implicitement savoir qu'il existe un statut de Président, qui suppose des pouvoirs, dans un système constitutionnel donné, je peux désirer ce poste parce qu'il me permettra de réduire la faim dans le monde, parce que je veux rendre mes parents fiers, etc. Pour avoir un effet, ces croyances et ces désirs n'ont pas besoin d'être articulés par la conscience.

En résumé, et conformémement à la seconde propriété que nous avons attribué aux valeurs, celles-ci sont des entités psychologiques qui détiennent une double dimension, cognitive et conative. Elles peuvent reposer, à l'arrière-plan, sur ces croyances et désirs inconscients. Il est d'ailleurs souvent difficile pour un agent de déterminer les valeurs qui conduisent son action. Elles constituent, comme l'écrivent Maio and Olson [1998] des truismes dont les ressorts sont difficiles à questionner. Comme cela a été mentionné en section 1, j'appelle « valeurs d'arrière-plan », ces valeurs auxquelles l'agent adhère sans en être conscient.

4.4 La thèse de l'inutilité de la conscience

En dépit des travaux évoqués en amont de cette section, la conscience joue un rôle marginal dans la théorie du choix rationnel. Ce manque d'intérêt tient au fait que, dans la théorie standard, l'agent est soit conscient de tout, soit conscient de rien [37]. La conscience est un concept superflu, parce que non discriminant pour la théorie. Nous appelerons cette idée, la thèse de l'« inutilité de la conscience ». Dans le cadre de la délibération partielle, cette thèse affirme qu'avoir conscience d'une valeur n'est ni une condition nécessaire, ni une condition suffisante pour qu'une valeur soit jugée pertinente. Elle n'est pas nécessaire dès lors que l'on accepte de considérer qu'un agent peut adhérer à des valeurs inconsciemment : l'agent a des *valeurs d'arrière-plan*. Elle n'est pas suffisante car il est parfaitement envisageable qu'un agent soit conscient d'une valeur potentielle sans y adhérer : l'agent juge certaines *valeurs irrecevables* [38].

36. L'article de Nisbett and Wilson [1977] fournit une référence incontournable sur le rôle de l'inconscient dans l'exercice du jugement et sur les préférences.

37. Cela n'implique pas qu'il ne puisse être incertain. En théorie bayésienne, les agents, potentiellement inconscients, ne peuvent être *unaware* Dekel, Lipman and Rustichini [1998], c'est-à-dire qu'ils ne peuvent pas savoir qu'ils ne savent pas quelque chose.

38. Prendre en considération un désir ne lui confère pas de statut motivationnel particulier. Aussi Pettit and Smith [1990, p.339] écrivent-ils que : « And similarly it seems that in deliberating the agent will consider the alleged fact that it is desirable in some way that r or s - it would be fun, it would be rewarding, it would be morally fine if r or s without considering the fact about himself, if it is a fact, that he desires that r or that s or even that he desires the relevant property. »

L'argument du *as if* illustre la thèse de l'inutilité de la conscience. Que l'agent maximise en conscience ou non ses gains n'a pas d'intérêt en soi, du moment qu'il se comporte *comme s'il* en était ainsi. Nul besoin de la conscience, cette capacité à faire preuve de réflexivité, qui offre à l'agent la possibilité de faire retour sur ses propres opérations mentales, pour analyser le comportement. La thèse de l'inutilité de la conscience mérite d'être prise au sérieux. J'entends montrer par la suite que, bien qu'elle soit défendable dans un contexte statique, cette thèse contredit notre intuition dès lors que la dynamique est en jeu.

On peut imaginer deux façons de défendre la thèse de l'inutilité de la conscience. La première consiste à affirmer que les causes du comportement n'importent guère ; seule la coïncidence entre comportement observé et comportement prédit fait office de critère. Il n'est pas utile de faire référence à ce qui se déroule dans l'esprit des agents du moment que, pour l'observateur, son comportement obéit aux canons de la maximisation. De même que l'arbre ne maximise pas en conscience la surface de son feuillage qu'il offre au soleil, de même il n'est nullement besoin de supposer un agent conscient de ce qu'il fait pour postuler qu'il maximise son utilité. Il s'agit moins de dire que les individus maximisent leur utilité, que la maximisation est légitime parce qu'elle permet d'imiter le phénomène étudié. La conscience n'est pas pertinente parce qu'elle est une cause hypothétique du modèle, alors que l'observateur s'intéresse à ses conséquences. Cette première façon de défendre la thèse de l'inutilité de la conscience va jusqu'à nier l'intérêt d'une théorie mentaliste. Même si elle valait sur le plan statique, elle n'est pas susceptible, conformément à ce que nous avons établi dans l'introduction de cette thèse, de rationaliser un changement de préférences. Aussi, ne peut-elle conduire qu'à une théorie *ad hoc* des changements de préférences, là où l'un des principaux objectifs de la délibération partielle consiste à pallier cette éventuelle *adhocité*.

La seconde façon de défendre la thèse de l'inutilité de la conscience reconnait que ce qui cause les préférences et, par voie de conséquence, le comportement, importe à l'analyste. Par contre, que l'agent ait conscience ou non de ces causes n'a pas d'intérêt : elles demeurent pas moins des causes qui déterminent à elles seules le comportement en tant qu'elles sont des causes et non en tant que l'agent en a conscience. Dire qu'un agent est poussé par une force qui le dépasse n'implique pas que cette force ne puisse se manifester par une opération de maximisation dont le résultat coïncide avec le comportement observé. La conscience n'apporte pas d'information comportementale supplémentaire. En situation de certitude, cela revient à établir, qu'en statique, les deux propositions suivantes sont équivalentes pour l'observateur :

P1 « Le comportement de Victor Hugo est intégralement causé par l'ensemble des causes C ».

P2 « Victor Hugo est conscient que son comportement est causé par l'ensemble des causes

Ce qui, dans cette citation, vaut pour le désir vaut également pour les valeurs, car dans un contexte statique, mon concept de valeur s'apparente à celui de désir.

C_1 » , « Victor Hugo n'est pas conscient que son comportement est causé par l'ensemble des causes C_2 » et « Le comportement de Victor Hugo est intégralement causé par $C = C_1 \cup C_2$ ».

Refuser cette équivalence c'est admettre qu'il y a quelque chose de plus dans **P2**, apporté par l'information sur la conscience que Victor Hugo a de C_1, qui change quelque chose à de son comportement. C'est dire que la conscience de Victor Hugo doit appartenir à l'ensemble C_1 des causes qui régissent son comportement, non plus en tant qu'état épistémique, mais en tant qu'état motivationnel. L'ensemble « Victor Hugo a conscience de C_1 » est le même que l'ensemble C_1. Mais alors la conscience que Victor Hugo a de C est bien présente dans **P1** non en tant que conscience cependant, mais en tant qu'état motivationnel. Il n'y aurait donc rien tant en matière d'observation qu'en matière de prédiction pour distinguer **P1** de **P2** du point de vue comportemental. En conséquence nous pouvons nous permettre d'être agnostique sur la conscience de l'agent car *in fine*, si seule la conscience de C cause son comportement, ce n'est pas en tant qu'élément cognitif, mais en tant qu'état motivationnel. Il est indifférent de savoir s'il délibère de façon consciente ou non car, en situation de certitude, soit la conscience n'est pas un état motivationnel susceptible d'agir comme cause du comportement, soit elle est un état motivationnel, mais en ce sens seulement qu'elle produit son effet. La conscience peut être une cause du comportement mais, en statique, il n'y pas de raison de lui accorder un statut motivationnel particulier.

En tant qu'il s'appuie sur le concept de conscience, le mécanisme de la délibération partielle est soumis à cette seconde objection. Il est donc nécessaire de répondre à la thèse de l'inutilité de la conscience. Une telle réponse repose sur le principe d'efficacité dynamique de la conscience : la conscience ne cause pas les préférences à proprement dit, mais leurs potentielles transformations.

4.5 Raison et dynamique des préférences

Affirmer que les propositions **P1** et **P2** ci-dessus sont équivalents va à l'encontre de l'intuition quotidienne qui suggère que **P2** apporte une information pertinente à l'observateur. Savoir si un criminel est conscient des valeurs qui le conduisent à agir constitue, par exemple, un élément crucial pour les juges, car l'idée que la conscience est un déterminant de l'action, par lequel la volonté se fait autonome, est si ancrée dans notre mode de pensée que son usage semble, à terme, inévitable. Mais peut-on se fier à une telle intuition ? Qu'un agent soit conscient de certaines valeurs qui déterminent son comportement ne revient pas à dire qu'il maitrise tout ce qui guide son action. Le mécanisme de la délibération partielle peut-il s'accommoder d'un tel scepticisme ?

La réponse à cette question est positive. Contrairement à une théorie sans conscience, la délibération partielle considère des agents non conscients de l'intégralité du processus par lequel leurs préférences se forment. Pour cette raison, le mécanisme dont il est question

dans cette thèse laisse ouvert la possibilité que le choix effectif d'une action par l'agent ne coïncide pas avec le choix qu'il ferait s'il n'était motivé que par des valeurs dont il a conscience [39].

Cet écart entre ce que la conscience de l'agent lui dicte de faire et son comportement réel est une situation commune [40]. Jaeggi [2014] donne l'exemple d'Hélène qui se surprend régulièrement à pouffer de rire comme une jeune adolescente lorsqu'elle discute avec son compagnon. Féministe éclairée, elle réprouve ce type de comportement caractéristique de l'épouse non émancipée et qui véhicule une représentation de la femme comme un « petit être adorable et vulnérable ». Malgré sa réprobation, elle ne peut s'empêcher d'agir ainsi. Hélène éprouve un écart profond entre son comportement effectif et le comportement hypothétique que devrait produire les valeurs féministes dont elle a conscience. Avec la délibération partielle, cet écart s'explique par le fait que la relation de préférencesinduite par les valeurs auxquelles elle adhère diffère de la relation de préférencesinduite par ses raisons motivantes (les valeurs auxquelles elle adhère et dont elle a conscience).

$$\preceq_V \neq \preceq_{V \cap A}$$

Toutefois, l'exemple d'Hélène confirme que, d'un point de vue statique, il n'est pas nécessaire d'attribuer un statut motivationnel particulier aux raisons du choix par rapport à ses valeurs d'arrière-plan. Qu'Hélène soit consciente de ses valeurs féministes ne donne pas à celles-ci un statut motivationnel particulier. Sur le plan statique, la conscience est bien non pertinente. Jusqu'ici la délibération partielle est en phase avec l'argumentation de la thèse de l'inutilité de la conscience. Cependant, il n'en est rien dans un contexte dynamique.

Supposons, comme le fait Jaeggi, qu'à l'issue d'une « introspection douloureuse », Hélène découvre que ses coquetteries font écho à des valeurs d'arrière-plan, profondément enfouies, comme le besoin de protection et de sécurité ; valeurs qui ne sauraient être conciliables avec une relation amoureuse strictement égalitaire. Une fois mis au jour, le conflit opposant ces deux ensembles de valeurs contradictoires doit pouvoir se résoudre par l'abandon de l'un de ces ensembles. Conformément à la caractérisation interne de la conscience, l'effort d'introspection d'Hélène peut conduire au rejet de ses besoins de sécurité et de protection, ou de son désir d'entretenir une relation amoureuse strictement égalitaire [41]. La conscience immédiate de ces deux ensembles de raisons contradictoires est donc nécessaire au changement des préférences d'Hélène.

39. Un tel écart est parfaitement illustré par ces meurtres commis dans lieux publics, sans que personne ne vienne en aide à la victime, alors même que la sollicitude est une valeur largement partagée [Latané and Darley, 1976, Rokeach, 1973]. Une importante littérature s'est intérréssée à ce type de problème. Voir par exemple [Maio et al., 2001].

40. Ce phénomène est à l'origine de la méfiance à l'égard des expériences fondées sur de simples questionnaires.

41. En toute rigueur cela la conduit à rejeter la valeur « la sécurité est bonne » ou la valeur « l'égalité entre les sexes est bonnes ».

Est-ce suffisant pour expliquer les préférences de l'agent ? Pour que ce soit le cas, il faudrait qu'Hélène ait la capacité de changer les valeurs auxquelles elle adhère. Mais par quel artifice la conscience aurait-elle le pouvoir d'interdire la cohabitation de raisons motivantes contradictoires ? En vertu de ce qui a été établi, rien n'interdit qu'Hélène vive consciemment avec cette contradiction. Pour empêcher cette éventualité, nous avons besoin d'une quatrième hypothèse.

5 Décider de ses valeurs

Nous avons établi jusqu'à présent que les ensembles de valeurs forment des systèmes susceptibles 1) de déterminer les préférences, 2) d'être classés par une relation ($\trianglelefteq$). Dans ce schéma, la conscience confère à l'agent la capacité de se représenter ses raisons, non seulement en tant qu'elles causent son comportement, mais aussi en tant qu'elles constituent une conception potentielle du préférable. Le terme de « potentielle » est ici essentiel. Il indique que l'agent peut envisager d'adhérer à d'autres valeurs que celles auxquelles il adhère aujourd'hui.

Conforme à la définition (forte) que j'ai donné de la conscience, cette hypothèse dote l'agent d'un retour critique sur ses valeurs et le met en position épistémique de résoudre les contradictions à l'oeuvre dans son ensemble de valeurs [42]. Elle met au jour *pour* l'agent les potentiels conflits à l'œuvre entre les raisons constituantes de cet ensemble de valeurs. Encore faut qu'il dispose des moyens effectifs de résoudre ce conflit. Pour ce faire, nous avons besoin de postuler que l'agent est capable de modifier par lui-même l'ensemble des valeurs auxquelles il adhère.

Hypothèse 4 : *L'agent a la capacité de modifier son adhésion à certaines valeurs, à condition que ces dernières appartiennent à l'ensemble des valeurs dont il a conscience.*

Si cette hypothèse nous rapproche tout aussi bien de la littérature philosophique sur le concept d'autonomie que de celle sur le concept de métapréférences, elle est également corroborrée par des études empiriques menées sur les valeurs.

5.1 Efficacité dynamique des métapréférences

Notre quatrième hypothèse se fonde sur une idée amplement développée dans le corpus philosophique et qui parcourt l'ensemble des conceptions hiérarchiques de l'autonomie. Etre capable de faire un retour réflexif sur ses valeurs, suppose que l'agent dispose d'une hiérarchie de préférences. Frankfurt [1988], illustre représentant de cette tradition, soutient que l'autonomie d'une personne dépend de la conformité de ses préférences de premier niveau à ses préférences de second niveau. Dans son vocabulaire, les préférences de premier

42. Ces capacités introspectives relèvent de ce que la psychologie cognitive appelle la métacognition.

niveau portent sur des objets effectifs, l'agent préfère x à y. Les préférences de second niveau, portent sur des préférences de premier niveau, l'agent préfère *préférer* x à *préférer* y. Ma conception emprunte à celle de Frankfurt car elle insiste sur le rôle des conflits de valeurs dans la transformation des préférences.

Selon Hirschman [1984], l'organisation conceptuelle proposée par Frankfurt correspond aux métapréférences, concept inséparable de la nécessité de penser la dynamique des préférences. Selon lui, les changements de comportements constituent une preuve absolument essentielle de l'existence des métapréférences. Leur analyse se trouve enrichie par ce concept qui permet d'intégrer des changements non impulsifs, mais délibérés.

> Les changements dans les comportements de choix sont donc essentiels pour valider le concept de métapréférences ; inversement, ce concept est susceptible d'éclaicir les différentes natures des changements de préférences.
>
> [Hirschman, 1984, p.14]

Cette interdépendance entre métapreférence et variation des préférences renforce notre argument en faveur de l'efficacité dynamique de la conscience. Selon Hirschman une telle efficacité proviendrait d'une incohérence entre les préférences de premier niveau et les métapréférences.

> Si préférences et métapréférences coincident perpétuellement de sorte que l'agent est perpétuellement en paix avec lui-même [...], alors ses métapréférences peinent à mener une existence indépendante et restent dans l'ombre de ses préférences. Si, *a contrario*, ces deux types de préférences sont en perpétuel conflit de sorte que l'agent agit à l'encontre de son « meilleur jugement », alors ce n'est pas seulement l'intérêt du concept de métapréférences qui peut être ignoré du fait de son inefficacité, mais c'est également leur existence même que devient douteuse.
>
> [Hirschman, 1984, p.14]

C'est parce que l'agent a des préférences qui ne correspondent pas à ses métapréférences qu'il est enclin à changer ses préférences. Car, de même que des métapréférences indépendantes des préférences sur lesquelles elles portent seraient inefficaces, de même, si les deux concepts étaient coextensifs, celui de métapréférences serait redondant. Le principe d'efficacité dynamique des métapréférences provient donc, selon Hirschman, d'un décalage entre métapréférences et préférences.

5.2 La partielle efficacité dynamique des valeurs

Dans mon interprétation, les métapréférences sont incarnées par des valeurs en tant que ces dernières constituent des jugements sur les valeurs qui méritent d'être adoptées (hypothèse 1) et qu'elles forment un système susceptible d'être ordonné par un critère axiologique (hypothèse 2). Mais ces hypothèses ne suffisent pas à rendre compte de l'idée

d'Hirschman selon laquelle la modification des préférences surgit du décalage entre préférences et métapréférences. En outre, ce décalage semble entrer en contradiction avec l'idée que la délibération partielle n'implique pas, qu'en statique, l'agent soit contradictoire.

Si, comme le suggère la première hypothèse, les valeurs déterminent les préférences, comment se peut-il qu'il y ait décalage entre les valeurs des agents, en tant qu'elles représentent leur conceptions du préférable, et leur préférence ? C'est que le décalage tel qu'il est formulé par Hirschman ne peut être interprété du point de vue objectif de l'observateur, mais comme une expérience vécue par l'agent. C'est à ce stade que le concept de conscience entre en jeu : le décalage dont parle Hirschman correspond à une incohérence entre les préférences de l'agent et celles qu'il devrait avoir, fussent-elles déterminées exclusivement par ses raisons. Autrement dit, le décalage provient du fait que :

$$\preceq_V \neq \preceq_{V \cap A}$$

Cette situation correspond à l'exemple d'Hélène, la féministe que nous avons évoquée dans la quatrième section. On comprend donc que c'est parce que le domaine de la conscience est partiel qu'un décalage peut avoir lieu. Dès lors, la prise de conscience étend cet ensemble et est susceptible de jouer un rôle dans la transformation des préférences de l'agent.

Dans le modèle de la délibération partielle, l'agent est capable de changer ses valeurs seulement si ces dernières appartiennent au domaine de sa conscience. Parce qu'elle constitue un sous-ensemble de valeurs sur lequel l'agent peut *délibérer*, la conscience est susceptible de produire une modification partielle du système de valeurs de l'agent. Cette idée est d'ailleurs soutenue empiriquement par les travaux Maio and Olson [1998] [43]. Ils mettent en évidence l'effet de l'introspection sur les valeurs. Après avoir demandé à des sujets d'évaluer plusieurs valeurs une première fois, ils les invitent à donner des raisons d'adhérer à ces valeurs, puis leur demandent de réévaluer leur système de valeurs. Autrement dit, Olson et Maio testent l'effet de la prise de conscience par les agents des valeurs en tant que conception du préférable. Ils observent alors que les agents changent leur système de valeurs, précisément là où leur conscience a été orientée. Leurs travaux confirment les hypothèses la délibération partielle.

Je viens de justifier pourquoi l'agent serait susceptible de changer ses valeurs, dès lors que ces dernières appartiennent au domaine de la conscience. Que l'agent soit capable de modifier ses valeurs ne nous indique pas, cependant, selon quel principe il doit le faire. Cette fonction est remplie par la cinquième hypothèse.

43. On peut également citer [Wilson et al., 1993]

6 Délibération partielle et principe de maximisation

Je dois maintenant dire comment l'agent choisit de changer les valeurs auxquelles il adhère, ce qui nous permettra de comprendre le type de rationalisation à l'oeuvre dans le processus de délibération partielle. Pour le comprendre, il est nécessaire de faire appel à la dimension systémique des valeurs (hypothèse 2). C'est parce qu'il est capable de classer les valeurs selon un critère axiologique que l'agent modifie l'ensemble de valeurs auquel il adhère. Je pose donc l'hypothèse suivante :

Hypothèse 5 : *L'agent choisit, dans l'ensemble des valeurs dont il a conscience, celles qui sont maximales au sens de son critère axiologique.*

L'hypothèse 4 implique que l'agent ne peut décider d'adopter (ou d'abandonner) ses valeurs qu'à condition d'en avoir conscience. Par conséquent, le principe de maximisation évoqué dans l'hypothèse 5 s'exerce sous la contrainte des valeurs d'arrière-plan. En d'autres termes, l'agent ne peut adopter (ou abandonner) ses valeurs d'arrière-plan, même si, à l'arrière plan, ces valeurs influencent le changement d'état motivationnel de l'agent. Je propose donc de reformuler l'hypothèse précédente comme suit :

hypothèse 5' : *L'agent choisit, dans l'ensemble des valeurs dont il a conscience, celles qui lui permettent de choisir le système de valeurs maximmal au sens de son critère axiologique, conditionnellement aux valeurs auxquelles il adhère sans en avoir conscience.*

Dans ce cadre, l'agent est pourvu d'une rationalité axiologique.

6.1 L'usage d'un critère axiologique

A l'aide de la relation d'ordre $\trianglelefteq$ définie sur la collection des ensembles des valeurs, on peut déterminer la manière dont l'agent effectue son choix. Cette relation d'ordre peut être vue comme l'assignation d'un poids aux raisons, mais on peut également le voir de façon plus générale, comme une relation permettant d'associer un poids aux ensembles de valeurs. Par son usage, il est possible de déterminer les raisons que l'agent a intérêt à rendre motivantes ou non, conditionnellement à l'ensemble de ses valeurs d'arrière-plan. L'agent ne peut modifier les valeurs auxquelles il adhère qu'à condition d'en être conscient et il sélectionne le sous-ensemble de raisons qui, uni aux sous-ensembles de ses valeurs d'arrière-plan, est $\triangleleft$-maximum.

Le message de la délibération partielle est donc le suivant : rendre quelqu'un conscient d'une valeur revient à lui imposer de la prendre en compte dans son arbitrage entre différentes valeurs contradictoires (ou complémentaires) et, par conséquent, l'oblige à changer l'ensemble des valeurs auxquelles il adhère. Plusieurs cas de figure sont concevables :

— L'individu prend conscience qu'il adhère implicitement à des groupes de valeurs

contradictoires. Dans ce cas, il pourra choisir de ne plus adhérer à certaines de ces valeurs afin de résoudre cette contradiction.

— L'individu devient conscient d'une valeur qui était précédémment dormante mais qui, combinée avec d'autres valeurs, a un effet positif (négatif) au sens de la relation $\trianglelefteq$.

Le résultat de la délibération partielle est alors déterminé par la hiérarchie des systèmes de valeurs induite par $\trianglelefteq$.

6.2 Délibération partielle et autonomie

La délibération nous conduit à modifier l'idée que nous nous faisons d'un agent rationnel. La conformité des préférences de premier et second niveau constitue, selon Frankfurt [1969], le principe de rationalité fondamental par lequel l'agent peut être qualifié d'autonome [44]. Toutefois, une telle approche ne permet pas de comprendre comment un agent autonome devrait changer de préférences. Elle suppose même implicitement des préférences fixes et ne s'intéresse pas au processus qui conduit à leur formation. La définition de l'autonomie de Frankfurt est purement statique.

Pourtant, penser l'autonomie nous invite à questionner la façon dont l'agent peut se changer lui-même. Christman [1991] insiste sur le fait que l'autonomie est un concept local. L'agent peut parfaitement faire preuve d'autonomie dans certains domaines de sa vie, et être sous le joug de l'hétéronomie dans d'autres. Rien ne s'oppose à ce que l'individu qui a une phobie par exemple, soit autonome dans les domaines où sa phobie ne s'exerce pas. Les agents sont le produit d'une histoire, certaines de leurs valeurs sont si profondément inscrites en eux qu'ils n'ont pas la possibilité de les modifier.

C'est pourquoi Christman [1991] propose de comprendre l'autonomie comme la capacité à influer le *processus* par lequel l'individu constitue son histoire et ses conditionnements. Autrement dit, l'agent autonome est celui qui peut intervenir sur le processus de formation de ses préférences. En ce sens, la forme de rationalité à l'oeuvre dans la délibération partielle peut être comprise comme l'exercice, par un agent, d'un autonomie partielle.

7 Les apports de la délibération partielle

Les principaux objectifs des sections précédentes étaient de présenter la délibération partielle, de montrer ses articulations avec la théorie du choix rationnel et à justifier son usage en s'appuyant sur un principe d'efficacité dynamique (de la conscience et des valeurs). Il convient maintenant de montrer que la délibération nous permet de répondre

44. A l'instar de Frankfurt, une partie des théories contemporaines de l'autonomie ont une conception hiérarchique de l'action autonome : pour être une personne, un agent doit être capable de faire correspondre ses désirs de premier niveau à ses désirs de second niveau.

au problème que nous nous sommes posé initialement : comment trouver une voie médiane entre conception *externe* et conception *interne* de l'évolution du comportement.

C'est à cette tâche que je m'intéresse dans cette section. En premier lieu, la délibération offre une voie médiane entre deux conceptions antagonistes de l'analyse des causes du comportement : conception externe et conception interne. Elle pourrait donc constituer un terrain de dialogue entre les tenants de ces différentes approches. En outre la troisième voie explorée par la délibération partielle se singularise par l'accent qu'elle met sur la *complémentarité* entre ces deux conceptions de la détermination du comportement, la plupart des tentatives de conciliation consistant à insister sur leur *substituabilité* ou leur *juxtaposition*. Ce faisant, elle permet d'envisager la formation de préférences nouvelles, mais dont la nouveauté ne procède pas exclusivement de l'aléa.

Cela ne signifie pas que la délibération partielle ait la prétention d'unifier tous les discours sur les changements de préférences. Car, en second lieu, je défends l'idée qu'il n'y a pas de nécessité à ce que la délibération partielle soit susceptible de rendre compte de tout ce qui relève d'un changement de préférences. S'il est difficile *a priori* de dire quel comportement peut être exclu des éventuelles applications de la délibération partielle, le chapitre suivant offre un certain nombre de conditions caractérisant les changements de préférences induits par la délibération partielle.

7.1 Un changement de préférences qui résulte de la combinaison des causes internes et de causes externes

L'un des avantages conceptuels de la délibération partielle consiste en ce qu'elle permet d'ouvrir une troisième voie entre de conception externes, selon laquelle les changements de préférences résultent d'une réception passive de l'influence du monde extérieur et une conception interne qui, *a contrario*, fait de l'agent l'unique acteur de ses changements de préférences.Temoin de l'importance d'une telle opposition Jon Elster écrit :

> Une des lignes de clivage les plus tenaces à l'intérieur du domaine des sciences sociales est celle qui oppose deux formes de pensée que l'on associe respectivement aux noms de Adam Smith et de Emile Durkheim : c'est l'opposition entre *homo oeconomicus* et *homo sociologicus*. Celui-là est guidé par une rationalité instrumentale, tandis que le comportement de celui-ci est dicté par des normes sociales. Le premier est « tiré » par la perspective d'avantages à venir alors que le second est « poussé » de derrière par des forces quasi inertielles.
>
> [Elster, 1989, p. 99]

Pour bien comprendre en quoi la délibération partielle offre une troisième voie qui intègre ces approches, je développe un tableau non-exhaustif de ces approches.

La conception externe

Avec la conception externe, la formation des préférences de l'agent résulte de l'influence de causes externes, que l'agent ne reçoit que de façon passive. [45] Cette conception sert de paradigme constitutif à de nombreux programmes de recherche des sciences sociales. C'est le cas de la sociologie structuraliste, qui voit dans l'existence de structures s'imposant aux individus « une condition de la scientificité du discours sociologique ». Durkheim affirme par exemple que « la sociologie ne pouvait naître que si l'idée déterministe, fortement établie dans les sciences physiques et naturelles, était enfin étendue à l'ordre social. » Mais ce type de courant n'est pas seulement l'apanage des sociologues. Inspirés par la biologie, d'importants courants de la théorie des jeux évolutionnaire tendent à faire de l'agent un individu passif soumis à des forces aveugles.

La force de ces approches tient à ce qu'elles permettent de rendre compte d'un grand ensemble de comportements qui obéissent à des règles irréflechies, que l'agent appliquent alors qu'elles echappent à sa conscience. Il n'est pas capable d'expliciter ces règles de lui-même et donc, d'en inverser les effets sur son comportement. L'explication qu'elles donnent des changements de comportement de l'agent est simple : l'agent change parce qu'il est aux prises avec des forces qui lui parviennent de l'extérieur et qui changent selon le contexte. Nul besoin d'ouvrir la boite noire de son esprit.

Le problème de ce type de conception tient à ce qu'elle suppose de faire l'économie de concepts mentaux dont il est difficile d'affirmer qu'ils n'interviennent à aucun moment dans la détermination du comportement de l'agent. Cela conduit, par exemple, à faire de la conscience un épiphénomène et à dénier l'importance de la psychologie du sens commun. Pourtant, dans la vie de tous les jours, pour comprendre le comportement de nos contemporains, nous utilisons des concepts comme ceux de croyance, de conscience ou de valeurs. Ces concepts, nous les utilisons pour rationaliser le comportement de nos semblables. Or, s'ils ne jouaient strictement aucun rôle, il serait difficile, d'un point de vue externe, de comprendre ce qui justifie leur usage dans un contexte quotidien. C'est bien parce qu'ils ont une certaine efficacité qu'ils semblent, d'un point de vue fonctionnel, nécessaires.

La conception interne

La conception interne soutient que le choix de l'agent résulte d'une maximisation délibérée de sa satisfaction [46]. Si cette interprétation suppose que l'agent choisisse en fonction de préférences qui sont stables, il est possible de l'étendre à la théorie des changements de préférences. C'est la proposition de Becker pour qui les changements de préférences relèvent du fait que l'agent investit dans son capital social et personnel en maximisant une

45. Une telle conception est l'apanage d'une littérature philosophique si diverse qu'il est impossible d'en restituer la richesse ici.

46. Une telle interprétation littérale de la théorie du choix rationnel ne saurait représenter qu'une partie réduite des défenseurs de la théorie du choix rationnel.

fonction d'utilité.

L'avantage est que, selon les propos de David Lewis, un tel type d'approche permet de rendre compte de l'usage que nous faisons des concepts de la psychologie de sens commun. Elle intégre l'idée que l'agent observé des sciences sociales est susceptible de réagir au contenu des théories déployées par ces mêmes sciences sociales. Cet avantage n'est pas négligeable. Il permet de distinguer les sciences sociales des sciences de la nature.

Dans le contexte qui est le nôtre, le problème posé par ce type d'approches est, qu'à elles seules, elles permettent difficilement de comprendre pourquoi l'agent change de préférences. Car si ses changements de préférences devaient résulter d'une maximisation parfaitement autonome, comment se fait-il que les agents ne choisissent pas d'emblée les préférences qui leur convienent le mieux ? Cela suggère que c'est nécessairement l'apparition d'éléments externes pour l'agent, qui conduisent au changement.

La délibération partielle comme voie mediane

La délibération partielle jette les bases pour résoudre la tension entre ces deux conceptions. Elle suppose que, lorsqu'il choisit ses actions, l'agent est effectivement déterminé par les valeurs auxquelles il adhère. En ce sens, la délibération partielle est déterministe sur le plan statique : l'agent est passivement soumis à l'exercice des causes qui déterminent son comportement à un moment donné.

En revanche, l'agent est susceptible de s'auto-déterminer, c'est-à-dire qu'il est capable de choisir les valeurs qui induisent son choix. Face à son histoire personnelle, l'agent n'est pas passif. Il est pourvu d'une capacité à choisir délibérément de changer sa trajectoire. Capacité qui, néanmoins, est limitée de deux manières par l'empire de la conscience de l'agent.

— D'abord, dans son choix délibéré de changer ses préférences, l'agent est orienté par des valeurs d'arrière-plan, sur lesquelles il n'a aucune prise.

— Ensuite, la délibération partielle est compatible de avec l'idée que l'agent n'est pas capable de choisir ce dont il prend conscience, qu'il soit soumis en la matière à des influences extérieures, comme des expériences qui s'imposent à lui, à un cadre social particulier ou à des activités contraintes.

La délibération partielle permet donc de penser une forme d'autonomie au sein de laquelle l'agent est activement engagé dans la modification de ses préférences, sans pour autant en être l'auteur isolé. Ce faisant, elle propose une voie médiane entre la conception externe et la conception interne.

Une voie médiane qui explique l'apparition de préférences nouvelles

La délibération partielle n'est pas la seule tentative de conciliation d'une conception interne et d'une conception externe. J'ai déjà cité, à la fin de l'introduction de cette thèse,

de nombreuses tentatives de conciliation. Je voudrais maitenant tenter de les catégoriser. Il existe trois façons de penser une articulation entre causes externes et causes internes de la détermination du comportement. Leur point commun est qu'elles reconnaissent que ces deux tendances s'expriment de concert, chacune induisant des normes différentes qui interragissent. Leur différence peut être envisagée à travers le produit de leur interaction, i.e., par les changements de comportements induits.

La première repose sur la *juxtaposition* de ces causes. Elle consiste à penser que, globalement, le comportement de l'agent obéit à une norme de rationalité, mais que l'exercice de cette norme est bruité, soumis à des biais de raisonnement. Le comportement se comprend alors comme une déviation vis-à-vis de cette norme. Une telle idée est par exemple l'apanage de la théorie choix stochastique.

La seconde repose sur la *substituabilité* et consiste à affirmer que, selon le contexte, l'agent est influencé par des causes externes, appliquant aveuglément des règles de comportement internalisées, ou bien à des causes internes, l'agent prenant le temps de déployer sa logique pour parvenir à adopter le comportement adéquat. On retrouve ce type proposition dans le fameux *Système 1, système 2* de Kahneman [2012].

Sous ces deux interprétations, le comportement effectif des agents est simplement balancé entre deux pôles hétérogènes. Compris comme équilibre entre ces causes hétérogènes, le comportement de l'agent ne peut être nouveau.

La troisième conçoit l'interaction entre causes externes et causes internes dans leur complémentarité au sens où c'est la coexistence de ces deux types de causes qui engendre le changement de préférences. Autrement dit, c'est parce que l'agent réflechit sur ses propres déterminations (les valeurs auxquelles il adhère), mais qu'il ne le fait de façon limitée que le comportement de l'agent peut changer. L'exercice des causes externes, ne vient donc pas se surajouter ou se substituer aux normes que devrait pousuivre un agent mû exclusivement par le raisonnement. Elles sont, dans leurs interactions avec le raisonnement, productrices du comportement.

Cette troisième voie permet d'envisager un agent qui serait à l'origine de comportements nouveaux. Car un individu conscient de ses capacités limitées a tendance à chercher à résoudre les tensions qu'il rencontre par la recherche de nouvelles persepctives, i.e., en essayant de prendre conscience de nouvelles valeurs.

La délibération partielle s'engage dans cette troisième voie. Elle propose en effet de voir la coexistence des causes internes et des causes externes comme la condition de possibilité des changements de préférences. Dans l'interprétation qu'elle propose, un individu capable d'user sans limite de ses capacités délibératives, de la force de son raisonnement, serait tout simplement conscient de toutes les valeurs possibles et envisageables. Il ne changerait pas de préférences puisqu'il choisirait toujours le système de valeur classé au plus haut dans la hiérarchie donnée par le critère axiologique. Il aurait la capacité de changer son système de valeur, mais n'aurait jamais intérêt à ce changement.

A l'inverse, un individu dépourvu de capacité délibérative serait conscient d'aucune

des valeurs par lesquelles il est traversé. Dans ces conditions, son système de valeur ne changerait pas non plus. L'agent aurait intérêt à changer son système de valeur, mais pas la capacité de le faire. C'est donc bien l'interaction entre causes internes et causes externes qui produit le changement de préférences.

Bien que le travail fourni dans cette thèse ne permet pas, à ce stade, de l'envisager comme telle, la délibération partielle peut être étendue de façon à modéliser l'émergence de nouvelles préférences. L'avènement de ces nouvelles préférences n'est pas dû, comme dans la théorie darwinienne, à l'intervention de l'aléatoire. Il procède de la conscience qu'à l'agent de son incomplétude. En un sens, la délibération partielle fait de pari de reconcilier la théorie de la rationalité avec l'idée que l'incomplétude de l'individu le pousse à devenir autre.

7.2 Tous les changements de préférences relèvent-ils de la délibération partielle ?

Avec la délibération partielle, je ne prétends pas représenter l'ensemble des changements de préférences possibles. Si tel était le cas, il faudrait que chaque changement de préférences :

1. soit précédé par une prise de conscience,
2. suppose que l'agent change ses valeurs en premier lieu, et que ses préférences, et donc son choix, changent dans un second temps.

Il n'est pas évident que tous les changements de préférences satisfassent conjointement ces critères.

D'abord, rien n'exlut qu'un changement de préférences advienne sans prise de conscience préalable. Il semble que le toxicomane voit son addiction pour une drogue se renforcer sans avoir réellement conscience de ce qui, dans son cerveau est à l'origine du manque qu'il ressent. La modification qui se produit résulte de mécanismes automatiques et modulaires ; elle ne suppose pas l'intervention d'un système central comme la conscience [Fodor, 1983]. Et si, de l'aveu des addictologues, une prise de conscience est nécessaire pour mettre un terme aux comportements addictifs, il n'est donc pas évident que le processus qui conduit à l'addiction puisse s'expliquer par délibération partielle.

Ce qui est vrai des altérations neuronales induites par certaines addictions, l'est également d'un point de vue strictement mental [47]. Certaines caractéristiques psychologiques des préférences se transforment de manière implicite, sans que la conscience entre en jeu. Comme le font remarquer Bisin and Verdier [2001*a*], les parents peuvent transmettre des valeurs de façon implicite, simplement parce que leurs enfants les imitent de façon inconsciente.

Ensuite, il est possible que le choix des options précède celui des valeurs. Bon nombre de travaux expliquent que les changements de préférences procèdent en premier lieu de

47. Si tant est que séparer les mécanismes physiologiques de mécanismes purement mentaux ait un sens.

comportements opportunistes, que l'agent rationalise *ex post* en modifiant ses valeurs. Connu sous le terme de dissonnance cognitive, ce phénomène est exemplifié par le renard de La Fontaine qui, non content de ne pouvoir accéder à une grappe de raisins, s'exclame que ces derniers sont « trop verts, et bons pour des goujats » [Elster, 1983]. [48] Ce type de phénomène, largement étudié par les sciences cognitives [49] est au coeur des explications des changements de préférences de nombreux économistes [Rabin, 1994]. Hirschman en fait par exemple une explication du « loyalisme inconscient » de ceux qui après avoir consenti un coût important pour intégrer un groupe social (financier, concours, cooptation, renommée, références des pairs), admettent plus difficilement qu'ils se sont trompées sur les valeurs portées par ce groupe que du fait que leur choix a été établi :

> [C]ette théorie affirme que si une personne, pour une raison quelconque, agit à l'encontre de ses valeurs, ou à l'encontre de ce qu'elle croit être ses valeurs, elle est en état de dissonance. Cet état étant déplaisant, la personne cherche à réduire la dissonance. Or comme elle est déjà engagée dans sa décision, et ne peut revenir dessus, elle préfèrera changer sa valeur. [Hirschman, 1984]

Ce type de comportement opportuniste, par lequel les individus modifient leurs valeurs pour rendre compte de leurs choix, constituent des comportements spinozistes [50], là où la délibération partielle se fait kantienne.

Chacune de ces eventualités suggèrent l'existence de changements de préférences qui ne répondent pas directement à la logique de la délibération partielle. Rien n'indique cependant qu'elles ne puissent être complémentaires, voire enrichies par l'analyse que de la délibération partielle. Il se pourrait que l'addiction procède d'une perturbation du mode de prise de conscience par l'individu des déterminants de son comportement. A ce titre, la plupart des thérapies contre les addictions promeuvent le recours à la prise de conscience. De même, il est possible d'expliquer l'existence de coyances motivées, par lesquelles l'agent modifie son système de croyances pour l'adapter à son choix par une connexion entre prise de conscience de l'agent et son système de valeur. En définitive, ce n'est qu'à l'issue d'une étude théorique complète de la dynamique conjointe de la conscience et des préférences, et de ses conséquences conceptuelles, que nous serons à même d'invalider de la délibération partielle pour décrire certains comportements.

Dans cette thèse, je n'aborde la question du lien entre dynamique de la conscience et des préférences que de façon partielle ; dans un ensemble très spécifique de situations. En ce sens, l'étude conduite n'offre qu'une réponse incomplète aux capacités explicatives de la délibération partielle.

48. Voir [Hill, 2009] pour traitement rigoureux de cet exemple.

49. Johansson et al. [2005] sugèrent par exemple que les agents sont capables de justifier des choix qu'ils n'ont pas fait, lorsque l'expérimentateur leur prétend qu'ils ont fait ces choix.

50. Je fais référence ici à la fameuse scolie du livre I de l'*Ethique* où Spinoza affirme que « Nous ne désirons pas une chose parce que nous la jugeons bonne, mais nous la jugeons bonne parce que nous la désirons. » Spinoza

8 Remarques finales

En résumé, dans la délibération partielle, le changement de préférences provient d'une prise de conscience, qui permet à l'agent d'observer une potentielle contradiction entre les valeurs auxquelles il adhère. De cette observation, et de sa capacité à changer les valeurs auxquelles il adhère, découle le passage d'un ensemble de valeurs auxquelles il adhère à un autre. On retrouve donc l'idée évoquée dans la première section de ce chapitre : la règle de transition est une fonction de V, l'ensemble des valeurs auxquelles il adhère en *ex ante*, ainsi que de A, l'ensemble des valeurs dont l'agent a conscience en *ex post*.

$$V' \in \rightarrow_{V,A}$$

Ce changement doit alors produire son effet sur les préférences de l'agent, assurant le passage de $\preceq_V$ à $\preceq_{V'}$.

Comme le suggère cette présentation, la fonction $\rightarrow$ n'a aucune raison de déboucher sur un unique ensemble V'. Il se peut, selon la relation $\trianglelefteq$ qu'un couple (V, A) corresponde à plusieurs images. Aussi convient-il d'étudier les propriétés du critère axiologique $\trianglelefteq$, aboutissant à un mécanisme de délibération partielle déterministe, i.e., associant à tout couple (V, A) une unique image. De même, selon la structure de la relation $\trianglelefteq$, cette application n'est pas nécessairement surjective au sens où elle peut rendre impossible certaines images. Le chapitre suivant discute les propriétés de $\trianglelefteq$ qui permettent de satisfaire ces propriétés.

Etant donné le rôle fondamental joué par la prise de conscience dans ce schéma, il pourrait être légitime de questionner la manière dont elle a lieu. Cela revient à interroger la dynamique de la conscience de l'individu comme le font par exemple Hill [2010] ou Van Benthem and Velázquez-Quesada [2010]. A cette question les réponses sont multiples. Comme je l'ai suggéré à la section 4, on peut par exemple supposer que le domaine d'extension de la conscience varie avec le contexte, ou bien qu'il varie avec les ressources linguistiques de l'agent, et donc avec sa capacité à poser des mots sur le réel. Il est également envisageable d'interroger la relation entre l'état motivationnel dans lequel se trouve l'agent et la dynamique de sa prise de conscience. Une telle éventualité pourrait s'avérer féconde pour traiter de mécanismes psychologiques très documentés et dont il a déjà été suggéré qu'ils jouent un rôle important dans la formation des préférences. On peut par exemple penser à la prise de ses désirs pour des réalités ou à la dissonance cognitive. Dans cette thèse, toutefois, je considère la prise de conscience comme exogène. C'est par exemple ce que produit l'action du locuteur lorsqu'il offre à la conscience de son interlocuteur la considération de nouvelles valeurs, situation qui est étudiée dans le chapitre 3 de cette thèse.

Chapter 2

RATIONALIZING PREFERENCES FORMATION BY PARTIAL DELIBERATION

Accounting for preference formation is one of the most promising ways to rationalize unexplained behavior and to reconcile rational choice theory with competing paradigms in other social sciences. Indeed, preference formation is at work in many economic contexts such as advertising or the evolution of political preferences. They have been used for instance to capture the evolution of preferences for redistribution [Alesina and La Ferrara, 2005]; the evolution of attitudes toward uncertainty [Netzer, 2009]; the evolution of norms like those of cooperation and competition; the way "moral reasoning" [Bénabou, Falk and Tirole, 2018, n.d.] can alter behaviors. However, so far economic models incorporate preference changes in order to study their consequences rather than their causes. By doing so, they cannot address two major issues of preference changes: a methodological issue and a conceptual issue.

First, incorporating preference changes within the toolbox of economic theory raises a methodological problem. As Grüne-Yanoff and Hansson [2009, p.7] put it, "it is possible to explain almost anything on the unrestricted hypothesis that consumers' preferences are changing". In other words, an unrestricted concept of preference formation may be theoretically useless since it would produce *ad hoc* explanations. This criticism has also been famously raised by Stigler and Becker [1977, p.89] who claimed that

> "[N]o significant behavior has been illuminated by assumptions... of unstable tastes. Instead, they... have been a convenient crutch to lean on when the analysis has bogged down. They give the appearance of considered judgement, yet really have only been ad hoc arguments that disguise analytical failures".

To avoid this pitfall, it is necessary to build a theory of preference changes that explains why some preference transformations are credible for a rational decision maker (DM),

while others are not. In the words of Grüne-Yanoff and Hansson [2009, p.7], "[t]he first thing that economists need in order to incorporate preference changes in their model is an appropriate theoretical structure". Such a theory must make explicit the connection between preference changes and choice reversals. Specifically, it should explain by which mechanism past choices can alter the DM's preference, inducing new choices. In this paper, I provide such a theory.

Second, a conceptual problem opposes the approaches grounded on rational choice behavior and those emphasizing external conditioning. While the rational choice approach tends to conceptualize decisions as resulting from a forward looking DM who *internally* maximizes a *given* preference relation, for many sociologists or biologists, preferences are determined *externally*, by forces - like social or biological determinations - which are not under the DM's control *per se*. The theory I provide in this paper helps to reconcile these approaches. It is based on the idea that the DM adapts himself *consciously* to the effect of external influences. It relies on what logicians have called a specific *mental trigger*, i.e., a psychological pattern that explains the change and conditions its logical structure [Van Benthem and Liu, 2007]. [1]. The notion of mental trigger is also essential for methodological concerns since it allows for *explaining* and characterizing a type of preference change.

Thus, to address these issues, it is necessary to build a theory that explains preference changes driven by a specific mental trigger (T) and that makes explicit the conditions which restrict the set of possible preference changes that can mentally be triggered by (T).

This article presents a theory providing these conditions using the fact that the DM can change her awareness of the values that may determine her preference. It models in a very general way an awareness-driven mechanism of preference formation: the process of *partial deliberation on values*. With such generality the model leaves room for many interpretations of the connection between preference changes and choice reversals and the potential normative constraints they imply. The paper also discusses formally narrower conditions. Partial deliberation is based on five hypotheses whose conceptual grounds have been philosophically discussed in the previous chapter.

1. *The preference of the DM is induced by sets of values.* These values can be thought as evaluative statements like "X is good, fair, or legitimate".
2. The DM is *partially aware* of some of these values. By aware, I mean that she is able to reflect either on these values. Hence, the preference of DM is grounded on both foreground values, of which the DM is aware, and on background values, of which she is not.
3. Sets of values form systems that can be ordered by what I refer to as an *axiological*

1. The fact that this trigger is specific implies that a theory of preference changes can hardly be exhaustive. There may be several triggers inducing preference changes.

criterion, denoted by the binary order $\trianglelefteq$ (thus, transitive, reflexive and complete). Conceptually, this relation indicates which one of two value sets is axiologically better than other. It is based on the idea that some values may contradict (in a broad sense) each other while others may complete each other.

4. The DM can only reject (or adhere to) values she is aware of. [2]
5. Based on her background values, the DM chooses the best value system she can reach within her awareness ("best" in the sense of the axiological criterion $\trianglelefteq$).

To summarize, partial deliberation is based on the idea that the DM deliberates about the values she may adhere to. However, her ability to deliberate is partial because it is based on the restricted set of values she is aware of. Technically, she chooses the value system *maximizing* the axiological criterion, her choice being bounded by her awareness.

Crucially, partial deliberation relies on the idea that the DM does not permanently question every value. This claim is consistent with what certain psychological experiments indicate. Maio and Olson [1998]'s experiment, for instance, suggests that values are truisms that the DM avoids questioning. However, these authors also show that people tend to change their values when encouraged to reflect on the arguments that support them. In other words, people do not deliberate systematically on their values, but when they do so, they are more likely to adopt (or to reject) them than when they don't.

Partial deliberation provides the framework to rationalize such observations. It is based on the idea that the deliberative abilities that yield preference formation rely on two components:

1. the ability to rank value systems given by the axiological criterion. One may see this criterion as representing the axiological rationality of the DM mentionned by the sociologist Boudon [1997],
2. the ability of the DM to reflect on some of her values and to (partially) use the axiological criterion, by choosing the best value system her awareness makes feasible.

Hence, to be rationalized by partial deliberation a set of preference changes that are driven by awareness changes has to be structured so that there exists an order (i.e., the axiological criterion denoted by $\trianglelefteq$) such that each preference change can be explained by the fact that the DM maximizes this order conditionally on her awareness. The first result of this article is to provide an axiomatic characterization that guarantees that this is the case. More concretely, a preference change follows a preference formation rule: a ternary relation between 1) the value systems that the DM adheres to *ex ante*, 2) the value system she adheres to *ex post* and 3) her awareness *ex post*. The axioms dealing with the theoretical structure of this ternary relation ensure that each preference change is induced by an axiological criterion capturing the five hypotheses of partial deliberation mentioned above.

2. As mentioned in the first chapter (p. 61), a reader that is skeptic about awareness may define awareness only by this ability.

In principle, the existence of such an axiological criterion allows to rule out a first kind of preference change. Since partial deliberation is based on the idea that the DM chooses, at each period, the value system that maximizes the axiological criterion, *a DM never turns her value system into another value system if the latter is strictly less axiologically ranked than the former.* The only way a DM can turn one system to another is by choosing the value system that is ranked best among the two value systems.

However, this is not the only reason why partial deliberation can rule out some preference changes. *A transition between two value systems may be impossible in any way.* In that kind of situation, if a DM adheres to any of these two value systems, then it would be irrational to turn her system into the other. This comes from the fact that, in general, an axiological criterion inducing a preference change is not unique: there may be no sequence of awareness changes that yields a change from a value system to another. As a second step, I investigate some of these situations and exhibit how they relate to the axiological criterion.

With this general characterization, I then relate the concept of *preference formation rule induced by partial deliberation* to a *choice reversal rule that is driven by partial deliberation.* This connection is established by using the concept of *weighing relation* borrowed from Dietrich and List [2013*b*]'s work. I show that even when the axiological criterion and the weighing relation are similar, the DM may end up choosing an option that is axiologically dominated by other options available to her. This simple result arises from the primacy that partial deliberation assigns to the choice of value systems over the choice of the options. Under partial deliberation, this is not because she has chosen an option that the DM should change her value system, but this is because she had to change her value system that she may change her behavior. In other word, partial deliberation implicitly assumes that the DM primarily adheres to values and chooses her behavior on the basis of this values. The DM does not chooses her values opportunistically to make her choice better axiologically speaking. Interestingly, this way of tackling the problem reverses the explanation of preference changes provided by cognitive dissonance. In these theories the DM changes the values (or narratives) she adheres to in order to justify her past choices. [3]

Since this characterization is very general, the results obtained are quite abstruse. Thus, I then explore two specific structures of preference changes and 1) I identify some of the preference formation paths that can be ruled out within them, and 2) I use them to draw applications with choices reversal: the endowment effect and addictive behavior. First, the *anything goes* structure relies on the principle that every value is worth adopting. I show that, with such a structure, the axiological criterion is any order following the structure of the inclusion relation. Nevertheless, this structure does not capture an appealing property of partial deliberation: sequential awareness (of value set) does not result in the same outcome as simultaneous awareness.

3. Rigorously they change their beliefs. However, in my terminology adhering to a value is like having a belief of the evaluative kind: adhering to the value v is to believe that v.

Second, the *partitioned structure* relies on the idea that values are clustered into antagonist groups. This structure gives rise to sequential changes and can be better approached when the axiological criterion incorporates a rule to rank cluster's elements.

The first section of this article gives the primitives of the model and indicates in what sense they account for the five hypotheses of partial deliberation. It results in a general definition of a preference formation rule that is induced by partial awareness. Based on such a definition, the second section characterizes the preference formation rule that are induced by an axiological criterion, thus fitting the process of partial deliberation. To do so, I exhibit six axioms structuring the preference formation rule that allows to recover the axiological criterion. Then, I relate the situations in which value systems are not connected at all with the uniqueness of the axiological criterion. The third section establishes the connection between choice and partial deliberation. The fourth section studies more specific preference formation rules and gives some insights about how constraining they are. The fifth section draws two basic applications using the structures of the fourth section.

Related literature: There have been many attempts in economics, philosophy or sociology to understand the causes and consequences of preference changes. While a complete review of the literature on preference changes is behind the scope of this article, it is worth mentioning some previous works. In economics, Cyert and DeGroot [1979] have proposed a model of *adaptive utility* in which the DM is aware that her preference may change and adapts her behavior. Some authors have provided a representation in which the DM anticipates that she might change her preference [Gul and Pesendorfer, 2005, Kreps, 1979] and speculate over this potential change. Such approaches aim at clarifying how a rational DM should deal with these changes. They focus on the consequences of these changes for rational choice theory and do not explain how and why they occur. Only a few papers focused on how preference changes are triggered. Becker [1996]'s famous *Accounting for taste* intents to explain preference changes. But the causes he emphasizes are opaque and hardly shed light on the psychological process that yields preference formation. While he rhetorically invokes habit formation and argues that imagination plays a role in preference changes, his account of how these mental states induce these changes cannot be founded on an epistemic representation of the DM. [4] State dependent preference [Hill, 2009] and reference-dependent [Kőszegi and Rabin, 2006] models have also been suggested to model preference changes. As for Dietrich and List [2016] mentioned below, these approaches are interesting to understand how the preference of the DM can change from one context to another, but they do not explain how the preference of the DM change within contexts.

By contrast, my goal is to investigate what Van Benthem and Liu [2007] call a pref-

4. In her view, imagination serves to anticipate the effect of habits formation."The analysis in this book allows people to maximize the discounted value of present and future utilities partly by spending time and other resources to produce "imagination" capital that helps them better appreciate future utilities". Becker 1996, p. 11

erence change "trigger". This trigger is based on the idea that preference changes occur when the DM becomes aware of new values. The notion of triggered preference changes is crucial in logic. Following the classical work of Von Wright [1963] on the logic of preference, logicians have recently investigate the question as to how a formal representation of preference changes can be given. Van Benthem and Liu [2007] propose two kinds of trigger that yield preference changes both implying distinct changes: suggestions and commands. In [Hansson, 1995], the DM changes her preference by updating her belief about the properties attached to options of choices. My own trigger is awareness changes. This allows me to define a framework in which the DM changes her preference internally, by maximizing an axiological criterion, but she is constrained externally by her awareness and the background values she adheres to.

1 Defining partial deliberation

Let $\hat{X}$ be a finite set of options and $\mathcal{P}(X)$ the set of all subspaces of X. These options can indifferently be seen as actions, consequences or policies. At each given point of time, the DM is provided with a preference relation over these options.

This preference relation is induced by the value system of the DM. For instance, Ana would rather vote for a politician *because* she (tacitly) believes that free market is efficient and inequalities are unfair.[5] Therefore, a value system does not only say that the DM prefers an option to another, it *explains* why she does.

Formally speaking, I denote by $\hat{V}$ the (finite) set of every of those values and by $\mathcal{V}$ the set of all subsets of $\hat{V}$. An element of $\mathcal{V}$ is called a *value system*. Each element of X can be thought as a subset of $\hat{V}$: an option corresponds to the set of every values it satisfies. Thus, I write $v \in x$ when an option $x \in X$ satisfies the value $v \in V$. In accordance with the first hypothesis of partial deliberation, the choice behavior is characterized by a family $(\preceq_V)_{V \in \mathcal{V}}$ of preference relations over X. One can also define $\sim_V$ and $\prec_V$ as the symmetric and the asymmetric part of $\preceq_V$ for each $V \in \mathcal{V}$.

I give a more elaborated account of the connection between preference relations, their attached value systems and choice in section 3. For now let me consider this intuitive example.

Example: Lets say that the value system of the DM is composed only of two values such that:

$$\hat{V} = \{\text{“}\textit{free market is good}\text{"}, \text{“}\textit{inequalities are unfair}\text{"}\}$$

Then a political project x which is consistent with free market and which deals with

5. The concept of belief should not be taken as a factual belief. Therefore it is not, in my framework, a probability distribution over a set of possible states. A value is an evaluative belief of the form “something is good". Thus saying that the DM adheres to a value is to say that the DM believes that something is good.

inequalities

$$\{\text{“}\textit{free market is good}\text{”}\ ,\ \text{“}\textit{inequalities are unfair}\text{”}\} \subseteq x$$

will be perceived as better by the DM, than a political project that is only consistent with free market.

Using this framework, I ensure that partial deliberation satisfies the first hypotheses, namely the fact that the preference of the DM is induced by values. I now need to describe how partial deliberation is consistent with the remaining four hypotheses.

To do so, I need to model a preference formation rule indicating how value systems are changing. Formally, I need a transition, $\rightarrow$, from the *ex ante* value system the DM adheres to, I generically denote V, and the *ex post* value system, I generically denote V'.

As the second hypothesis of partial deliberation puts it, this transition is driven by the fact that the DM's awareness is changing. Thus, I need to consider subsets of values the agent is aware of *ex post*. This subset is generically denoted A and it belongs to $\mathcal{V}$. I denote a preference formation rule by a family $(\rightarrow_A)_{A\in\mathcal{V}}$. $V \rightarrow_A V'$ literally means that when the DM is aware of values in A *ex post*, then her *ex ante* value system V turns into V'.[6] I denote by $V \triangle V'$ the symmetric difference between value systems V and V', i.e., the set containing the values that the DM either decided to adopt ($V' \backslash V$) during the preference formation process, or the values she decided to reject ($V \backslash V'$). In order to make the vocabulary as clear as possible, let state the following definition.

Definition 1. *A value system V lead to V' through A if $V \rightarrow_A V'$. Moreover, when considering the preference change $V \rightarrow_A V'$, the DM is said to change the motivational status of the value v whenever $v \in V \triangle V'$.*

In what follows, I assume that for any (A, V) there exists V' (potentially equal to V) such that $V \rightarrow_A V'$. This assumption is almost tautological. It simply implies that, for a given level of awareness and an *ex ante* value system, a preference formation rule leads somewhere. In the same fashion, I assume that for any (V, A) such that $\emptyset \rightarrow_A V$. This means that the DM is always better off with a value system than with no values at all. Intuitively, this is a reasonable assumption as even very relativistic or nihilist value systems seems themselves to consist in value systems.

Now that the basic notations have been given, it is possible to account for the way the preference formation rule can satisfy the remaining hypotheses of partial deliberation. The mechanism of partial deliberation proceeds as follow. *Ex ante*, the DM's value system is V. But, *ex post*, her awareness becomes A so that she can change the motivational status of the values belonging to A and, by doing so, she alters the set of values she adheres to *ex post*. Note that according to the fourth hypothesis of partial deliberation, she can only change the motivational status of values she is aware of. Therefore, the part of V the DM

6. Hence, the relation $\rightarrow \in \mathcal{V} \times \mathcal{V} \times \mathcal{V}$ can be thought as a ternary relation. However, from my knowledge, this way of seeing the problem is not helpful in this context.

is not aware of must not change and the DM cannot decide to adhere to values she is not aware of. This means that $V \rightarrow_A V$' must imply that there exists B, contained in A such that $V' = B \cup (V \backslash A)$. In other words, the DM must be able to reach V' from V with awareness A.

Definition 2. *A value system V' is said to be reachable from (V, A) if $V' = B \cup (V \backslash A)$ for some $B \subseteq A$.*

However, the fact that V' is reachable does not explain why the DM finally changes the set of values that induces her preference. According to the fifth hypothesis this change results from a (maximization) decision to change. However, this leaves open the question of the criterion that helps the DM to make her decision. In the framework of partial deliberation, this criterion is modeled by an *axiological criterion*, a relation that I denote by $\trianglelefteq \in \mathcal{V} \times \mathcal{V}$. Let's also denote $\triangleleft$ its asymmetric part and $\triangledown$ its symmetric part. The existence of an axiological criterion is based on the third hypothesis of partial deliberation: values systems can be ranked by their axiological criterion $\trianglelefteq$. This relation is on sets of values. Conceptually, it is supposed to rank the *consistency* of value sets, meaning that while some groups of values contradict each other, some other groups are, by contrast, complementing each other.

One way to support this idea consists in arguing that some value systems are more consistent than others. For instance, consider the following value system:

$$\hat{V} = \{\text{“racism is good”}, \text{“tolerance is good”}\}$$

In that case, since "racism is good" seems to contradict the fact of "tolerance good", it is reasonable to argue that an axiological criterion ranks $\{\text{“racism is good”}, \text{“tolerance good”}\}$ lower than $\{\text{“racism is good”}\}$ or $\{\text{“tolerance good”}\}$.

With the axiological criterion the fifth hypothesis of partial deliberation can be stated by saying that the DM chooses the best value system when compared to value systems she can reach with her awareness. Formally, this condition is expressed by the fact that for all B contained in A, we have $B \cup (V \backslash A) \trianglelefteq V'$.

Examples:

1. To see how partial deliberation operates on the example I have just given, suppose that, *ex ante*, the DM adheres to the value system

   $$\{\text{“racism is good”}, \text{“tolerance good”}\}$$

   And that she suddenly becomes aware of *“racism is good”*. Then, partial deliberation implies that, *ex post*, she adheres to $\{\text{“tolerance good”}\}$ since

   $$\{\text{“racism is good”}, \text{“tolerance good”}\} \triangleleft \{\text{“tolerance good”}\}$$

and the DM is aware of *"racism is good"* so that she can drop this value from her value system. What suggests this example is that even if the DM is not aware of *"tolerance good"*, this value plays a role in her process of preference formation: the DM adheres implicitly to *"tolerance good"* and she questions her adhesion to the value *"racism is good"* on the basis of her implicit adhesion to that value. These values being mutually inconsistent, she chooses to reject *"racism is good"*. In more abstract words, the DM becomes aware of some patterns determining her behavior and she decides to reject these patterns because they are not in line axiologicaly speaking the background values she adheres to. But depending on the structure induced by the axiological criterion, the DM may also decide to adhere to new values.

2. For example, take

$$\hat{V} = \{\text{"free market is fair"}, \text{"free competition is fair"}\}$$

Suppose these values are complementary:

$$\{\text{"free market is fair"}\} \lhd \{\text{"free market is fair"}, \text{"free competition is fair"}\}$$

and

$$\{\text{"free competition is fair"}\} \lhd \{\text{"free market is fair"}, \text{"free competition is fair"}\}$$

In such a situation, the DM whose value system is compounded of one of these values and who becomes aware of the other, adheres *ex post* to both of them.

3. As a third example, take a more abstract situation in which there are three values $\hat{V} = \{a, b, c\}$.

$$\{a, b, c\} \lhd \{a, c\} \lhd \{a\} \lhd \{a, b\} \lhd \{b\} \lhd \{c\} \lhd \{b, c\}$$

Assume that the DM is aware of $\{b, c\}$. This is represented in figure 2.1 a) and figure 2.2 a) by the fact these values are surrounded by a circle. Moreover, in both figures, values that are within rectangles are value the DM (could) adhere to.

Suppose first that the DM's value system *ex ante* is $\{a, b\}$. Then, as suggested by figure 2.1 b), since she is aware of $\{b, c\}$, she can reach the value systems $\{a, c\}, \{a\}, \{a, b\}$ or $\{a, b, c\}$. But $\{a, b\}$ is axiologically dominating these value systems, so that, as showed by part c) of figure 2.1 her value system remains $\{a, b\}$.

Now, assume that the DM value system *ex ante* is simply $\{b\}$. Then, with her awareness of $\{b, c\}$, the DM can reach the value systems $\{b\}, \{c\}$ or $\{b, c\}$ (figure 2.2, b). Given the axiological criterion she thus have to adhere to the system $\{b, c\}$.

The comparison between figure 2.1 and 2.2 makes it clear the role that is played by

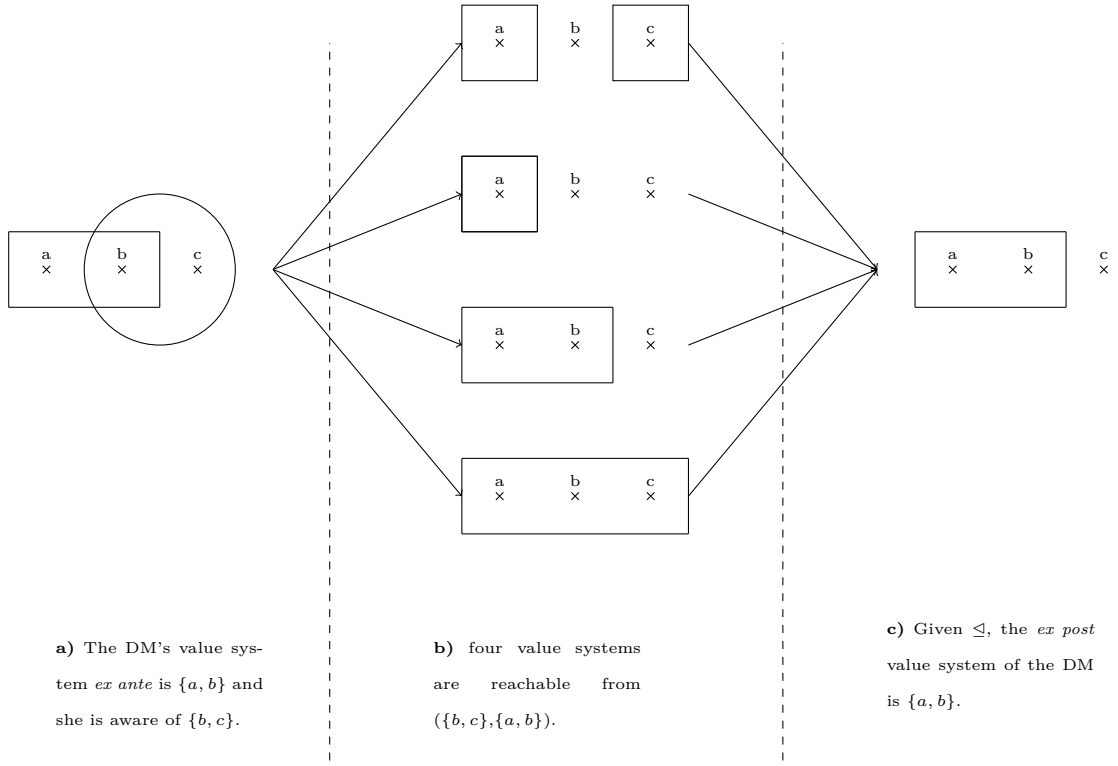

Figure 2.1 – The process of partial deliberation for example 3 with a as a background value.

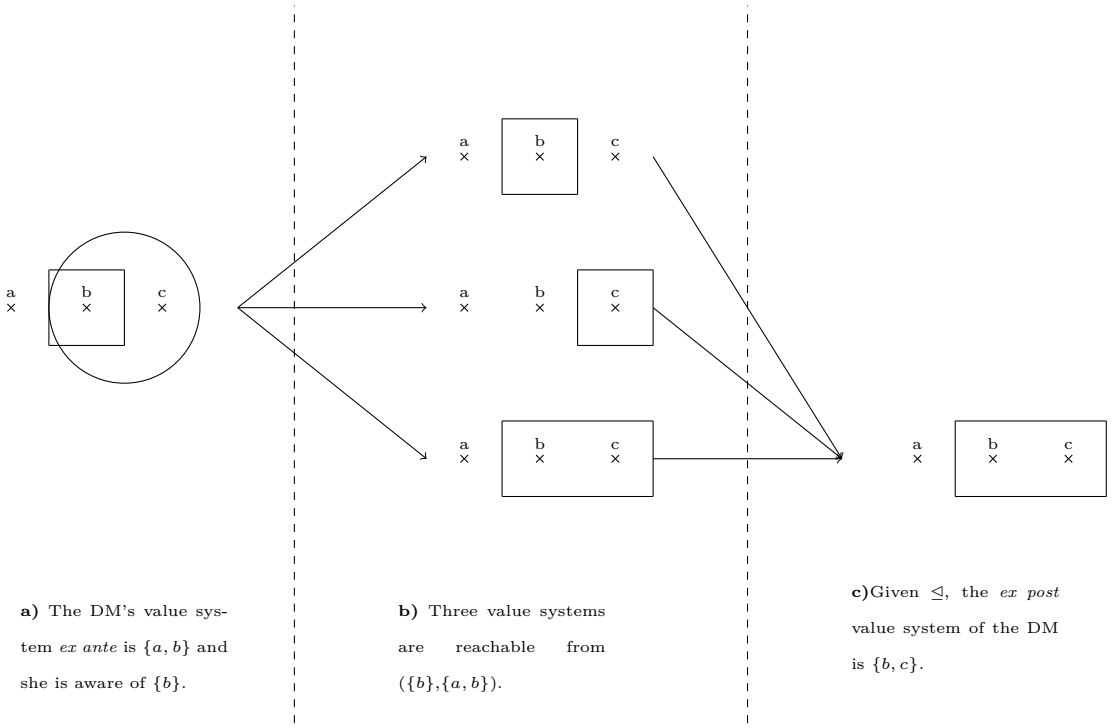

Figure 2.2 – The process of partial deliberation for example 3 without background value.

values the DM adheres to without being of aware of them. In both figures, the DM has the same awareness, but she adheres to the value a in figure 2.1 while she is not in figure 2.2. This difference has two consequences:

1. It changes the outcome *ex post* of partial deliberation since at the end of the process described by figure 2.1 it induces the DM to adhere to $\{a, b\}$, while in figure 2.2 the outcome is $\{b, c\}$.
2. The fact that the DM adheres to a implies that even if she could have adhered to c in figure 1 she does not. However, when the DM does not adheres to a, adhering to c becomes valuable.

This is an essential feature of partial deliberation. Conceptually, this can be interpreted by the fact that the DM evaluates value c through the lens of a when she adheres to a, but she does not when she does not adhere to a.

The following definition summarizes formally how partial deliberation proceeds.

Definition 3. *A preference formation rule $(\rightarrow_A)_{A\in\mathcal{V}}$ is said to be induced by partial deliberation if there exists a reflexive, transitive and complete relation $\trianglelefteq$ such that,*

$$V \rightarrow_A V' \iff \begin{cases} \exists B \subseteq A, V' = B \cup (V \backslash A) \\ \forall B' \subseteq A, B' \cup (V \backslash A) \trianglelefteq V' \end{cases} \tag{2.1}$$

Note that for a given couple (A, V), a formation rule induced by partial deliberation does not necessarily lead to an unique outcome. Two value systems may be equivalent axiologically speaking and reachable with an awareness set A. In other words, V' belongs to the following set:

$$\{B \cup (V \backslash A) : B \subseteq A \text{ and } \forall B' \subseteq A, B' \cup (V \backslash A) \trianglelefteq V'\}$$

Of course, preference formation rules are not generally induced by an axiological criterion. My first goal is to exhibit when they are, by providing an axiomatic structure that characterizes the class of preference formation rules that can be induced by such a relation. In other words, *I intend to give the conditions for the five hypotheses to be satisfied.* Providing these conditions constitutes the first step in my attempt to show which kind of preference changes the theory of partial deliberation can rule out and to tackle the criticisms arguing that preference changes are ad hoc.

2 Characterizing the axiological criterion

In this section, I give a very general characterization of the preference formation rules that are consistent with the definition of partial deliberation. Then, I investigate in a very general setting the kind of preference changes that can be ruled out.

2.1 Axioms and characterization

The first axiom simply translates the fourth hypothesis of partial deliberation. It says that partial deliberation is driven by awareness and by nothing else, i.e., the DM can only change the values she is aware of.

Axiom 1 : ***"Preference changes are driven by awareness."*** For all V, V' and A, if $V \rightarrow_A V'$ then $V \triangle V' \subseteq A$.

Consequently, the (symmetric) difference between the *ex ante* value system of the DM and her *ex post* value system has to be included in A, either because she adheres to new values in $V' \backslash V$, or because she no longer adheres to values in $V \backslash V'$.

One may argue that some preference changes do not fulfill this property. For instance, someone developing an addiction is changing her preference but she is not aware of the physiological determinants that drive this change. The smoker seeks for pleasure and she does not aim to change her preference. In this case, the preference change would not be due to awareness. It simply corresponds to a physiological response. Some preference changes can be seen, however, as driven by awareness changes. For instance, feminist groups engage in "awareness campaign" in order to change the behavior of the citizens. This axiom gives the set of value the DM can manipulate.[7] From the point of view of the axiological criterion, we need to build what makes of the *ex post* outcome of partial deliberation a *reachable maximum*.

Axiom 2 says that when the DM turns her *ex ante* values system V to another value system V' with an awareness A, then every value systems that is reachable from (A, V) either leads to V or to V'. In other words, there should be an awareness set A' such that either $B \cup (V \backslash A) \rightarrow_{A'} V'$ or $V' \rightarrow_{A'} B \cup (V \backslash A)$.

Axiom 2 : ***"Awareness connexion between mutually feasible system".*** For all V, V' and A, if $V \rightarrow_A V'$ then $\forall B \subseteq A$ then $B \cup (V \backslash A) \rightarrow_{A'} V'$ or $V' \rightarrow_{A'} B \cup (V \backslash A)$ for some A'.

Conceptually, this axiom suggests that, since V and V' are related through an awareness set, the DM must be able to compare every value systems that are reachable through this set $(B \cup (V \backslash A))$ with V'. From the point of view of the relation of the axiological criterion, this axiom can be understood as a local completeness one: it says that every feasible value system must be related to the *ex post* value system V'. After all, according to the fourth hypothesis of partial deliberation, awareness is an ability to change the motivational status of reachable values. From the point of view of the axiological criterion,

7. Another justification would consist in arguing that, in my model, it is not necessary to see the set A as the set of values the DM is aware of, but simply as the set the DM is able to manipulate.

this axiom is also the first to include the idea that the DM maximizes the consistency of the value system she adheres to, since it requires the *ex post* value system (the maximum) to be related to others reachable value systems.

In order to keep on building this principle of maximal consistency, I first need to impose a key property to partial deliberation: the axiological criterion does not vary with awareness. The slogans of this principle can be stated by saying that "awareness does not value values" or "awareness brings nothing but flexibility". Conceptually, this principle is based on the idea that awareness only makes values reachable to the DM, but it does not affect the relation that allows the DM to rank these value systems. Another way to say so is that the awareness of the DM constrains her ability to use the relation, but it does not change the axiological criterion by itself. If this property were to be violated, the axiological criterion would depend on the awareness of the DM. Thus, to satisfy equation (2.1) such a preference change needs to be ruled out.

To see the kinds of situations I have in mind, consider a preference formation rule in which these slogans are not satisfied. Suppose that $\hat{V} = \{a, b, c\}$ and that we have $\{a, b, c\} \rightarrow_{\{a,b,c\}} \{b, c\}$ and *not*-$\{a, b, c\} \rightarrow_{\{a,b\}} \{b, c\}$. While in each of these situations the DM has the same value systems *ex ante*, the value systems she adheres to *ex post* differ. This difference seems due to the fact that the DM is aware of c in the first case while she is not in the second. But in both cases the motivational status of c has not been altered. This would suggest that being aware of c changes the outcome of preference formation by itself, making the value a less valuable. The DM changes her system not only because her awareness makes the *ex post* value system reachable, but because awareness values this system. In other words, awareness brings something more than flexibility. Consequently, the axiological criterion ($\trianglelefteq$) I want to build cannot be a binary order since $\{a, b, c\} \rightarrow_{\{a,b,c\}} \{b, c\}$ seems to imply that $\{a, b, c\} \trianglelefteq \{b, c\}$ while, since $\{b, c\}$ is reachable through $\{b, c\}$ the fact that it is false that $\{a, b, c\} \rightarrow_{\{a,b\}} \{b, c\}$ implies that $\{a, b, c\} \triangleright \{b, c\}$.

The axiom that allows to rule out these situations can be stated as follows.

Axiom 3 : ***"Restricting awareness to relevant values does not change the outcome".*** If for some A, A' and B such that $B \subseteq A \subseteq A'$, $V \rightarrow_{A'} B \cup (V \backslash A)$ then $V \rightarrow_A B \cup (V \backslash A)$.

Axiom 3 simply says that since $B \cup (V \backslash A)$ is reachable from (A, V) and since it is also reachable from (A', V), the fact V leads to $B \cup (V \backslash A)$ through A' must imply that V also leads to $B \cup (V \backslash A)$ through A. Having the additional values belonging to $A' \backslash A$ in mind does not change anything *ex post* because the DM does not change the motivational status of values in $A' \backslash A$ when she is aware of them. Changing the motivational status of these values is not relevant to him.

However, this axiom does not rule out every situation in which the axiological criterion I intend to build would depend on the DM's awareness. To see this, take a preference form-

ation rule implying the following preference transformations: $\{a,b\} \rightarrow_{\{b,c,d\}} \{a,b,c\}$ and $\{a\} \rightarrow_{\{a,b,c\}} \{a,b,c\}$ but *not*-$\{a\} \rightarrow_{\{b,c,d\}} \{a,b,c\}$. Note first that since there is no inclusion between sets of awareness in each of these transitions, the third axiom cannot be applied to rule out this situation. However, we still are in a situation in which the axiological criterion that would rationalize these situations depends on the awareness of the DM. The first two statements suggest that $\{a,b,c\}$ is axiologically better than $\{a,b\}$ and $\{a\}$. However, $\{a\}$ does not lead to $\{a,b,c\}$ whereas this value system is reachable from $(\{b,c,d\},\{a\})$. Yet, every value system that are reachable from $(\{b,c,d\},\{a\})$ are also reachable from $(\{b,c,d\},\{a,b\})$. Therefore, the only reason to explain that $\{a\}$ does not lead to $\{a,b,c\}$ through $\{b,c,d\}$ while it does through $\{a,b,c\}$ is the fact that the awareness of the DM is different in those situations. This is possible only if the awareness of the DM affects the axiological criterion. Once again, from the prospect of the axiological criterion this would imply that, depending of the awareness of the DM, $\{a\} \trianglelefteq \{a,b,c\}$ and $\{a\} \triangleright \{a,b,c,\}$: the axiological criterion would not be an order. A fourth axiom is therefore required to rule out those situations.

Axiom 4 : ***"Awareness invariance of the outcome for pairwise transformation that lead to the same outcome"*** Let V and V'. If for some A, A' and $B \subseteq A$, $V \rightarrow_A V'$ and $B \cup (V \backslash A) \rightarrow_{A'} V'$ then $B \cup (V \backslash A) \rightarrow_A V'$.

To summarize in more abstract terms the preceding example, Axiom 4 says that when V leads to V' through A and a value system which is reachable from (A,V) also leads to V' but through A' it also needs to lead to V' through A. Since V' is reachable from $(A, B \cup (V \backslash A))$ and (A,V) there is no reason that it does not lead to the same outcome *ex post*. Contrary to axiom 3, it does not require that $A \subseteq A'$, but it involves a twofold relation leading to V', adding $V \rightarrow_A V'$ as a necessary condition. Indeed, the fact that $V \rightarrow_A V'$ implies that even if $A \not\subseteq A'$, the extra values belonging to $A \backslash A'$ are not relevant since both V and $B \cup (V \backslash A)$ lead to the same outcome *ex post*.

To complete the construction of the notion of maximum, I now require a last axiom, which says that the DM needs only one stage to reach the best value system as soon as this system is reachable through her awareness.

Axiom 5 : ***"Adhesion to the best value system".*** If for some A, A', $V \rightarrow_A V' \rightarrow_{A'} V''$ with $A' \subseteq A$ then $V'' \rightarrow_{A''} V'$ for some $A'' \in \mathcal{V}$.

This axiom says that, at a given level of awareness A, if she does not become aware of new values ($A' \subseteq A'$), there is no intermediate value system V' to which the DM would adhere to if this system is not as good as the best system V''. In other words, if $V \rightarrow_A V' \rightarrow_{A'} V''$ then $V \rightarrow_A V''$ directly. Thus the fact that $V \rightarrow_A V'$ indicates that V' and V'' must be axiologically equivalent. Yet, since $V' \rightarrow_{A'} V''$ we have that V' is

reachable from (A, V'') so we must have $V'' \to_A V'$. Note that this implies that the DM indeed maximizes the consistency of her system of value that she can take into through her awareness. This is another kind of independence between the awareness of the DM and the axiological criterion.

With these five axioms it is possible to demonstrate that, when changing her value system, the DM chooses the best system her awareness makes reachable to him. The next lemma accounts for such property.

Lemma 1. *If axioms* $1-5$ *are satisfied, then, for all* V, V' *and* A*, if* $V \to_A V'$ *then* $\forall B \subseteq A, B \cup (V \backslash A) \to_A V'$.[8]

Proof. We first prove the following claim: *If for some* $A \in \mathcal{V}$, $V \to_A V'$ *then* $V \cap V' \to_A V'$.
Proof of the claim: To prove the claim, suppose that for some $A \in \mathcal{V}$, $V \to_A V'$. By axiom 1 $V \backslash A \subseteq V'$ so that $V \backslash A \subseteq V \cap V'$. Thus

$$V \cap V' = B \cup (V \backslash A) \tag{2.2}$$

for some $B \subseteq A$. By axiom 2, we know that there exists A' such that either $V \cap V' \to_{A'} V'$ or $V' \to_{A'} V \cap V'$. In the first case we can apply axiom 4 and we have that $V \cap V' \to_A V'$. In the second case, because $V' \triangle (V \cap V') = V' \backslash V \subseteq A'$, by axiom 3 we have that $V' \to_{V' \backslash V} V \cap V'$. Moreover, since $V' \backslash V \subseteq V \triangle V'$ we can apply axiom 5 and thus $V \cap V' \to_{A''} V'$ for some $A'' \in \mathcal{V}$ which, by (2.2) and axiom 4 implies that $V \cap V' \to_A V'$. Which completes the proof of the claim.
Proof of lemma 1: To prove lemma 1, let $V, V', A \in \mathcal{V}$ and $B \subseteq A$ such that

$$V \to_A V' \tag{2.3}$$

If there exists A' such that $B \cup (V \backslash A) \to_{A'} V'$, then by axiom 4, there is nothing else to prove. So suppose by contradiction that there exists no such A'. Therefore by axiom 2 and (2.3), we must have that:

$$V' \to_{A'} B \cup (V \backslash A) \tag{2.4}$$

for some $A' \in \mathcal{V}$.

By axiom 1, (2.4) and (2.3), we have that $V' \backslash A' \subseteq B \cup (V \backslash A)$ and $V' \backslash A = V \backslash A \subseteq B \cup (V \backslash A)$. So $V' \backslash (A \cap A') \subseteq B \cup (V \backslash A)$. Thus $B \cup (V \backslash A) = B' \cup (V' \backslash (A \cap A'))$ for some $B' \subseteq B \subseteq A$. Moreover, by axiom 1 $B' \subseteq A'$ so $B' \subseteq A \cap A'$. Thus by axiom 3 and (2.4),

$$V' \to_{A \cap A'} B \cup (V \backslash A) \tag{2.5}$$

8. In fact it is possible to show that when axiom 1 is satisfied, axioms 2-5 are equivalent to this property.

Furthermore, by the claim we have that:

$$V \cap V' \rightarrow_A V' \tag{2.6}$$

It is clear that $(V \cap V')\backslash A \subseteq V'$ and $(V \cap V')\backslash A' \subseteq V'$. So V' can be written $V' = B'' \cup ((V \cap V')\backslash(A' \cap A))$ for some $B'' \subseteq A$. Note that $B'' \subseteq A \cap A'$, otherwise axiom 1 and (2.6) would lead to a contradiction. Thus we can apply axiom 3 to (2.6) and we have:

$$V \cap V' \rightarrow_{A \cap A'} V' \tag{2.7}$$

By (2.6), (2.5), and axiom 5 we have that $B \cup (V\backslash A) \rightarrow_{A \cap A'} V'$. Which contradicts the hypothesis and completes the proof of Lemma 1. □

Lemma 1 states that that when the value system of the DM moves from V to V' while the DM is aware of A, then every value system reachable to the DM ($B \subseteq A$), containing the set of her background values ($V\backslash A$), must also lead to V'. Conceptually it implies a twofold locality. It implies local completeness in the sense that when V' is linked to V, every set that is available from (A, V) should also be linked to V'. It implies local maximization in the sense that the DM alway chooses the "best" system. Note, once again, that this lemma implies that the axiological criterion does not depend on the awareness of the DM. In other words, being aware of a value does not imply the DM to adopt (or reject) it, but it gives the ability to adopt it.

The axioms I have given for now rely on direct transitions between value systems. However, since I want the axiological criterion to be transitive, direct transitions are not enough, and additional definitions are required.

Definition 4. *We say that there is a transition path from V to V' when there exists n, $(V_k)_{k\in\{1,\dots,n-1\}}$ and $(A_k)_{k\in\{1,\dots,n\}}$ such that:*

$$V \rightarrow_{A_1} V_1 \rightarrow_{A_2} \dots \rightarrow_{A_{n-1}} V_{n-1} \rightarrow_{A_n} V' \tag{2.8}$$

Moreover in what follows I denote:

1. *$V \rightarrow_{(A_n)} V'$ as a path of length n from V to V'.*
2. *$V \bowtie_k V'$ when there is no transition path of length lower than k between V and V'*
3. *$V \bowtie V'$ when there is no transition path at all between V and V'.*

As suggested in this definition, two value systems can be connected by a transition path while there is no direct transition between them ($\bowtie_1 \neq \bowtie$). Formally, this means that for no $A \in \mathcal{V}$, $V \rightarrow_A V'$ but there exists a sequence of awareness changes (A_n) such that $V \rightarrow_{(A_n)} V'$. The reason for this is that, if I want the axiological criterion to be transitive, I need that the existence of a transition path between two value systems to implies these values to be ordered $V \trianglelefteq V'$, even if there is no direct transition between these value

systems. To ensure this, I need the following axiom.

Axiom 6 : ***"Axiological neutrality within closed paths."*** If V, V' and A, $V \rightarrow_{(A_n)} V' \rightarrow_A V$ then $V \rightarrow_A V'$.

This axiom says that if there is a cycle linking V to V', then the two value systems should be axiologically equivalent. It ensures that the DM does not move from a value system V to V' if the former is "better" than the latter. The family $(\rightarrow_A)_{A\in\mathcal{V}}$ is therefore understood as a hierarchy that provides a ranking between different value systems.

With these axioms, it is now possible to give the following characterization of partial deliberation.

Proposition 1. *The family $(\rightarrow_A)_{A\in\mathcal{V}}$ is a preference formation rule that satisfies Axioms 1-6 if and only if it is induced by partial deliberation.*

Proof. We start with the *"only if"* part. Define the relation $\trianglelefteq^*$ as follow. For all $V, V' \in \mathcal{V}$,

$$V \trianglelefteq^* V' \iff \exists n, (V_k)_{k\in\{1,2,\dots,n-1\}} \text{ and } (A_k)_{k\in\{1,2,\dots,n\}}, V \rightarrow_{A_1} V_1 \rightarrow_{A_2} \dots \rightarrow_{A_{n-1}} V_{n-1} \rightarrow_{A_n} V' \tag{2.9}$$

It is easy to check that, $\trianglelefteq^*$ is transitive. Suppose we have.

$$V \rightarrow_A V' \tag{2.10}$$

By definition we have that, $V \trianglelefteq^* V'$. By axiom 1 it is necessary that $V\backslash A \subseteq V'$, $V'\backslash V \subseteq A$ and $V\backslash V' \subseteq A$, so $V' = B' \cup (V\backslash A)$ for some $B' \subseteq A$. Thus, we can apply Lemma 1, so that for all $B \subseteq A$, $B \cup (V\backslash A) \rightarrow_A V'$. Thus, by definition of $\trianglelefteq^*$, we have that $B \cup (V\backslash A) \trianglelefteq^* V'$ for all $B \subseteq A$.

Let V and V' be such that for all $B \subseteq A$, $B \cup (V\backslash A) \trianglelefteq^* V'$ with $V\backslash A \subseteq V' \subseteq A \cup V$. Suppose by contradiction that it is wrong that $V \rightarrow_A V'$. Therefore, by assumption we have that $V \rightarrow_A V''$ for some $V'' \neq V'$ which, by Lemma 1 and the fact that $V\backslash A \subseteq V' \subseteq A \cup V$, implies that

$$V' \rightarrow_A V'' \tag{2.11}$$

By axiom 1, $V'' = B'' \cup (V\backslash A)$ for some $B \subseteq A$, so by definition of $\trianglelefteq^*$ we have that

$$V'' \rightarrow_{A_1} V_1 \rightarrow_{A_2} \dots \rightarrow_{A_{n-1}} V_{n-1} \rightarrow_{A_n} V' \rightarrow_A V'' \tag{2.12}$$

for some $(n, (V_k)_{k\in\{1,\dots,n\}, (A_k)_{k\in\{1,\dots,n\}}})$. Therefore by axiom 6, (2.12) and (2.11) we have that $V'' \rightarrow_A V'$. Thus by axiom 5, $V \rightarrow_A V'$.

To complete the proof we need to find a complete relation $\triangleleft$ such that $\triangleleft^* \subseteq \trianglelefteq$. Since $\mathcal{V}$ is finite and $\triangleleft^*$ is transitive, there is no problem in doing so.

Which complete the *"only if"* part of the proof.

Reciprocally, suppose there exists $\trianglelefteq$ such that (2.1) is satisfied. Let V, V', and A such that $V \rightarrow_A V'$. By (2.1), $V' = (V' \cap A) \cup (V \backslash A)$ and $V \triangle V' \subseteq A$. So axiom 1 holds. Suppose $V \rightarrow_A V'$. This is equivalent to $V \trianglelefteq V'$ and the fact that for all $B \subseteq A, B \cup (V \backslash A) \trianglelefteq V'$. So for all $B \subseteq A$, $B \cup (V \backslash A) \rightarrow_A V'$ and $B \cup (V \backslash A) \bowtie_1 V'$ is false. So axiom 2 holds. Let A, A' and B such that $B \subseteq A \subseteq A'$, $V \rightarrow_{A'} B \cup (V \backslash A)$. Thus, since axiom 1 holds, $B' \cup (V \backslash A') \trianglelefteq B \cup (V \backslash A)$ for all $B' \subset A'$. Since $V \backslash A \subseteq V$, we can write that $V = B'' \cup (V \backslash A)$, with $B'' \subseteq A \cup A' = A'$. So it is true that for all $B'' \subseteq A$, $V = B'' \cup (V \backslash A) \trianglelefteq B \cup (V \backslash A)$, which implies that $V \rightarrow_A B \cup (V \backslash A)$. Therefore, axiom 3 holds. Let A, A' and B such that $B \cup (V \backslash A) \rightarrow_{A'} V'$ and $V \rightarrow_A V'$. Then for all $B' \subseteq A$, $B' \cup (V \backslash A) \trianglelefteq V'$ and $B \subseteq A$ so $B \cup (V \backslash A) \rightarrow_A V'$. Thus axiom 4 holds. Let V, V', V'', and A, A' with $A' \subseteq A$ such that $V \rightarrow_A V' \rightarrow_{A'} V''$. Note that because $A' \subseteq A$, we have that $V' \backslash A \subseteq V' \backslash A'$ and $(V' \backslash A') \backslash (V' \backslash A) = V' \cap (A')^{\complement} \cap A \subseteq A$. Thus, there exists $C \subseteq A$ such that $C \cup (V' \backslash A) = (V' \backslash A')$. Moreover, from the fact that axiom 1 holds and $V \rightarrow_A V'$, we have that $V \backslash A = V' \backslash A$, so that $C \cup V \backslash A = V' \backslash A'$. Since $V' \rightarrow_{A'} V''$ we have that $V'' = B \cup (V' \backslash A')$ for some $B \subseteq A'$. Thus, $V'' = B \cup C \cup (V \backslash A)$ with $B \cup C \subseteq A \cup A' = A$. Furthermore, from the fact that $V \rightarrow_A V'$ we have that $B' \cup (V \backslash A) \trianglelefteq V'$ for all $B' \subseteq A$. In particular, for $B' = B \cup C$, we have that $V' = B' \cup (V \backslash A) \trianglelefteq V'$. So axiom 5 holds. Finally, suppose that V, V' and A, $V \rightarrow_{(A_n)} V' \rightarrow_A V$. Therefore, by transitivity of $\trianglelefteq$, $V \triangledown V'$. Since, $A \cap V \subseteq A$ and $A \cap V' \subseteq A$ and, by axiom 1, $V \backslash A = V' \backslash A$, we have that $V' \rightarrow_A$. Therefore axiom 6 holds.

□

By definition, the existence of an axiological criterion inducing a preference formation rule indicates the preference changes that can be ruled out. A value system that is axiologically strictly dominating another value system can never lead to this system. This, however, does not mean that the latter leads to the former. First, there might be no direct transition linking the two systems ($V \bowtie_1 V'$). Second, there might be no transition path at all ($V \bowtie V'$). Understanding when this latter situation occurs is, therefore, a key issue in our investigation. The next subsection establishes the link between such a situation and the axiological criterion.

2.2 When value systems don't communicate?

Proposition 1 does not state that a given preference formation rule is induced by an unique axiological criterion. This is due to the fact that there may be no transition path between two value systems. To see this, simply consider $\hat{V} = \{a, b\}$, and two axiological criterion $\trianglelefteq^1$ and $\trianglelefteq^2$ such that $\{a\} \triangleleft^1 \{b\} \triangleleft^1 \{a, b\}$ and $\{b\} \triangleleft^2 \{a\} \triangleleft^2 \{a, b\}$. In both cases there can be no A such that $\{a\} \rightarrow_A \{b\}$ or $\{a\} \rightarrow_A \{b\}$ because, on the one hand, axiom 1 implies that such an A contains both a and b but, on the other hand, if $A = \{a, b\}$ then in both cases $\{a\} \rightarrow_A \{a, b\}$ and $\{b\} \rightarrow_A \{a, b\}$ and not $\{a, b\} \rightarrow_A \{b\}$ and $\{a\} \rightarrow_A \{a\}$.

One can interpret this by saying that a preference formation rule that is driven by awareness may under-determine what would be a complete relation. So to get a complete axiological criterion, we need to complete it artificially. As the last example illustrates, the preference formation rule may not give enough information because of the DM's rationality. Since a DM whose value system is $\{b\}$ can reach the system $\{a\}$ if and only if she can reach $\{a, b\}$, there is reason to adhere to $\{a\}$ for him.

The rationality at work in partial deliberation implies that some value systems cannot be related. This intuitively suggests that some value systems may be not commensurable. However, this term should be carefully used. Take two DM adhering to two value systems that are not commensurable. This means that none of them can persuade the other to adhere to her own system. However, this comes from the fact that there is a better value system to which they would both adhere if they were to exchange their awareness in a convenient way.

In the light of these remarks, a general characterization can be established of no feasible transition paths as the couple from $\mathcal{V} \times \mathcal{V}$ in which the relation of consistency is not unique. To do so, let $E^{\rightarrow}$ the set of $\trianglelefteq$ inducing the preference formation rule $(\rightarrow_A)_{A \in \mathcal{V}}$.

Proposition 2. *Let $(\rightarrow_A)_{A \in \mathcal{V}}$ be a preference formation rule induced by partial deliberation. Then for any $\trianglelefteq$ inducing the preference formation rule $(\rightarrow_A)_{A \in \mathcal{V}}$*

$$\bowtie = \left(\trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \right)^{sym}$$

where R^{sym} is symmetric closure of the binary relation R.

Proof. The remaining proofs of this chapter are available in the appendix.

□

It is not straightforward to give a more explicit account of preference formation rules in which there can be two value systems that are not linked by a transition path of awareness; in such a general setting, it is also not obvious to exhibit the kind of structure implying that $V \bowtie V'$. In fact, some axiological criterion allow to link all the value systems. Take $\hat{V} = \{a, b, c\}$ and an axiological criterion $\trianglelefteq$ such that:

$$\{a\} \triangleleft \{a, b\} \triangleleft \{b\} \triangleleft \{b, c\} \triangleleft \{c\} \triangleleft \{a, c\} \triangleleft \{a, b, c\}$$

This gives the following preference formation rule in which every subset of $\hat{V}$ is connected by a transition path:

$$\{a\} \rightarrow_b \{a, b\} \rightarrow_a \{b\} \rightarrow_c \{b, c\} \rightarrow_b \{c\} \rightarrow_a \{a, c\} \rightarrow_b \{a, b, c\}$$

However, the axiological criterion does not allow to make a direct transition the value system $\{a, b\}$ to the system $\{b, c\}$ since one would need the DM that adheres to $\{a, b\}$

to become aware of $\{c\}$, which would lead him to adhere to $\{a, b, c\}$ instead of $\{b, c\}$. Such a phenomenon emphasizes an interesting property of partial deliberation: there is a difference between becoming aware sequentially of a set of values and becoming aware of all these values at the same time. The reason is that by becoming aware sequentially of values the DM has less flexibility in the choice of her value system.

2.3 Additional properties on direct transition

Because the given characterization of partial deliberation and the kind of transition path it can rule out is very general, it is worth mentioning a more specific result on $\bowtie_1$. This result is based on set inclusion.

Proposition 3. *Consider a preference formation rule $(\rightarrow_A)_{A\in\mathcal{V}}$ driven by awareness. Let V, V', and V'' such that $V \subseteq V' \subseteq V''$. If for no A, either $V' \rightarrow_A V''$ or $V' \rightarrow_A V$ then $V \bowtie_1 V''$.*

To explain this result note that since for no A, either $V' \rightarrow_A V''$ or $V' \rightarrow_A V$ implies that $V' \triangleright V''$ and $V' \triangleright V$. It is to say that V' axiologically dominates V'' and V'. But since $V \subseteq V' \subseteq V''$, it is not possible to be able to reach V'' from V' or V' from V'' without being able to reach V' which is axiologically better than both of these value systems.

One may wonder if the reverse of this proposition holds. To see that it is not the case take $V = \{a\}$, $V' = \{a, b, c\}$ and $V'' = \{a, b, c, d\}$ and suppose the axiological hierarchy is such that $V' \triangleleft V'' \triangleleft V \triangleleft \{a, b\}$. Here we have that $V' \rightarrow_d V''$, $V \bowtie_1 V''$. Indeed, since $V'' \triangleleft V$ for no A , $V'' \rightarrow_A V$. Suppose then that $\exists A$, such that $V'' \rightarrow_A V$. By axiom 1, $\{b, c, d\} \subseteq A$. But since $V \triangleleft \{a, b\}$, we have that $V \rightarrow_A \{a, b\}$ for any$A \supseteq \{b, c, d\}$, and not $\{a, b\} \rightarrow_A V$. Therefore, for no A, $V'' \rightarrow_A V$ otherwise axiom 2 would be contradicted. The next section is dedicated to this task.

Because the results I have found for now are very general, it might be insightful to investigate preference formation rules that are induced by more structured axiological criterion.

3 Partial deliberation and behavioral change

So far, partial deliberation does not say how value changes and behavioral changes are connected. Yet, as stated above, such a connection should be discussed by every preference formation theory. This section aims at introducing some interpretative patterns about this connection.

3.1 Defining choice reversals

To do so, it is necessary 1) to introduce the concept of choice function and 2) to establish how it relates to value systems. In the standard choice theoretic framework, a

choice function is a mapping from a set of (feasible) options to a set of these options. Denoting this mapping by C we have for any $K \in \mathcal{K}$:

$$C(K) \subseteq, K$$

where $\mathcal{K}$ is the collection of feasible sets of options. Moreover, it is said that a preference relation $\preceq$ rationalizes C if for all $K \subseteq X$:

$$C(K) = \{x \in K : y \preceq x \text{ for all } y \in X\}$$

Since in the framework of partial deliberation I consider a family $(\preceq_V)_{V \in \mathcal{V}}$ of preference relations indexed by $V \in \mathcal{V}$, I also write that V rationalizes C if $\preceq_V$ rationalizes C. To adapt my framework to the standard formulation of choice, it is convenient to consider the set $\mathcal{C}^{\mathcal{V}}$ of choice functions that can be rationalized by $(\preceq_V)_{V \in \mathcal{V}}$.

$$\mathcal{C}^{\mathcal{V}} = \bigcup_{V \in \mathcal{V}} \{C : \mathcal{X} \to \mathcal{X} : V \text{ rationalizes } C\}$$

3.2 Revealing preference changes through behavioral changes

Note that $C \in \mathcal{C}^{\{V\}}$ means that the choice function C is rationalized by V. We can now define what I mean by a behavioral change induced by partial deliberation.

Definition 5. *A behavioral change, denoted $(\mapsto_A)_{A \in \mathcal{A}}$, is said to be induced by partial deliberation, if there exists a preference formation rule $(\to_A)_{A \in \mathcal{A}}$ induced by partial deliberation, such that for all $A \in \mathcal{V}$ the three following assertions are satisfied:*

1. *$\exists V, V' \in \mathcal{V}$, such that $V \to_A V'$*
2. *for all $K \in \mathcal{X}$, all $y \in C(K)$ and all $x \in K$ we have $x \preceq_V y$,*
3. *for all $K \in \mathcal{X}$, all $y' \in C'(K)$ and all $x' \in K$ we have $x' \preceq_{V'} y'$*

In this case we write that for all A, we have that $C \mapsto_A C'$

Literally, $C \mapsto_A C'$ says that a DM whose choice function ex ante is rationalized by the value system V, changes her behavior after becoming aware of A, adopting a choice function rationalized by V'. Note that if $(\mapsto_A)_{A \in \mathcal{A}}$ is induced by partial deliberation, by definition, we have that for any $V, V' \in \mathcal{V}$, $A \in \mathcal{A}$, any $C \in \mathcal{C}^{\{V\}}$ and any $C' \in \mathcal{C}^{\{V'\}}$:

$$V \to_A V' \implies C \mapsto_A C'$$

The reverse implication is false in general since a choice function can be rationalized be several value systems. For instance, we may have that $C \in \mathcal{C}^{\{V_1\}}$ and any $C' \in \mathcal{C}^{\{V_2\}}$ and $V_1 \to_A V_2$, implying that $C \mapsto_A C'$, while it is false that $V \to_A V'$. This is problematic since this implies that we cannot infer, in any condition, a change of value systems simply by eliciting the choices function of an individual at different periods. This suggests that

the general conditions under which a preference formation rule is fully characterized by a choice reversal would be worthwhile to study. It would allow to progress into the behavioral investigation of partial deliberation. However, these conditions are not straightforward to get, and it is beyond the scope of this paper to provide conditions characterizing the reverse implication.

However, in the context of paper, I establish the weaker condition insuring that the reverse implication holds. I first need to specify 1) how value systems determines the preference relation of the DM and, consequently, 2) how it affects choice. To do so let's remind first that options belonging to X are conceived as sets of values: $v \in x$ means that the option x satisfies v. [9]

A way to deal with this is to follow Dietrich and list's reason-based approach by arguing that the DM's preference relation can be induced by a *weighing relation* , i.e., another binary relation, $\geq \subseteq \mathcal{V} \times \mathcal{V}$, over sets of values. This relation gives the behavioral response of the DM: it states that her preferred option depends on the set of values she adheres to. This relation induces preferences in the following way

$$x \preceq_V y \iff V \cap x \leq V \cap y \text{ for all } V, V'$$

However, the meaning of this later relation differs from the axiological criterion in that it gives the behavioral response of the DM while the axiological criterion gives the way the DM should change her value system. In other words, the DM prefers an option y to the option x if the subset of her value system V that are consistent with y is better (in the sens of $\leq$) than the subset of her value system with which x is consistent. [10] Figure 2.3 summarizes how partial deliberation deals with behavioral changes.

I can now establish a condition that insure the reverse application.

Definition 6. *The weighing relation ($<$) is said to be highly discriminating if whenever $V \neq V'$ then either $V > V'$ or $V < V'$.*

This condition implies that $<$ is complete and strict.

9. I use this strategy get notation as simple as possible. According to Dietrich and List (2016) this strategy is not flawless. However, their main argument relies on the fact that they intend to deal with preference changes between contexts, while I intend to account how within the same context preferences change.

10. To account the full generality of Dietrich and List's latest contribution on that matter, I could have define value systems as sets containing sets of motivationaly relevant properties and weighting relations allowing to rank these relation ($V = (W, V)$ where W is a set of weighing relations). The results of this paper would holds anyway. This would allow the DM to change her weighing relation. However such generalization is useless for what I intend to do in this paper. Since this would require further interpretations I focus on a less general framework in which only an unique weighting relation applies to every value system.

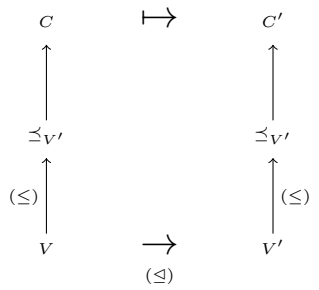

Figure 2.3 – The different relations leading to choice changes.

Proposition 4. *Assume that* $\mathcal{K} = \mathcal{P}(\hat{V})$ *and* $<$ *is highly discriminating. Then if* $(\mapsto_A)_{A\in\mathcal{A}}$ *is induced by partial deliberation, for any* $V, V' \in \mathcal{V}$, $A \in \mathcal{A}$, *any* $C \in \mathcal{C}^{\{V\}}$ *and any* $C' \in \mathcal{C}^{\{V'\}}$*:*

$$V \to_A V' \iff C \mapsto_A C'$$

for some $(\to_A)_{A\in\mathcal{A}}$.

3.3 Axiological rationality and behavioral response

What is the connection between what determines the preferences changes and the choice conducted? In other word, does the DM chooses the best available option axiologicaly speaking? Even when the axiological criterion coincide with the weighing relation it is actually possible that the DM chooses an option that is axiologicaly dominated by another option that is available to her.

Example: Let $\hat{V} = \{a, b, c\}$ and suppose that the axiological criterion and weighing relation, which are similar (so $\preceq = \trianglelefteq$), are given by :

$$\emptyset \triangleleft \{c\} \triangleleft \{a, b\} \triangleleft \{a\} \triangleleft \{b, c\}$$

Note that before becoming aware of $\{a, c\}$ we have that:

$$\{c\} \preceq_{\{a,b\}} \{a\} \text{ so that } C(\{a\}, \{c\}) = \{a\}$$

whereas ex post:

$$\{a\} \preceq_{\{b,c\}} \{c\} \text{ so that } C'(\{a\}, \{c\}) = \{c\}$$

And thus,

$$C'(\{a\},\{c\}) \triangleleft C(\{a\},\{c\}) \text{ while } C(\{a\},\{c\}) \mapsto_{\{a,c\}} C'(\{a\},\{c\})$$

Therefore, while partial deliberation relies on the idea that the DM improves axiologically her value system, it does not imply that the choice she makes is an axiological improvement when making her choice. Therefore, even if the axiological criterion and the weighing relation are similar, the DM does not choose the best option according to her ranking of values. This comes from the fact that the DM first aims an ideal option with the axiological criterion and then chooses the best option available. But it may be the case that this available option is in fact ranked worst than the best option she could have chosen if she were able to condition the choice of her value system on the options that were available to her. In more abstract terms, with partial deliberation, choice reversals are conceived as a sequential constrained maximization:

— first, the DM maximizes the axiological criterion by choosing the value system she adheres to,

— then she chooses the options on the basis on the value system she adheres to.

Thus, the primary constraint is the constraint imposed by the value system of the DM. In this case the DM is an idealist that privileges her principles of her action. This property seems, at first glance, to contrast with the lessons from the literature on cognitive dissonance or on the effects of narratives Falk and Tirole [2016] on preference. But it is possible to reverse this order, arguing that the DM first chooses the best (axiologically speaking) options available and the value system that best justifies this choice. This would implies that the axiological criterion would depend on previous choices made by the DM. In this second case the DM is pragmatic.

It is behind the scope of this paper to develop this idea but I would suggest that it may be possible to formalize Weber [2013]'s distinction between the *ethic of responsability* and the *ethic of conviction* by pursuing it.

4 Some specific axiological structures:

In this section, I investigate what I believe to be the two most natural kinds of structures inducing partial deliberation: *the anything goes structure* and *partitioned structure*. Both impose restrictions on the axiological criterion. These restrictions rely on the acceptable relations between values, requiring the relation of complementarity between values to be transitive.

For these specific axiological structures, I also study how strong is the constrain they impose to the evolution of choice.

4.1 The anything goes structure

A first structure is based on the principle that each value is good in itself. In other words, background values play no role in the process of preference formation. Thus, each time the DM becomes aware of a value, she should adhere to it. Therefore, from the point of view of the preference formation rule, an anything goes structure can be defined as follow.

Definition 7. *A preference formation rule $(\rightarrow_A)_{A\in\mathcal{V}}$ is induced by an anything goes structure if, $V \rightarrow_A V \cup A$.*

It is easy to check that a preference formation rule that relies on an anything goes structure also relies on partial deliberation. Therefore, it is induced by an axiological criterion. What kind of restriction does the anything goes structure imply on this criterion? The next proposition simply states that the DM is better off when adhering to more values.

Proposition 5. *The family $(\rightarrow_A)_{A\in\mathcal{V}}$ is a preference formation rule induced by an anything goes structure, if and only if it relies on a monotonic axiological criterion $\lhd$, i.e.:*

$$V \subseteq V' \Rightarrow V \trianglelefteq V'$$

What kind of preference changes does such structure rule out? As suggested by proposition 5, there exists a transition between two value systems only if one of them includes the other. When they are not, the axiological criterion has to be completed and proposition 2 applies.

Proposition 6. *If a preference formation rule is induced by an anything goes structure, then the following propositions are equivalent:*

1. $V \bowtie_1 V$
2. $V \bowtie V'$
3. $V \not\subseteq V'$ *and* $V' \not\subseteq V$

The consequences of proposition 6 are twofold. First, when they contain specific values, two value systems cannot be compared from the point of view of the axiological criterion and, consequently, it is impossible that a transition path conducts from one value system to the other. In fact, there is an upper transition path between these systems and their union and a lower one with their intersection. One may wonder if every structure in which $\bowtie=\bowtie_1$ is an anything goes structure. The answer is no. To see that, simply take $\hat{V} = \{a, b\}$ with $\{a\} \lhd \{a, b\} \lhd \{b\}$. Here you have that $\bowtie=\bowtie_1=\emptyset$ but it is not an anything goes structure.

Second, for preference formation rules that rely on an anything goes structure there is no difference between a sequential acquisition of values and a simultaneous one. Interpreting this is straightforward. Because there is no role for background values in the process of preference formation there is no room for inertia in the formation of preferences. To see that background values have no effect on the preference formation process, I simply remark that:

$$B \cup (V \backslash A) \rightarrow_A (P \cap A) \cup (V \backslash A) \iff B \rightarrow_A (P \cap A)$$

These two consequences are conceptually very demanding. It is therefore convenient to study a more general kind of axiological structure.

Concerning choice, it is easy to show that the anything goes structure implies that for any options x and y, the DM can reach a value system so that their is no awareness raising that can yields a choice reversal. Formally, there exists V such that there is no A such that $C \mapsto_A C'(x, y)$ with $C \in \mathcal{C}^{\{\mathcal{V}\}}$, $C(x, y) = x$ and $C'(x, y) = y$. Normatively, this feature can also be seen as very demanding since it implies that some choices are more "reasonable" axiologically speaking than others since they rely on value systems that have been reached with more awareness.

4.2 Partitioned structure

The second axiological structure of interest is based on the principle that values are structured into antagonist clusters. This means that the total set of values can be partitioned into sets belonging to $\mathcal{V}$ (1) in which all the values are complementing each other and (2) in which values of different classes are contradicting each other.

Definition 8. *Two values v and v' are said to be complementary if $v \trianglelefteq vv'$ and $v' \trianglelefteq vv'$. In this case, we write $v \sim v'$. Two values are contradictory when it is wrong that $v \sim v'$.*

From the point of view of the preference formation rule, this structure can be defined by saying that the DM adopts the value from a set of the partition P she is aware and that, given her background values, is improving her value system. But since the values from the other sets of the partition are contradicting the values in P she must also reject the values she is aware of and that do not belong to P. Therefore, we must have that:

$$B \cup (V \backslash A) \rightarrow_A (P \cap A) \cup (V \backslash A) \tag{2.13}$$

for some P.

But this property is not sufficient to account for the idea of partitioned structures. First, when (2.13) is satisfied, P is the most axiologically consistent set of the partition that the DM can adopt. This means that if she were to become aware of a new value belonging P, the DM must adopt it. Another restriction is thus required. Second, there is no reason for a set of the partition to dominate the others if the DM is neither aware

nor motivationally affected by any value of this P. A preference formation rule relying on a partitioned structure can be stated as follows.

Definition 9. *The preference formation rule $(\rightarrow_A)_{A\in\mathcal{V}}$ relies on a partitioned structure if there exists a partition of $\hat{V}$ denoted $\mathcal{P}$ such that for all (V, A) and for all $B \subseteq A$:*

$$B \cup (V\backslash A) \rightarrow_A (P \cap A) \cup (V\backslash A) \tag{2.14}$$

for some $P \in \mathcal{P}$ with :

i.) for all $v \in P$,

$$B \cup (V\backslash A) \rightarrow_{A+v} (P \cap A) \cup (V\backslash A) + v$$

[11]

ii.) $P \cap (A \cup V) \neq \emptyset$.

Note that the anything goes structure can be seen as a specific case of partitioned structure in which the partition is made of only one class. However, contrary to an anything goes structures, the background values play a key role in partitioned structures and thus it is now wrong that :

$$B \cup (V\backslash A) \rightarrow_A (P \cap A) \cup (V\backslash A) \iff B \rightarrow_A (P \cap A)$$

Since the values in B might lead to another set of the partition $P \cap A$ when $V\backslash A$: the preference change is conditioned by the background values of the DM, $V\backslash A$. Another consequence of this is that sequential changes of awareness induce different preference changes. It allows for the DM to adopt contradictory values through time. To see this let's take two examples.

Examples : Take $\{1_1, 2_1, 3_1, 4_1, 1_2, 2_2, 3_2, 4_2\}$ as the set $\hat{V}$ partitioned by $\mathcal{B} = \{\{1_1, 2_1, 3_1, 4_1\}, \{1_2, 2_2, 3_2, 4_2\}\}$. Assume that for $k, k' \in \{1, 2\}$, $\{i_k\} \triangleleft \{j_{k'}\}$ if and only if $i < j$. To understand the evolution of the value system, we need a rule that determines which of the partition dominates the other. Without that kind of rule, how to say whether $\{1_1, 2_1\}$ ranks better than $\{3_2\}$?

1. Suppose as a first example that a set V ranks better than V' when its maximal value is ranked better than the maximal value of V'. First, note that while $\{3_1, 1_2\} \rightarrow_{\{1_2,2_2\}} \{3_1\}$, we have $\{1_2\} \rightarrow_{\{1_2,2_2\}} \{1_2, 2_2\}$. In this example, since in the first case, the maximal value ex ante is 3_1, becoming aware of values dominated by 3_1 and which belong to different parts of the partition (1_2 and 2_2), leads to drop these values rather than adopting them. Conversely, when the DM has a value system $\{1_2\}$, and is aware of the same values (i.e., 1_2 and 2_2), the second part of the partition is now dominant so that the value 2_2 is now to be adopted by the DM. This means

11. We abuse notation by writing $V + vv'$ for $V \cup \{v, v'\}$ and $V - vv'$ for $V\backslash\{v, v'\}$.

that with partitioned structure, background values play a key role to determine how preferences change.

But it is also worth mentioning that with partitioned structure, the DM can adhere to contradictory values. To see this, note that:

$$\{1_1\} \rightarrow_{\{1_1,1_2,2_2\}} \{1_2, 2_2\} \rightarrow_{\{3_1,2_2\}} \{3_1, 1_2\}$$

This example show that even if, initially, the DM was neither adhering to 1_2 nor to 3_1, her final value system ends up being $1_2, 3_1$, two values that contradicts each other. Therefore, the DM can adhere to contradicting values even if these values were not in her initial background. Moreover, if we had a simultaneous change, we would get:

$$\{1_1\} \rightarrow_{\{1_1,1_2,2_2,3_1\}} \{1_1, 3_1\}$$

the *ex post* outcome being different. This suggests that a DM which awareness is growing sequentially won't end up in the same situation.

2. As a second example, one can assume that V ranks better than V' when the sum of its values is bigger. In such a case we have:

$$\{1_1\} \rightarrow_{\{1_1,1_2,2_2\}} \{1_2, 2_2\} \rightarrow_{\{2_2,3_1\}} \{1_2, 2_2\}$$

Note that with such a rule the sequentiality is satisfied with the same awareness and the same initial value system.

While partitioned structures are much more flexible than anything goes structures, they still imply a conception of value systems that is demanding: value systems are conceived as strongly antagonistic. This can be an interesting property to describe the opposition of communist values and capitalistic one for instance. However, the intuition suggests that their can be a transition path to connect value systems that can be competing in some areas.

I first need to establish the connection between partitioned structures and partial deliberation. However, to do so I need to consider only the preference formation rules that satisfy axiom 6. [12]

Lemma 2. *If a preference formation rule $(\rightarrow_A)_{A\in\mathcal{V}}$ is induced by partitioned structure and satisfies axiom 6 then it is induced by partial deliberation.*

Because a preference rule that is induced by a partitioned structure is also induced by partial deliberation, we know that there exists $\trianglelefteq$ such that (2.1) holds. It is now interesting to wonder what kind of restrictions this type of preference formation rule implies on $\trianglelefteq$. The

12. I have not been able to show that a preference formation rule that is induced by a partitioned structure satisfies this axiom. However, there are good reasons to believe that it does.

next proposition characterizes the axiological criterion that induce this type of preference formation rule.

Proposition 7. *The preference formation rule $(\rightarrow_A)_{A\in\mathcal{V}}$ is induced by a partitioned structure if and only if it relies on a clustering axiological criterion $\trianglelefteq$, i.e. satisfying for all $v \notin V$ and $v' \in V$:*

1. *If $v \sim v'$ and $v' \sim v''$ then $v \sim v''$.*
2. *If $V \trianglelefteq V + v$ and $vv' \triangleleft v$ then $V + v \triangleleft V - v' + v$.*
3. *If $V + v \trianglelefteq V$ and $v \sim v'$ then $V \triangleleft V - v'$.*
4. *For all V, if for all $v \in K$, with $K \cap V = \emptyset$, $V + v \trianglelefteq V$ then there exists $v' \in V$ such that if $v'' \sim v'$ for some $v'' \notin V$ then $v'' \notin K$.*

This proposition gives the restrictions ensuring that the axiological criterion induces a partitioned structure. The first restriction states that complementarity between values is a transitive relation. Because it is, by definition, a symmetric relation, this ensures that the reflexive closure of this relation is an equivalence relation. Therefore, complementary values can form partitions. The next restrictions intend to capture in an abstract way the diversity of the rules that determine which part of the partition dominates the others.

Restriction *2.* says that when a value v constitutes an axiological improvement of a value system V (i.e., $V \trianglelefteq V + v$), then it is always strictly better to drop values that contradict this value ($vv' \triangleleft v$) from the improved system ($V + v \triangleleft V - v' + v$). The idea is that if a value improves, axiologically speaking, a system of value, this means that it should belong to a set of the partition (not necessarily unique) that dominates the others. In other words, values from other elements of the partition contradict the dominating partition and it is better for the DM to drop them when becoming aware of them. This improvement is strict since we may have that $V \triangledown V + v$, meaning that the element of the partition to which v belongs to may be axiologically equivalent to the one v' belongs to. However, by dropping v' the axiological equivalence no longer holds and the system should be strictly improved.

Restriction *3.* is based on a similar idea. It says that when it is worth dropping every value $V + v \trianglelefteq V$ system, it is worth dropping every value that are complementary with this value ($v \sim v'$). This is due to the fact that if it is worth dropping a value from $(V + v)$ the element of the partition to which v belongs to is not strictly dominating all the other elements of the partition. Therefore it is strictly better to drop a value complement v. The reason why this domination is strict follows the same argument of the restriction *2.*

Restriction *4.* is trickier. It says that if it is always better not adding a value from a given set (K) to a system of value (V). Then, there should be a set of the partition intersecting V in which no value complements a value from K. Note that there must even be at least two sets of the partition obeying otherwise it would be false that $\forall v \in K$,

$V + v \trianglelefteq V$. This assumption is necessary to deal with the case in which the DM only rejects values.

However, the examples above make clear that the rules allowing to describe which set of the partition is preferred are necessary to determine which kind of preference change can be ruled out.

5 Applications

In this section, I suggest two potential applications of partial deliberation. Each of them incorporates an idea that has not been developed yet: the fact that awareness of the DM is oriented in a certain way by her preference. These applications take the word value in a broader sense than the definition given in chapter 1. The reader may prefer using the concept of motivating properties as Dietrich and List do.

These applications are driven by separate motivations. The first application deals with the endowment effect. It explains both the phenomenon and the fact that it depends on the DM's attitudes toward property. By applying partial deliberation to such a phenomenon, I intend to suggest that partial deliberation may provide explanation to well documented lab experiments. The relevance of these explanations can be tested, either allowing to assess the scope of behaviors that can be explained by partial deliberation, or simply demonstrating that it can be falsified.

The second application deals with addictive behavior. It is meant to suggest that while addiction does not seem to fit very well with the concept of partial deliberation, there is actually a room for giving a psychological account of addiction with a simple model of partial deliberation. This suggests once again that careful attention should be paid to the process of addiction to assess whether or not partial deliberation can shed light on this phenomenon. Such application would require further developments and, it would be beyond the scope of this paper to make explicit the whole mathematical apparatus it requires to be described in full generality.

5.1 The endowment effect

Many experimental studies highlight that the price people are willing to pay for a particular good is often less than the price they are willing to give up this good. Currently called the *endowment effect*, this phenomenon has been explained in many ways. [13] Specifically, many psychologists point out the psychological effect of ownership, arguing that when the DM becomes the owner of the good, she identifies herself with the good [Beggan, 1992]. Interestingly, this effect seems to depend on the DM deliberative capacities, suggesting that awareness plays a role in the way ownership induces the endowment effect [Carmon, Wertenbroch and Zeelenberg, 2003]. For instance, Ashby, Dickert and

13. It was initially proposed by Kahneman, Knetsch and Thaler [1990] that loss aversion explains the endowment effect. Since the person that owns an item, may feel like a loss the fact of giving up this item.

Glöckner [2012] suggest that the disparity between willingness to accept and willingness to pay increases with increases in deliberation time.

Moreover, the magnitude of this bias seems to be affected by the way individual's culture induces them to think about themselves, by specific experiences or the attitude of the DM toward consumption and selling. For instance, [Maddux et al., 2010] suggest that "the more self-enhancing a person's culture, the more likely she is to exhibit an endowment effect" [Morewedge and Giblin, 2015, p. 343]. In the same spirit, Engelmann and Hollard [2010] show that market experience tends to alleviate the effect.

Thus, there are two ways by which the endowment effect is related to preference changes. First, the fact of becoming the owner of the good may change the DM's preference, inducing her to give higher value to the good. Second, this effect of ownership may depend on the DM's attitude towards her self concept and this attitude can itself change. Although it is not the only preference change that can capture this phenomenon, partial deliberation can be used to formalize both of these effects of ownership, explaining the endowment effect, and the variation of its magnitude.

The explanation crucially relies on the hypothesis that the value *"ownership is part of one's identity"* influences the awareness of the DM. Since ownership is conceived as part of one's self, a DM simulates the possible versions of himself using the good and, by this process, she becomes aware of values attached to this good rather than values attached to the good proposed ex post. Formally, let x and y two goods. Assume also that the axiological criterion is induced by an anything goes structure and that the weighing relation has the following property.

$$y \preceq_V x \text{ and } V' = V \cup A, \text{ for some } B \subseteq x, \text{ implies that } y \preceq_{V'} x \tag{2.15}$$

By becoming the owner of x the DM changes her preference from V to $V \cup A$ with $A \subseteq x$. Thus from (2.15), $y \preceq_V x$ implies that $y \preceq_{V \cup A} x$. Consequently, every DM initially preferring x to y, should also prefer x to y, after having become the owner of x. Conversely, some DM preferring y to x may have now changed her preference. Therefore, the model predicts that it is likelier that the DM prefers a good after becoming the owner of the good. This simple idea provides an explanation of the endowment effect itself. This explanation is in line with the rather influential paper of [Strahilevitz and Loewenstein, 1998].

It is also possible to model the reason why the endowment effect depends on the cultural attitude of the DM. Suppose the DM's value system is made of the following values:

$$\{\text{ownership is part of one's identity}, \text{exchange can be worthwhile}, x, y\}$$

Where x and y are the values fulfilled by the goods. The main hypothesis, then, is that the DM's awareness depends on whether she adheres, additively, to the value *"exchange*

can be worthwhile": when she does, she tends to become aware not only of the values of the goods she owns, but also of goods she could sell. Since she values the fact of selling and she is used to simulate the values that another person could see when considering the alternative good. The endowment effect is, in this case, alleviated since her awareness embrace the values of both goods.

5.2 The addiction trap

Can partial deliberation helps to understand addictive behaviors? The idea is that the DM's preference might be determined by so many values attached to her addictive behavior that she cannot easily change the motivational status of values that drive her toward this addictive behavior. Such a mechanism is reinforced by the fact the DM's awareness can be attracted to values related to her addiction when she has behaved addictively in the past. A simple way to formalize this idea would consist in making four assumptions: an assumption about the axiological criterion, an assumption about the set of available options, an assumption about the weighing relation and an assumption about the probability that the DM becomes aware of new value.

The first assumption is that the axiological criterion relies on a specific partitioned structure compounded of two class of values. Values that drive her toward her addictive behavior, indexed by d, and values that drives her toward stopping her addictive habits, indexed nd.

$$\{P_d = \{x_d, y_d, z_d, ...\}, P_{nd} = \{x_{nd}, y_{nd}, z_{nd}, ...\}\}$$

with $P_d^{\#} = P_{nd}^{\#}$

I assume also that $\trianglelefteq$ is represented by the following function:

$$\Phi(V) = |(P_i \cap V)^{\#} - (P_{-i} \cap V)^{\#}|$$

where $A^{\#}$ is the cardinal of the set A, $i \in \{d, nd\}$ and $-i \neq i$. Thus the part of the partition that dominates the other is simply the one that contains more value and, in this case, it is always better to drop some values from the other part of the partition. In other words the addict is someone whose value system V contains more values related to her addiction: $(P_d \cap V)^{\#} > (P_{nd} \cap V)^{\#}$.

The second assumption is that, for the DM, there is no available option that overlaps these two classes of the partition. So if $x \in X$ is feasible for the DM then it should be included either in P_d or in P_{nd}. Moreover, to get things simple we can assume $X = \mathcal{P}(P^d) \cup \mathcal{P}(P^{nd})$.

The third assumption is that the weighing relation $\geq$ ranks better options that contain more values the DM adheres to. So for all $x, y \in X$:

$$x \preceq_V y \iff V \cap x \leq V \cap y \iff (V \cap x)^{\#} \leq (V \cap y)^{\#}$$

This implies that every value satisfied by an option tends to make this option more valuable to the DM when she adheres to it.

The fourth assumption is that the probability that the DM becomes aware of a value she was not adhering to is conditioned on her previous choice. Moreover, the DM is likelier to discover a new value v she was not adhering to when this value is in the part of the partition of the option she chose in the previous period.

$$p = \mathbb{P}\Big(v_i \in A^n \backslash V^{n-1} | C^{n-1}(K) \in P_i\Big) > \mathbb{P}\Big(v_i \in A^n \backslash V^{n-1} | C^{n-1}(K) \in P_{-i}\Big) = q$$

These four assumptions establish the basic structure of the model of addiction.

At period n, take a DM whose value system, V^n, contains more values that are related to drugs than values that don't. Let respectively denote by V_d^n and V_{nd}^n the values in $V^n \cap P_d$ and in $V^n \cap P_{nd}$. If $K^t = X$, we have that then:

$$V^n \cap V_{P_d}^n \geq V^n \cap y \text{ for all } y \in K \text{ and thus } C^{V^n}(K) \subseteq P_d$$

Since $p > q$ and by the hypothesis of independence and it can be shown that:

$$\mathbb{P}\Big((Argmax\Phi(V) \cap P_d)^{\#} > (Argmax\Phi(V) \cap P_{nd})^{\#} | (V_d^n)^{\#} > (V_{nd}^n)^{\#}\Big) > 1/2$$

Since it is certain that the best value system available to the DM is either $(V^n \backslash A_{nd}) \cup A_d$ or $(V^n \backslash A_d) \cup A_{nd}$, it is likelier that the value system ex post of the DM contains more values related to her addiction.

$$\mathbb{P}\Big((V_d^{n+1})^{\#} > ((V_{nd}^{n+1})^{\#} | (V_d^n)^{\#} > (V_{nd}^n)^{\#}\Big) > 1/2$$

Thus, it is likelier that the DM adopts an addictive behavior in period $n+1$. Moreover, her value system is altered in such a way that she tends to drop more and more values that are not related to her addiction.

Therefore, partial deliberation can explain how a DM is getting more and more addict. How can one explains how she can get rid of her addiction. A potential answer simply consists in restricting the set of available options of the DM for many periods. This restriction can be imposed by a third party as in the case of drug rehabilitation. Thus, let assume that there exists T such that for all $t \in \{n+1, ..., n+T\}$ we have that $K^{t+1} \subseteq P_{nd}$.

Then, the DM remains addict for many periods but her addiction is less likely to be reinforced by her awareness since she cannot choose any option related to her drug addiction. However she still is addict and need to become aware of sufficiently enough values from $P_{nd} \backslash V_{nd} \cup V_d$ to stop being addict. This simple model therefore provides interesting features. One can easily make it much more powerful with finer assumptions on the axiological criterion, the weighing relation of the DM or on the way her awareness

is moved.

6 Concluding remarks

I have presented a very general mechanism explaining preference formation, the mechanism of partial deliberation, and the type of preference changes it can rule out. The general setting can be characterized. It gives some insight on the fact that, in partial deliberation, some path of preference changes are not possible because of the DM's rationality. It is also compatible with appealing properties like the fact that between some value systems there may be no direct transition, but only by a sequence of awareness changes in which the DM loses awareness of some values and become aware of some others. This property translates the fact that partial deliberation can capture some consequences of what sociologists have called internalization processes.

However, because it can be induced by very chaotic axiological criterion, the general setting of partial deliberation gives a very abstruse representation of the kind of preference changes one can rule out. This is why I have also investigated preference formation rules that are induced by more structured axiological criterion. While the anything goes structure is easy to handle, its generalization by partitioned structures requires much more involved analysis and additional assumptions on the rule that allows to rank the partitions of values.

However, since they are much simpler these latter structures offer interesting intuitions about how restrictive partial deliberation is, either positively or normatively. They also give applications that need further investigations. These applications are based on the idea that awareness and preferences can interact retroactively. This suggests that my account of partial deliberation is incomplete. A theory of how preference can influence awareness is also required (I discuss this in in the section 2.4 of the conclusion).

Chapter 3

A THEORY OF PREFERENCE MANIPULATION

One of the main motivation to model partial deliberation is that it corresponds to what Lewis [1978] calls a platitude of folk psychology. The argument is that even if human being do not change their preference when they become aware of new values, at least, they behave as if they were able to transform the preference of others simply by making them aware of new values. It is thus interesting to wonder how an individual can base her decision on partial deliberation when she tries to manipulate the preference of another individual.

More concretely, a sender (she) often tries to change a receiver's (he) preference in order to propose him a project that corresponds to both of their preferences. For instance, a policy maker designing a tax policy may have a preference for a given tax schedule. But she cannot simply choose what she prefers. She also needs citizens to consent to the tax policy she proposes. Otherwise she would have to bear political costs, which would weaken the efficiency of her tax policy. Rather than choosing a tax policy that meets consensus, the sender may seek to manipulate the citizens' preference by using her preference manipulation technology. A close situation has been studied by Rodrik [2014] in his model of ideas versus vested interest for instance.

More formally, the general timeline to deal with this problem is the following. At period 0, the sender commits to a strategy compounded of 1) a project $\mathbf{x}$ and 2) a sequence of preference manipulation technology $(A^1, A^2, ..., A^n)$ that will be implemented at each period and that can change the receiver's preference. In each of the following periods i, the receiver reacts to A^i by changing his preference and, depending on this change, he chooses an action $\mathbf{a}$. At the final period, the sender receives her payoff that depends 1) on the project she choses at period 0, 2) on the receiver's action at each period and 3) on her own preference. Thus, the problem for the sender consists in choosing the best project coupled with the best disclosure strategy given her anticipation of the receiver's reaction to her preference manipulation technology. In this paper, I limit the analysis to the cases

of three and four periods. The specific timelines of each of these cases are given at the end of this introduction by figures 3.1 and 3.2.

Modelling this kind of situations requires accounting for the technology that allows the sender changing the receiver's preference and how this technology constrains the sender's strategy. Without any theoretical grounds specifying how this change can occur, modelling this situation is straightforward. When assuming that the preference of the receiver is fixed, a well-informed sender has simply to choose a half-way project. Conversely, when she can manipulate the receiver's preference as she wants, her best strategy is simply choosing her preferred project and ensuring that the receiver's preference fits her own preference. However, the situation is more complex when the analysis is based on a theory in which preference changes are not ad hoc, i.e., a theory in which some (and not any) preference changes are possible. On the ground of such a theory, the sender is constrained in her attempt to manipulate the receiver's preference. This paper proposes to exhibit some of these constraints.

The analysis is based on an euclidean approach to *partial deliberation*. Partial deliberation gives a psychological account of preference changes. However, to model the specific situation of preference manipulation, a concept of distance between the decision makers' preferences is now required. This is the reason why I develop an *euclidean* approach to partial deliberation. The euclidean approach to partial deliberation rests on the same five hypotheses that have been developed in chapters 1 and 2, but its mathematical apparatus is based on euclidean spaces rather than abstract sets. Similarly to the previous formulation of partial deliberation, the preference of a decision maker is induced by values. The decision makers can now assign *axiological intensities* to these values. Thus, in this framework, the utility functions of both the sender and the receiver depend upon their axiological states $\mathbf{v}$, i.e., a vector that assigns to each value i an axiological intensity $v(i)$. By the term "axiological intensity", I formally refer to a number between 0 and 1 assessing how much the decision makers adhere to a given value. Thus, when $v(i) = 0$, the decision makers do not adhere to value i at all, whereas when $v(i) = 1/2$, they adhere to value i with an intensity of $1/2$. This framework enables accounting for the distance between the preference of the sender and the preference of the receiver on the one hand, and for the distance between the decision makers' preferences and the project proposed by the sender on the other hand.

In the paper, I first show that the concept of axiological intensity between values can be based on the set theoretic approach of partial deliberation developed in chapter 2. This confirms that the euclidean framework of partial deliberation is built upon similar ideas. As in chapter 2, values form systems that can be ranked using an axiological criterion. This criterion can be represented by a function Φ that maps each vector $\mathbf{v}$ into $\mathbb{R}$. Moreover, the decision maker may be unaware of the values inducing her preference. This constrains her ability to change the axiological intensity she assigns to a value and, consequently, to maximize the function Φ representing the axiological criterion. As for partial deliberation,

in the euclidean approach, preference changes result from the decision maker's choice of altering the axiological intensity she assigns to the values she is aware of. Thus, her choice is constrained by her bounded awareness. The decision maker changes her preference but, when so doing, she may be influenced by background values, i.e, the values she adheres to unconsciously.

Since the euclidean approach to partial deliberation provides a conceptual ground to exhibit the constraints faced by a sender trying to manipulate a receiver's preference, it is natural to wonder how the former can take advantage of partial deliberation, disclosing new values to the latter. This may not be the only way a sender can manipulate a receiver's preference. But this kind of manipulation is normatively appealing since it relies on the receiver's reasoning.

As mentioned above, if the sender can influence the receiver in such a way that he adopts her preference with complete certainty, then the situation is trivial. The sender would simply have to choose her favorite project and to disclose all of the values she is aware of. However, there may be extra cost to disclose values. The theoretical ground of partial deliberation allows accounting for some of these costs. In this paper, I focus on two of them. They are based on the two interesting properties of partial deliberation.

First, the way in which the receiver changes his preference through partial deliberation may depend on his background values. Assume that the receiver is aware that he may have background values without being aware of them. This has two consequences:

1. unless she is able to disclose these background values, the sender is uncertain about the receiver's reaction.

2. she is unable to disclose these background values.

Consequently, whenever the sender is aware of the existence of values in the receiver's background, she is uncertain about how the receiver will react to her disclosure strategy. I refer to this situation as *imperfect empathy*. The result of imperfect empathy is that the sender is uncertain about the receiver's reaction. This uncertainty constrains her strategy, and implies that the sender faces a trade-off when deciding whether or not to disclose values. On the one hand, she has an interest in disclosing as much values as possible since they (may) bring the receiver's preference closer from her own preference. On the other hand, the more values she discloses, the more uncertainty is generated and the less optimal her chosen project may be. I model this situation within three periods only. The timeline is given by figure (3.1).

Second, becoming sequentially aware of values does not necessarily yield the same preference change as becoming simultaneously aware of them. The issue is to find out the conditions under which the sender may benefit to use a sequential disclosure strategy rather than a silmutaneous one. Indeed, assuming that he looses her awareness from a period to the other, when the receiver becomes aware sequentially of two values, he can change the axiological intensity of one value at the time. Thus, his choice is more constrained

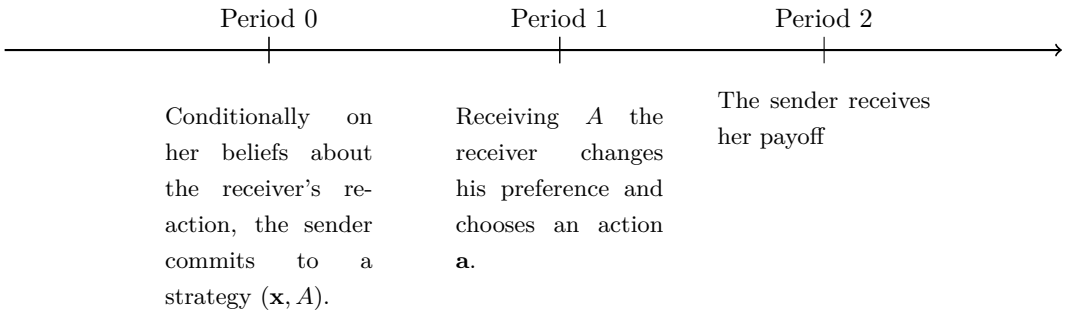

Figure 3.1 – The decision sequence when empathy is imperfect.

than when his awareness enables him changing the axiological intensity of both values. He may end up choosing an axiological state that does not maximize the axiological criterion represented by Φ. Whereas when being simultaneous aware of both values the receiver is less constrained and he can choose the axiological state that maximizes Φ. The timeline of this decision situation is given by figure (3.2).

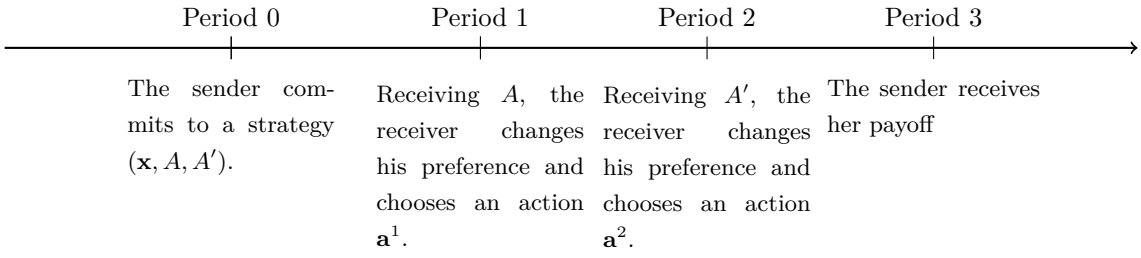

Figure 3.2 – The decision sequence with sequential disclosure.

This paper extends the study of partial deliberation to a context in which there are several agents. While it can be seen as a complement to the literature on persuasion, it does not consider the key issue of strategic communication. Indeed in this literature, the sender trying to change the receiver believes might be constrained to take into account the receiver's sophistication [Milgrom and Roberts, 1986]. The latter, indeed, is willing to maximize the amount of revealed information in order to be able to maximize his expected utility. This reasoning, however, applies in a context in which the preference of the decision maker is fixed or, in the worst case scenario, predictable for the agent whose preference is changing. Given that our approach focuses on the ability of the sender to change the receiver's preference, it is more consistent not to pay attention to the introspective reasoning that is implied in the context of interactive believes. In this sense, this paper is closer to Kamenica and Gentzkow [2011] analysis of bayesian persuasion. Nonetheless, once again, given that they consider beliefs modification, their revision rule is based on bayesian rationality whereas I consider a revision rule based on partial deliberation.

In the first section of the paper, I give the theoretical primitives of the model and establish how it captures partial deliberation. In the second section, I focus on a situation

in which the sender is uncertain about the receiver's reaction to her disclosure strategy and wonder whether she must disclose values or not. Then in the third section, I focus on a situation in which she may have an interest in using a sequential disclosure strategy, disclosing one value per period.

1 The euclidean framework of partial deliberation

This section presents the basic elements of the model and shows how they can be related to the framework of partial deliberation developed in the previous chapter. I assume that the sender and the receiver (potentially) adhere to three values. Lets denote $\hat{V} = \{1, 2, 3\}$, the set containing these three values. As it will be clear in section 2, while the two first values are relevant for evaluating of the project proposed by the sender, the third value is only relevant for assessing how values 1 and 2 relate axiologically. These restrictions are not required on the ongoing section, but they clarify the analysis for the two next sections. Most of the results of this paper can be generalized to situations with more values and a similar structure.

1.1 Axiological states and value system

In the previous chapter, I formalized value systems as sets V containing values and I studied the preference formation rules, $\rightarrow_A$, that are compatible with the principles of partial deliberation. In this chapter, I now consider that each set V contains the *ordered range* of a function v belonging to a set of functions $v \in \mathcal{F}$ that assigns an axiological intensity to each of these values.

By ordered range, I mean that for any element of V we can identify the axiological intensity of each value. So, a typical V can be expressed by $\{v(1) + 1, v(2) + 2, v(3) + 3\}$ for some $v \in \mathcal{F}$. Moreover, each V must contain one and only one element assigning an axiological intensity for each value. For example, we cannot have $\{v(1)+1, v(2)+2, v(3)+2, v(3)+3\}$ or by $\{v(1)+1, v(2)+2\}$ since the first set assigns two intensities to the second value, while the second set does not assign any axiological intensity to the third value. Let us denote $V^{\mathcal{F}}$, the set of all $V's$ of this form.

I assume that $\mathcal{F}$ is finite. This enables us applying the results from the previous chapter. I assume that $\mathcal{F}$ is stable under the substitution of ranges between these functions, i.e., for all $v, v' \in \mathcal{F}$, and all $B \subseteq \hat{V}$, there exists $v'' \in \mathcal{F}$ such that

$$v''(i) = \begin{cases} v(i) \text{ if } i \in B \\ v'(j) \text{ if } j \in B^{\complement} \end{cases}$$

This implies that for all v, v' and v" in $\mathcal{F}$, every set of the form $V = \{v(1) + 1, v'(2) + 2, v"(3) + 3\}$ belongs to $V^{\mathcal{F}}$. In other words, when the receiver *can* assign an axiological intensity $v(1)$ to value 1 and an axiological intensity $v(2)$ to value 2, he can also assign an

axiological intensity of $v(1)$ to value 1 when he assigns $v'(2)$ to value 2.

With this new formulation, a value system is compounded of a (cardinal) hierarchy between values in $\hat{V}$. Cardinality is a necessary requirement: in the decision situations modeled here, the sender explicitly makes an interpersonal comparison.

There exists a bijection from $V^{\mathcal{F}}$ to the set vectors of the form $\left(v(i)\right)_{i\in\hat{V}}$ with $v \in \mathcal{F}$. Thus, a value system can be seen as a vector $\mathbf{v}$ whose coordinates i give the axiological intensity $v(i)$ assigned to value i. This is what I refer to as an *axiological state.*[1] The next definition summarizes the primitive elements of an euclidean framework for partial deliberation

Definition 10. *A $(\hat{V},\mathcal{F})-$axiological state is a vector $\mathbf{v}$ whose coordinates are the intensities a decision maker assigns to each value, i.e., $\mathbf{v} = \left(v(i)\right)_{i\in\hat{V}}$ with $v \in \mathcal{F}$ and $\mathcal{F}$ is stable under substitution of ranges. An euclidean framework of values is a tuple $< \hat{V}, \mathcal{F} >$ such that there exists a bijection ψ from $V^{\mathcal{F}}$ to the set of $(\hat{V},\mathcal{F})-$axiological states denoted by $\mathbb{V}^{\hat{V},\mathcal{F}}$.*

1.2 Axiological state and projects

I assume that the sender does not change her preference. Thus, it is not necessary to distinguish her axiological state at each period. I denote her axiological state $\mathbf{v}^s \in [0,1]^3 = X$. The axiological state of the receiver at period i is denoted by $\mathbf{v}^i$ and belongs to X. These axiological states induce the preferences of both decision makers in a way that will be exhibited below. I denote by $v^i(j)$, the axiological intensity of the $j-th$ value in the axiological state $\mathbf{v}^i$ and by $\mathbf{v}_{|A}$ the restriction of the vector $\mathbf{v}$ to a vector whose coordinates are the axiological intensities assigned to values in A i.e., $\mathbf{v}_{|A} = \left(v^i(j)\right)_{j\in A}$. I assume that the receiver does not adhere to value 1 and 2. Thus, $\mathbf{v}^0 = (0,0,v(3))$ and the sender knows it.

Both decision makers evaluate the project proposed by the sender with their own axiological state. Thus, each project is characterized by the axiological intensity it assigns to each value. In the spirit of Lancaster [1966]'s "new consumer's theory", a project is a vector $\mathbf{x} \in [0,1]^3 = X$ that assigns an axiological intensity between 0 and 1 to values 1, 2 and 3. For instance, consider a tax policy, $\mathbf{x}$, that could be associated with the value "*being efficient is good*", but not with the value "*being fair is good*". Let us say that the project is given by $\mathbf{x} = (1,0,x(3))$, with the first coordinate of $\mathbf{x}$ representing the axiological intensity with which the project satisfies the value "*being efficient is good*", while the second represents the axiological intensity with which it satisfies "*horizontal equity*". Furthermore, I denote $x(i)$, the intensity with which the project $\mathbf{x}$ fulfills the value i. As already mentioned, I assume that value 3 is irrelevant to the project. Thus,

1. Take for instance $\psi : V \to \sum_{a\in V}\left(i_{\{1\leq a<2\}}(a-1), i_{\{2\leq a<3\}}(a-2), i_{\{3\leq a<4\}}(a-3)\right)$. By definition of the value systems V this is a bijection.

for any $\mathbf{x} = (x(1), x(2), x(3))$, I can assume that the set of projects is:

$$\{\mathbf{x} \in X : x(3) = 0\}$$

1.3 Awareness of values

The sender is able to induce a change in the receiver's preference by making him aware of either value 1 or value 2 or both. Let us generically denote by $A \subseteq \hat{V}$, the set of values the receiver may be aware of and by $\mathbf{v}(A; \mathbf{v}^i)$, the axiological state to which the receiver adheres to when his initial axiological state is $\mathbf{v}^i$ and he becomes aware of values $A \subseteq \{1, 2\}$. Thus when, as in the third section below, there are two periods, we have $\mathbf{v}(A; \mathbf{v}^0) = \mathbf{v}^1$ and $\mathbf{v}(A; \mathbf{v}^1) = \mathbf{v}^2$. Note that the sender is not able to make the receiver aware of value 3. I also assume that the receiver is not aware of any values in $\hat{V}$. This hypothesis is demanding. It clearly means that there is an asymmetry between the receiver's and the sender's awareness.

Since axiological states differ on the axiological intensity they assign to values, I also need to define the set of values' axiological intensities the receiver is aware of when he is aware of values in a set A. So let us denote $\hat{A}$, the set of potential intensities that can be assigned to values in A, i.e.,

$$\hat{A} = \{v(i) + i : i \in A, v \in \mathcal{F}\} \text{ and } if\, A \in \mathcal{A} \text{ then } \hat{A} \in \hat{\mathcal{A}}$$

1.4 An euclidean framework for values partial deliberation

How does this change of axiological state occur? In the previous chapter, I used a preference formation rule, $\rightarrow' \in \mathcal{V}^2 \times \mathcal{A}$ from a value system to another one, and I characterized the conditions under which it is *induced by partial deliberation*. Under these conditions there exists an order $\trianglelefteq$ on value systems such that the decision maker always chooses the best value system axiologically speaking (in the sense of $\trianglelefteq$).

Definition 11. *(see chapter 2) A preference formation rule $\rightarrow' \in \mathcal{V}^2 \times \mathcal{A}$ is induced by partial deliberation if there exits an order $\triangleleft$ on $\mathcal{V}^2$ such that for all $V, V' \in \mathcal{V}$ and all $\hat{A} \in \mathcal{A}$*

$$V \rightarrow'_{\hat{A}} V' \iff \begin{cases} \exists B \subseteq \hat{A}, V' = B \cup (V \backslash \hat{A}) \\ \forall B' \subseteq \hat{A}, B' \cup (V \backslash \hat{A}) \trianglelefteq V' \end{cases} \tag{3.1}$$

In the previous chapter, I established the necessary conditions for a preference change to satisfy this definition. These conditions can still apply in an euclidean framework. To see this, let us define the euclidean version a partial deliberation.

Definition 12. *A preference formation rule* $\rightarrow \in \mathbb{V} \times \mathbb{V} \times \mathcal{A}$ *is said to be induced by an euclidean version of partial deliberation if*

— *there exist* $\hat{V}$ *and* $\mathcal{F}$ *such that* $< \hat{V}, \mathcal{F} >$ *is an euclidean framework of values where* ψ *is its corresponding bijection and* $\mathbb{V} = \psi(\hat{V}^{\mathcal{F}}) \subseteq [0,1]^3$.

— *there exists a preference formation rule induced by partial deliberation* $\rightarrow' \in V^{\mathcal{F}} \times V^{\mathcal{F}} \times \hat{\mathcal{A}}$ *such that*

$$V \rightarrow'_{\hat{A}} V' \iff \mathbf{v} = \psi(V) \rightarrow_A \psi(V') = \mathbf{v}'$$

A preference change is induced by partial deliberation if there exists a function such that the receiver chooses the best axiological state that his awareness makes it reachable.

Proposition 8. *A preference formation rule is induced by an euclidean version of partial deliberation if and only if there exists a function* Φ *such that:*

$$\mathbf{v} \rightarrow_A \mathbf{v}' \iff \mathbf{v}' \in \left(\mathbf{v}_{|\neg A}, \underset{\mathbf{v}''_{|A}}{Argmax}\ \Phi(\mathbf{v}_{|\neg A}, \mathbf{v}''_{|A}) \right) \tag{3.2}$$

where $\mathbf{v}_{|A} = (v(i))_{i \in A}$

Thus, in the framework proposed here, preference changes are induced by a function that maps the vector of axiological states into $\mathbb{R}$. Finally, we have :

$$\mathbf{v}(A, \mathbf{v}^i) \in \left(\mathbf{v}^i_{|\neg A}, \underset{\mathbf{v}''_{|A}}{Argmax}\ \Phi(\mathbf{v}^i_{|\neg A}, \mathbf{v}''_{|A}) \right) \tag{3.3}$$

Crucially, note that $\mathbf{v}(A, \mathbf{v}^i)$ depends on A and $\mathbf{v}^i_{|\hat{V} \setminus A}$, the decision maker being free to choose appropriately the $\mathbf{v}_{|A}$ that maximizes Φ.

In my specific framework, since the receiver is only able to become aware of values 1 and 2, it is possible to write every possible preference changes induced by partial deliberation as follows:

$$\mathbf{v}(\{j\}; \mathbf{v}^i) \in \left(\underset{v \in [0,1]}{Argmax}\ \Phi(v, v^i(-j)), v^i(-j) \right) \text{ and } \mathbf{v}(\{1,2\}; \mathbf{v}^i) \in \left(\underset{(v,v') \in [0,1]^2}{Argmax}\ \Phi(v, v', v^i(3)), v^i(3) \right) \tag{3.4}$$

1.5 Axiological connection between values

A key issue is to know how the shape of Φ represents the axiological relation between values. The idea is that the receiver may be more disposed to assign a high axiological intensity to a value i when he already assigns a high axiological intensity to another value j.

Let us consider two examples:

1. First, consider the situation in which the axiological criteria is represented by the following function:

$$\Phi\Big(v(1),v(2)\Big)=-\Big(v(1)-\alpha\Big)^2-\Big(v(2)-\beta\Big)^2$$

In this example, it is clear that the axiological intensity that the receiver assigns to value 1 does not influence the axiological intensity he assigns to value 2. Then we have :

$$\mathbf{v}(\{1,2\};\mathbf{0})=\begin{pmatrix}\alpha\\ \beta\end{pmatrix}=\mathbf{v}(\{1\},\mathbf{0})+\mathbf{v}(\{2\},\mathbf{0})$$

2. Second, assume that the axiological criteria is the following:

$$\Phi\Big(v(1),v(2),v(3)\Big)=-\Big(v(1)-\alpha_1\Big)^2-\Big(v(2)-\alpha_2\Big)^2-v(3)v(1)v(2)$$

In such a situation, assuming that $\frac{\alpha_1+v(3)\alpha_2}{1-v(3)^2}\in(0,1)$ and $\frac{\alpha_2+v(3)\alpha_1}{1-v(3)^2}\in(0,1)$, we have that:

$$\mathbf{v}(\{1\};\mathbf{0})=\begin{pmatrix}\alpha_1\\ 0\end{pmatrix};\mathbf{v}(\{2\};\mathbf{0})=\begin{pmatrix}0\\ \alpha_2\end{pmatrix};\mathbf{v}(\{1,2\};\mathbf{0})=\frac{1}{1-v(3)^2}\begin{pmatrix}\alpha_1+v(3)\alpha_2\\ \alpha_2+v(3)\alpha_1\end{pmatrix}$$

With this in mind, we can define what it means for two values to be axiologically independent when the initial axiological state is $\mathbf{v}$.

Definition 13. *We say that values 1 and 2 are* $\mathbf{v}$*-axiologically independent when if :*

$$\mathbf{v}(\{1,2\},\mathbf{v})=\mathbf{v}(\{1\},\mathbf{v})+\mathbf{v}(\{2\},\mathbf{v})$$

This definition says that, when the decision maker's initial axiological state is $\mathbf{v}$, two values are axiologically independent if becoming aware of these two values does not imply changing the relative axiological intensities of value 1 or value 2. Another formulation is that, when there are two periods, being simultaneously aware of the two values gives, after the last period, the same outcome as being sequentially aware of these two values.

Proposition 9. *If values 1 and 2 are* $\mathbf{v}$*-axiologically independent then*

$$\mathbf{v}(\{1,2\},\mathbf{v})=\mathbf{v}\Big(\{i\},\mathbf{v}(\{-i\},\mathbf{v})\Big)$$

The case of axiological independence is illustrated by the point (I) in figure (1.5). Axiological independence simply means that, whatever the background values of the decision

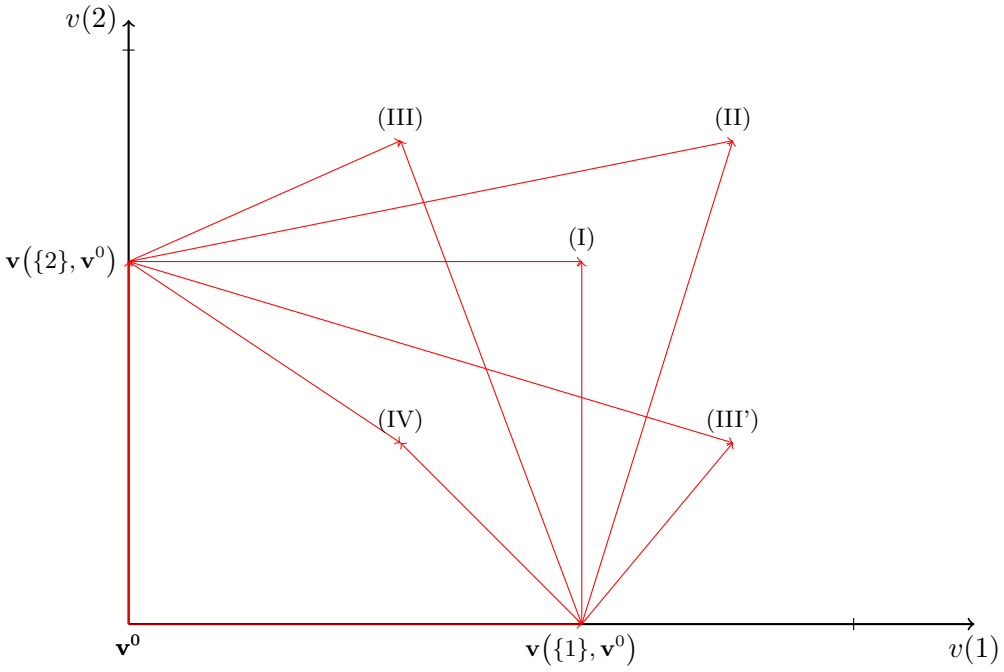

Figure 3.3 – The kinds of axiological depedence

makers, adhering to a value does not influence the propensity to adhere to the other value. Conversely, it is possible to define negative and positive dependence.

Definition 14. *We say that value i:*

— $\mathbf{v}$*-depends positively on* $-i$ *whenever we have* $v(i)(\{i,-i\};\boldsymbol{v}) \geq v(i)(\{i\},\boldsymbol{v}^0)$.

— $\mathbf{v}$*-depends negatively on* $-i$ *whenever we have* $v(i)(\{i,-i\},\mathbf{v}) \leq v(i)(\{i\},\boldsymbol{v}^0)$.

Moreover, we are in situation of

— $\mathbf{v}-$*positive (negative) dependence if both values* $\mathbf{v}-$*depend positively (negatively) on each other.*

— $\mathbf{v}-$*mixed dependence if a value* i $\mathbf{v}-$*depends positively on* $-i$ *while* $-i$ $\mathbf{v}-$*depends negatively on* i.

In order to ease the notation, I simply refer to "$\mathbf{v}^0-$positive" as positive dependence.

The situation of positive dependence is illustrated by the point (II) of figure (1.5), negative dependence is illustrated by the point (IV) and mixed dependence is illustrated by points (III) and (III').

1.6 Preference changes

We now need to state how axiological states induce the utility functions of the sender and the receiver. I assume that the receiver does not speculate on the intention of the

sender. He takes the values she has disclosed for granted and modifies passively his axiological state. This assumption is guided by conceptual concerns. In my framework, the receiver is unable to conceive the values 1 and 2 before the sender makes him aware of these values. Consequently, he cannot anticipate the possible changes that this disclosure strategy could induce on his preference. This point crucially relies on the conceptual nature of unawareness: unawareness is an inability to conceive a value, i.e., the receiver is also unaware that he is unaware of values 1 and 2. In a sense, this approach has some similarities with Kahneman and Snell's (1992) idea that consumers are unable to predict how their taste will change. It may be worth studying a situation in which the receiver reacts to the disclosure strategy of the sender by trying to figure out additional values that have not been disclosed to him. But this is beyond the scope of this paper.

The receiver chooses an action $\mathbf{a}$ that maximizes his preference. The action $\mathbf{a}$ is a vector that belongs to the set

$$\{(a(1), a(2), a(3) \in X : a(3) = 0\}$$

The fact that $a(3) = 0$ is due to the irrelevance of the third value for the project.

To model how the receiver's preference $\preceq_{\mathbf{v}^i}$ over these actions depends on his axiological state, I assume that

$$\mathbf{a} \preceq_{\mathbf{v}^i} \mathbf{a}' \iff ||\mathbf{a} - \mathbf{v}^i|| \geq ||\mathbf{a}' - \mathbf{v}^i||$$

Thus, the following utility function represents the receiver's preference at period i

$$u_{\mathbf{v}^i}(\mathbf{a}) = -||\mathbf{a} - \mathbf{v}^i||^2$$

where $\mathbf{a} \in X$ is the action of the receiver. Thus, his best action is simply $\mathbf{a}^* = \mathbf{v}^i$.

In this context, the sender simply anticipates the receiver's reaction and does not need to pay attention to what he believes. Moreover, the sender has her own preference, but she also needs her proposed project to fit with the receiver's action. This is the reason why, in addition to proposing a project $\mathbf{x}$, she may choose to disclose value 1 or value 2. Thus, the strategy of the sender is a pair $(A, \mathbf{x})$, and she chooses it so that it maximizes:

$$u(\mathbf{x}, A) = -\beta||\mathbf{x} - \mathbf{v}_P||^2 - ||\mathbf{x} - \mathbf{a}^*||^2 \tag{3.5}$$

s.t.

$$\mathbf{v}(A, \mathbf{v}^0) \in \underset{\mathbf{v}''_{|A}}{Argmax}\ \Phi(\mathbf{v}^i_{|\neg A}, \mathbf{v}''_{|A}) \tag{3.6}$$

$$\mathbf{a}^* = \mathbf{v}(A, \mathbf{v}^0) \tag{3.7}$$

with β representing the weigh she assigns to the receiver adopting her own preference. In

a strategy $(\mathbf{x}, A)$, A is referred to as the *disclosure strategy*.

To interpret this utility function, note that, in the first term of this equation, the sender tends to choose a political project that is close from her own axiological state. However, in the second term, we see that this project should not be too far from the receiver's axiological state. Given that she is able to alter the receiver's axiological state, the sender has an interest in running a disclosure strategy that makes the new axiological state of the receiver closer to her own axiological state.

In an extended model with several senders, the value of β may differ among the senders. For instance, Margaret Thatcher's famous concept of *conviction politics* clearly aims at justifying a high value for this parameter. [2] One may argue that β depends upon the preference of the sender and that it is a function of the sender's values.

If the sender is able to make sure that the receiver will adopt her own axiological state, then her decision would simply consist in 1) choosing her preferred project $\mathbf{x}$ and 2) choosing the disclosure strategy A that brings the receiver's axiological state to $\mathbf{x}$, i.e., $(\mathbf{x} = \mathbf{v}(A, \mathbf{v}^0)$. In other words, there is no trade-off in this situation. In the rest of the paper, I relax such an hypothesis in order to model imperfect empathy and to consider the gain from sequential disclosure of values.

I assume that initially, while the sender assigns an axiological intensity of $v^s(1) > 0$ to value 1 and an axiological intensity $v^s(2) > 0$ to value 2, the receiver assigns an axiological intensity of 0 to both values 1 and 2. Thus $\mathbf{v}^0 = \big(0, 0, v(3)\big)$.

The next section of the paper investigates a situation in which the sender is uncertain about the effect of her disclosure strategy. She is likely to change the receiver's payoff so that he adopts the same preference. In other words, the sender is imperfeclty empathetic but there is a probability that her preference aligns with the receiver's preference. Thus, the sender first chooses, *ex ante*, a project and a disclosure strategy, then the receiver chooses an action and, eventually, both decision makers receive their payoffs.

In the third section of the paper, I assume that the sender can choose sequentially two disclosure strategy. Thus there are three periods. In the first period, she chooses a project and a disclosure strategy. In the second period, she can choose another disclosure strategy. In the third period, the receiver chooses an action and they both receive their payoff.

2 Disclosure strategy with imperfect empathy

In this section, I assume that the sender does not know exactly how the receiver will react to her disclosure strategy. By doing so, I formalize the concept of imperfect empathy in a way that is compatible with partial deliberation.

2. According to Thatcher, rather than consensus, a politician should base his campaign on his own fundamental values or ideas. She declared in 1979, "I am not a consensus politician. I am a conviction politician". http://content.time.com/time/magazine/article/0,9171,916773-6,00.html. For an analysis of Thatcher's position on that matter see Metcalfe, 1993.

2.1 Imperfect empathy

The sender is aware that there may exist a background value - i.e., value 3 - that determines how the receiver will react to her disclosure strategy. But she is not aware of this third value. Moreover, a decision maker assigning a certain axiological intensity to value 3 may never assign axiological intensity to values 1 or 2. Therefore, she can neither disclose value 3, nor anticipate how it may affect the receiver's reaction. In other words, there is no guaranty that the receiver will react to every value that the sender discloses. As the sender is uncertain about the receiver's reaction, she forms beliefs about this reaction. Thus the timeline of this situation is given by figure 3.4.

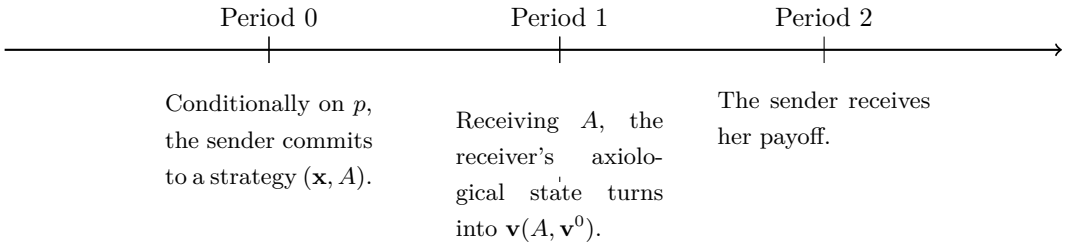

Figure 3.4 – The decision sequence when empathy is imperfect

I refer to the sender's beliefs on the receiver's preference as *imperfect empathy*. Imperfect empathy relies on the following idea :

1. when making her decision, the sender must try to evaluate which values determine the receiver's preference,
2. she can do it only imperfectly, by using her awareness and her own preference.

Before giving a formal account of imperfect empathy, it is worth mentioning that this concept has been used in the context of cultural transmission to mean something quite different. In their famous model of cultural transmission, Bisin and Verdier [2001*b*] use the concept of imperfect empathy to refer to a "myopic or paternalistic altruism" [p. 306]. They assume that parents want the best for their children but, in evaluating what is the best, they are biased by their own preference. [3] Interestingly, they refer to Smith's theory of sympathy in which a lack of imagination makes it difficult for a decision maker to "place herself in others situations". [4] [5]

3. "We interpret such assumptions as a form of myopic or paternalistic altruism (hence the name, "imperfect empathy"). Parents are aware of the different preference traits children can adopt, and are able to anticipate the socio-economic choice a child with preference trait $i \in [a, b]$ will (optimally) make. Parents are not able, though, to altruistically evaluate their children's actions with the children's utility function (to "perfectly empathize" with the children), but they are biased by their own (the parents') preference evaluations." [Bisin and Verdier, 2001*b*, p. 306].

4. "As we have no immediate experience of what other men feel, we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in the like situation. Though our brother is upon the rack, ... our senses will never inform us of what he suffers. .... By the imagination we place ourselves in his situation ...."

5. For another interpretation of the role of background beliefs in the interactive contexts of incomplete

[Smith, 2014 (1759, p. 1] Thus, I believe that the definition of Bisin and Verdier [2001*b*] contains three things:

1. the fact that our imagination allows us to imperfectly evaluate what determines another person's preference,
2. the fact that in evaluating others' preference we are biased by our own preference,
3. the fact of being altruist.

In this paper, I take imperfect empathy in the narrowest sense, restricting its definition to 1. I assume that when the sender tries to evaluate the receiver's preference, her evaluation is imperfect because she is not aware of some background values determining the reaction of the receiver to her disclosure strategy. However, I also assume that the sender is aware of this imperfection and that she forms beliefs about the receiver's reaction to her disclosure strategy. Moreover, imperfect empathy makes this reaction uncertain. To capture this idea in a simple way, I make two assumptions.

First, I assign a specific shape to the function representing the axiological criterion.

Assumption 1: *The axiological criterion is represented by a function* Φ *that has the following form:*

$$\Phi(v(1),v(2),v(3)) = -\Big(v(1)-(\alpha_1+\alpha_1' i_{B_1\cap B_2}(v(3)))\Big)^2 i_{B_1}(v(3)) - \Big(v_2-(\alpha_2+\alpha_2' i_{B_1\cap B_2}(v(3)))\Big)^2 i_{B_2}(v(3))$$

Where $i_B(v)$ *is the indicative function, i.e., it is equal to 1 whenever* $v \in B$ *and to 0 otherwise; and for* $i \in \{1,2\}$, $B_i \subseteq [0,1]$.

Thus, when the sender makes the receiver aware of value(s) in A, the receiver's axiological state must be in the set:

$$\underset{\{\mathbf{v}:\mathbf{v}_{|\neg A}=\mathbf{v}^0_{|\neg A}\}}{Argmax} \quad \Phi(v(1),v(2),v(3))$$

Note that when the disclosure decision is partial, Φ may not have an unique maximum. For instance, if the sender discloses values 1 and 2 and if $v(3) \notin B_2$, then for every v, the vector $(\alpha_1, v, v(3))$ would maximize Φ. For the sake of clarity, I also assume that the receiver is conservative in the sense that, in the set of value systems maximizing Φ (conditionally on his awareness), the receiver chooses the closest one from his original value system. Formally, this property is given by the following assumption:

Assumption 2: *If the initial value system of the receiver is* $v^0 = (0,0,v(3))$ *and the sender discloses values in* A*, then the value system of the receiver is unique and is given*

contract, see [Chaserant, 2000].

ex post by:

$$\mathbf{v}(A, \mathbf{v}^0) = \underset{\mathbf{v} \in M}{Argmin} ||\mathbf{v} - \mathbf{v}^0||$$

s.t.

$$M = \underset{\{\mathbf{v}: \mathbf{v}_{|\neg A} = \mathbf{v}^0_{|\neg A}\}}{Argmax} \Phi(v(1), v(2), v(3))$$

It is easy to compute the value system that maximizes Φ conditionally upon the axiological intensity that the receiver assigns to value 3 and upon the set of values he becomes aware of. These maximal values are given in the following tab.

	$\emptyset$	$\{1\}$	$\{2\}$	$\{1,2\}$
$v(3) \in B_1^{\complement} \cap B_2^{\complement}$	$\mathbf{v}^0$	$\mathbf{v}^0$	$\mathbf{v}^0$	$\mathbf{v}^0$
$v(3) \in B_1 \cap B_2^{\complement}$	$\mathbf{v}^0$	$(\alpha_1, 0, v(3))$	$\mathbf{v}^0$	$(\alpha_1, 0, v(3))$
$v(3) \in B_1^{\complement} \cap B_2$	$\mathbf{v}^0$	$\mathbf{v}^0$	$(0, \alpha_2, v(3))$	$(0, \alpha_2, v(3))$
$v(3) \in B_1 \cap B_2$	$\mathbf{v}^0$	$(\alpha_1, 0, v(3))$	$(0, \alpha_2, v(3))$	$1/2(\alpha_1 + \alpha_1', \alpha_2 + \alpha_2', v(3))$

Moreover, I assume that the sender does not know which axiological intensity $v(3)$ the receiver assigns to value 3. This assumption may be interpreted in terms of awareness of unawareness; the sender may be aware that there may exist a value 3 which prevents the receiver from adopting either value 1 or value 2. Since the sender is uncertain about the axiological intensity that the receiver assigns to value 3, then she does not know how he will react to her disclosure strategy. This uncertainty is modeled by a probability distribution over the possible intensities that the receiver may assign to the third value.

Assumption 3 : *The sender's beliefs are given by a probability measure μ on $([0,1], \mathcal{B})$), where $\mathcal{B}$ is the $\sigma-$field generated by $\{B_1, B_2\}$.*

In order to write the program that the sender solves, it is convenient to translate μ into the family of probabilities $(p_A)_{A \in \mathcal{A}}$ giving to each disclosure strategy A the probability that the receiver reacts to the axiological intensity of values in K. Thus p_A is a measure on 2^V giving the probability that, when the sender discloses A, the receiver reacts to values in K. This probability is given by the following formula:

$$p_A(K) = \begin{cases} \mu\Big((\bigcap_{i \in K} B_i) \cap (\bigcap_{i \in A \setminus K} B_i^{\complement}) \Big) & \text{if } K \subseteq A \\ 0 & \text{otherwise} \end{cases}$$

By the law of total probability, for any $j \in A$ and $K \subseteq A\backslash\{j\}$, we have that: [6]

$$p_{A\backslash\{j\}}(K) = p_A(K) + p_A(K \cup \{j\}) \tag{3.8}$$

Note also that, by definition, for any i, if $v(3) \notin B_i$ then:

$$\mathbf{v}(A\backslash\{i\}; \mathbf{v}^0) = \mathbf{v}(A; \mathbf{v}^0)$$

When the sender makes the receiver aware of A, the receiver's value system becomes $\mathbf{v}(K; \mathbf{v}^0)$ with probability $p_A(K)$.

Therefore, I implicitly assume that the uncertainty on the receiver's reaction increases with the number of values disclosed by the sender. From the sender's uncertainty specification, and assuming that she is risk neutral, we deduce that the program she solves is the following:

$$\mathbb{E}u(\mathbf{x}; A) = -\beta||\mathbf{x} - \mathbf{v}^s||^2 - \sum_{K \subseteq A} p_A(K)||\mathbf{x} - \mathbf{a}(K; \mathbf{v}^0))||^2 \tag{3.9}$$

Where

$$\mathbf{a}(K; \mathbf{v}^0) = \mathbf{v}\big(K; (0, 0, v(3))\big) \text{ with } v(3) \in \Big(\bigcap_{i \in K} B_i\Big) \cap \Big(\bigcap_{i \in A\backslash K} B_i^{\complement}\Big)$$

Given that the receiver does not necessarily react to the sender's disclosure strategy, it may be better for the sender not to disclose some values. The main reason is that, by disclosing more values, the sender makes the receiver's reaction more unpredictable. Since she is risk neutral, this uncertainty implies choosing a project that takes into account the different possible reactions of the receiver. Thus, the trade-off faced by the sender is the following. On the one hand, her goal is to disclose as much values as she can since it would choosing a project closer to her axiological state. This consists in minimizing the first term of (3.9). On the other hand, disclosing values creates uncertainty in the receiver's reaction and taking into account this uncertainty implies choosing a project that is more consensual. This consists in minimizing the second term of (3.9). In other words, since the sender knows which axiological intensity the receiver originally assigns to values 1 and 2, the axiological intensity increases with the number of values disclosed. Yet, in the choice of the optimal project, the sender has to take into account this uncertainty by choosing a project that neither corresponds to her project nor to the project of a sender who is perfectly aware. Consequently, making the receiver aware yields costs: disclosing

6. Indeed we have that

$$p_{A\backslash\{j\}}(K) = \mu\Big(\Big(\bigcap_{i \in K} B_i\Big) \cap \Big(\bigcap_{i \in A\backslash(K\cup\{j\})} B_i^{\complement}\Big)\Big) = \mu\Big(\Big(\bigcap_{i \in K} B_i\Big) \cap \Big(\bigcap_{i \in A\backslash(K\cup\{j\})} B_i^{\complement}\Big) \cap \big(B_j \cup B_j^{\complement}\big)\Big)$$
$$= \mu\Big(\Big(\bigcap_{i \in K\cup\{j\}} B_i\Big) \cap \Big(\bigcap_{i \in A\backslash(K\cup\{j\})} B_i^{\complement}\Big)\Big) + \mu\Big(\Big(\bigcap_{i \in K} B_i\Big) \cap \Big(\bigcap_{i \in A\backslash K} B_i^{\complement}\Big)\Big) = p_A(K) + p_A(K \cup \{j\})$$

creates an undesirable uncertainty implying that the chosen project is a middle solution that satisfies nobody.

In the rest of this section, I study this trade-off between disclosure and uncertainty in a context in which the preference of the sender can be aligned with the preference of a fully aware sender , i.e., there is a non-null probability that $\mathbf{v}^s = \mathbf{v}(\{1,2\}, \mathbf{v}^0)$.

2.2 Preliminary results

I first establish some trivial results that guaranty the model to give coherent intuitions.

It is straightforward to see that, if the sender knows that the axiological intensity of the receiver's third value prevents him to adhere to a value i whatever the sender's disclosure strategy may be, then the sender has no interest in disclosing this value. In this situation, the sender is better off proposing a project in which the axiological intensity of i is the following convex combination between $v^0(i)$ and $v(i)^s$: $(\frac{\beta}{1+\beta}v^s(i))$. In other words, the best project is the middle one between what satisfies the sender and what satisfies the receiver. Denoting by $\mathbf{x}_A$ the optimal project when the disclosure strategy is A we can state this result with the following lemma.

Lemma 3. *If there exists i such that $p_{\{1,2\}}(K) = 0$ for all K such that $i \in K$ then:*

$$\mathbb{E}u(\mathbf{x}_A, A) \geq \mathbb{E}u(\mathbf{x}_{A\backslash\{i\}}, A\backslash\{i\})$$

As mentioned above, disclosing values creates an undesirable uncertainty. Thus, when the sender has no interest in her own preference ($\beta = 0$), her best option is to choose the project fitting the initial axiological state of the receiver $\mathbf{v}^0$. This is established by lemma 4.

Lemma 4. *When $\beta = 0$ the sender has no interest in disclosing any value.*

In the next subsection, I focus on a situation of axiological independence between values.

2.3 Axiologically independent values

Given the shape of Φ, axiological independence is satisfied whenever $\alpha_1 = \alpha_1'$ and $\alpha_2 = \alpha_2'$. In this context we obtain the following result.

Proposition 10. *Consider a sender with belief p. If values 1 and 2 are axiologically independent then the sender discloses value i if and only if $p_{\{i\}}(\emptyset) < \beta$.*

Unsurprisingly, proposition (10) states that when the interest of the sender to see the receiver adopting her value system (β) is greater than the probability that the receiver does not react to value i $p_{\{i\}}(\emptyset)$, it is in her interest to disclose value i. In other words, the disclosure strategy only depends on whether the sender believes that the background

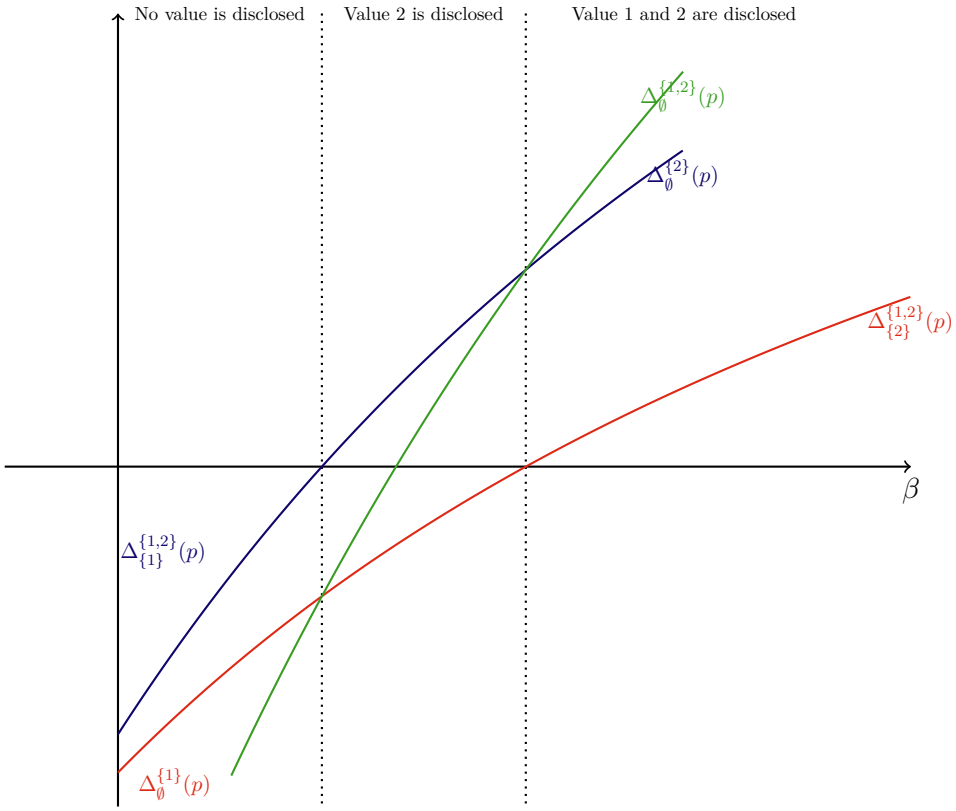

Figure 3.5 – The sender's problem

value of the receiver belongs to B_i (the set in which value i affects the preference changes of the receiver).

As shown in Figure 3.5, the result of this proposition can be seen in a simple way by computing the differences between the sender's optimal expected payoff when disclosing values belonging to the set A and her optimal expected payoff when disclosing values belonging to A'.

$$\Delta_{A'}^{A}(p) = \mathbb{E}(\mathbf{x}_A; A) - \mathbb{E}(\mathbf{x}_{A'}; A')$$

Whenever $\Delta_{A}^{A'}(p) > 0$, it is better to disclose values belonging to A than to disclose values belonging to A'. Thus, from figure 3.5 we have that:

— If $\beta \in [0, p_{\{-i\}}(\emptyset))$, then $\Delta_{\{1\}}^{\{1,2\}}(p) = \Delta_{\emptyset}^{\{2\}}(p) < 0$. It is thus better to disclose only $\{1\}$ (respectively $\emptyset$) than disclosing $\{1,2\}$ (respectively $\{2\}$). So, whenever $\beta \in [0, p_{\{-i\}}(\emptyset))$, $\{1\}$ and $\emptyset$ are the best candidate for the sender's disclosure's strategy. But since $\Delta_{\emptyset}^{\{1\}}(p) < 0$, disclosing $\emptyset$ is better than disclosing $\{1\}$. Thus, $\emptyset$ is the best disclosure strategy whenever $\beta \in [0, p_{\{-i\}}(\emptyset))$.

— For $\beta \in (p_{\{-i\}}(\emptyset), p_{\{i\}}(\emptyset))$, the same reasoning applies and leads to disclose $\{2\}$.

— For $\beta \in (p_{\{i\}}(\emptyset), 1]$, $\{1,2\}$ is the best disclosure strategy.

Stated in more general terms, we have that $\Delta_{A\setminus\{i\}}^{A}(p)$ increases with β: as the sender's willingness to see the receiver adopting her own preference increases, she is willing to

take a higher risk in order to influence the receiver's preference. Therefore, the optimal disclosure strategy of the sender can be represented by a partitioned interval in which if $p_{\{i\}}(\emptyset) > p_{\{-i\}}(\emptyset)$ with $-i = \{1,2\}\backslash\{i\}$. Then:

— if $\beta \in [0, p_{\{-i\}}(\emptyset))$ then sender has no interest in disclosing any value.

— if $\beta \in (p_{\{-i\}}(\emptyset), p_{\{i\}}(\emptyset))$ then sender has an interest in disclosing value i.

— if $\beta \in (p_{\{i\}}(\emptyset), 1]$ then sender has an interest in disclosing both values.

This shows that in case of axiological independence, the sender can simply consider each value separately when deciding whether to disclose it or not. This implies that she is never indifferent between adopting either only value 1 or only value 2.

Moreover, it is worth wondering what happens when the sender has no reason to believe that the receiver is likely to take a given belief into consideration. After all, being aware of her unawareness, the sender may have no clue to predict the reaction of the receiver. I refer to this situation as symmetric axiological expectancy.

Definition 15. *It is said that an imperfectly empathetic sender has a symmetric axiological expectancy whenever:*

$$p_{\{1,2\}}(1) = p_{\{1,2\}}(2) \text{ or equivalently } p_{\{1\}}(1) = p_{\{2\}}(2)$$

The following result states that when the sender has a symmetric axiological expectancy, she never has an interest in disclosing values.

Corollary 1. *Given a belief distribution p, if value 1 and value 2 are axiologically independent, then the sender has no interest in disclosing only one value.*

Summing up, when values are axiologically independent, the sender can consider them separately. Her interest in disclosing both values depends on her interest in disclosing value 1 and on her interest in disclosing value 2.

2.4 Disclosing values in the perspective of the second theorem of welfare

To make things concrete, let us apply this result to the inaugural example of the policy maker designing a tax policy. I assume that the sender is a fiscal adviser who believes that the second theorem of welfare economics holds. This belief results from a background value $v(3)$, which she may not be aware. For instance, she does not question the hypothesis of the second welfare theorem. She may also have ethical concerns explaining her adhesion to $v(3)$. Indeed, $v(3)$ can be thought as a concatenation of different values.

The task of the fiscal adviser is to propose a tax policy to the policy maker. This tax policy can satisfy two values. The first value is "efficient allocation of resources is good". The second value is "equal distribution of resources is fair". The fiscal adviser is aware of these two values and she assigns a positive axiological intensity ($v(1)$ and $v(2)$) to each of them. Moreover, since she is a supporter of the second theorem of welfare economics,

she believes that any Pareto efficient allocation can be achieved as an equilibrium. A first best equilibrium is achievable through a lump sum transfer policy and a tax policy can combine equality and allocation efficiency. Thus there is no trade-off between efficient allocation of resources and equality. This basically means that $v(1)$ and $v(2)$ are assumed to be axiologically independent (at least) in the mind of the policy adviser.

However, the policy advisor is aware that $v(3)$ may exist and that it may not be shared by the policy maker. For instance $v(3)$ may incorporate either a value like NEA ="efficient allocation is based on a bad conception of humans preference", contradicting $v(1)$ or a value like NED ="the rich deserve what they get", contradicting $v(2)$. Rigorously, the fiscal advisor is not supposed to be aware of NEA and NED and she interprets $v(3)$ in terms of possible reactions. For the sake of simplicity assume that she believes that:

— if $v(3) \in (0, 1/4)$ then the policy maker adheres to NEA and NED,

— if $v(3) \in (1/4, 1/4)$ then the policy maker adheres to NEA but not to NED,

— if $v(3) \in (1/2, 3/4)$ then the policy maker adheres to NED but not to NEA,

— if $v(3) \in (1/2, 3/4)$ then the policy maker neither adheres to NED nor to NEA.

The goal of the fiscal adviser is to design a tax policy 1) she adheres to and 2) that will be applied by the policy maker. In other words, her utility function is given by equation (3.5) where β can be thought as a background value stating how much the fiscal advisor believes that scientists should compromise with politicians. The goal of the policy maker is to choose an action that best fits his preference. His utility is given by equation (1.6).

It is clear that, if μ is uniform on $[0, 1]$, then

$$p_{\{1,2\}}(1) = p_{\{1,2\}}(2) \text{ or equivalently } p_{\{1\}}(1) = p_{\{2\}}(2)$$

and the fiscal adviser has symmetric axiological expectancy.

In this case, we obtain from Corollary (1) that when she is not willing enough to compromise her thought with the politician's program ($\beta > 1/4$), then the fiscal adviser discloses both values and proposes a project situated on the convex combination between her preference and the initial axiological state of the policy maker. Conversely, when she is willing to compromise ($\beta < 1/4$), then she discloses no value and proposes the policy maker's project.

However, the fiscal adviser may have some clues about how the policy maker will react to her disclosure strategy. For instance, she can feel from other aspects of the policy maker's political program that he is more inclined to adhere to a value NEA than NED. In this case, it is more likely that the policy maker adheres to value 2 than to value 1. Then, the fiscal adviser discloses only value 2 when her willingness to compromise is moderate ($p_{\{1\}}(1) < 1 - \beta < p_{\{2\}}(2)$). Given that she believes the policy maker to not be convinced by the value of efficient allocation, she does not want to propose a project that incorporates the uncertain reaction of the policy maker about value 1 and thus she does

not disclose value 2.

In any case, a fiscal advisor believing in the second welfare theorem can simply take separately the value of equality and efficiency when choosing her disclosure strategy.

2.5 The case of axiologically dependent values

It is worth mentioning that corollary 1 is no longer true when values are axiologically dependent. To see this, consider the following example. The sender wants to implement a tax policy satisfying two values: "fostering social mobility is good" and "being redistributive is good". These two values can be seen, to some extent, as complementary. Thus, let us assign the following axiological intensities to the four disclosure strategies:

$$\mathbf{v}_0 = (0, 0, v(3)); \mathbf{v}(1) = (1/2, 0, v(3)); \mathbf{v}(2) = (0, 1/2, v(3)); \mathbf{v}(1,2) = (2/3, 2/3, v(3)))$$

Note that this situation is perfectly symmetric: no value is superior to the other one.

Assume also that the sender has *symmetric beliefs* about the reaction of the receiver, i.e., she assigns the same probability to every possible reaction of the receiver.

Definition 16. *The sender has symmetric beliefs about the receiver's reaction if*

$$p_{\{1,2\}}(i) = p_{\{1,2\}}(1,2) = p_{\{1,2\}}(\emptyset) = 1/4$$

Then, if

— $\beta < \frac{3}{10}$ she discloses no value

— $\frac{3}{10} < \beta < \frac{33}{64}$ she discloses either only value 1 or only value 2, *but not both*

— $\frac{33}{64} < \beta$ she discloses both values

Interestingly, when $3/10 < \beta < 33/64$, disclosing the first value alone is equivalent to disclosing the second one alone, but only one of these values has to be disclosed. Such a situation was not possible with independent values. Indeed, because the two values are now positively dependent, there is an additional cost of disclosing them simultaneously rather than one at the time. Indeed, when she discloses both values, the sender needs to take into account the risk that the receiver does not adopt them.

Moreover, as compared to the situation of independent values, β needs to be larger in order for the sender to disclose both values. This is due to the extra cost of disclosing two values rather than one at a time. This means that the sender should give more weigh to her own preference in order to disclose both values. However, β has to be lower to get the sender disclosing any value since there is an extra gain from disclosing both values: the axiological state of the sender is further away from the axiological states that the receiver can adhere to.

The algebra to study situations of axiological dependence is tiedous and does not add anything to intuition, but the lessons from this example may be extended in a more

general setting with axiological dependence. To deal with this issue, I keep on assuming that the beliefs of the sender are symmetric. This assumption allows to isolate the effect of axiological dependence by suppressing the effect of specific beliefs. It can be seen as a situation in which the sender is completely ignorant about the receiver's reaction. She thus assigns the same probability to every possible reaction.

In the same spirit, and for the sake of clarity, I assume another kind of symmetry implying that the changes in the axiological intensity of values 1 and 2 are of the same order.

Definition 17. *Preference changes satisfy axiological symmetry if*

$$\alpha_1 = \alpha_2 \tag{3.10}$$

and

$$|\alpha'_1| = |\alpha'_2| \tag{3.11}$$

Equation (3.10) implies that changing value 1 alone has the same effect as changing value 2 alone. Equation (3.11) implies that the axiological relation between values 1 and 2 are of the same axiological intensity. Note that they can be opposite in the case of mixed axiological dependence: one values depending positively on the other, while the latter depending negatively on the former. With this assumption, we can account in a simple way for the effect of the different kinds of axiological dependence. The following proposition states that the intuitions from the example above are true either when both values are positively dependent or when they are negatively dependent.

Proposition 11. *Suppose that the sender has symmetric beliefs and that preference changes satisfies axiological symmetry. Suppose also that we are either in a situation of negative axiological dependence or in a situation of positive axiological dependence (thus $\alpha'_1 = \alpha_2$), then there exist β_1 and β'_2 with $\beta_1 < \frac{1}{2} < \beta_2$ such that :*

- *if $\beta < \beta_1$ the sender discloses no value*
- *if $\beta_1 < \beta < \beta_2$ either she discloses value 1 alone, or she discloses only value 2 alone,* but not both.
- *if $\beta_2 < \beta$ she discloses both values.*

Thus, when she moderately weighs her own preference ($\beta_1 < \beta < \beta_2$), the sender benefits from disclosing either value 1 alone or value 2 alone. In a sense, a moderate sender will randomize her disclosure strategy. This is due to the fact that the axiological dependence has an effect by itself. Moreover, axiological dependence makes disclosure of only one value more interesting for the sender in comparison to the situation of axiological independence. Indeed, the disclosure of any value implies β to be inferior to $1/2$. Yet, given that under the assumption of symmetric axiological expectancy we have that $1/2 = p_1(\emptyset)$,

as stated by proposition 10, in a situation of axiological dependence the sender discloses a value if, and only if $\beta > 1/2$. In this case, the choice of disclosing only one value rests on the effect of the axiological dependence on the sender's preference.

Conversely, the sender is less likely to disclose two values. Indeed, the sender dislcoses two values if β is superior to a number that exceeds 1/2. There is a combined effect of the sender's preference and of the receiver's preference.

When the axiological dependence between values is mixed, the situation has some similarities, but it is more complex. The result can be stated by the following proposition.

Proposition 12. *Suppose that the sender has symmetric beliefs and that preference changes satisfies axiological symmetry. Suppose also that value i depends positively on value $-i$ and value $-i$ depends negatively on value i. Then if $\alpha'_1 = -\alpha'_2$, there exist β_1 and β_2 with $\beta_1 < \frac{1}{2} < \beta_2$ such that :*

— *if $\beta < \beta_1$ the sender discloses no value*

— *if $\beta_1 < \beta < \beta_2$ she discloses only value i*

— *if $\beta_2 < \beta$ she discloses both values.*

Otherwise, there is no gain from disclosing only one value.

First, the sender is always better disclosing the value that is positively dependent than disclosing the value that is negatively dependent. Second, there may be no gain at all to disclose only one value. Third, when there is a gain, the situation is similar to positive and negative dependence.

2.6 Disclosing values in the perspective of the violation of the second theorem of welfare

To illustrate the effect of axiological dependence, consider the subsection 2.4's example of the fiscal adviser, except that now she does not believe that the second theorem of welfare economics holds. There may be distortions that make lump sum transfers impossible. Thus, (at least) from the point of view of the fiscal adviser, there is a trade-off between efficiency and equality. This trade-off can be stated in terms of Rawls' political philosophy suggesting that inequalities are justified as long as they benefit to the poors. The idea is simply that incentives play a key role in the citizens' reaction to tax policy. Again, this may be seen as a background value that is not questioned rather than to a fact. For instance, in October 2017, Edouard Philippe, the french Prime Minister said he "would like the rich people to stop leaving the country" in order to argue in favor of wealth taxation suppression. This idea may be presented as a fact but, considering the lack of consensus about the reality of incentives, it is reasonable to see it as a value. For instance, Thomas Piketty argued that "there is no such hemorrhage".

Thus, in this example, I assume that values 1 and 2 are negatively dependent. I also assume that symmetric axiological expectancy holds and that $v_1 = v_2$. In this case, if the

fiscal adviser is moderate enough, she discloses either only value 1 or only value 2 two but not both. Indeed, she is willing to influence the policy maker, but the negative dependence reduces the gain from the disclosure.

3 Sequential disclosure

In this section, I study a situation in which the sender may choose to sequentially rather than simultaneously disclose values.

3.1 Using partial deliberation sequentially

This section studies the effect of an interesting property of partial deliberation that has been highlighted in the previous chapter: sequential disclosure may differ from simultaneous disclosure. The sender now has two periods for disclosing values 1 and 2 to the receiver. The question is to know whether she has an interest in doing so. The timeline is the following. At period 0, the sender chooses her project. At period 1, she chooses a disclosure strategy and the receiver reacts to this disclosure strategy. At period 2, she chooses a second disclosure strategy and the sender reacts again. The formal version of the timeline is given by figure (3.6).

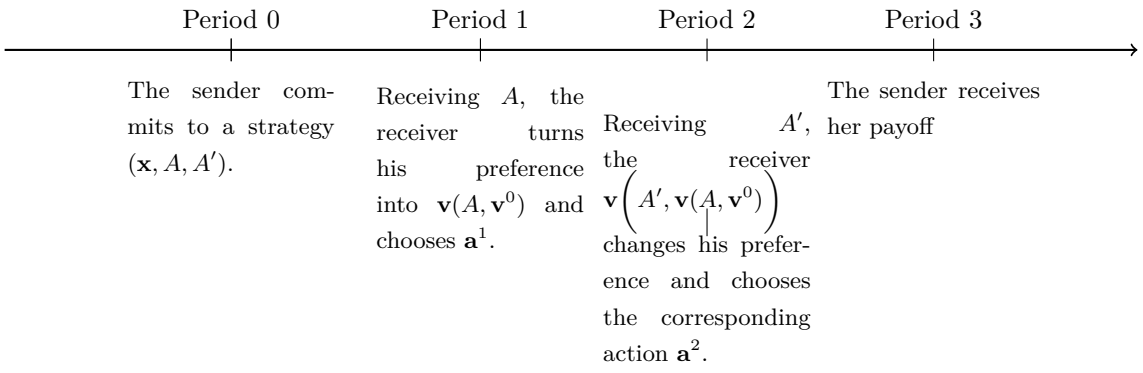

Figure 3.6 – The decision sequence with sequential disclosure

In addition, I assume that the receiver looses his awareness from period 1 to period 2. Therefore, if the sender was to disclose only value i at the first period and value $-i$ at the second period, the outcome would differ from a situation in which she would have disclosed both values simultaneously. This comes from the fact that by becoming aware of value i at period 1, the receiver cannot change the axiological intensity he assigns to value $-i$.

To put this formally, we need to define new payoffs for the receiver and for the sender. The receiver now chooses, at each period, an action that maximizes his payoff at this period. Let $\mathbf{a}^i$ be the action of period i. The payoff of the receiver at period i is simply

given by:

$$u^i(\mathbf{a}^i) = -||\mathbf{a}^i - \mathbf{v}^i||$$

The sender's payoff depends on 1) the distance between the receiver's action and the proposed project and 2) the distance between the sender's preference and the project. I assume that the sender commits to the project at period 0 and that :

$$u^s(\mathbf{x}, A, A') = -\beta||\mathbf{x} - \mathbf{v}^s||^2 - ||\mathbf{x} - \mathbf{a}^1||^2 - ||\mathbf{x} - \mathbf{a}^2||^2$$

s.t

$$\mathbf{a}^i = \underset{\mathbf{a}\in X^2}{Argmax}\ u^i(\mathbf{a})$$

At each period, the best action of the receiver is $\mathbf{v}^i$. Since $\mathbf{v}^i$ depends on both values disclosed at period i (denoted by A^i) and on the receiver's axiological state of the previous period $\mathbf{v}^i$, the sender's payoff can be written as follows:

$$u^s(\mathbf{x}, A, A') = -\beta||\mathbf{x} - \mathbf{v}^s||^2 - ||\mathbf{x} - \mathbf{v}(A, \mathbf{v}^0)||^2 - ||\mathbf{x} - \mathbf{v}(A', \mathbf{v}(A, \mathbf{v}^0)||^2 \tag{3.12}$$

The main question is to determine the conditions under which the sender has an interest in sequentially disclosing values. When this is the case, I say that sequential disclosure has a value.

Definition 18. *There is a value of sequential disclosure whenever there exists a* $\beta \in [0, 1]$ *such that the sender is better off using a sequential disclosure strategy at period 1 and 2, i.e. when :*

$$u(\mathbf{x}_{\{i,-i\}}, \{i\}, \{-i\}) > u(\mathbf{x}_{A,A'}, A, A') \text{ for all } A, A' \subset \hat{V} \text{ for } i \in \{1, 2\}$$

In what follws, I assume that , whatever the awareness of the receiver may be, Φ is single peaked. This ensures that the outcome of any disclosure strategy is unique.

3.2 Benefits from sequential disclosure

Let denote by $\widehat{\mathbf{v}^{i,-i}}$ the sequential disclosure of value i by:

$$\widehat{\mathbf{v}^{i,-i}} = \mathbf{v}\Big(\{-i\}, \mathbf{v}(\{i\}, \mathbf{0})\Big)$$

and $\mathbf{v}^{i,-i}$ the simultaneous disclosure of values i and $-i$ by:

$$\mathbf{v}^{i,-i} = \mathbf{v}\Big(\{i, -i\}, \mathbf{v}(\{i, -i\}, \mathbf{0})\Big)$$

It is easy to understand that when the two decision makers maximize the same axiological criterion, there is no value of sequential disclosure. Indeed, there is no uncertainty in the receiver's reaction to the sender's disclosure strategy. The sender thus knows that she can change the receiver's preference in order to perfectly match her own preference. Thus, she can propose a project $\mathbf{x}$ that perfectly fits her preference ($\mathbf{x} = \mathbf{v}^s$) and discloses both values. She then gets $u^s(\mathbf{v}^s, \{1,2\}, \{1,2\}) = 0$.

Note that when the preferences of both decision makers can be aligned by full disclosure, then the sender has no interest in sequentially disclosing values. In other words, when $\mathbf{v}^s = \mathbf{v}(\{1,2\})$, then $\{1,2\}$ is the only strategy that can achieve a payoff of 0, and 0 is the best possible payoff since u^s is negative. We thus have the following lemma.

Lemma 5. *Whenever* $\mathbf{v}(\{1,2\}) = \mathbf{v}^s$*, there is no value for sequential disclosure.*

It is also worth mentioning that if the sender discloses a set of values A at the first period, there will be no gain to disclose a subset of these values at the second period. Becoming aware of A, the receiver has already assigned the best axiological intensity to A at the first period. He has nothing to change during the second period.

Lemma 6. *For all* $A, A' \subseteq \hat{V}$ *with* $A' \subseteq A$*, we have that* $u^s(\boldsymbol{x}, A, A') = u^s(\boldsymbol{x}, A, \emptyset)$

Moreover, the outcome of a sequential disclosure strategy (A, A') is equivalent to a simultaneous disclosure strategy $(A \cup A', A \cup A')$ when two values are axiologically independent. Therefore, there is no special gain in first disclosing value 1, then disclosing value 2 in comparison to disclose both values together at period 1, even if the sender's and receiver's preferences are not aligned. In other words, because a sequential disclosure strategy gives the same outcome as a full disclosure strategy, there is no reason for the sender to choose it.

Lemma 7. *When values are axiologically independent, sequential disclosure has no value.*

However, this is no longer true when we consider axiologically dependent values. To see this, note that a sequential disclosure strategy will not entail the receiver to adopt the same axiological state with a sequential disclosure strategy than with a full disclosure strategy. When disclosing sequentially i and $-i$, the receiver can change his preference only on the axis of value $-i$ since he is not aware of i anymore. Thus, when the sequential disclosure strategy brings the receiver's preference closer to the sender's preference, there may be a gain from sequentially disclosing values 1 and 2. This fact is illustrated in figure 3.7. On the figure, in case of sequential disclosure, the receiver's axiological state $\mathbf{v}^{\hat{1,2}}$ is closer to the sender's axiological state ($\mathbf{v}^s$) than the receiver's axiological states axiological state is, in case of simultaneous disclosure strategy $\mathbf{v}^{1,2}$. Formally, we have:

$$||\mathbf{v}^{\hat{1,2}} - \mathbf{v}^s|| < ||\mathbf{v}^{1,2} - \mathbf{v}^s|| \tag{3.13}$$

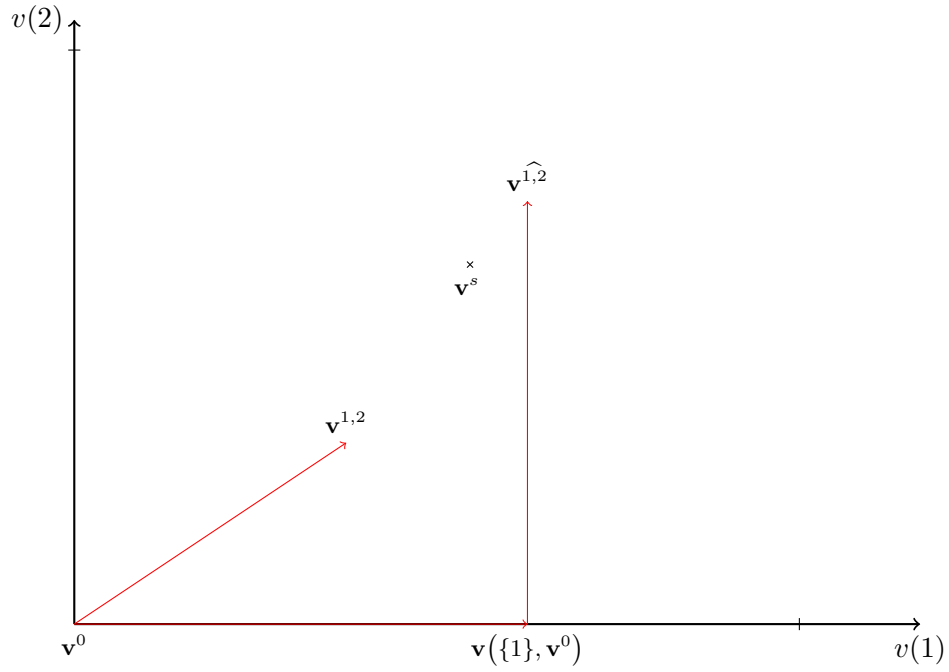

Figure 3.7 – Sequential disclosure strategy versus simultaneous disclosure strategy

This means that, by choosing a sequential disclosure, the sender may be able to choose a project closer to her preference without loosing too much in terms of the receiver's adhesion to her project.

However, satisfying (3.13) is not sufficient. This is established in the following proposition.

Proposition 13. *There is a value of sequential disclosure strategy if and only if there exists i and $-i$ such that,*

$$\left(\mathbf{v}^s|\mathbf{v}^{i,-i}-\mathbf{v}^i\right)^2 \leq ||\mathbf{v}^{\widehat{i,-i}}-\mathbf{v}^s||^2-||\mathbf{v}^{i,-i}-\mathbf{v}^s||^2 \tag{3.14}$$

In this case, the sender uses the sequential disclosure strategy whenever

$$\frac{||\mathbf{v}^{i,-i}-\mathbf{v}^i||^2}{2||\mathbf{v}^{\widehat{i,-i}}-\mathbf{v}^s||^2-2||\mathbf{v}^{i,-i}-\mathbf{v}^s||^2+||\mathbf{v}^{i,-i}-\mathbf{v}^i-\mathbf{v}^s||^2-||\mathbf{v}^s||^2} \leq \beta$$

The intuition behind this result is the following. Since the left side of (3.14) is positive, the right side indicates that the axiological state of the sender ought to be closer to the result of a sequential disclosure strategy than to the result of a simultaneous disclosure strategy. This simply means that, at the second period, there must be a gain from choosing the sequential disclosure strategy rather than the simultaneous one. But this is not sufficient. The gain from second period should also compensate the loss coming from period one. This loss is due to the fact that the sender chooses her project in order

to maximize her utility at both periods. However, given that the disclosure strategy is sequential, the axiological states of the receiver differs from one period to the other and so does his utility. Therefore, there is an extra cost for sequential disclosure. This cost is given by the scalar product between the axiological intensity the sender assigns to the secondly disclosed value ($\mathbf{v}^s$) and the change of the axiological intensity he assigns to this second value between the first and the second period ($\mathbf{v}^{i,-i} - \mathbf{v}^i$).

Note that proposition 13 gives a necessary condition on the distance between the final result of both simultaneous and sequential disclosure strategies, in order the sequential disclosure strategy to be valuable.

Corollary 2. *If sequential disclosure strategy has a value, then there exists $i, -i$ such that:*

$$\left(\mathbf{v}^s|\mathbf{v}^{i,i} - \mathbf{v}^{-i}\right) \leq ||\widehat{\mathbf{v}^{i,-i}} - \mathbf{v}^{i,-i}|| \tag{3.15}$$

This implies that whenever a sequential disclosure strategy has a value, the distance with the simultaneous disclosure strategy should be higher than the loss of the first period.

3.3 Sequential disclosure in fiscal disclosing

Let me illustrate the effect of sequentiality with the fiscal advisor example. Assume that we have a fiscal adviser similar to the fiscal adviser of section 2 but, additionally, she has a special interest in a project that promotes efficiency rather than equality. This may be the case if she represents a specific group of interest like a lobby or an elite class. A symmetric case can be given by an agent defending the interest of labour classes and that has a special interest in a project promoting equality. Then, her utility is still given by equation (3.5) but now β represents a value like "favoring your social class is good". This additional value cannot be confessed to the policy maker, because he can only change the axiological intensity he assigns to values of efficiency and inequality. This means that the preferences of both decision makers are not aligned.

According to lemma **??**, whenever the values are independent, the fiscal adviser has no interest in using a sequential disclosure strategy. Thus, when the fiscal adviser is a supporter of the second theorem of welfare economics, there is no gain in sequential disclosure. However, when values are axiologically dependent, there may be a gain from disclosing these values sequentially.

4 Concluding remarks

In this chapter, I apply partial deliberation to a case in which a sender tries to manipulate the preference of a receiver. To do so, I develop an euclidean framework for partial deliberation according to which the receiver changes the axiological intensity that he assigns to the values he becomes aware of. With this apparatus, I propose to model two situations.

In the first application, the sender is uncertain on the receiver's reaction because she she is aware that she is unaware of his background values. I show that depending on how the sender prefers that the receiver adopts her value system, there is room for any kind of disclosure strategy. I also show that when values are axiologically independent, the sender prefers a partial disclosure strategy when the weigh she assigns on her own preference is located between her belief on the reaction of the receiver to the dislosure of value i and her belief on his reaction to the dislosure of value 2. When values are not axiologically independent the situation is trickier. However the logic is similar: the higher the sender's motivation to see the receiver adopting her axiological state, the more values she is likely to disclose. Moreover, axiological dependence implies that the sender may have an interest in randomizing her dislosure strategy. This is due to the fact that the axiological dependence has an effect by itself.

In the second application, I study the conditions in which the sender has an interest in using a sequential disclosure strategy. I show that the result from the sequential disclosure strategy must necessarily be closer to the axiological state of the sender, than the result from the simultaneous strategy. Moreover, it must be closed enough to compensate the loss from sequentially disclosing the values.

These results give a first insight on how partial deliberation could be applied in a context including several decision makers. Future research will aim at using this framework in an interactive context.

Chapitre 4

CONCLUSION

Dans cette thèse, j'ai construit et étudié un mécanisme psychologique permettant de rendre compte des changement de préférences qui résultent d'une prise de conscience : la délibération partielle. Pour motiver cette étude, j'ai principalement invoqué des raisons abstraites : la délibération partielle offre une troisième voie permettant de concilier causes externes et causes internes dans l'explication des comportements. Bien que, comme nous l'avons vu dès l'introduction de cette thèse, certains travaux aient également motivé leur apport par la recherche de cette troisième voie, la mienne se singularise par le fait que les changement de préférences y reposent sur la complémentarité entre causes internes et externes. Dans une certaine mesure, cependant, il reste à prouver que répondre à ce type d'interrogation est susceptible d'introduire une différence utile avec ce qui a déjà été établi par la littérature, et qui jette une lumière nouvelle sur certains phénomènes des sciences sociales.

Au cours de cette conclusion je suggère que la délibération partielle peut expliquer et modéliser d'une façon relativement simple et instructive la relation entre changement de préférences et d'autres phénomènes évolutifs importants. Elle permet de conceptualiser avec précision, et dans le vocabulaire de la psychologie du sens commun, des intuitions qui sont implicitement partagées, sans pourtant être bien comprises et ou modélisées sous cet angle. Notamment elle permet :

1. d'articuler dynamique des prises de conscience et changements de préférence,
2. d'articuler évolution du capital humain et évolution des préférences,
3. d'articuler dynamique des groupes sociaux et évolution des préférences,
4. d'articuler le rôle de l'empathie imparfaite, ainsi que de la tolérance dans les changements de préférences.

Or, si l'apport conceptuel de la délibération partielle peut être prolongé dans toutes ces directions, c'est précisément dans la mesure où elle repose sur une frontière entre ce qui cause le comportement de façon interne (ce qui vient à la conscience de l'agent) et ce qui la cause de façon externe (ce qui le détermine indépendemment de sa volonté).

Toutefois, que la délibération partielle offre quelque avantage conceptuel ne doit pas nous conduire à négliger la cruciale question de son appréciation empirique. Car la délibération partielle repose sur des concepts obscurs de conscience et de valeur. Autrement dit, le gain conceptuel apporté par le fait de rendre compte de la dichotomie entre différentes causes du comportement (causes internes et causes externes), pourrait aller de paire avec des difficultés sur le plan empirique. La dernière partie de cette conclusion s'intéresse à ce problème, revient sur ces enjeux et défend l'idée qu'une epistémologie conciliant différents niveaux de réfutation, et qui rende possible le dialogue entre différentes approches des sciences sociales, est nécessaire. Or c'est précisément à ce type d'épistémologie que la délibération partielle entend contribuer.

1 Synthèse des résultats de la thèse

Avec la délibération partielle, je me suis donné pour objectif de formaliser une platitude de la psychologie du sens commun qui veut que les changement de préférences procèdent de la découverte par un agent de nouvelles valeurs. A elle seule, cette raison suffit à rendre la délibération partielle digne d'intérêt. Car même s'ils ne parvenaient jamais à changer les préférences de leurs semblables en leur faisant prendre conscience de nouvelles valeurs, il n'en demeure pas moins intuitif que les individus agissent comme s'ils en étaient capables. La délibération partielle propose donc, *a minima*, de formaliser ce sur quoi repose le raisonnement des individus lorsqu'ils tentent de changer les préférences des individus qui les entoure.

Dans le premier chapitre, je discute l'armature conceptuelle de la délibération partielle en la caractérisant par cinq hypothèses :

1. les préférences sont fondées sur des valeurs,
2. un agent est potentiellement conscient de certaines de ces valeurs,
3. il existe un critère axiologique susceptible de classer les ensembles de valeurs,
4. la conscience offre à l'agent la capacité d'adhérer à (ou d'abandonner) une valeur,
5. l'agent choisit à chaque instant le meilleur système de valeur que sa conscience met à sa portée au sens du critère axiologique.

A l'appui de ces hypothèses, je montre ensuite en quoi la délibération partielle permet effectivement de concevoir une troisième voie telle que causes internes et causes externes de l'évolution du comportement sont complémentaires. Je définis cette complémentarité par le fait que dans le cadre de la délibération partielle, si seule l'une des formes de causes (internes ou externes) était à l'oeuvre il n'y aurait pas lieu pour l'agent de changer ses préférences.

Au cours du second chapitre, j'étudie les conditions nécessaires et suffisantes que doit vérifier un ensemble de changement de préférences pour que ces cinq hypothèses soient

conjointement compatibles. C'est-à-dire qu'à l'aide de six axiomes je montre quels changement de préférences peuvent être rationalisés par une relation d'ordre sur les ensembles de valeurs, telle que l'agent choisit de changer de valeurs en 1) maximisant cette relation d'ordre, 2) sous la contrainte de ce que sa conscience met à sa portée. Avec cette caractérisation, certains changement de préférences sont théoriquement exclus, ce qui permet de répondre au problème de l'adhocité des préférences.

Sous cette forme générale, les changement de préférences induits par la délibération partielle sont susceptibles de vérifier deux propriétés intéressantes :

1. d'abord, les valeurs d'arrière-plan, c'est-à-dire les valeurs dont l'agent n'a pas conscience mais auxquelles il adhère, jouent un rôle dans la façon dont il change ses préférences,
2. ensuite, prendre séquentiellement conscience de nouvelles valeurs n'induit pas nécessairement les mêmes changement de préférences qu'une prise de conscience simultanée de ces mêmes valeurs.

Ces résultats constituent l'analyse de la délibération partielle dans sa pleine généralité. La suite du chapitre 2 s'attaque à deux problèmes.

En premier lieu, à l'aide du modèle de [Dietrich and List, 2013*b*], j'y établis une liaison entre délibération partielle et changements de comportement. De cette analyse, je tire deux enseignements significatifs et qui me semblent suffisamment instructifs pour inciter à davantage d'investigation en la matière :

— d'abord, axiomatiser la délibération partielle directement en termes de fonction de choix supposerait des restrictions supplémentaires sur l'ensemble des options possibles pour le décideur. Cela supposerait qu'il soit capable de construire des paniers de biens tels que l'observation des choix permette d'identifier le système de valeurs induisant le choix. Ce résultat suggère des développements possibles, notamment pour établir des conditions permettant de révéler les systèmes de valeurs en termes de choix,

— ensuite, la délibération offre une perspective sur les changement de préférences qui va à rebours des phénomènes de dissonance cognitive.

En second lieu, j'y caractérise différents critères axiologiques munis d'un peu plus de structure, pour finalement les utiliser pour deux applications. Le premier critère axiologique consiste en une structure de type « tout est permis ». Elle suppose que toutes les valeurs sont bonnes à prendre pour l'agent. Je montre que ce critère axiologique est tout simplement la relation d'inclusion. Cette simplicité a cependant pour conséquence qu'une telle structure n'exhibe aucune des propriétés mentionnées plus haut :

— avec la structure « tout est permis » les valeurs d'arrière-plan n'influencent pas le processus de changement des préférences de l'agent ;

— avec la structure « tout est permis » il n'y a pas d'effet de la séquentialité de la prise de conscience de l'agent.

A l'inverse, la seconde structure que je caractérise rend possible ces propriétés. Cette seconde structure est « partitionnelle », et suppose les valeurs réparties en classes antagonistes.

Le dernier chapitre modélise une situation où un agent (un envoyeur) tente d'en manipuler un autre (un receveur). Pour ce faire, il s'appuie sur une approche euclidienne de la délibération partielle. Cette approche permet de définir des concepts de distance entre les préférences des agents.

Dans un second temps, je modélise une situation d'*empathie imparfaite*, où l'envoyeur est incertain quant à la façon dont le receveur pourrait réagir à sa stratégie de divulgation. Conceptuellement, le concept d'empathie imparfaite se fonde donc sur l'existence de valeurs d'arrière-plan dont l'envoyeur ignore si le receveur leur accorde une valeur. Je m'intéresse tout particulièrement à distinguer les situations où les valeurs que peut divulguer l'envoyeur entretiennent des relations axiologiques. Je montre que l'indépendance implique simplement pour l'envoyeur de s'intéresser à chacune des valeurs une à une, alors qu'en cas de dépendance, le fait de ne divulguer qu'une valeur peut avoir *en soi* un intérêt.

Enfin, j'aborde les situations dans lesquelles le receveur peut chercher à tirer partie du fait que, lorsqu'il prend séquentiellement conscience des valeurs, le receveur change ses valeurs d'une manière différente que lorsqu'il en prend conscience simultanément. Je montre que la divulgation séquentielle peut profiter au receveur dès lors que ses préférences ne peuvent être alignées avec celles du receveur, et que la stratégie de divulgation partielle rapproche suffisamment leurs préférences mutuelles pour compenser la perte due à la séquentialité de la divulgation.

2 Articuler délibération partielle et d'autres phénomènes évolutifs

La délibération partielle présuppose le recours au concept de prise de conscience. Or, je le suggère dans cette section, il est commode d'utiliser ce concept pour rendre compte de nombreux phénomènes évolutifs comme l'acquisition de capital humain, la formation des groupes sociaux et la représentation d'autrui. Pour cette raison très simple, il est possible d'articuler la délibération partielle avec ces phénomènes évolutifs.

Dans l'analyse conduite au cours de cette thèse, je suppose la conscience de l'agent exogène. Toutefois, il est parfaitement possible que les changements de préférencesexpliquent la dynamique de la prise de conscience de l'agent et ce pour trois raisons complémentaires mais distinctes.

— La première est que certaines options de choix sont susceptibles de produire plus de prise de conscience que d'autres. On peut suspecter que préfèrer voyager pour découvrir de nouvelles cultures est susceptible d'apporter davantage de conscience de nouvelles valeurs que la consommation de psychotropes à toute heure du jour.

— La seconde tient au fait que l'agent peut être conscient d'être inconscient de certaines valeurs. Autrement dit, il est conscient qu'il pourrait être ignorant. Se pose alors la question de son aversion (ou bien au contraire de son appétence) à cette éventualité qui est spécifiée par ses préférences, et des conséquences d'une telle aversion. Le pousse-t-elle à tenter de prendre conscience de nouvelles valeurs ? Le pousse-t-elle au contraire à éviter de rechercher de nouvelles valeurs ? Là encore, ces réactions pourraient dépendre de ses préférences.

— Dans l'esprit de Dewey and Tufts [1932] on peut également arguer que l'existence d'un conflit latent entre les valeurs de l'agent peut le pousser à chercher la résolution de ce conflit par l'extension de sa conscience.

2.1 Délibération partielle et acquisition de capital humain

Quelle relation peut-il y avoir entre l'acquisition de capital humain et l'évolution des préférences individuelles ? L'intérêt d'une telle question est aussi bien théorique que pratique. L'acquisition de capital humain constitue l'objectif central de nombreuses institutions comme le système éducatif, les formations internes aux firmes ou le tissu associatif. Pour évaluer ces institutions, il est nécessaire de comprendre comment elles permettent d'atteindre deux types d'effets recherchés. Le premier consiste à former les individus au marché du travail, le second à promouvoir une certaine cohésion sociale. Avec la délibération partielle il est possible de mieux comprendre comment ces deux objectifs peuvent être conceptuellement imbriqués.

Le premier effet recherché consiste à permettre d'augmenter les capacités productives des individus pour que l'offre de travail qualifié corresponde à sa demande tout en favorisant une croissance durable et équilibrée.

Une telle question prend une portée nouvelle avec le développement de la robotique et de l'intelligence artificielle qui, pour bon nombre d'économistes, préfigure de profondes mutations des rapports que nous entretenons au travail. Cette révolution supposée tend à rendre les tâches généralistes plus compétitives sur le marché du travail que les tâches plus spécialisées, davantage menacées de remplacement par des machines [1]. Or, il est possible de distinguer travailleurs généralistes et travailleurs specialisés en utilisant les concepts de conscience et de connaissance. Le travailleur généraliste est conscient de beaucoup de choses, mais il ne les connait pas nécessairement. Dans son esprit, plusieurs états du monde sont envisageables mais il sait avec moins d'acuité que le spécialiste si ces états sont vérifiés. Le spécialiste, à l'inverse est conscient de peu de choses, mais il connait le peu de choses dont il est conscient. Par conséquent, les changements sur le marché du travail que laisse augurer l'intervention des nouvelles technologies suggèrent qu'un nouveau rapport

1. En toute rigeur, ce sont les tâches standardisées qui sont les plus à même d'être remplacées. Ici, je suggère qu'il y a une corrélation entre spécialisation et standardisation. On sait par exemple que les tâches les plus à même d'être remplacées dans le domaine de la médecine sont non pas les moins qualifiées, mais les plus spécialisées, comme la chirurgie.

à la conscience des individus soit favorisé par rapport à un autre. Les individus sont davantage généralistes et donc plus conscients (sans être connaisseurs). Ce phénomène recoupe une idée, largement mise en avant par la littérature, qu'avec le développement des technologies du numérique, la connaissance est beaucoup moins couteuse à acquérir et que, par conséquent, les compétences des individus consistent davantage en le fait d'être capable d'aller chercher l'information si nécessaire, que d'être capable de la mobiliser à tout moment.

Si donc la délibération partielle permet bien de rendre compte de l'effet des prises de conscience sur les préférences des individus, on peut supputer que ces mutations sur le marché du travail sont susceptibles d'avoir d'importantes conséquences en matières de préférences sociales, idéologiques et politiques.

Cela nous conduit au second effet recherché de l'acquisition de capital. Celui-ci souligne les effets du mode d'aqcuisition des compétences choisi sur les comportements sociaux en général. Par exemple, si l'éducation permet assurément d'offrir de nouvelles compétences aux individus, elle a rempli un rôle déterminant dans l'émergence d'un sentiment national au sein des Etats nations du XIXe siècle. Il semble aujourd'hui clair que la lutte contre les discriminations sous toutes leurs formes doit passer par l'éducation. Elle est susceptible d'orienter le comportement des familles, comme dans l'Etat indien du Kérala, où la croissance démographique a pu être maitrisée simplement par la généralisation de l'instruction publique. Ces propriétés du système scolaire ont été empiriquement soulignées, entre autres, par Robinson and Acemoglu [2012]. Autrement dit, dans la mesure où elle peut engendrer des changements de préférences, l'acquisition de capital humain est susceptible de produire des externalités, dont il est indispensable de comprendre la portée pour évaluer des institutions dont l'objectif principal est l'acquisition de compétences.

A ce titre, l'éducation est sans doute l'une des institutions qui recueille le plus de sympathie. Chacun reconnait son rôle capital dans les changements de comportements sociaux. Il s'agit donc d'une platitude de la psychologie du sens commun que d'affirmer que l'acquisition de capital humain implique des changements de préférences. Cette idée est si banale qu'il est difficile de rencontrer des projets politiques de long terme qui fassent l'économie d'un discours sur la façon dont la politique éducative doit être conduite et sur les contenus qu'elle doit être à même de transmettre. Elle est un secteur stratégique de toute entreprise politique de long terme, comme en témoigne de nombreux exemples historiques comme celui des jeunesse hitlériennes, du rôle prépondérant qu'elle devait jouer dans l'avènement d'un « homme nouveau » communiste ou des scouts anglais.

Je soutiens que ces idées sont susceptibles de trouver une expression simple avec la délibération partielle. La délibération partielle permet de relier évolution du capital humain et évolution des préférences car ces deux phénomènes supposent l'intervention d'une prise de conscience. Si l'acquisition de capital humain engendre la prise en compte de nouvelles valeurs, elle peut être responsable de changements de préférences. Ainsi pourraient être pensées les conséquences des mutations induites par la révolution numérique - le premier

effet recherché - sur les comportements politiques - le second effet recherché. En un sens, il est possible d'arguer que ces conséquences se perçoivent déjà dans les bouleversements du rapport que les individus entretiennent avec la politique. En occident, les citoyens sont moins enclins à se revendiquer d'une idéologie figée qu'ils ne l'étaient au sortir de la seconde guerre mondiale, leurs attitudes politiques sont plus mouvantes.

2.2 Articuler formation des groupes sociaux et évolution des préférences

De la sociologie à l'anthropologie en passant par la psychologie sociale, ainsi que de nombreuses sphères empiriques de l'économie contemporaine, nombreuses sont les disciplines à insister sur le rôle des groupes ou communautés d'appartenance dans la formation des préférences individuelles. La question dès lors est celle des mécanismes psychologiques susceptibles d'expliquer commment des groupes sociaux constitués autour de certains types de préférences peuvent se former et s'auto-perpétuer.

La délibération partielle donne un cadre suffisamment général pour interpréter et théoriser différentes manière d'aborder ce problème. Si l'on admet que les agents sont capables de se rendre mutellement conscients des valeurs dont ils ont conscience, et qu'ils ont tendance à mettre en exergue les valeurs auxquelles ils adhèrent, on comprend aisément comment des individus qui appartiennent à un même groupe ont tendance à voir leurs préférencesconverger vers celle du groupe. Ce mécanisme se renforçant d'autant plus que leur appartenance commune résulte du fait qu'ils aient les mêmes préférences.

Dans sa plus grande simplicité, un tel schéma conduirait à une polarisation extrême des groupes sociaux. C'est d'ailleurs par ce type de raisonnement que les tenants de la théorie huntingtonienne rendent raison du prétendu *choc des civilisations*. Sous la plume de ces auteurs, des prévisions apocalyptiques vont souvent de paire avec une hiérarchisation plus ou moins explicite des civilisations. Seulement, dans toute sa généralité, la délibération partielle ne conduit pas nécessairement à une telle conclusion. Comme l'écrit très justement Sen dans *Identité et violence*, ce n'est pas parce que l'appartenance à des groupes sociaux joue un rôle considérable dans l'évolution des comportements humains, qu'il faut considérer que les groupes sociaux forment des entités homogènes et étanches. A l'instar d'Akerloff et Kranton, il est peut-être utile de voir les individus comme des ensembles d'identités. En raisonnant sur des ensembles de valeurs, c'est exactement ce qu'est à même de faire la délibération partielle. Par ailleurs, elle n'implique pas de hiérarchisation sur les valeurs que peuvent porter les différentes communautés mais permet de réflechir sur le type de critère axiologique que supposerait une hiérarchisation particulière. C'est l'une des raisons pour laquelle le travail entamé au cours du chapitre 2 sur les différentes structures que peuvent revêtir le critère axiologique est digne d'intérêt.

2.3 Articuler le rôle de l'empathie imparfaite, ainsi que de la tolérance dans les changements de préférence

Ces dernières réflexions posent également la question du type de relation que les citoyens d'une société hétérogène peuvent entretenir face à leur différence. Du point de vue du décideur politique, il peut être utile de se demander comment favoriser un certain rapport aux préférences qu'un individu peut entretenir avec les préférences de ses semblables. Dans le chapitre 3, j'ai commencé à élaborer une théorie de l'empathie imparfaite. Ces réflexions méritent toutefois d'être poussées dans plusieurs directions.

Un exemple d'application de ce type d'analyse porte sur l'élaboration d'une théorie de la tolérance. La tolérance est sans doute l'un des traits de comportements le plus souhaitable. Comme l'écrivent Corneo and Jeanne [2009, p. 692] « [u]ne théorie de la tolérance doit être en mesure d'identifier les circonstances sous lesquelles les parents peuvent avoir intérêt à éduquer leur progéniture à être ouvert d'esprit, i.e., a transmettre un système de valeurs qui attribue une valeur equivalente à différents traits et styles de vie ». Or la tolérance d'un individu peut être interprétée comme une tendance à voir dans le comportement d'autrui, lorsqu'il est inexplicable, le résultat de l'influence de valeurs dont il se suppose inconscient.

Par ailleurs, dans le chapitre 3, j'ai eu recours à l'idée que les agents se conçoivent comme potentiellement inconscients de certaines valeurs. A force de constater que le comportement de ses semblables ne correspond pas au comportement qu'ils devraient choisir s'ils avaient ses préférences, il est envisageable qu'un agent rationnel devienne de plus en plus en mesure de juger probable qu'il découvrira de nouvelles valeurs. Cela implique que l'observation du comportement des autres peut être susceptible de pousser un agent à chercher des nouvelles valeurs susceptibles de rationaliser le comportement des autres. Un tel processus d'apprentissage ne garantit pas qu'il découvre de fait les valeurs qui expliquent le comportement de ses congénères. Mais il peut le conduire également à en inventer des nouvelles, et changer ainsi de préférences par l'observation du comportement des autres, sans pour autant adopter les mêmes préférences. L'intérêt de cette idée est qu'elle pourrait rendre compte du fait que les agents sont incités à inventer de nouvelles valeurs du fait de l'observation d'autrui.

3 Analyse empirique de la délibération partielle

Il convient pour finir de donner mon sentiment sur un problème important et que j'ai largement négligé tout au long de cette thèse : y a-t-il un traitement empirique de la délibération partielle ? Je me contenterais ici d'aborder cette question à propos des valeurs et des relations qu'elles peuvent entretenir, bien que de nombreuses questions empiriques se posent à propos de la conscience également.

3.1 La mesure des valeurs

L'analyse empirique des valeurs, en particulier leur mesure, n'est pas aisée. La plupart des travaux empiriques à ce sujet reposent sur des questionnaires, dont la validité dépend d'un cadre psycho-théorique chargé. C'est par exemple le cas des mesures de [Schwarz and Strack, 1991]. En outre, ces mesures sont fondées sur des conceptions très restrictives des valeurs et qui, partant, diffèrent de ce que j'entends par le concept de valeur. Contrairement à la délibération, il est admis dans cette littérature que les valeurs potentielles sont très peu nombreuses et sont connues de tous.

Par ailleurs, et c'est le problème sur lequel je voudrais insister, les mesures déclaratives employées par ces auteurs laissent encore de nombreuses zones d'ombres sur leur capacité à servir d'indicateur comportemental. Une enquête assez célèbre, le GLOBE, met en évidence une corrélation négative entre les valeurs des individus d'un même pays et les pratiques que les individus semblent révéler dans ces pays. Le même type de défauts caractérise les données utilisées par une imposante partie de l'économie politique qui s'intéresse aux valeurs, aux idéologies ou à l'évolution des préférences.

Ce problème du décalage entre ce que les agents disent d'eux-mêmes et leur comportement est considérable. Y répondre suppose d'adopter l'une ou l'autre des deux attitudes suivantes :

— soit considérer qu'il est insurmontable et conclure qu'il faut se restreindre aux préférences révélées. Auquel cas l'ensemble des travaux réalisés en économie politique qui font usage des concepts de valeurs, d'identité ou d'idéologie sont condamnés à reposer sur des fondements douteux,

— Soit nous tentons de spécifier théoriquement comment les valeurs peuvent fonder le comportement. Tout en proposant une explication du décalage entre les déclarations des agents et leur comportement.

Ma position est que la seconde attitude est difficile mais nécessaire et c'est celle que j'ai essayé d'adopter avec la délibération partielle. Pour s'en convaincre, on remarquera que j'ai souligné au chapitre premier qu'être conscient de valeurs, c'est-à-dire être en mesure de les déclarer, n'implique pas d'y adhérer comportementalement. La délibération partielle offre donc une interprétation du décalage entre déclaration et comportement.

Par ailleurs, bien que des développements ultérieurs soient nécessaires, j'ai entamé une réflexion sur la connexion entre changement de système de valeur et changement de choix. Ces travaux sont à poursuivre à n'en pas douter.

3.2 Révélation des systèmes de valeurs

Comment révéler le type de relations possibles entre les valeurs induites par le critère axiologique ? Je voudrais défendre une approche pragmatique de cette question, selon laquelle, il y a plusieurs manières de mener de telles études, chacune avec ses avantages et

ses inconvénients. Une étude systématique de la façon dont le valeurs évoluent consiste à combiner ces études.

La première est celle conduite par les philosophes « dans leur fauteuil » et consiste en des expériences de pensée. En permettant l'analyse conceptuelle, elles offrent la possibilité d'étudier le type de relation envisageable entre les valeurs et de clarifier celles qui correspondent à nos intuitions morales. Elle permet notamment d'expliciter nos intuitions. Un exemple du succès d'une telle approche peut être illustré par l'expérience de pensée de la machine à plaisir. Son auteur, Robert Nozick, cherche à contrer l'affirmation utilitariste selon laquelle le plaisir serait l'unique valeur digne d'intérêt pour un agent moral. Pour ce faire, Nozick demande à son lecteur d'imaginer une « machine à plaisir », susceptible d'accueillir des individus et de leur donner un plaisir infini tout au long de leur existence. Pouvons-nous dire d'un individu qui passerait sa vie dans une telle machine que sa dignité morale a été respectée ? Nozick répond par la négative. Nous ne voulons pas vivre une vie où toutes nos satisfactions ont été simulées. Nous ne voulons pas d'une machine qui simule l'amour, nous voulons tomber amoureux. Nous ne voulons pas d'une machine qu'elle simule la réalisation de nos projets, nous voulons réaliser nos projets. Bien entendu, cette approche repose en dernière instance sur l'intuition, ce qui ne lui donne pas la validité d'une étude expérimentale. Elle permet d'établir des relations crédibles entre diffèrents systèmes de valeurs, mais aussi de se comparer avec celles reportées par les agents. Elle peut être en outre complétée par des travaux de philosophie expérimentale.

Un second type d'approche est constitué par 1) l'histoire de l'émergence de nouvelles valeurs et 2) leur diffusion. Ces deux processus méritent d'être distingués car les valeurs nouvelles peuvent tout aussi bien naître de sphères sociologiquement distinctes [2], que s'imposer « matériellement »à l'ensemble d'une population [3]. Dans chacun de ces cas, le mécanisme de diffusion est diffèrent. Alors que dans le premier, la mise en évidence d'une valeur ne suppose pas l'adhésion automatique tant des sphères dirigeantes, des experts ou du grand public ; dans le second, il est nécessaire de relier la transformation des conditions matérielles d'existence des individus à l'avènement progressif d'une valeur nouvelle. Le type de liaison qui s'opère entre ces sphères est un objet d'étude en soi. L'histoire des idées révèle la difficulté de certaines valeurs à pénétrer le terrain intellectuel d'une époque particulière. Certains contextes intellectuels sont plus propices à la compréhension de valeurs données, et à leur adhésion. On retrouve ici les ingrédients de ce que Foucault qualifiait d'*épistémé*. En soulignant ces difficultés, l'histoire des idées suggère *a posteriori* des transitions possibles entre certains systèmes de valeurs. Les outils employés par cette

2. Je ne prétends pas que seul les chercheurs, artistes ou autres créateurs soient susceptibles de mettre en exergue de nouvelles valeurs. En revanche, je prétends que les nouvelles valeurs sont créées par une communauté qui rencontre des problèmes particuliers et dont les préoccupations ne sont pas automatiquement celles de l'ensemble de la population. Par ailleurs, il ne s'agit pas de dire que les créateurs de nouvelles valeurs soient susceptibles d'être extrait du contexte social et intellectuel qui est le leur. A ce titre, histoire interne et histoire externe se complètes.

3. Bien que j'emploie le terme d'histoire des idées, je ne crois que la délibération partielle nous impose d'adhérer à une forme d'idéalisme radical.

histoire doivent être diverse : du recueil d'archives à l'analyse des données, elle fournit un ensemble d'indices non-négligeables susceptibles de corroborer certaines liaisons entre les systèmes de valeurs. Une telle histoire doit également donner la part belle à des comparaisons internationales.

Enfin, un troisième type d'approche repose sur l'étude expérimentale de l'évolution des valeurs. S'il existe quelques expériences en laboratoire [Maio and Olson, 1998], le recueil des données sur les réseaux sociaux pourrait également constituer une ressource empirique considérable. En particulier, les algorithmes de recommandation disposent à la fois de données sur l'attention des individus, leurs choix et comment ils semblent évoluer.

A ce sujet, les outils de l'intelligence artificielle sont oscurs. Ils ne donnent pas d'explication de ce pourquoi les individus devraient choisir telle ou telle recommandation. Développer davantage la délibération partielle pourrait donc aider à interpréter les changements de comportements tels que les machines nous les révèlent.

Chapitre 5

Appendix

1 Proofs of Chapter 2

1.1 Proofs of Section 2

Proof of proposition 2 : For this proof we write $(V, V') \in \trianglelefteq$ for $V \trianglelefteq V'$. Let be $(\rightarrow_A)_{A\in\mathcal{V}}$ a preference formation rule induced by partial deliberation and $E^{\rightarrow}$ the set containing every $\trianglelefteq$ inducing $(\rightarrow_A)_{A\in\mathcal{V}}$.

We first prove in three steps that

$$\bowtie \subseteq \left(\trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \right)^{sym}$$

Step 1 : *Notations*

Let $V, V' \in \mathcal{V}$ such that $V \bowtie V'$ and let $\trianglelefteq \in E^{\rightarrow}$. Since $\trianglelefteq$ is complete we have either $(V, V') \in \trianglelefteq$ or $(V', V) \in \trianglelefteq$. Suppose w.l.o.g that $(V, V') \in \trianglelefteq$. Let $\trianglelefteq' = \Big(\trianglelefteq \cup (V', V) \Big) \setminus (V, V')$. Denote also $(\rightarrow'_A)_{A\in\mathcal{V}}$, $\bowtie'$ and $\bowtie'_1$ the preference formation rule induced by $\trianglelefteq'$.

Step 2 : *Let's prove that $\trianglelefteq'$ induces $(\rightarrow_A)_{A\in\mathcal{V}}$.*

To prove this, we have to show that $\forall (V, V') \in \mathcal{V}^2$, $V \rightarrow_A V' \iff V \rightarrow'_A V'$ and $\bowtie = \bowtie'$. So, note that we have :

— $\forall (V'', V''') \neq (V, V')$ and A

$$V'' \rightarrow_A V''' \iff \begin{cases} \exists B \subseteq A, V''' = B \cup (V'' \setminus A) \\ \forall B' \subseteq A, B' \cup (V'' \setminus A) \trianglelefteq V''' \end{cases}$$

$$\iff \begin{cases} \exists B \subseteq A, V''' = B \cup (V''\backslash A) \\ \forall B' \subseteq A, B' \cup (V''\backslash A) \trianglelefteq' V''' \end{cases} \iff V'' \rightarrow'_A V'''$$

— (V, V') and all A since $\bowtie \subseteq \bowtie_1$ the fact that $(V, V') \notin \bowtie$ implies that for all A such that $V' = B \cup (V\backslash A)$ with $B \subseteq A$, there exists $B' \subseteq A, B' \cup (V\backslash A) \triangleright V' \trianglerighteq V$ and $B'' \subseteq A, B'' \cup (V'\backslash A) \triangleright V \triangleleft V'$. Thus there exists $B' \subseteq A, B' \cup (V\backslash A) \triangleright' V' \trianglelefteq' V$ and $B'' \subseteq A, B'' \cup (V'\backslash A) \triangleright' V \triangleright' V'$. Thus if there is no A such that $V \rightarrow_A V'$, there is no A such that $V' \rightarrow'_A V$. So $\bowtie_1 = \bowtie'_1$. But since for all $(V'', V''') \neq (V, V')$, we have that $V'' \rightarrow_A V'' \iff V'' \rightarrow'_A V'''$ we also have that $\bowtie = \bowtie'$.

Therefore $\trianglelefteq'$ induces $(\rightarrow_A)_{A \in \mathcal{V}}$ and $\trianglelefteq' \in E^{\rightarrow}$.

Step 3 : *Conclusion*

Hence, since $(V, V') \notin \trianglelefteq'$ we have that $(V, V') \notin \bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq'$. :

$$(V, V') \in \trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \subseteq \left(\trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \right)^{sym}$$

Since we can build $\trianglelefteq'$ in the same fashion for any V, V' such that $V \bowtie V'$, we have that :

$$\bowtie \subseteq \left(\trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \right)^{sym}$$

Conversely, let $\trianglelefteq \in E^{\rightarrow}$ take

$$(V, V') \in \trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right)$$

If we had that $(V, V') \notin \bowtie$, then there would be a transition path between V and V'. W.l.o.g assume this transition path is $V \rightarrow_{(A_n)} V'$. Thus by transitivity $V \trianglelefteq V'$. Let $\trianglelefteq' \in E^{\rightarrow}$ such that $(V, V') \notin \trianglelefteq'$. By completeness of $\trianglelefteq'$ we, thus, have that $(V', V) \in \triangleleft'$ which contradicts the fact that $V \rightarrow_{(A_n)} V'$ and $\trianglelefteq' \in E^{\rightarrow}$. So $(V, V') \in \bowtie$. So

$$\trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \subseteq \bowtie$$

And since $\bowtie$ is symmetric, and the symmetric closure of a set is the smallest extension of this set that is symmetric,

$$\left(\trianglelefteq \setminus \left(\bigcap_{\trianglelefteq' \in E^{\rightarrow}} \trianglelefteq' \right) \right)^{sym} \subseteq \bowtie$$

□

Proof of proposition 3 : We first prove the two following assertions :

1. If $V \rightarrow_A V''$ then $V' \rightarrow_A V''$
2. If $V'' \rightarrow_A V$ then $V' \rightarrow_A V$

To prove this, suppose that $V \to_A V''$. From axiom 1 and the fact that $V \subseteq V' \subseteq V''$, we have $V'\backslash A \subseteq V''\backslash A = V\backslash A \subseteq V'\backslash A$. Thus $V''\backslash A = V\backslash A = V'\backslash A$. Suppose that $V \to_A V''$, then by Lemma 1 $\forall B \subseteq A$, $B \cup (V\backslash A) \to_A V''$ and thus $B \cup (V'\backslash A) \to_A V''$. Moreover, by axiom 1 and the fact that $V \subseteq V'$, we have that $V''\backslash V' \subseteq V''\backslash V \subseteq A$. So we can take $B = V' \cap A$ and we obtain that $V' \to_A V''$, which proves *1.* The proof of *2.* is similar and left to the reader. Suppose now that there is no A such that either $V' \to_A V''$ or $V' \to_A V$. Then by contraposition of *1.* and *2.* there is no A such that either $V \to_A V''$ or $V'' \to_A V$ and $V \bowtie_1 V''$, which completes the proof.

□

1.2 Proof of section of section 3

Proof of proposition 4 : Let $V, V' \in \mathcal{P}(\hat{V})$, $A \in \mathcal{A}$ and C, C' such that $C \in \mathcal{C}^{\{V\}}$, $C' \in \mathcal{C}^{\{V'\}}$

$$C \mapsto_A C' \tag{5.1}$$

By (5.1) we have that there exists V_1 and V_2 such that 1) $V_1 \to_A V_2$, 2) $C \in \mathcal{C}^{\{V_1\}}$ and 3) $C' \in \mathcal{C}^{\{V_2\}}$ and $C' \in \mathcal{C}^{\{V_2\}}$. Since we also have that $C \in \mathcal{C}^{\{V\}}$ and $C' \in \mathcal{C}^{\{V'\}}$ it clear that for an $K, K' \in \mathcal{K}$ such that $K = \{x, y\}$ and $K = \{x', y'\}$ we have that

$$x \cap V < y \cap V \iff x \cap V_1 < y \cap V_1 \text{ and } x' \cap V' < y' \cap V' \iff x' \cap V_2 < y' \cap V_2$$

Since $\mathcal{K} = \mathcal{P}(\hat{V})$ let $K = \{V \cup V_1, V\}$. Since $(V \cup V_1) \cap V = V = V \cup V$ we have that $V_1 = (V \cup V_1) \cap V_1 = V_1 \cap V$ which, since $<$ is highly discriminating, implies that $V = V_1$.

□

1.3 Proof of section of section 4

Proof of proposition 5 : Suppose that $(\to_A)_{A \in \mathcal{V}}$ is induced by a monotonic axiological criterion $\trianglelefteq$. We need to check the property is satisfied. Because the hierarchy $\trianglelefteq$ is monotonic and $V \subseteq V \cup A$, for every A, one $V \subseteq V \cup A$ has $V \trianglelefteq V \cup A$. Moreover, for all $B \subset A$, $V\backslash A \cup B \trianglelefteq V \cup C$, so $V \to_A V \cup A$.

Conversely, suppose that $V \subseteq V'$ and that $(\to_A)_{A \in \mathcal{V}}$ relies on an anything goes structure. Because every preference formation rule that is induced by an anything goes structure is also induced by partial deliberation $(\to_A)_{A \in \mathcal{V}}$ there exists an axiological criterion $\triangleleft$ that satisfies 2.1. So by the property $V \to_{V'} V \cup V' = V'$ and $V \trianglelefteq V'$.

□

Proof of proposition 6 : By definition, we have that *2.* $\implies$ *1.* Suppose that *3.* is not satisfied. Then either $V \subseteq V'$ or $V' \subseteq V$. W.l.o.g suppose $V \subseteq V'$. Let $A = V'\backslash V$, then since $(\to_A)_{A \in \mathcal{V}}$ is induced by an anything goes structure, we have that $V \to_A V'$. Thus *2.* $\implies$ *3.*. Suppose that *3.* is satisfied and, by contradiction, assume that there exists A

such that $V \rightarrow_A V'$. Thus since $(\rightarrow_A)_{A\in\mathcal{V}}$ is induced by an anything goes structure then $V \subset V \cup A = V'$. But by *3.* neither $V \triangle V' \not\subseteq V$ or $V \triangle V' \not\subseteq V'$. This is a contradiction.

□

Proof of lemma 2 : Suppose $(\rightarrow_A)_{A\in\mathcal{V}}$, a preference formation rule that is induced by a partitioned structure, axioms 1-6 hold. For the first axiom, it is clear that for all $B \subseteq A$, $B \cup (V\backslash A) \triangle (P \cap A) \cup (V\backslash A) = B \cup (P \cap A) \subseteq A$. So axiom 1 holds. Since $V \rightarrow_A (P \cap A) \cup (V\backslash A)$ implies that for all $B \subseteq A$, $B \cup (V\backslash A) \rightarrow_A (P \cap A) \cup (V\backslash A)$ axiom 2 holds. If, $V \rightarrow_{A'} B \cup V\backslash A$ for some $B \subseteq A \subseteq A'$, then there exists $P \in \mathcal{P}$ such that $B = P \cap A' = P \cap A$ so $V \rightarrow_A B \cup V\backslash A$. So axiom 3 holds. Suppose for some A, A' and $B \subseteq A$, $V \rightarrow_A V'$ and $B \cup (V\backslash A) \rightarrow_{A'} V'$. then there exists $P \in \mathcal{P}$ such that $V' = A \cap P \cup (V\backslash A) = A' \cap P \cup (V\backslash A')$. But $V\backslash A' = V\backslash A$ so $A \cap P = A' \cap P$ and $V \rightarrow_A V'$. So axiom 4 holds. Suppose V, V', V'' and A such that $V \rightarrow_A V' \rightarrow_A V''$. In this case, there simply exists P such that $V' = (P \cap A) \cup (V\backslash A) = V''$. So $V'' \rightarrow_A V'$. So axiom 5 holds. Now let V, V' and A, $V \rightarrow_{(A_n)} V' \rightarrow_A V$. $V' \rightarrow_A V$ so there exists $P \in \mathcal{P}$ such that $V = (A \cap P) \cup (V'\backslash A)$.

□

Proof of proposition 7 :

To show the *if part*, let A, V and $B \subseteq A$ such that $V \rightarrow_A B \cup (V\backslash A)$. I proceed in three steps.

Step 1 : *Building* $\mathcal{P}$

Suppose $(\rightarrow_A)_{A\in\mathcal{V}}$ is induced by a clustering axiological criterion $\trianglelefteq$. For all $v \in \hat{V}$, $P_v = \{v' \in \hat{V} : v \sim v'\} + v$. Note that $v \in P_v$. Moreover, if $v' \in P_v$ then by *1.*, $v \in P'_v$. Thus, by symmetry, for all $v, v' \in \hat{V}$, we have that either $P_v = P_{v'}$ or $P_v \cap P_{v'} = \emptyset$. Therefore, $\mathcal{P} \equiv \{P_v : v \in \hat{V}\}$ is a partition.

We now have to prove that $\mathcal{P}$ is the candidate satisfy (2.14). Consider (V, A) and $B \subseteq A$ we must prove that :

Step 2 : *For* $B = \emptyset$

Suppose first that $B = \emptyset$. Because $(\rightarrow_A)_{A\in\mathcal{V}}$ it is induced by a partitioned structure, by Lemma 2 it is induced by partial deliberation. Thus, $B' \subseteq A$,

$$B' \cup (V\backslash A) \trianglelefteq V\backslash A \tag{5.2}$$

Thus, by assumption $V\backslash A$ needs to be nonempty. Moreover if A is empty, then take any P such that $(V\backslash A) \cap P \neq \emptyset$ and (2.14) holds. Suppose there exists $P \in \mathcal{P}$, such that $P \cap A = \emptyset$ and $P \cap (V\backslash A) \neq \emptyset$. Then let $B = P \cap A$ and (2.14) holds. Therefore we can suppose that $\forall P \in \mathcal{P}$ such that $P \cap (V\backslash A) \neq \emptyset$, we have that $P \cap A \neq \emptyset$. From (5.2)

we have that for all $v \in A$, $(V\backslash A) + v \trianglelefteq V\backslash A$. Thus, we can apply *4.* and there exists $v' \in V\backslash A$ such that if $v'' \sim v'$ for some $v'' \notin V\backslash A$ then $v'' \notin A$. This contradicts the fact that $\forall P \in \mathcal{P}$ such that $P \cap (V\backslash A) \neq \emptyset$ we have $P \cap A \neq \emptyset$. Thus, there exists $P \in \mathcal{P}$, such that $P \cap (V\backslash A) \neq \emptyset$ and $P \cap A = \emptyset$ $P \cap ((V\backslash A) \cup A) \subseteq V\backslash A$. Let $B = P \cap A$ and (2.14) holds. Moreover, we have that $V\backslash A \rightarrow_{A+v''} V\backslash A + v''$. Indeed we have that $v'' \in A + v''$ with $v'' \sim v'$ and $v'' \notin V$. So by contraposition of *4.* there exists $v''' \in A + v''$ such that $V \triangleleft V + v'''$. But since $V + v \trianglelefteq V$ for all $v \in A$ we must have that $v''' = v''$.

Step 3 : *For* $B \neq \emptyset$

Suppose now that $B \neq \emptyset$. Let's prove first that there exists $P \in \mathcal{P}$ such that $B \subseteq P \cap A$. By contradiction, assume that $\forall P \in \mathcal{P}$, $B \not\subseteq P$. Since a preference formation rule that is induced by a partitioned structure is also induced by partial deliberation, $\forall B' \subseteq A$, $B' \cup (V\backslash A) \trianglelefteq B \cup (V\backslash A)$ and $\exists v, v' \in B$ such that $vv' \triangleleft v$ and $B \cup (V\backslash a) - v \triangleleft V \cup (V\backslash A)$. Then, according to *3.*, $B \cup (V\backslash A) - v' \triangleright B \cup (V\backslash A)$ and since $B \cup (V\backslash A) - v'$ is reachable from (A, B), $V \rightarrow_A B \cup (V\backslash A)$ is wrong. Therefore, there exists $P \in \mathcal{P}$ such that $B \subseteq P \cap A$. Denote P^* such a P. Suppose there exists $v \in P^* \cap A$ but $v \notin B$. Thus $B \cup (V\backslash A) + v \trianglelefteq B \cup (V\backslash A)$ with $v \trianglelefteq vv'$ for all $v' \in B \subseteq P$. So by *3.*, $B \cup (V\backslash A) \triangleleft B \cup (V\backslash A) - v'$. Which is a contradiction. So $B = P \cap A$ for some $P \in \mathcal{P}$. It is clear that $P \cap (A \cup (V\backslash A)) \neq \emptyset$. Moreover, for all $v \in P$, $(P \cap A) \cup (V\backslash A) \rightarrow_{A+v} (P \cap A) \cup B\backslash A$. Indeed, let $v' \in P \cap A$ and $v \in P$, if we had $(P \cap A) \cup (V\backslash A) \triangleright (P \cap A) \cup (V\backslash A) + v$ then by *3.* we would have that $(P \cap A) \cup (V\backslash A) \triangleleft (P \cap A) \cup (V\backslash A) - v'$, which contradicts what we have just established. This completes the *if* part of the proof.

To show the *only if* part. Suppose that $(\rightarrow_A)_{A \in \mathcal{V}}$ is induced by a partitioned structure. Then it is induced by partial deliberation and there exists a partition $\mathcal{P}$ such that (2.14) is satisfied.

Let $v, v', v'' \in \hat{V}$ $v \sim v'$ and $v' \sim v''$. Then $v \rightarrow_{vv'} vv'$ and thus there is $P \in \mathcal{P}$ such that $v, v' \in P$. In the same way, $v' \rightarrow_{v'v''} v'v''$ and $v, v' \in P'$ for some $P' \in \mathcal{P}$. Since $\mathcal{P}$ is a partition $P = P'$. Thus, since we cannot have $v \rightarrow_{vv''} \emptyset$, either $v \rightarrow_{vv''} v$, $v \rightarrow_{vv''} v''$ or $v \rightarrow_{vv''} vv''$. But since $v, v'' \in P$ and $(\rightarrow_A)_{A \in \updownarrow \dashv \sqcup \langle \rfloor \dashv \updownarrow \mathcal{V}}$ is induced by a partitioned structure, $v \rightarrow_{vv''} vv''$ is the case and, thus, $v \trianglelefteq vv''$. In the same way we can show that $v'' \trianglelefteq vv''$ and $v \sim v''$ and *1.* holds.

Now let $V \in \mathcal{V}$ and $v \notin V$ such that $V \trianglelefteq V + v$ and $vv' \triangleleft v$. Thus $vv' \rightarrow_{vv'} v$ and $P_v \neq P_{v'}$. If we had $V + v - v' \trianglelefteq V + v$ then $V + v - v' \rightarrow_{vv'} V + v = P_v \cap \{v, v'\}$ since $V \trianglelefteq V + v$. But this contradicts the fact that $P_v \neq P_{v'}$. So *2.* holds.

Let $V \in \mathcal{V}$ and $v \notin V$ such that $V + v \trianglelefteq V$ and $v \sim v'$. Thus $V + v \rightarrow_v V$ and there exists $P \in \mathcal{P}$ such that $V + v \rightarrow_v V = P \cap \{v\} \cup V$ so that $v \notin P$. Moreover, there exists $P' \in \mathcal{P}$ such that $v, v' \in P'$ and since $\mathcal{P}$ is partition, $v' \notin P$. Thus, we cannot have that $V - v' \trianglelefteq V$ since this would imply that $V - v' \rightarrow_{vv'} V$ while $v, v' \in P$. So *3.* holds.

Let $V \in \mathcal{V}$, such that for all $v \in V$, $V \triangleleft V - v$. Since $\trianglelefteq$ induces the partitioned

structure $(\rightarrow_A)_{A\in\mathcal{V}}$ and $V \rightarrow_v V - v$, there exists $P \in \mathcal{V}$ with $v \notin P$. Therefore if there exists, $v' \in P \cap V^{\complement}$, then we have that V

Let $V \in \mathcal{V}$. Suppose that there exists $K \in \mathcal{V}$ such that $K \cap V = \emptyset$ and for all $v \in K$, $V + v \triangleleft V$. Then for all $v \in K$, $V + v \rightarrow_v V = P \cup (V \backslash A)$ for some $P \in \mathcal{P}$. Suppose $\exists v' \in P$ and suppose such that $v' \sim v''$ for some $v'' \notin V$. We must have that $V \rightarrow_{vv'} V + v'$ thus $V \triangleleft V + v'$, $v' \notin K$.

□

2 Proofs of Chapter 3

2.1 Proof of section 1

=

Proof of proposition 8 : Consider a preference formation rule $\rightarrow\in \mathbb{V} \times \mathbb{V} \times \mathcal{A}$ that is induced by an euclidean version of partial deliberation. Thus, there exists $\hat{V}$ and $\mathcal{F}$ such that $< \hat{V}, \mathcal{F} >$ is an euclidean framework of value, where ψ is its corresponding bijection and $\mathbb{V}$. Moreover, there exists $\rightarrow'\in V^{\mathcal{F}} \times V^{\mathcal{F}} \times \hat{\mathcal{A}}$ such that

$$V \rightarrow'_{\hat{A}} V' \iff \mathbf{v} = \psi(V) \rightarrow_A \psi(V') = \mathbf{v}' \tag{5.3}$$

We have to show that there exists Φ, such that for any pair $\mathbf{v}, \mathbf{v}' \in \mathbb{V} \times \mathbb{V}$

$$\mathbf{v} \rightarrow_A \mathbf{v}' \iff \mathbf{v}' \in \left(\mathbf{v}_{|\neg A}, \underset{\mathbf{v}''_{|A}}{Argmax}\ \Phi(\mathbf{v}_{|\neg A}, \mathbf{v}''_{|A})\right) \tag{5.4}$$

Suppose $\mathbf{v} \rightarrow_A \mathbf{v}'$, then by (5.3) and the fact that ψ is a bijection, we have that $V \rightarrow'_{\hat{A}} V'$ where $\psi^{-1}(\mathbf{v}) = V$ and $\psi^{-1}(\mathbf{v}') = V'$. Thus we have that there exists an order $\trianglelefteq$ such that

$$\begin{cases} \exists B \subseteq \hat{A}, V' = B \cup (V \backslash \hat{A}) \\ \forall B' \subseteq \hat{A}, B' \cup (V \backslash \hat{A}) \trianglelefteq V' \end{cases} \tag{5.5}$$

Since $\trianglelefteq$ is an order there exists Φ' such that

$$\begin{cases} \exists B \subseteq \hat{A}, V' = B \cup (V \backslash \hat{A}) \\ \forall B' \subseteq \hat{A}, \Phi'(B' \cup (V \backslash \hat{A}) < \Phi'(V') \end{cases} \iff \begin{cases} \forall i \in \hat{V}, v(i) \neq v'(i) \implies i \in A \\ \forall i \in A, \Phi'(\psi^{-1}((v_i)_{i\in A}, \mathbf{v}_{|A^{\complement}}) < \Phi'(\psi_{-1}(\mathbf{v}')) \end{cases} \tag{5.6}$$

Let $\Phi = \Phi \circ \psi^{-1}$. We have that :

$$\begin{cases} \forall i \in \hat{V}, v(i) \neq v'(i) \implies i \in A \\ \forall i \in A, \Phi((v_i)_{i\in A}, \mathbf{v}_{|A^{\complement}}) < \Phi(\mathbf{v}') \end{cases}$$

This is equivalent to

$$\mathbf{v} \to_A \mathbf{v}' \iff \mathbf{v}' \in \left(\mathbf{v}_{|\neg A}, \underset{\mathbf{v}''_{|A}}{Argmax}\ \Phi(\mathbf{v}_{|\neg A}, \mathbf{v}''_{|A})\right) \tag{5.7}$$

Since in the proof every implication is in fact an equivalence we have that (3.2) is satisfied.

In order to show the opposite, take $\hat{V}$ as the a set enumerating each coordinates of vectors v. $\mathcal{F}$ a set of functions that gives the value of each coordinate. With Ψ and Φ we can build an axiological criterion that induces partial deliberation.

□

Proof of proposition 9 : Consider an euclidean framework of partial deliberation $(\hat{V}, \mathcal{F})$ and denote by Φ its axiological criterion. From proposition (8) we have that $\mathbf{v}(\{1,2\}n\mathbf{v})$ maximizes Φ on the set

$$\{\mathbf{v}' : \mathbf{v}'_{|\hat{V}\setminus\{1,2\}} = \mathbf{v}_{|\hat{V}\setminus\{1,2\}}\}$$

Moreover $\mathbf{v}(\{-i\}, \mathbf{v})$ maximizes Φ on

$$\{\mathbf{v}' : \mathbf{v}'_{|\hat{V}\setminus\{i\}} = \mathbf{v}_{|\hat{V}\setminus\{i\}}\}$$

But since values 1 and 2 are independent we have that :

$$\mathbf{v}_i(\{i\}, \mathbf{v}) = \mathbf{v}_i(\{1,2\}, \mathbf{v})$$

Thus,

$$\mathbf{v}_i(\{1,2\}, \mathbf{v}) \in \{\mathbf{v}' : \mathbf{v}'_{|\hat{V}\setminus\{-i\}} = \mathbf{v}_{|\hat{V}\setminus\{-i\}}(\{i\}, \mathbf{v})\}$$

Since, $\mathbf{v}(\{-i\}, \{i\}, \mathbf{v})$ maximizes Φ on this set this set we have that :

$$\Phi\Big(\mathbf{v}(\{-i\}, \{i\}, \mathbf{v})\Big) \geq \Phi\Big(\mathbf{v}(\{-i1, 2i\}, \mathbf{v})\Big)$$

But, we also clearly have

$$\Phi\Big(\mathbf{v}(\{-i\}, \{i\}, \mathbf{v})\Big) \leq \Phi\Big(\mathbf{v}(\{1,2\}, \mathbf{v})\Big)$$

Because, we get $\mathbf{v}(\{1,2\}, \mathbf{v})$ from a maximization program with less constraints than $\mathbf{v}(\{-i\}, \{i\}, \mathbf{v})$ is. Thus,

$$\Phi\Big(\mathbf{v}(\{-i\}, \{i\}, \mathbf{v})\Big) = \Phi\Big(\mathbf{v}(\{1,2\}, \mathbf{v})\Big)$$

And since Φ is single peaked,

$$\mathbf{v}(\{-i\}, \{i\}, \mathbf{v}) = \mathbf{v}(\{1,2\}, \mathbf{v})$$

□

2.2 Proofs of Section 2 : Disclosure with imperfect empathy

We start by determining the utility at the optimum, conditionally on each disclosure strategy. In order to ease notations, in this subsection, I will write $\mathbf{v}(1,2)$ for $\mathbf{v}(\{1,2\},\mathbf{v}^0)$, $\mathbf{v}(i)$ for $\mathbf{v}(\{i\},\mathbf{v}^0)$ and $\mathbf{v}(\emptyset)$ for $\mathbf{v}(\emptyset,\mathbf{v}^0)$. And I note $\mathbf{v}_i$ for $\mathbf{v}_{\{1\}}$.

— When the disclosure strategy is $\emptyset$:

The expected utility of the sender is given by :

$$\mathbb{E}U(\mathbf{x};\emptyset) = -\Big(||\mathbf{x}||^2 + \beta.||\mathbf{x}-\mathbf{v}_P||^2\Big)$$

Note that because we have $\mathbf{v}^s = \mathbf{v}(1,2)$ the optimal project is :

$$\mathbf{x}_\emptyset = (\frac{\mathbf{v}_1(1,2).\beta}{\beta+1}, \frac{\mathbf{v}_2(1,2).\beta}{\beta+1})$$

So, the optimal utility of the sender is given by :

$$-\frac{\beta}{\beta+1}.||\mathbf{v}(1,2)|| \tag{5.8}$$

— When the disclosure strategy is i :

The expected utility of the sender is given by :

$$\mathbb{E}U(\mathbf{x};i) = -\Big((1-p_1(1)).||\mathbf{x}|| + p_1(\emptyset).||\mathbf{x}-\mathbf{v}(i)|| + \beta.||\mathbf{x}-\mathbf{v}_P||\Big)$$

The optimal project is thus :

$$\mathbf{x}_i = (x_i, x_{-i})$$

with $x_i = \frac{\mathbf{v}_i(i).p_i(i)+\mathbf{v}_i(1,2).\beta}{\beta+1}$ and $x_{-i} = \frac{\mathbf{v}_i(i).\beta}{\beta+1}$. And the optimal utility payoff is :

$$\mathbb{E}U(\mathbf{x};i) = -\Big(\frac{(\beta+p_i(\emptyset)).p_i(i).\mathbf{v}_i(i)^2 - \big(2.\beta.\mathbf{v}_i(1,2).\mathbf{v}_i(i)(p_i(\emptyset)) + \beta.||\mathbf{v}(1,2)||}{\beta+1}\Big)$$

— When the disclosure strategy is $(1,2)$:

The expected utility of the sender is given by :

$$\mathbb{E}U(\mathbf{x};\{1,2\}) = -\Big(p_{1,2}(\emptyset).||\mathbf{x}||^2 + p_{1,2}(1).||\mathbf{x}-\mathbf{v}(1)||^2 + p_{1,2}(2).||\mathbf{x}-\mathbf{v}(2)||^2$$
$$+p_{1,2}(1,2).||\mathbf{x}-\mathbf{v}(1,2)||^2 + \beta.||x-\mathbf{v}^s||\Big)$$

The optimal project is then :

$$\mathbf{x}_{1,2} = \frac{\mathbf{v}(1).p_{1,2}(1) + \mathbf{v}(2).p_{1,2}(2) - (p_{1,2}(1,2) - \beta).\mathbf{v}(1,2)}{\beta + 1} \tag{5.9}$$

With this we get that the expected utility of the sender is a four degrees polynomial function $\mathbb{E}U(\mathbf{x}; \{1,2\})(p) = -\frac{1}{1+\beta}\Big(a_4.||\mathbf{v}^s||^2 + a_3.\mathbf{v}_1^s.\mathbf{v}_1(1) + a_3.\mathbf{v}_2^s.\mathbf{v}_2(2) + a_2.(\mathbf{v}_1(1))^2 + a_1.(v_2(2))^2 + a_0\Big)$ with

$$a_4 = p_{1,2}(1,2).(\beta + p_{1,2}(1,2))$$

$$a_3 = -2.p_{1,2}(1).(\beta + p_{1,2}(1,2))$$

$$a_2 = -2.p_{1,2}(2).(\beta + p_{1,2}(1,2))$$

$$a_1 = p_{1,2}(1).(\beta + 1 - p_{1,2}(2))$$

$$a_0 = p_{1,2}(2).(\beta + 1 - p_{1,2}(2))$$

Proof of proposition 10 : To prove proposition 10, assume that we are in a case of axiological dependence.

In this context, we have that :

$$\Delta^i_\emptyset(p) = \mathbb{E}U(\mathbf{x}_i; \{i\}) - \mathbb{E}U(\mathbf{x}_\emptyset; \emptyset) = \frac{1}{1+\beta}\Big(\mathbf{v}_{-i}(-i).(p_{1,2}(-i) + p_{1,2}(1,2)).(\beta - p_{1,2}(i) - p_{1,2}(\emptyset))\Big)$$

$$\Delta^{1,2}_i(p) = \mathbb{E}U(\mathbf{x}_i; \{1,2\}) - \mathbb{E}U(\mathbf{x}_i; \{i\})\frac{1}{1+\beta}\Big(\mathbf{v}_i(i).(p_{1,2}(i) + p_{1,2}(1,2)).(\beta - p_{1,2}(-i) - p_{1,2}(\emptyset))\Big)$$

$$\Delta^{1,2}_\emptyset(p) = \mathbb{E}U(\mathbf{x}_{1,2}; 1,2) - \mathbb{E}U(\mathbf{x}_\emptyset; \emptyset) = \Delta^{1,2}_i(p) - \Delta^i_\emptyset(p)$$

Thus, $\Delta^i_\emptyset(p)$ and $\Delta^{1,2}_i(p)$ increase with β and

— $\Delta^i_\emptyset(p) = 0$ for $\beta = p_{1,2}(-i) + p_{1,2}(\emptyset) = p_i(\emptyset)$ by 3.8

— $\Delta^{1,2}_i(p) = 0$ for $\beta = p_{1,2}(i) + p_{1,2}(\emptyset) = p_{-1}(\emptyset)$ by 3.8

Thus, we have that $\beta \in [0, p_i(\emptyset))$ if and only if $\Delta^i_\emptyset(p) < 0$. So, the sender is better off disclosing no value rather than value i $\beta \in [0, p_i(\emptyset))$. Moreover, $\beta \in [0, p_{-i}(\emptyset))$ if and only $\Delta^{1,2}_i(p) = 0 < 0$. Thus, for $\beta \in [0, p_{-i}(\emptyset))$ the sender is better off disclosing i than disclosing 1 and 2. Thus, for $\beta \in [0, min(p_i(\emptyset), p_{-i}(\emptyset))$, the sender discloses no value. For $\beta \in \Big(min(p_i(\emptyset), p_{-i}(\emptyset)); max(p_i(\emptyset), p_{-i}(\emptyset))\Big)$ she discloses the value i such that $max(p_i(\emptyset), p_{-i}(\emptyset) = p_i(\emptyset)$. And, finally she disloses the two values whenever $\beta \in \Big(max(p_i(\emptyset), p_{-i}(\emptyset)); 1\Big)$

□

Proof of proposition 11 : I only prove the case of positive axiological dependence. The proof

of negative axiological dependence follows the same argument. Under the assumptions that symmetric beliefs hold and preference changes satisfy axiological symmetry, we have that,

$$\Delta_i^{1,2}(p) = \frac{1}{8.(1+\beta)}.\Big((4\beta-3).(\mathbf{v}_i(1,2))^2 + 2.\mathbf{v}_i(i).\mathbf{v}_i(1,2) - (\mathbf{v}_i(i))^2\Big)$$

Thus for

$$\beta = \frac{3.(\mathbf{v}_i(1,2))^2 - 2.\mathbf{v}_i(1,2)\mathbf{v}_i(i) + (\ \mathbf{v}_i(i))^2}{4.(\mathbf{v}_1(1,2))^2} = \frac{1}{2} + \frac{\mathbf{v}_i(1,2) - \ \mathbf{v}_i(i)}{\mathbf{v}_i(i))^2} > 1/2$$

We have $\Delta_i^{1,2}(p) = 0$ and $8.(1+\beta).\Delta_i^{1,2}(p)$ increases with β.

Similarly, we can show that, for

$$\beta = \frac{\mathbf{v}_i(i)}{4.\mathbf{v}_i(1,2) - 2\mathbf{v}_i(i)} = \frac{1}{2} - \frac{2\mathbf{v}_i(1,2) - 2\mathbf{v}_i(i)}{4.\mathbf{v}_i(1,2) - 2\mathbf{v}_i(i)} < 1/2$$

we have that $\Delta_\emptyset^i(p) = 0$ and increases with β.

Take

$$\beta_1 = \frac{1}{2} - \frac{2\mathbf{v}_i(1,2) - 2\mathbf{v}_i(i)}{4.\mathbf{v}_i(1,2) - 2\mathbf{v}_i(i)} \text{ and } \beta_2 = \frac{3.(\mathbf{v}_i(1,2))^2 - 2.\mathbf{v}_i(1,2)\mathbf{v}_i(i) + (\ \mathbf{v}_i(i))^2}{4.(\mathbf{v}_1(1,2))^2}$$

Thus, the disclosure policy is given by the following conditions :

— If $\beta < \beta_1 < 1/2$ we have that $\Delta_\emptyset^i(p) < 0$ and $\Delta_i^{1,2}(p) < 0$ and, thus, the sender discloses no values.

— If $\beta_1 < \beta < \beta_2$ we have that $\Delta_\emptyset^i(p) > 0$ and $\Delta_i^{1,2}(p) < 0$ and, thus, the sender discloses either value i or value $-i$ but not both.

— If $1/2 < \beta_2 < \beta$ we have that $\Delta_\emptyset^i(p) > 0$ and $\Delta_i^{1,2}(p) > 0$ and, thus, the sender discloses both values.

□

2.3 Proof of section 3 : Sequential disclosure

The lemmas of section 3 are easily derived and there are left to the reader.

Noting that corrolary 2 simply results from proposition 3.14 and Cauchy-Schwarz unequality. I prove proposition 13.

Proof of proposition 13 :

— The utility of sequential disclosure is :

$$u(\mathbf{x}, \{i\}, \{-i\}) = -\Big(||\mathbf{x} - \mathbf{v}(i)|| + ||\mathbf{x} - \mathbf{v}(-i, \mathbf{v}(i))|| + \beta.||\mathbf{x} - \mathbf{v}^s)||\Big)$$

Thus, the optimal project is given by :

$$\mathbf{x}_{i,\hat{-i}} = \Big(\frac{\beta.\mathbf{v}_i^s + 2.\mathbf{v}_i(i,-i)}{\beta+2}, \frac{\beta.\mathbf{v}_{-i}^s + .\mathbf{v}_{-i}(i,-i)}{\beta+2}\Big)$$

And the utility with this project :

$$u(\mathbf{x}_{i,\hat{-i}},\{i\},\{-i\}) = -\frac{1}{\beta+2}.\Big(2.\beta.(||\mathbf{v}^s - \mathbf{v}(i,\hat{-i})||^2) + \mathbf{v}_{-i}(i,\hat{-i}).\mathbf{v}^s_{-i} + 1/2.\big(\mathbf{v}_{-i}(i,\hat{-i})\big)^2\Big) + \big(\mathbf{v}_{-i}(i,\hat{-i})\big)^2\Big) \tag{5.10}$$

— The utility of simultaneous disclosure is :

$$u(\mathbf{x},\{1,2\},\{1,2\}) = -\Big(||\mathbf{x}-\mathbf{v}(1,2)||^2 + ||\mathbf{x}-\mathbf{v}(1,2))||^2 + \beta.||\mathbf{x}-\mathbf{v}^s)||^2\Big)$$

Thus, the optimal project is given by :

$$\mathbf{x}_{1,2} = \Big(\frac{\beta.\mathbf{v}^s_1 + 2.\mathbf{v}_1(1,2)}{\beta+2}, \frac{\beta.\mathbf{v}^s_2 + .\mathbf{v}_2(1,2)}{\beta+2}\Big)$$

And the utility with this project :

$$u(\mathbf{x}_{1,2},1,2) = -\frac{2.\beta}{\beta+2}||\mathbf{v}^s - \mathbf{v}(1,2)||^2$$

Thus,

$$\Delta^{1,2}_{i,-i} = -\frac{1}{\beta+2}.\Bigg(2.\beta.\Big(||\mathbf{v}^s-\mathbf{v}(1,2)||^2-||\mathbf{v}^s-\mathbf{v}(i,\hat{-i})||^2)-\mathbf{v}_{-i}(i,\hat{-i}).\mathbf{v}^s_{-i}+1/2\big(\mathbf{v}_{-i}(i,\hat{-i})\big)^2\Big)+\big(\mathbf{v}_{-i}(i,\hat{-i})\big)^2\Bigg)$$

Denoting by,

$$\beta_1 = \frac{\big(\mathbf{v}_{-i}(i,\hat{-i})\big))^2}{2.||\mathbf{v}^s-\mathbf{v}(1,2)||^2 - 2.||\mathbf{v}^s-\mathbf{v}(i,\hat{-i})||^2 - 2.\mathbf{v}_{-i}(i,\hat{-i}).\mathbf{v}^s_{-i} + \big(\mathbf{v}_{-i}(i,\hat{-i})\big))^2}$$

we have to consider two cases :

— **Case 1 :** If

$$2.||\mathbf{v}^s-\mathbf{v}(1,2)||^2 - 2.||\mathbf{v}^s-\mathbf{v}(i,\hat{-i})||^2 - 2.\mathbf{v}_{-i}(i,\hat{-i}).\mathbf{v}^s_{-i} + \big(\mathbf{v}_{-i}(i,\hat{-i})\big))^2 > 0$$

then $\Delta^{1,2}_{i,-i}$ decreases with β, and for

$$\beta = \frac{\big(\mathbf{v}_{-i}(i,\hat{-i})\big))^2}{2.||\mathbf{v}^s-\mathbf{v}(1,2)||^2 - 2.||\mathbf{v}^s-\mathbf{v}(i,\hat{-i})||^2 - 2.\mathbf{v}_{-i}(i,\hat{-i}).\mathbf{v}^s_{-i} + \big(\mathbf{v}_{-i}(i,\hat{-i})\big))^2} > 0$$

We have

$$\Delta^{1,2}_{i,-i} = 0$$

Thus, whenever $\beta < \beta_1$, we have $\Delta^{1,2}_{i,-i} > 0$ and the simultaneous disclosure strategy

is better than the sequential disclosure strategy. Conversely, however, $\beta > \beta_1$, implies that $\Delta^{1,2}_{i,-i} < 0$ and the sequential disclosure strategy is better. There may be a value for partial disclosure. We simply need to check that $\beta_1 < 1$, otherwise there won't be any β such that the sequential disclosure strategy has a value.

We have that $\beta_1 < 1$ if and only iff (3.14) is satisfies.

— **Case 2 :** if If

$$2.||\mathbf{v}^s - \mathbf{v}(1,2)||^2 - 2.||\mathbf{v}^s - \mathbf{v}(i,\hat{-}i)||^2 - 2.\mathbf{v}_{-i}(i,\hat{-}i).\mathbf{v}^s_{-i} + (\mathbf{v}_{-i}(i,\hat{-}i)))^2 < 0$$

then $\Delta^{1,2}_{i,-i}$ increases with β and, β_1 being negative, the simultaneous disclosure strategy is better for all $0 < \beta < 1$.

This completes the proof of proposition 13. □

Bibliographie

Alesina, Alberto, and Eliana La Ferrara. 2005. "Preferences for redistribution in the land of opportunities." *Journal of public Economics*, 89(5-6) : 897–931.

Alesina, Alberto, and Nicola Fuchs-Schündeln. 2007. "Good-bye Lenin (or not ?) : The effect of communism on people's preferences." *The American Economic Review*, 97(4) : 1507–1528.

Alesina, Alberto, and Silvia Ardagna. 2010. "Large changes in fiscal policy : taxes versus spending." *Tax policy and the economy*, 24(1) : 35–68.

Alexander, J McKenzie. 2009. "The structural evolution of morality."

Algan, Yann, and Pierre Cahuc. 2013. "Trust and growth." *Annu. Rev. Econ.*, 5(1) : 521–549.

Ambrus, Attila, and Kareen Rozen. 2014. "Rationalising Choice with Multi-self Models." *The Economic Journal*, 125(585) : 1136–1156.

Anscombe, G. E. M. 1957. *Intention.* Harvard University Press.

Ashby, Nathaniel JS, Stephan Dickert, and Andreas Glöckner. 2012. "Focusing on what you own : Biased information uptake due to ownership." *Judgment and Decision Making*, 7(3) : 254.

Ashley, Richard, Clive WJ Granger, and Richard Schmalensee. 1980. "Advertising and aggregate consumption : An analysis of causality." *Econometrica : Journal of the Econometric Society*, 1149–1167.

Barbera, Salvador, Walter Bossert, and Prasanta K Pattanaik. 2004. "Ranking sets of objects." In *Handbook of utility theory.* 893–977. Springer.

Baujard, Antoinette. 2007. "Conceptions of freedom and ranking opportunity sets. A typology." *Homo Oeconomicus*, 24(2) : 1–24.

Becker, Gary S. 1996. *Accounting for tastes.* Harvard University Press.

Beggan, James K. 1992. "On the social nature of nonsocial perception : The mere ownership effect." *Journal of personality and social psychology*, 62(2) : 229.

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and extrinsic motivation." *The Review of Economic Studies*, 70(3) : 489–520.

Benabou, Roland, and Jean Tirole. 2004. "Willpower and Personal Rules." *Journal of Political Economy*, 112(4) : 848–886.

Bénabou, Roland, Armin Falk, and Jean Tirole. 2018. "Narratives, Imperatives and Moral Reasoning." Working Paper.

Bénabou, Roland, Armin Falk, and Jean Tirole. n.d.. "Eliciting Moral Preferences."

Bergstrom, Theodore C, and Oded Stark. 1993. "How altruism can prevail in an evolutionary environment." *The American Economic Review*, 83(2) : 149–155.

Bernheim, B Douglas, and Antonio Rangel. 2009. "Beyond revealed preference : choice-theoretic foundations for behavioral welfare economics." *The Quarterly Journal of Economics*, 124(1) : 51–104.

Bisin, Alberto, and Thierry Verdier. 2001*a*. "Agents with imperfect empathy may survive natural selection." *Economics Letters*, 71(2) : 277–285.

Bisin, Alberto, and Thierry Verdier. 2001*b*. "The economics of cultural transmission and the dynamics of preferences." *Journal of Economic Theory*, vol 97 : pp. 298–319.

Bisin, Alberto, and Thierry Verdier. 2004. "Work ethic and redistribution : a cultural transmission model of the welfare state." *Unpublished Manuscript, New York University.*

Boehm, Christopher. 1982. "The evolutionary development of morality as an effect of dominance behavior and conflict interference."

Boudon, Raymond. 1997. "Le « paradoxe du vote » et la théorie de la rationalité." *Revue française de sociologie*, 38(2) : 217–227.

Boudon, Raymond. 1999. *Le sens des valeurs.* Vol. 280, Presses Universitaires de France-PUF.

Boudon, Raymond, and François Bourricaud. 1983. "Dictionnaire critique de la sociologie (Paris : PUF, 1982)." *Délits des jeunes et jugement social, recherche comparative internationale, Paris, Maison des Sciences de l'Homme*, 116–123.

Bowles, Samuel. 1998. "Endogenous Preferences : The Cultural Consequences of Markets and other Economics Institutions." *Journal of Economic Litterature*, vol XXXVI : pp. 75–111.

Bowles, Samuel. 2001. "Individual interactions, group conflicts, and the evolution of preferences." *Social dynamics*, 155 : 190.

Broome, John. 1999. "Can a Humean be Moderate ? in his Ethics out of Economics."

Buchanan, James M. 1991. *The economics and the ethics of constitutional order.* University of Michigan Press.

Bénabou, Roland, and Jean Tirole. 2006. "Belief in a Just World and Redistributive Politics." *The Quarterly Journal of Economics*, 121(2) : 699–746.

Carmon, Ziv, Klaus Wertenbroch, and Marcel Zeelenberg. 2003. "Option attachment : When deliberating makes choosing feel like losing." *Journal of Consumer research*, 30(1) : 15–29.

Cavalli-Sforza, Luigi Luca, and Marcus W. Feldman. n.d.. *Cultural transmission and evolution : a quantitative approach.*

Chalfant, James A, and Julian M Alston. 1988. "Accounting for changes in tastes." *Journal of Political Economy*, 96(2) : 391–410.

Chaserant, Camille. 2000. "Rationalité et gestion de l'incomplétude dans les relations contractuelles." PhD diss. Paris 10.

Cherepanov, Vadim, Timothy Feddersen, and Alvaro Sandroni. 2013. "Rationalization." *Theoretical Economics*, 8(3) : 775–800.

Christman, John. 1991. "Autonomy and personal history." *Canadian Journal of Philosophy*, 21(1) : 1–24.

Cohen, Michèle, and Jean-Marc Tallon. 2000. "Décision dans le risque et l'incertain : L'apport des modèles non additifs/Decision under risk and uncertainty : the non-additive approach." *Revue d'économie politique*, 631–681.

Corneo, Giacomo, and Hans Peter Grüner. 2002. "Individual preferences for political redistribution." *Journal of public Economics*, 83(1) : 83–107.

Corneo, Giacomo, and Olivier Jeanne. 2009. "A theory of tolerance." *Journal of public economics*, 93(5) : 691–702.

Cyert, Richard M, and Morris H DeGroot. 1975. "Adaptive utility." In *Adaptive Economic Models*. 223–246. Elsevier.

Cyert, Richard M, and Morris H DeGroot. 1979. "Adaptive utility." In *Expected Utility Hypotheses and the Allais Paradox*. 223–241. Springer.

Davidson, Donald. 1980. "Toward a Unified Theory of Meaning and Action." *Grazer Philosophische Studien*, 11 : 1–12.

Davidson, Donald. 1993. *Actions et événements.* Presses Universitaires de France - PUF.

De Clippel, Geoffroy, and Kfir Eliaz. 2012. "Reason-based choice : A bargaining rationale for the attraction and compromise effects." *Theoretical Economics*, 7(1) : 125–162.

Dekel, Eddie, Barton L Lipman, and Aldo Rustichini. 1998. "Standard state-space models preclude unawareness." *Econometrica*, 66(1) : 159–173.

Delaney, Jason, Sarah Jacobson, and Thorsten Moenig. 2017. "Preference Discovery." Department of Economics, Williams College Department of Economics Working Papers 2017-02.

De Mol, Jan, Gilbert Lemmens, Lesley Verhofstadt, and Leon Kuczynski. 2013. "Intergenerational transmission in a bidirectional context." *Psychologica Belgica*, 53(3) : 7–23.

Dewey, John, and James Hayden Tufts. 1932. "Ethics (rev. ed.)." *New York : H. Holt and company.*

Diaye, Marc-Arthur, and André Lapidus. 2012. "Pleasure and belief in Hume's decision process." *The European Journal of the History of Economic Thought*, 19(3) : 355–384.

Dietrich, Franz. 2017. "Savage's Theorem Under Changing Awareness."

Dietrich, Franz, and Christian List. 2011. "A model of non-informational preference change." *Journal of theoretical politics*, 23(2) : 145–164.

Dietrich, Franz, and Christian List. 2013*a*. "A reason-based theory of rational choice." *Noûs*, 47 : 104–134.

Dietrich, Franz, and Christian List. 2013*b*. "Where do preferences come from ?" *International Journal of Game Theory.*

Dietrich, Franz, and Christian List. 2016. "Mentalism versus behaviourism in economics : a philosophy-of-science perspective." *Economics and Philosophy*, 32(02) : 249–281.

Donald Davidson. 1991. "Paradoxes de l'irrationalité." In *Paradoxes de l'irrationalité.* 21–43. éditions de l'éclat.

Dowding, Keith. 2002. "Revealed preference and external reference." *Rationality and Society*, 14(3) : 259–284.

Eliaz, Kfir, and Ran Spiegler. 2011. "Consideration sets and competitive marketing." *The Review of Economic Studies*, 78(1) : 235–262.

El-Safty, Ahmad E. 1976. "Adaptive behavior, demand and preferences." *Journal of Economic Theory*, 13(2) : 298–318.

Elster, Jon. 1983. *Sour Grapes : Studies in the Subversion of Rationality.* New York :Cambridge University Press.

Elster, Jon. 1989. "Social norms and economic theory." *Journal of economic perspectives*, 3(4) : 99–117.

Engelmann, Dirk, and Guillaume Hollard. 2010. "Reconsidering the effect of market experience on the "endowment effect"." *Econometrica*, 78(6) : 2005–2019.

Falk, Armin, and Jean Tirole. 2016. "Narratives, imperatives and moral reasoning." *Unpublished manuscript.*

Fehr, Ernst, and Karla Hoff. 2011. *Tastes, castes, and culture : the influence of society on preferences.* The World Bank.

Ferey, Samuel. 2011. "Paternalisme libéral et pluralité du moi." *Revue économique*, 62(4) : 737–750.

Fletcher, Jeffrey A, and Martin Zwick. 2007. "The evolution of altruism : game theory in multilevel selection and inclusive fitness." *Journal of theoretical biology*, 245(1) : 26–36.

Fodor, Jerry A. 1983. *The modularity of mind.* MIT press.

Frankfurt, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy*, 66 : 829–39.

Frankfurt, Harry G. 1988. "Freedom of the Will and the Concept of a Person." In *What Is a Person ?*. 127–144. Springer.

Frank Jackson, Graham Oppy, and Michael Smith. 1994. "Minimalism and Truth Aptness." *Mind*, 103 : 287–302.

Friedman, Milton. 1962. *Price Theory : A Provisional Text.*

Fritz, Peter, and Harvey Lederman. 2016. "Standard state space models of unawareness." *arXiv preprint arXiv :1606.07520.*

Fudenberg, Drew, and David K Levine. 2012. "Timing and self-control." *Econometrica*, 80(1) : 1–42.

Galbraith, John Kenneth. 1958. "The affluent society." *Nova York, New American Library.*

Gintis, Herbert. 1974. "Welfare criteria with endogenous preferences : the economics of education." *International Economic Review*, 415–430.

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr. 2003. "Explaining altruistic behavior in humans." *Evolution and human Behavior*, 24(3) : 153–172.

Gorman, William M. 1967. "Tastes, habits and choices." *International economic review*, 8(2) : 218–222.

Grabisch, Michel, and Christophe Labreuche. 2010. "A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid." *Annals of Operations Research*, 175(1) : 247–286.

Grüne-Yanoff, Till, and Sven Ove Hansson. 2009. *Preference change : Approaches from philosophy, economics and psychology.* Vol. 42, Springer Science & Business Media.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2004. "The role of social capital in financial development." *American economic review*, 94(3) : 526–556.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2006. "Does culture affect economic outcomes ?" *Journal of Economic perspectives*, 20(2) : 23–48.

Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and self-control." *Econometrica*, 69(6) : 1403–1435.

Gul, Faruk, and Wolfgang Pesendorfer. 2005. "The revealed preference theory of changing tastes." *The Review of Economic Studies*, 72(2) : 429–448.

Güth, Werner, and Hartmut Kliemt. 1998. "The indirect evolutionary approach : Bridging the gap between rationality and adaptation." *Rationality and Society*, 10(3) : 377–399.

Hansson, Sven Ove. 1995. "Changes in Preference." *Theory and Decision*, 38(1) : 1–28.

Hauk, Esther, and Maria Saez-Marti. 2002. "On the cultural transmission of corruption." *Journal of Economic theory*, 107(2) : 311–335.

Hill, Brian. 2009. "Three analyses of sour grapes." In *Preference Change*. 27–56. Springer.

Hill, Brian. 2010. "Awareness dynamics." *Journal of Philosophical Logic*, 39(2) : 113–137.

Hirschman, Albert O. 1983. *Bonheur privé, action publique.* Fayard.

Hirschman, Albert O. 1984. "Against parsimony : Three easy ways of complicating some categories of economic discourse." *Bulletin of the American Academy of Arts and Sciences*, 37(8) : 11–28.

Holbrook, Morris B, and John O'shaughnessy. 1988. "On the scientific status of consumer research and the need for an interpretive approach to studying consumption behavior." *Journal of consumer research*, 15(3) : 398–402.

Houlding, Brett. 2008. "Sequential decision making with adaptive utility." PhD diss. Durham University.

Inglehart, Ronald. 1971. "The silent revolution in Europe : Intergenerational change in post-industrial societies." *American political science review*, 65(4) : 991–1017.

Jackson, Frank. 1985. "Internal conflicts in desires and morals." *American Philosophical Quarterly*, 22(2) : 105–114.

Jaeggi, Rahel. 2014. *Alienation.* Columbia University Press.

Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson. 2005. "Failure to detect mismatches between intention and outcome in a simple decision task." *Science*, 310(5745) : 116–119.

Jonas, Hans. 2017. *Une éthique pour la nature.* Arthaud.

Jones, Ernest. 1908. "Rationalization in every-day life." *Journal of abnormal Psychology*, 3(3) : 161–169.

Kahneman, Daniel. 2012. *Système 1/Système 2 : Les deux vitesses de la pensée.* Flammarion.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory : An analysis od Decision under Risk." *Econometrica*, 47(2) : 263–292.

Kahneman, Daniel, and Jackie Snell. 1992. "Predicting a changing taste : Do people know what they will like ?" *Journal of Behavioral Decision Making*, 5(3) : 187–200.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler. 1990. "Experimental tests of the endowment effect and the Coase theorem." *Journal of political Economy*, 98(6) : 1325–1348.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian persuasion." *The American Economic Review*, 101(6) : 2590–2615.

Kant, Immanuel, and Jean Mondot. 1991. *Qu'est-ce que les Lumières ?* Publications de l'Université de Saint-Étienne.

Kapteyn, Arie, and Tom Wansbeek. 1982. "Empirical evidence on preference formation." *Journal of Economic Psychology*, 2(2) : 137–154.

Karni, Edi, and David Schmeidler. 1990. "Fixed Preferences and Changing Tastes." *American Economic Review*, 80(2) : 262–67.

Karni, Edi, and Marie-Louise Vierø. 2013. "" Reverse Bayesianism" : A choice-based theory of growing awareness." *American Economic Review*, 103(7) : 2790–2810.

Kavka, Gregory S. 1991. "Is individual choice less problematic than collective choice ?" *Economics & Philosophy*, 7(2) : 143–165.

Kőszegi, Botond, and Matthew Rabin. 2006. "A model of reference-dependent preferences." *The Quarterly Journal of Economics*, 121(4) : 1133–1165.

Kreps, David M. 1979. "A representation theorem for" preference for flexibility"." *Econometrica : Journal of the Econometric Society*, 565–577.

Lallement, Jérôme. 2002. "A la recherche des objets de l'économie." *Sciences de la sociét*, 55 : 9–20.

Lancaster, Kelvin J. 1966. "A new approach to consumer theory." *The journal of political economy*, 132–157.

Landsburg, Steven E. 1981. "Taste change in the United Kingdom, 1900-1955." *Journal of Political Economy*, 89(1) : 92–104.

Latané, Bibb, and John M Darley. 1976. *Help in a crisis : Bystander response to an emergency.* General Learning Press.

Lerner, Melvin J. 1980. "The belief in a just world." In *The Belief in a just World.* 9–30. Springer.

Levi, Isaac. 1990. *Hard choices : Decision making under unresolved conflict.* Cambridge University Press.

Levy, Moshe. 2015. "An evolutionary explanation for risk aversion." *Journal of Economic Psychology*, 46 : 51 – 61.

Lewis, David. 1978. "Truth in Fiction." *American Philosophical Quarterly*, 15(1) : 37–46.

Lewis, David. 1983. "Postscript to Truth in Fiction." In *Philosophical Papers.* 276–280. Oxford University Press.

Liu, Fenrong. 2011. *Reasoning about preference dynamics.* Vol. 354, Springer Science & Business Media.

Maddux, William W, Haiyang Yang, Carl Falk, Hajo Adam, Wendi Adair, Yumi Endo, Ziv Carmon, and Steven J Heine. 2010. "For whom is parting with possessions more painful ? Cultural differences in the endowment effect." *Psychological Science*, 21(12) : 1910–1917.

Maio, Gregory R, and James M Olson. 1998. "Values as truisms : Evidence and implications." *Journal of Personality and Social Psychology*, 74(2) : 294.

Maio, Gregory R, James M Olson, Lindsay Allen, and Mark M Bernard. 2001. "Addressing discrepancies between values and behavior : The motivating effect of reasons." *Journal of Experimental Social Psychology*, 37(2) : 104–117.

Manzini, Paola, and Marco Mariotti. 2007. "Sequentially rationalizable choice." *The American Economic Review*, 97(5) : 1824–1839.

Manzini, Paola, and Marco Mariotti. 2014. "Stochastic choice and consideration sets." *Econometrica*, 82(3) : 1153–1176.

Marschak, T. A. 1978. "On the Study of Taste Changing Policies." *The American Economic Review*, 68(2) : 386–391.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y Ozbay. 2016. "Revealed attention." In *Behavioral Economics of Preferences, Choices, and Happiness*. 495–522. Springer.

May, Kenneth O. 1954. "Intransitivity, utility, and the aggregation of preference patterns." *Econometrica : Journal of the Econometric Society*, 1–13.

Metcalfe, Les. 1993. "Conviction politics and dynamic conservatism : Mrs. Thatcher's managerial revolution." *International Political Science Review*, 14(4) : 351–371.

Milgrom, Paul, and John Roberts. 1986. "Relying on the information of interested parties." *The RAND Journal of Economics*, 18–32.

Mongin, Philippe. 2002. "Modèles d'information et théorie de la connaissance." *Laboratoire d'économétrie, Ecole Polytechnique*.

Morewedge, Carey K, and Colleen E Giblin. 2015. "Explanations of the endowment effect : an integrative review." *Trends in cognitive sciences*, 19(6) : 339–348.

Nalebuff, Barry J, Adam Brandenburger, and Agus Maulana. 1996. *Co-opetition.* HarperCollinsBusiness London.

Netzer, Nick. 2009. "Evolution of time preferences and attitudes toward risk." *American Economic Review*, 99(3) : 937–55.

Nisbett, Richard E, and Timothy D Wilson. 1977. "Telling more than we can know : Verbal reports on mental processes." *Psychological review*, 84(3) : 231.

Nozick, Robert. 1993. *The Nature of Rationality.* Princeton University Press.

Ok, Efe A. 2002. "Utility representation of an incomplete preference relation." *Journal of Economic Theory*, 104(2) : 429–449.

Parsons, Talcott. 1934. "Some reflections on "The nature and significance of economics"." *The Quarterly Journal of Economics*, 48(3) : 511–545.

Pessemier, Edgar A. 1978. "Stochastic Properties of Changing Preferences." *The American Economic Review*, 68(2) : 380–385.

Pettit, Philip. 1991. "Decision theory and folk psychology."

Pettit, Philip, and Michael Smith. 1990. "Backgrounding desire." *The Philosophical Review*, 99(4) : 565–592.

Polanyi, Karl, and Robert Morrison MacIver. 1944. *The great transformation.* Vol. 2, Beacon press Boston.

Pollak, Robert A. 1978. "Endogenous Tastes in Demand and Welfare Analysis." *The American Economic Review*, 68(2) : 374–379.

Proust, Joëlle. 2005. *La nature de la volonté.* Gallimard.

Quine, W. V. O. 1960. *Word and Object.* Cambridge, MA :MIT Press.

Rabin, Matthew. 1994. "Cognitive dissonance and social change." *Journal of Economic Behavior & Organization*, 23(2) : 177–194.

Robbins, Baron Lionel Robbins. 1932. *An essay on the nature & significance of economic science.* Macmillan & co., limited.

Robinson, James, and R Acemoglu. 2012. *Why nations fail.* Crown Publishing Group.

Rodrik, Dani. 2014. "When ideas trump interests : Preferences, worldviews, and policy innovations." *Journal of Economic Perspectives*, 28(1) : 189–208.

Rokeach, Milton. 1973. *The nature of human values.* Free press.

Rozen, Kareen. 2010. "Foundations of intrinsic habit formation." *Econometrica*, 78(4) : 1341–1373.

Rubinstein, Ariel, and Yuval Salant. 2011. "Eliciting welfare preferences from behavioural data sets." *The Review of Economic Studies*, 79(1) : 375–387.

Salant, Yuval, and Ariel Rubinstein. 2008. "(A, f) : choice with frames." *The Review of Economic Studies*, 75(4) : 1287–1296.

Savage, Léonard Jimmy. 1954. "The Foundation of Statistics."

Scanlon, Thomas. 1998. *What we owe to each other.* Harvard University Press.

Schipper, Burkhard C. 2014. "Preference-based unawareness." *Mathematical Social Sciences*, 70 : 34–41.

Schumpeter, Joseph Alois, and François Perroux. 1935. "Théorie de l'évolution économique."

Schwartz, Shalom H. 2012. "An overview of the Schwartz theory of basic values." *Online readings in Psychology and Culture*, 2(1) : 11.

Schwarz, Norbert, and Fritz Strack. 1991. "Context effects in attitude surveys : Applying cognitive theory to social research." *European review of social psychology*, 2(1) : 31–50.

Searle, John R. 1983. *Intentionality : An Essay in the Philosophy of Mind.* Cambridge University Press.

Searle, John R. 1995. *The Construction of Social Reality.* Free Press.

Sen, Amartya. 1993. "Internal consistency of choice." *Econometrica : Journal of the Econometric Society*, 495–521.

Sen, Amartya K. 1977. "Rational fools : A critique of the behavioral foundations of economic theory." *Philosophy & Public Affairs*, 317–344.

Shafir, Eldar, Itamar Simonson, and Amos Tversky. 1993. "Reason-based choice." *Cognition*, 49(1-2) : 11–36.

Slovic, Paul. 1995. "The construction of preference." *American psychologist*, 50(5) : 364.

Smith, Adam. 2014 (1759). *Théorie des sentiments moraux.* . 3ème édition ed., Paris :Presses Universitaires de France - PUF.

Smith, John Maynard. 1982. *Evolution and the Theory of Games.* Cambridge university press.

Stigler, George J, and Gary S Becker. 1977. "De gustibus non est disputandum." *The american economic review*, 67(2) : 76–90.

Strahilevitz, Michal A, and George Loewenstein. 1998. "The effect of ownership history on the valuation of objects." *Journal of consumer research*, 25(3) : 276–289.

Strotz, Robert Henry. 1955. "Myopia and inconsistency in dynamic utility maximization." *The Review of Economic Studies*, 23(3) : 165–180.

Sugden, Robert. 2005. "Why rationality is not a consequence of Hume's theory of choice." *The European Journal of the History of Economic Thought*, 12(1) : 113–118.

Tabellini, Guido. 2010. "Culture and institutions : economic development in the regions of Europe." *Journal of the European Economic association*, 8(4) : 677–716.

Ticchi, Davide, Thierry Verdier, and Andrea Vindigni. 2013. "Democracy, dictatorship and the cultural transmission of political values."

Tourangeau, Roger. 1992. "Context Effects on Responses to Attitude Questions : Attitudes as Memory Structures." *Context Effects in Social and Psychological Research*, , ed. Norbert Schwarz and Seymour Sudman, 35–47. New York, NY :Springer New York.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under uncertainty : Heuristics and biases." *science*, 185(4157) : 1124–1131.

Tversky, Amos, and Daniel Kahneman. 1983. "Extensional versus intuitive reasoning : The conjunction fallacy in probability judgment." *Psychological review*, 90(4) : 293.

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. "The causes of preference reversal." *The American Economic Review*, 204–217.

Tversky, Amos, Shmuel Sattath, and Paul Slovic. 1988. "Contingent weighting in judgment and choice." *Psychological review*, 95(3) : 371.

Universals in the Content and Structure of Values : Theoretical Advances and Empirical Tests in 20 Countries. n.d.. "Universals in the Content and Structure of Values : Theoretical Advances and Empirical Tests in 20 Countries." In . , ed. Mark P. Zanna.

Van Benthem, Johan, and Fenrong Liu. 2007. "Dynamic logic of preference upgrade." *Journal of Applied Non-Classical Logics*, 17(2) : 157–182.

Van Benthem, Johan, and Fernando R Velázquez-Quesada. 2010. "The dynamics of awareness." *Synthese*, 177(1) : 5–27.

Van de Stadt, Huib, Arie Kapteyn, and Sara Van de Geer. 1985. "The relativity of utility : Evidence from panel data." *The review of Economics and Statistics*, 179–187.

von Weizsäcker, Carl Christian. 1971. "Notes on endogenous change of tastes." *Journal of Economic Theory*, 3(4) : 345 – 372.

Von Wright, Georg Henrik. 1963. *The logic of preference.* University of Edinburgh Press Edinburgh Scotland.

Weber, Max. 2003. *The Protestant Ethic and the Spirit of Capitalism.* Courier Dover Publications.

Weber, Max. 2013. *Le savant et le politique.* Presses Électroniques de France.

Weber, Max, Guenther Roth, Claus Wittich, Eric de Dampierre, Julien Freund, and Jacques Chavy. 1995. *Économie et société.*

Weckstein, Richard S. 1962. "Welfare Criteria and Changing Tastes." *The American Economic Review*, 52(1) : 133–153.

Whorf, Benjamin Lee. 1956. "Language, thought, and reality : selected writings of. . . .(Edited by John B. Carroll.)."

Williams, Bernard. 1981. "Internal and external reasons." *This page intentionally left blank*, 60.

Wilson, Timothy D, Douglas J Lisle, Jonathan W Schooler, Sara D Hodges, Kristen J Klaaren, and Suzanne J LaFleur. 1993. "Introspecting about reasons can reduce post-choice satisfaction." *Personality and Social Psychology Bulletin*, 19(3) : 331–339.

UNE APPROCHE DES CHANGEMENTS DE PRÉFEÉRENCE PAR LA DÉLIBÉRATION

RÉSUMÉ : L'objectif de cette thèse est d'enrichir la boîte à outils formelle et conceptuelle de la théorie des transformations de préférence. Pour ce faire, je modélise une procédure, la délibération partielle, par laquelle les agents changent l'ensemble des valeurs qui induisent leur relation de préférences. La délibération partielle se fonde sur l'idée qu'un tel changement résulte d'une prise de conscience (growing awareness) de ces certaines de ces valeurs. En prenant conscience de nouvelles valeurs, l'agent réalise qu'il adhère (respectivement qu'il n'adhère pas) à de mauvaises valeurs (resp. de bonnes), soit que les valeurs auxquelles il adhère sont contradictoires, soit qu'elles se renforcent mutuellement. Au travers d'une analyse critique de la littérature sur les transformations des préférences, l'introduction s'attache à dégager les principales motivations philosophiques de la délibération partielle. Le premier chapitre pose les fondations conceptuelles et formelles de cette procédure et présente, tout en les justifiant, les cinq grandes hypothèses psychologiques que suppose cette procédure. Le second chapitre axiomatise cette procédure et fournit une première interprétation du lien qu'elle entretient avec les changements de choix. Il spécifie de deux structures axiologiques sur laquelle se fonde cette procédure : une structure monotonique et une organisation partitionnelle. Le troisième chapitre, s'appuie sur la délibération partielle pour construire une théorie de la manipulation des préférences. L'idée est qu'un envoyeur choisit conjointement un projet et une stratégie de divulgation valeurs afin de manipuler les préférences d'un receveur. J'établie donc un lien entre le modèle développé au chapitre 2, je développe une théorie de la l'empathie imparfaite et de la divulgation séquentielle des valeurs.

Mots clefs : *changements de préférences - conscience - délibération - rationalité axiologique*

APPROACHING PREFERENCE CHANGE BY PARTIAL DELIBERATION

ABSTRACT : This Ph.D. dissertation aims at providing new conceptual and formal tools in order to model preference changes. To do so, I model a mechanism that I refer to as partial deliberation. Partial deliberation is based on the idea that individuals change their preference by becoming aware of new values. Indeed, their awareness allows them either to reject values they were adhering to or to adopt new values. By critically analysis of the literature on preference changes and on rational choice theory, the introduction emphasizes the main philosophical issues of partial deliberation. The first chapter justifies the five psychological hypotheses on which this mechanism relies and discusses their relation with rational choice theory. The second chapter axiomatically deals with partial deliberation and it formalizes two specific structures : a monotonic structure and partitional structure. Then, the third chapter models a situation of preference manipulation, in which a sender

jointly chooses a project and a disclosure strategy in order to manipulate the preference of a receiver. With this model, I account for imperfect empathy and sequential disclosure of values.

Keywords : *preference changes - awareness - deliberation - axiological rationality*