Critique de l'utilitarisme

By Ralph

Définition formelle de l'utilité

Le préordre de préférence

Formellement, soit A l'ensemble des résultats. Chaque agent a ses préférences sur cet ensemble A . Les préférences d'un agent sont modélisées par une relation binaire sur A notée $\geq$.

$a \geq b$ signifie « l'individu préfère a à b ou est indifférent ».

Pour représenter des préférences, on impose les conditions de cohérence suivantes :

- Réflexivité : $a \geq a$
- Transitivité : $a \geq b$ et $b \geq c \Rightarrow a \geq c$

$\geq$ est ainsi un préordre. Ces conditions sont suffisantes.

La relation de préférence $\geq$ détermine entièrement tous les choix possibles et imaginables que puisse faire un individu. Par exemple, comment modéliser un choix sous contrainte ? Rien de plus simple, on exprime la contrainte par le fait qu'elle limite l'ensemble des résultats possibles à un sous-ensemble de A . Le choix optimal pour l'agent est alors tout simplement un maximum de $\geq$ sur le sous-ensemble en question.

Fonctions d'utilité

Le préordre de préférence est la seule modélisation véritablement intrinsèque d'un classement sur des choix possibles. On pourrait introduire une autre façon de modéliser ce classement. On affecte un indice numérique à chaque choix possible de telle sorte qu'un choix est préférable à un autre si et seulement si son indice numérique est plus élevé. Mathématiquement, on a introduit une fonction à valeurs réelles appelée *fonction d'utilité* (de l'agent considéré). Notons U cette fonction. Elle satisfait à la condition suivante :

$$U(a) \geq U(b) \Leftrightarrow a \geq b,$$

Bien sûr, à la gauche de l'expression, $\geq$ représente l'ordre usuel sur les réels, alors qu'à droite, $\geq$ désigne le préordre de préférence de l'agent considéré.

Le point crucial est qu'en introduisant des indices numériques pour modéliser les préférences d'un agent, on a ajouté un élément d'arbitraire. Par définition, n'importe quel autre choix d'indices numériques qui conserve le classement aurait convenu. Or, il n'y a pas unicité du choix : si une fonction d'utilité U représente un préordre de préférences, alors toute transformation strictement croissante de U représente ce préordre (car à l'évidence, le classement induit est inchangé). On peut donc énoncer que :

U représente $\geq \Leftrightarrow f(U)$ représente $\geq$ pour toute fonction f strictement croissante.
--

U et $f(U)$ représentent les *mêmes* préférences. Par conséquent, tout énoncé formulé sur U , doit avoir la même signification si on remplace U par $f(U)$, ou alors il est dépourvu de sens.

En termes mathématiques, on parle de classe d'équivalence. Toute énoncé portant sur $\geq$ peut être formulé en utilisant un représentant U de $\geq$, mais l'énoncé doit être indépendant du représentant. Une fonction d'utilité n'a donc pas d'existence intrinsèque contrairement au préordre de préférence. Pourquoi alors utiliser des fonctions d'utilité ? La réponse est que c'est un objet mathématique beaucoup plus simple, qui se prête notamment au calcul différentiel.

Exposé de la thèse utilitariste

La thèse utilitariste consiste à considérer que l'objet intrinsèque pour représenter les préférences d'un agent est la fonction d'utilité. Ainsi que nous l'avons vu, les utilitaristes sont objectivement nuls en maths ! Cette erreur a pour part une raison historique. En effet, historiquement, les fonctions d'utilité ont été introduites avant les relations de préordre, à une époque où la formalisation mathématique n'était pas encore ce qu'elle est.

Plus généralement, les thèses utilitaristes considèrent qu'un préordre de préférence est insuffisant pour modéliser les préférences d'un agent. Dans le cadre de notre analyse, une théorie sera dite utilitariste si et seulement si la modélisation des préférences d'un agent excède la donnée de son préordre de préférence.

L'impossible mesure de l'intensité

On justifie souvent le fait que la relation de préférence ne décrit pas toute la préférence en prenant un exemple avec deux choix possibles (A et B). En effet, dans ce cas, il semble logique que la seule information donnée par $\geq$ ($A > B$, $A \sim B$, ou $A < B$) est relativement pauvre et qu'il faut introduire une intensité (de combien A est préférée à B) pour modéliser les préférences.

Cet exemple est spécieux en ce qu'il limite arbitrairement l'ensemble sur lequel est définie la relation de préférence par rapport à l'ensemble des décisions possibles. Prenons un exemple : le choix possible A correspond à « manger une pomme », le choix B correspond à « manger une poire ». Que peut signifier qu'un individu préfère « trois fois une pomme à une poire » (où l'on remarque que l'on a défini une intensité) ? La seule réponse acceptable est que cet individu est indifférent entre manger une pomme et trois poires. On remarque alors qu'on a défini un nouveau choix possible C « manger trois poires ».

Ainsi, on a ramené un problème posé en termes d'intensité à un problème posé en termes de relation de préférence posé sur un ensemble de choix plus grand. On voit en quel sens c'est la restriction des choix possibles (ou potentiels, imaginables) qui permet de justifier fallacieusement le raisonnement en termes d'intensité.

On peut aller plus loin et dire que toute expression en terme d'intensité doit pouvoir être exprimée également en terme de préférence d'occurrences par rapport à d'autres. Dans le cas contraire, on peut se demander ce que signifie cette intensité. *Parce qu'aucune information sur des choix imaginables ne pourra jamais en être déduite.* C'est là le point crucial.

Une conséquence : la manipulabilité des modes de décisions collectives

Une conséquence immédiate de ce qui précède est la manipulabilité de tout mode de décision collective dans le cadre utilitariste, et donc en fin de compte l'échec de l'utilitarisme dans sa prétention à agréger les préférences.

En effet, comme on ne peut avoir aucune « information » excédant le préordre de préférence, un individu pourra toujours prétendre que la fonction d'utilité U qui lui est associée, lorsqu'elle le désavantage, est fautive, et choisir une autre fonction d'utilité

cohérente avec $\geq$ (par exemple $\exp(U)$!). D'un point de vue logique, il aura parfaitement raison, et il ruinera les calculs de l' « ingénieur social ».

Les comparaisons interpersonnelles dans le cadre utilitariste

Ironiquement, on peut aller plus loin, et constater que même en acceptant le concept déjà métaphysique d'intensité des préférences, ce que nous ne faisons pas, les comparaisons interpersonnelles restent impossibles. En effet, *les mesures d'intensité restent relatives à chaque individu !*

Comment inférer du fait qu'une personne (1) « préfère trois fois une pomme à une poire » et qu'une autre (2) « préfère deux fois une pomme à une poire », laquelle « préfère le plus les pommes » ? Bien sûr, cette phrase n'a pas de sens. Pour l'utilitariste, cependant, elle doit en avoir un. Pour cela, il faut ajouter à l'arbitraire de l'intensité, *l'arbitraire de l'étalon commun*.

Dans l'exemple donné, si on fixe comme étalon commun une poire (plus précisément consommer une poire équivaut à une utilité de 1 pour tout individu), alors (1) « préfère plus les pommes » que (2). Notons qu'on aurait pu choisir la pomme comme étalon commun. A ce moment-là, (1) préfère autant les pommes que (2), par hypothèse, et (2) préfère plus les poires que (1). Heureusement, dans un monde utilitariste, des ingénieurs sociaux veilleraient pour choisir le « bon » étalon.

Intensité et théorie du choix social

Admettons donc pour la forme que les mesures d'intensité « intra-individuelle » sont valides, mais qu'il n'y a pas d'étalon commun accepté. Formellement, cela se traduit par le fait que deux fonctions d'utilité ne différant que par une transformation affine (strictement croissante), sont indistingables, correspondent aux mêmes préférences. En effet, le choix d'un étalon commun correspond à fixer les niveaux 0 (absence du bien étalon) et 1 d'utilité (présence du bien étalon).

Dans ce cadre, *les théorèmes d'impossibilité d'Arrow et de Gibbard-Satterthwaite restent valides¹*. Ceci est intuitif parce que ces impossibilités consistent en l'impossibilité d'agréger des préférences individuelles en une préférence collective, et parce qu'admettre que l'intensité « intra-individuelle » a un sens ne nous permet pas d'inférer quoi que ce soit sur les comparaisons interpersonnelles.

Le choix du critère d'agrégation des utilités

Le choix du critère d'agrégation des utilités constitue une troisième source d'arbitraire ! En effet pourquoi l'utilité collective (par définition ce qu'on cherche à maximiser dans le cadre utilitariste) serait-elle la somme des utilités individuelles ? Pourquoi pas le leximin si cher à Rawls ? Pourquoi pas n'importe quel autre critère ?

Conclusion : l'utilitarisme comme triplet d'arbitraires

Par conséquent toute théorie utilitariste doit faire trois choix arbitraires concernant :

- Ø La définition d'une intensité pour chaque individu ;
- Ø Un moyen de comparer les différentes intensités individuelles ;
- Ø Un critère d'agrégation des utilités.

Nous pouvons maintenant définir une théorie utilitariste : c'est la spécification des trois choix ci-dessus, c'est un triplet d'arbitraires purs !

¹ Voir notre **article** «Théorèmes d'impossibilité» pour plus de précision

Incompatibilité du libéralisme avec l'utilitarisme classique

Bien que l'essentiel ait été dit dans les paragraphes précédents, nous développons maintenant une critique conséquentialiste de l'utilitarisme. Nous mettons en avant des traits de l'utilitarisme qui sont radicalement incompatibles avec le libéralisme. Bien que cette partie s'adresse en priorité aux libéraux, à notre avis les arguments développés sont suffisamment impressionnants pour détourner de l'utilitarisme même des personnes s'estimant hostiles au libéralisme.

Rq. : nous nous limitons ici à l'utilitarisme classique (où l'utilité collective est définie comme la somme des utilités individuelles).

La « loi des avantages comparatifs » dans la version utilitariste ou l'esclavage des talentueux

Lorsque l'on s'intéresse à une économie de production, une nouvelle caractéristique de l'utilitarisme émerge. Considérons une économie composée de deux individus. Les deux individus ont la même fonction d'utilité mais leurs capacités productives diffèrent. Sous les hypothèses utilitaristes, l'un des deux individus est donc relativement plus apte à produire qu'à consommer, tandis que l'autre est relativement plus apte à consommer qu'à produire. Quelle est donc l'allocation optimale ? Un principe semblable à la loi des avantages comparatifs de Ricardo s'applique, et la conclusion est sans appel : une division des tâches doit s'instaurer dans laquelle le (relativement) plus productif produit, tandis que l'autre consomme ! Cet effet, mis en avant pour la première fois par Mirrlees en 1974, a été appelé (à juste titre !) l'esclavage des talentueux.

Hervé Moulin a établi formellement le résultat suivant : parmi des individus ayant la même fonction d'utilité, mais ayant une productivité différente, le critère utilitariste conduit à une utilité plus grande *pour les moins productifs*, si le loisir est un bien normal². Il donne un exemple raisonnable dans lequel un individu deux fois plus productif gagne au total deux fois moins tout en ayant quatre fois moins de loisir³. Remarquons en effet que l'allocation optimale devra être telle que la désutilité marginale de l'heure de travail soit deux fois plus grande pour la personne qualifiée, ce qui ne peut se produire que si elle travaille beaucoup plus. Le fait qu'elle bénéficie d'une moindre grande quantité de bien résultant d'un effet substitution loisir-bien. Et c'est seulement parce qu'on suppose une certaine substituabilité loisir-bien chez les plus productifs que ces derniers pourront profiter d'un petit peu de loisir.

Droits de propriété et utilité

Un autre aspect du libéralisme en délicatesse avec l'utilitarisme est l'existence même des droits de propriété. D'une manière plus générale, nous pourrions mettre en avant l'inadéquation même de l'existence de règles dans la conception utilitariste. Contentons-nous de commenter la tentative du libertarien-utilitariste David Friedman pour concilier les droits de propriété avec le critère utilitariste (en l'occurrence ici, le critère marshallien) :

Supposons que la pomme vaille deux dollars pour vous et quatre dollars pour moi. Au lieu de l'acheter pour trois dollars, je m'introduis furtivement dans votre verger, la nuit venue, pour la dérober, ce qui me coûte un dollar en temps et en effort. Vous perdez deux dollars (valeur que vous attribuez à la pomme), et je gagne trois dollars (valeur que

² Hypothèse guère contestable dans la mesure où le loisir n'a pas de substitut proche

³ Hervé Moulin – *Axioms of Cooperative Decision Making* p22-28

j'attribue à la pomme, moins les frais encourus pour me la procurer) : il y a donc un gain net d'un dollar ; lorsque je vole la pomme, cela représente une amélioration économique par rapport au fait de ne pas avoir cette pomme. Mais ne pas me procurer la pomme n'est pas le seul autre choix ; au lieu de cela, j'aurais pu l'acheter. Voler la pomme est pire que l'acheter, puisque l'achat de la pomme aurait abouti à un gain net de deux dollars. Une règle de droit efficace fera en sorte que, pour les gens qui veulent des pommes, il y aille de leur intérêt de les acheter plutôt que de les voler. C'est la raison pour laquelle nous punissons les voleurs.⁴

L'argument est assez catastrophique. C'est parce que le vol implique des frais de mise en œuvre qu'il est jugé inefficace et doit être réprimé. Mais l'on pourrait objecter que les frais globaux seraient encore plus faibles si l'on interdisait aux victimes de porter plainte, ce qui donnerait naissance au principe suivant : « le droit de propriété est respecté, mais tout voleur devient le légitime acquéreur de ce qu'il a volé ». Pour diminuer encore les frais de vol, on pourrait interdire aux propriétaires de défendre leur propriété, de sorte que le vol ne soit a priori pas plus coûteux que n'importe quel coût de transaction émanant d'un échange volontaire. Bref, dans la conception utilitariste pure et dure, la détention par une personne d'un objet ne peut être qu'arbitraire au sens où, a priori, la détention par une autre personne est tout aussi légitime.

Autres points

L'utilité en environnement incertain

La théorie de Von Neumann – Morgenstern (VNM) prouve le résultat suivant en environnement incertain : pour chaque individu, il existe une fonction appelée utilité de VNM et notée u , unique à transformation croissante affine près, telle que l'espérance de cette fonction représente les préférences de l'individu (i.e. $E[u]$ est une fonction d'utilité représentant les préférences de l'individu). Du fait que l'arbitraire du choix de u est réduit par rapport au choix de U en environnement certain, u est défini non plus à une transformation croissante près mais à une transformation croissante *affine* près, ce résultat a suscité de l'espoir chez les utilitaristes. Harsanyi s'est notamment engagé dans cette voie pour défendre l'utilitarisme.

Néanmoins, cette tentative est vaine. Le fait que les mêmes préférences puissent être représentées par l'espérance de deux utilités de VNM différant par une transformation affine, ruine tout espoir de fonder un étalon commun (et souligne en fait l'absence de cet étalon commun). D'autre part la fonction d'utilité $U = E[u]$ ne joue aucun rôle particulier, $V = \exp(E[u])$ représente tout aussi bien les mêmes préférences. Ainsi, et c'est là un point qui détruit entièrement les pesants travaux d'Harsanyi, l'environnement incertain ne réduit pas la classe des fonctions d'utilité représentant un même préordre de préférences.

Répetons-le, une fonction d'utilité est un artifice mathématique. En tant qu'artifice mathématique, $U = E[u]$ s'impose sans aucun doute sur ses rivales, mais n'a *aucune* dimension éthique particulière.

⁴ David Friedman – *The Machinery of Freedom* (version française : *Vers une société sans Etat* p289)

La position de Hammond

La position de Hammond est intéressante en ce que ce dernier est un utilitariste qui a admis l'arbitraire de la comparaison interpersonnelle des utilités. Contrairement à ce que le titre d'un de ses articles pourrait laisser croire⁵, Hammond admet qu'il n'y pas de critère objectif pour effectuer des comparaisons interpersonnelles, et c'est ce qui le pousse à souhaiter de développer des critères éthiques de comparaisons interpersonnelles :

Interpersonal comparisons of utility (ICU's) have to be made if there is to be any satisfactory escape from Arrow's impossibility theorem, with its implication that individualistic social choice has to be dictatorial (or at least oligarchic),

[...]

These considerations then suggest the need to consider the relationship between ICU's and explicit ethical choices regarding numbers of different types of people in the population. The result is an enriched social choice theory, capable of handling a broader range of ethical decision problems, in which the ICU's are explicit and play a clear role in the analysis.⁶

Ceci nous inspire deux remarques. D'une part, ce qui motive Hammond, à savoir dépasser le théorème d'impossibilité en introduisant les comparaisons interpersonnelles, repose sur une confusion. Le théorème d'Arrow discrédite effectivement toutes les théories éthiques essayant d'agrèger les préférences individuelles en une mythique préférence collective. En tant que tel, il discrédite notamment la démocratie, dont la légitimité réside dans la possibilité d'agrèger au moins sous une certaine forme, les préférences individuelles. Cependant, une théorie éthique ne doit pas nécessairement rechercher le type de régularité collective visé par le théorème d'Arrow. Par exemple, le cœur d'un jeu coopératif fournit une solution satisfaisante pour tout le monde à un problème de coopération, sans que cette solution ne puisse être formulée en termes de préférences collectives. Dans la mesure où une théorie éthique n'a donc pas nécessairement vocation à dépasser l'impossibilité du théorème d'Arrow – et de toute façon, elle ne le pourrait pas – la prétendue nécessité d'établir des comparaisons interpersonnelles ne s'impose pas.

D'autre part, le choix éthique nécessaire pour compléter une théorie utilitariste va à l'encontre du welfarisme inhérent à l'utilitarisme. Par welfarisme, nous entendons une doctrine qui prend en compte uniquement les préférences des agents et non des critères externes aux préférences, ce qu'est à l'évidence un choix éthique a priori. Le welfarisme est l'un des points forts de l'utilitarisme ; en l'abandonnant, on se heurte aux problèmes habituels au sujet des fondements des normes.

⁵ Hammond - *Interpersonal Comparisons of Utility : Why and How they are and should be made*

⁶ Hammond – *Ibid.*