Le statut épistémologique des « mécanismes » face aux paradoxes de la rationalité chez Jon Elster

Mamadi Aïssata CAMARA,
Département de Philosophie, Université de Kindia, Guinée, assaïtamadicamara@gmail.com
Souglouman BAMPINI,
Département de Philosophie, Université de Kindia, Guinée, Maître Assistant, bampinijosephsouglouman@gmail.com
Bolivard Koïkoï GROVOGUI,
Département de philosophie, Université Julius Nyerere de Kankan, Guinée, Maître Assistant, bolivardgro@gmail.com

Résumé :

Cette étude analyse le changement de paradigme de Jon Elster, dans Le Désintéressement (2009) et L'Irrationalité (2010). Face au modèle dominant de l'homo economicus (calculateur parfait guidé par son seul intérêt), Jon Elster montre que l'altruisme, l'auto-tromperie et la faiblesse de la volonté obéissent à des logiques psychologiques précises. Aussi, pour dépasser les limites de la théorie du choix rationnel, Jon Elster propose-t-il de remplacer la recherche de lois universelles par une boîte à outils de « mécanismes ». Ceux-ci sont des déclencheurs fréquents qui permettent d'expliquer un comportement a posteriori, bien que leur issue demeure imprévisible a priori en raison de nos contradictions émotionnelles. Malgré les critiques soulignant la difficulté de vérification empirique ou un excès de psychologisation, cette approche offre un réalisme précieux. En admettant une rationalité imparfaite, elle permet aux sciences humaines d'embrasser la complexité humaine.

Mots-clés : *Altruisme, logique de l'action, irrationalité, mécanisme psychologique, philosophie des sciences.*

Abstract :

This study analyses Jon Elster's paradigm shift in The Cement of Society / Reason and Rationality (referencing Le Désintéressement [2009] and L'Irrationalité [2010]). Challenging the dominant model of the homo economicus, a perfect calculator driven solely by self-interest, Elster demonstrates that altruism, self-deception, and weakness of will follow precise psychological logics. To overcome the limitations of rational choice theory, Elster proposes replacing the search for universal laws with a toolkit of

« mechanisms ». These are frequent triggers that make it possible to explain behavior a posteriori, although their outcome remains unpredictable a priori due to our emotional contradictions. Despite criticisms highlighting the difficulty of empirical verification or an excess of psychologisation, this approach offers invaluable realism. By acknowledging imperfect rationality, it enables the humanities and social sciences to embrace human complexity.

Keywords: *Altruism, logic of action, irrationality, psychological mechanism, philosophy of science*.

Introduction

En philosophie des sciences sociales, la logique de l'explication a traditionnellement oscillé entre deux paradigmes antagonistes : le modèle nomologique-déductif, emprunté aux sciences de la nature, qui cherche à subsumer les faits sous des lois universelles, et la démarche herméneutique, centrée sur la compréhension des significations subjectives. Au carrefour de ces tensions, la théorie du choix rationnel a tenté d'imposer une unification méthodologique stricte en érigeant le postulat de l'*homo economicus,* cet agent fictif guidé par la maximisation de son intérêt personnel et par une cohérence parfaite de ses choix, en loi fondamentale du comportement humain. L'avènement de ce modèle axiomatique a ainsi relégué les écarts de conduite, les élans altruistes et les passions au rang de scories statistiques ou d'anomalies négligeables. C'est précisément à l'endroit de cette faille conceptuelle que se déploie l'œuvre de Jon Elster. En effet, à travers les deux premiers volumes de son *Traité critique de l'homme économique*, intitulés respectivement *Le Désintéressement* (2009) et *L'Irrationalité* (2010), Jon Elster orchestre une déconstruction systématique de cet impérialisme économique. L'étude de sa pensée ne se justifie pas seulement par un intérêt pour la psychologie morale, mais par une exigence proprement épistémologique : elle interroge les conditions de validité et la structure interne des modèles explicatifs lorsque leur hypothèse centrale se trouve invalidée par l'expérience. Dès lors, face à l'omniprésence de comportements qui échappent au calcul utilitaire, une question de recherche fondamentale surgit : quel statut épistémologique Jon Elster accorde-t-il à la notion de « mécanisme » pour préserver l'intelligibilité des conduites irrationnelles et désintéressées, sans pour autant faire basculer

les sciences sociales dans le relativisme méthodologique ou le pur descriptivisme ?

À cette interrogation, nous formulons l'hypothèse que l'introduction des concepts de désintéressement et d'irrationalité chez Jon Elster ne vise pas à dissoudre la scientificité des sciences humaines, mais à opérer une rupture épistémologique majeure : le passage d'une science fondée sur des lois prédictives à une science axée sur des mécanismes causaux non déductibles. Car, loin d'être un résidu d'explication, l'irrationalité est analysée comme une modalité structurée de l'esprit, rendue intelligible grâce à un individualisme méthodologique élargi qui intègre les apports de la psychologie cognitive et de la philosophie analytique de l'action. L'objectif de notre travail est d'évaluer la validité scientifique du paradigme de Jon Elster en étudiant l'impact du désintéressement et des biais cognitifs sur la théorie du choix rationnel.

Notre démarche s'inscrit dans la philosophie analytique et d'une analyse conceptuelle stricte de la logique explicative de J. Elster. Nous examinons dans un premier temps la fissure du modèle de l'intérêt personnel, en montrant comment le désintéressement s'impose comme une anomalie épistémologique face au réductionnisme économique. Dans un deuxième temps, nous analysons l'anatomie de l'irrationalité en tant qu'objet de science, en disséquant les structures logiques des croyances motivées et des passions. Pour finir, nous évaluons la portée épistémologique du paradigme des mécanismes causaux, en interrogeant la redéfinition de la scientificité que J. Elster impose lorsque l'on renonce délibérément au pouvoir de prédiction au profit de la seule intelligibilité rétrospective.

1. Le désintéressement comme anomalie épistémologique chez J. Elster

Le propre d'une science engagée dans une trajectoire de mathématisation et d'unification paradigmatique est de chercher à réduire la diversité des phénomènes observés à un nombre minimal de principes fondamentaux. Dans le domaine des sciences sociales, et plus particulièrement sous l'égide de la théorie du choix rationnel, ce travail de réduction a trouvé son point d'ancrage absolu dans le postulat de l'intérêt personnel. L'*homo economicus* n'est pas simplement une hypothèse de

travail commode. Il constitue le pivot axiomatique qui garantit la prédictibilité des comportements, la formalisation algébrique des préférences et la clôture logique des modèles. C'est précisément à la racine de cette architecture épistémologique que J. Elster (2009) porte le fer dès le premier volume de son *Traité critique de l'homme économique*. Il introduit la notion de désintéressement qu'il traite non pas comme un simple supplément d'âme ou comme une concession faite à la morale ordinaire, mais plutôt comme une rupture épistémologique majeure qui arrache la science sociale à son dogme réductionniste.

Pour comprendre la portée de ce geste, il convient d'en examiner la formulation initiale, là où J. Elster (2009, p. 14) pose les jalons de sa dissidence théorique : « contre la tendance à chercher l'intérêt personnel partout, je soutiens que le désintéressement existe, qu'il est substantiel et qu'il ne peut pas être réduit à des motivations égoïstes déguisées. » Cette affirmation, d'apparence simple, constitue en réalité une charge frontale contre l'impérialisme économique qui a colonisé la sociologie, la science politique et la philosophie de l'action tout au long de la seconde moitié du XXe siècle. Pour tout spécialiste en épistémologie, le geste de J. Elster s'apparente à la mise en évidence d'une « anomalie » au sens popperien (K. Popper, 2007) et kuhnien (T. Kuhn, 2008), c'est-à-dire un fait récalcitrant qui refuse de se laisser absorber par le paradigme dominant et qui, par sa simple persistance, force à interroger les fondations mêmes de la théorie.

Le premier niveau de la critique elstérienne est d'ordre logique et méthodologique. L'un des tours de force de la théorie du choix rationnel a été sa capacité à réinterpréter toute action humaine, y compris les manifestations les plus évidentes d'altruisme, de sacrifice ou de dévouement public, comme des formes sophistiquées d'égoïsme rationnel. Dans cette perspective, l'individu qui donne sa fortune à une œuvre de charité ou qui risque sa vie pour sauver un inconnu ne ferait que maximiser une utilité subjective invisible : il achèterait sa propre paix intérieure, éviterait le coût psychologique de la culpabilité, ou investirait dans sa réputation sociale pour en retirer des bénéfices ultérieurs. J. Elster (1984) démontre avec une grande acuité conceptuelle que ce réductionnisme omniprésent repose sur un sophisme

tautologique. Si l'on pose comme définition préalable que toute action choisie est l'expression d'une préférence, et que toute préférence vise par nature la satisfaction de l'agent, alors il devient logiquement impossible d'agir autrement que par intérêt personnel. Une telle théorie n'explique plus rien, puisqu'elle explique tout et son contraire. En termes popperiens, elle souffre d'un défaut rédhibitoire d'infalsificabilité. Si le comportement d'un homme qui thésaurise de l'argent et celui d'un homme qui s'en dépouille sont tous deux validés comme des preuves de la maximisation de leur intérêt personnel, la théorie du choix rationnel cesse d'être une hypothèse scientifique testable pour devenir une métaphysique infalsifiable, un dogme immunisé contre les démentis de l'expérience.

En restaurant le désintéressement dans sa pureté conceptuelle, J. Elster (2007) brise cette circularité. Le désintéressement elstérien n'est pas un égoïsme à long terme, ni un calcul inconscient. Il se définit négativement comme l'absence de retour de l'action sur le bien-être ou l'utilité de l'agent. Cela signifie que Jon Elster réintroduit de la contingence et de la dualité là où l'économie standard exigeait de la linéarité et de l'homogénéité. L'action humaine n'est plus réductible à une seule variable (l'utilité), elle est le produit de forces hétérogènes et potentiellement incommensurables. Ce fractionnement de la structure motivationnelle de l'agent amène J. Elster (2010, p. 67) à formuler une distinction cruciale entre les actions guidées par l'intérêt et celles gouvernées par les normes sociales ou les principes moraux. L'*homo economicus* est un être dont les choix sont strictement orientés vers le futur : il évalue les conséquences de ses actes et ajuste son comportement pour optimiser ses gains futurs. À l'inverse, Jon Elster montre que le désintéressement se manifeste souvent à travers des comportements orientés vers le passé ou le présent, insensibles aux calculs de conséquences. C'est notamment le cas lors de l'obéissance stricte à des normes sociales intériorisées.

Lorsqu'un individu refuse de tricher alors qu'il a la certitude absolue de ne pas être pris, ou lorsqu'il participe à un vote de masse alors que son bulletin individuel a une probabilité mathématiquement nulle d'infléchir le résultat du scrutin, il n'agit pas en calculateur d'utilité. Il obéit à un impératif catégorique local ou à une norme de réciprocité. De ce point de vue, l'erreur

des économistes a été de vouloir fondre ces deux logiques distinctes dans un seul et même moule conceptuel. J. Elster (2003) soutient que les normes sociales possèdent une force causale autonome qui ne dérive pas de l'utilité individuelle, mais qui entre fréquemment en collision avec elle. L'être humain elstérien est donc un être structurellement divisé, capable de suspendre son intérêt au nom d'une règle. Pour l'épistémologue, cela implique que le modèle explicatif des sciences sociales doit renoncer à l'élégance de l'unification mathématique pour adopter un pluralisme explicatif. On ne peut pas déduire l'action humaine d'une équation unique. Il faut reconstituer la grammaire des motivations qui a présidé au choix, en admettant que les motivations morales ou normatives ont un statut ontologique aussi réel et aussi contraignant que les motivations matérielles.

L'analyse elstérienne du désintéressement, si elle réussit brillamment à assainir la logique interne de la philosophie de l'action en la libérant du carcan tautologique de l'utilitarisme, soulève néanmoins d'importantes difficultés épistémologiques lorsqu'on tente de la confronter à l'exigence de la preuve empirique. Le cœur du problème réside dans l'asymétrie méthodologique flagrante qui sépare l'identification de l'intérêt et celle du désintéressement. Identifier l'intérêt personnel d'un agent est une opération scientifique relativement aisée : elle s'appuie sur des indicateurs matériels, observables et mesurables (gains financiers, évitement de peines, accumulation de capital symbolique ou de ressources). En revanche, prouver l'existence d'un désintéressement *pur* s'avère être un défi épistémologique presque insurmontable pour l'observateur extérieur, et parfois pour l'agent lui-même. Comment isoler de manière indiscutable une motivation désintéressée dans le tissu complexe de l'action humaine ?

Même dans les situations de laboratoire les plus contrôlées (comme les jeux de l'ultimatum ou du dictateur en économie comportementale), l'hypothèse d'un « résidu d'égoïsme » subtil, sous la forme d'une autosatisfaction narcissique, d'un plaisir intrinsèque à être vertueux ou d'un calcul inconscient de réciprocité, ne peut jamais être totalement éliminée par le chercheur. Dès lors, la posture de J. Elster s'expose à une critique symétrique à celle qu'il adresse à ses adversaires. En effet, en affirmant que le désintéressement existe dans une pureté

irréductible, ne bascule-t-il pas à son tour dans une forme d'affirmation métaphysique difficilement vérifiable ? Si l'impérialisme économique pèche par un excès d'infalsificabilité en prétendant voir de l'intérêt partout, le dualisme motivationnel de Jon Elster souffre d'un défaut de testabilité directe. Le désintéressement pur risque de devenir un concept régulateur ou un idéal type wébérien, fort utile pour critiquer les modèles existants, mais singulièrement complexes à opérationnaliser dans le cadre d'une science empirique rigoureuse.

De surcroît, en insistant sur la rareté, mais la réalité du désintéressement, Jon Elster fragmente le sujet de la science sociale sans offrir de critère prédictif permettant de savoir *quand* un agent basculera de l'égoïsme calculateur au désintéressement normatif. L'explication gagne indéniablement en réalisme psychologique et en profondeur philosophique ce qu'elle perd en puissance nomologique et en capacité de prévision. C'est le prix épistémologique que Jon Elster accepte de payer : celui d'une science sociale qui troque l'illusion de la certitude mathématique contre la vérité des paradoxes humains. En montrant également que le désintéressement brise le dogme de l'intérêt personnel, nous avons mis en évidence une première faille dans le modèle de l'économie standard. Cependant, l'insuffisance de l'*homo economicus* ne se limite pas à l'existence de motivations altruistes. Elle réside également dans l'incapacité de l'agent à maintenir une cohérence logique interne. Il convient donc de passer de l'examen des valeurs de l'acteur à l'analyse des dysfonctionnements de sa propre machinerie cognitive. C'est ce saut conceptuel qui nous mène à disséquer l'anatomie de l'irrationalité, non plus comme une absence de logique, mais comme un ensemble de mécanismes psychologiques et passionnels structurés.

2. Les mécanismes cognitifs et les passions comme objets de science

Si le premier volume du *Traité critique de l'homme économique* s'attache à briser l'unification motivationnelle des sciences sociales par l'introduction du désintéressement, le second volume, *L'Irrationalité*, s'attaque à un dogme encore plus ancré : celui de la cohérence interne du sujet. Pour la théorie du choix rationnel,

l'agent peut certes poursuivre des buts variés, mais sa machinerie cognitive doit fonctionner de manière optimale (J. Elster, 2010, p. 42). Ses croyances doivent dériver logiquement des informations disponibles et ses actions doivent s'aligner sur ses croyances. Face à ce modèle idéal, l'épistémologie classique s'est souvent trouvée démunie, reléguant l'irrationalité au rang de « bruit » inexplicable, de folie passagère ou de défaillance purement négative, rebelle à toute modélisation scientifique.

L'apport proprement révolutionnaire de Jon Elster (2010, p. 45) consiste à extraire l'irrationalité de ce non-lieu épistémologique pour en faire un objet de science à part entière, doté d'une structure interne et d'une logique causale rigoureuse. C'est ce qu'il formule explicitement dans le cœur de son analyse en ces termes : « L'irrationalité n'est pas une simple absence de rationalité. C'est un déraillage de la rationalité, un processus par lequel l'esprit est activement conduit à former des croyances erronées ou à prendre des décisions contraires à ses propres jugements en raison de l'impact des passions. » Cette définition déplace radicalement le problème. L'irrationalité n'est pas un simple manque, un vide théorique ou une bêtise brute. Elle est le produit d'un processus *actif* et structuré. L'esprit humain possède une propension inhérente à se duper lui-même, et ce déraillage obéit à des micro-mécanismes récurrents que le philosophe des sciences peut isoler, répertorier et expliquer.

Le premier grand mécanisme analysé par J. Elster (2010, p. 47) sous l'angle de l'irrationalité cognitive est celui des croyances motivées, c'est-à-dire le fait de prendre ses désirs pour des réalités. Dans le modèle standard de la rationalité, le processus de décision suit une trajectoire linéaire et descendante : l'agent recueille des informations sur le monde, ces informations forment des croyances objectives, et ces croyances, combinées à ses désirs, déterminent l'action. La rationalité exige ici une étanchéité stricte entre les désirs et les croyances : ce que je *veux* ne doit en aucun cas influencer ce que je *crois* être vrai. La croyance doit être uniquement indexée sur les preuves empiriques disponibles. Jon Elster montre que l'irrationalité s'introduit précisément lorsque cette frontière s'effondre sous l'effet d'un court-circuit causal. Dans le cas de la croyance motivée, le désir de l'agent n'attend pas l'aval de l'évaluation rationnelle. Il remonte le courant de la décision pour aller configurer directement la

croyance. L'agent se met à croire que le monde est conforme à ce qu'il désire qu'il soit, non pas parce qu'il manque de données, mais parce que son désir biaise activement le traitement des données. Il surévalue les indices qui confirment son souhait et ignore superbement les preuves massives qui le contredisent.

Pour l'épistémologue, ce mécanisme est d'un intérêt capital, car il redéfinit le statut de l'erreur. L'erreur irrationnelle n'est pas un défaut d'information (une ignorance), mais un traitement biaisé et motivé de l'information. L'esprit rationnel n'est pas celui qui sait tout, mais celui dont le système de formation des croyances est immunisé contre l'interférence de ses propres préférences. En cartographiant ce court-circuit, J. Elster (2010, p. 52) montre que l'esprit irrationnel conserve une forme d'organisation logique : il utilise la raison pour rationaliser *a posteriori* une croyance dictée par la passion. Le second pilier de cette anatomie de l'irrationalité est la faiblesse de la volonté, un concept que Jon Elster réactualise en dialogue direct avec la philosophie analytique de l'action, notamment de D. Davidson (1993) et la psychologie cognitive. Cette faiblesse de volonté se définit comme le fait d'agir délibérément contre son propre jugement le mieux pondéré. En fait, pour l'économie classique, ce comportement est purement et simplement impossible : si un agent choisit l'option X plutôt que l'option Y, c'est la preuve irréfutable qu'il préférait X à Y au moment du choix. Parler de la faiblesse de la volonté ne serait qu'une hypocrisie morale.

Mais Jon Elster refuse ce réductionnisme comportementaliste. Il démontre que la faiblesse de la volonté est une réalité psychologique majeure, provoquée par l'irruption des émotions et des passions à court terme. Les émotions, par exemple, la colère, la peur, le désir immédiat, la jalousie, etc., agissent comme des forces gravitationnelles qui altèrent temporairement l'horizon temporel de l'agent. Sous l'effet d'une émotion vive, l'avenir lointain subit un escompte hyperbolique : le présent immédiat est démesurément surévalué au détriment du bien-être futur. L'agent ne change pas d'avis sur ce qui est objectivement le meilleur pour lui à long terme, mais sa volonté est momentanément capturée par l'urgence émotionnelle. Ainsi dit, la volonté brise le postulat de la transparence et de l'unité du sujet. Le moi n'est pas un bloc monolithique et rationnel. Il est un système dynamique traversé par des forces conflictuelles où le

jugement et la volonté peuvent se désaligner. L'explication de l'action humaine doit donc intégrer ces ruptures de continuité interne, ce qui exige de substituer à l'analyse purement logique des préférences une véritable mécanique des forces psychologiques.

L'effort de Jon Elster pour fonder une science de l'irrationalité se heurte toutefois à une aporie épistémologique classique, que la philosophie de l'esprit n'a cessé de formuler à l'encontre des théories de l'auto-illusion : le paradoxe de la dupe et du dupeur. Pour qu'un individu puisse se duper lui-même, il faut, semble-t-il, qu'il soit à la fois l'auteur du mensonge et sa victime. Or, d'un point de vue strictement logique, pour construire une croyance fausse afin de masquer une réalité déplaisante, il faut d'une certaine manière que l'agent connaisse déjà cette réalité déplaisante pour chercher à s'en cacher. Comment peut-on savoir une chose (pour l'éviter) tout en l'ignorant (pour être dupe) ? Si le processus est entièrement conscient, il s'agit d'un simple mensonge à soi-même, ce qui est logiquement contradictoire. S'il est entièrement inconscient, nous basculons dans la psychanalyse freudienne, avec ses instances séparées (le Ça, le Moi, le Surmoi), une approche que Jon Elster rejette catégoriquement en raison de son manque de rigueur méthodologique et de son caractère non falsifiable. Il tente de résoudre cette contradiction en décrivant des « mécanismes sans sujet », des processus causaux infra-intentionnels où le désir produit la croyance de manière automatique, sans qu'une intention délibérée de se tromper soit nécessaire.

Cependant, sur le plan épistémologique, cette solution fragilise la cohérence de l'individualisme méthodologique dont Jon Elster se réclame. Car, en postulant l'existence de mécanismes internes qui agissent « dans le dos » de l'agent, il réintroduit une forme de déterminisme causal au sein d'une philosophie qui se veut centrée sur l'action choisie. De plus, une critique majeure peut lui être adressée quant à la posture de l'observateur scientifique : si l'esprit humain est structurellement sujet à ces déraillages internes, à ces croyances motivées et à ces rationalisations *a posteriori*, comment l'épistémologue ou le sociologue peut-il garantir que sa propre théorie des mécanismes n'est pas elle-même le produit d'une croyance motivée ? En refusant la boussole de la rationalité parfaite comme description

du comportement, Jon Elster prive les sciences sociales d'un étalon de mesure stable. L'irrationalité devient intelligible, certes, mais au prix d'une fragilisation générale des fondements de la certitude empirique. Le chercheur ne fait plus face à des agents logiques dont on peut anticiper les trajectoires, mais à des systèmes psychologiques complexes dont on ne peut comprendre les dérives que de manière rétrospective.

L'analyse de l'irrationalité a révélé que les déraillages de l'esprit, tels que les croyances motivées ou la faiblesse de la volonté, possèdent une logique interne accessible à la science. Pour autant, cette découverte pose un défi méthodologique majeur à l'épistémologie : si les choix humains sont structurellement soumis à l'interférence des passions, comment bâtir une science sociale rigoureuse ? Face à l'échec des lois prédictives universelles mises en échec par ces biais psychologiques, il s'agit désormais d'analyser la réponse épistémologique de Jon Elster. Nous verrons ainsi comment il substitue au modèle nomologique une nouvelle scientificité fondée sur une boîte à outils de « mécanismes » causaux indéterminés.

3. Le paradigme des mécanismes contre le modèle nomologique

L'aboutissement de la critique elstérienne du modèle standard ne réside pas dans un constat d'impuissance sceptique qui condamnerait les sciences humaines à l'anecdote ou au pur descriptivisme littéraire. Le véritable enjeu du *Traité critique de l'homme économique* est de fonder une scientificité de rechange. Dès lors que l'on a démontré que le cadre rigide de la maximisation de l'intérêt personnel éclate sous le poids du désintéressement et que la cohérence interne du sujet est constamment subvertie par l'anatomie de l'irrationalité, le philosophe des sciences fait face à un dilemme : faut-il abandonner tout idéal de scientificité ou redéfinir radicalement ce que signifie « expliquer » le comportement humain ? Jon Elster choisit la seconde voie. Il opère un coup de force épistémologique en destituant le modèle nomologique-déductif (G. C. Hempel, 1965), qui domine les sciences de la nature et l'économie mathématique, au profit d'un paradigme fondé sur une boîte à outils de « mécanismes » causaux. Pour saisir la portée de cette

substitution paradigmatique, il faut se référer à la définition exacte que J. Elster (2010, p. 212) donne du mécanisme, qui constitue le cœur de son alternative méthodologique : « Un mécanisme est un déclencheur causal fréquemment reconnu et facilement identifiable qui s'opère dans des conditions inconnues ou avec des conséquences indéterminées. Ils nous permettent d'expliquer, mais non de prédire. »

Cette formule jette les bases d'une épistémologie post-nomologique. Elle assume une rupture conceptuelle majeure en séparant deux notions que la philosophie des sciences classiques, de René Descartes à Carl Hempel, avait toujours considérées comme indissociables : l'explication et la prédiction. Pour Jon Elster, on peut parfaitement expliquer un comportement *a posteriori* sans pour autant être capable de le prédire *a priori*. Avant d'analyser la structure logique de ces mécanismes, il convient de comprendre comment l'agent elstérien réagit à sa propre vulnérabilité. L'homme n'est pas simplement irrationnel. Il est, de manière plus subtile, « imparfaitement rationnel », c'est-à-dire qu'il est capable de prendre conscience de ses propres biais cognitifs et de la faiblesse de sa volonté, et d'anticiper ses déraillages futurs. C'est ici qu'intervient le concept cardinal de contraignabilité préalable, que J. Elster (1984) illustre fameusement à travers le mythe d'Ulysse et les Sirènes. Ulysse sait qu'en entendant le chant des Sirènes, sa raison va vaciller et qu'il précipitera son navire sur les récifs. Pour rester rationnel à long terme, il choisit de se boucher les oreilles avec de la cire et de se faire attacher au mât, privant son « moi futur » de la liberté d'agir de manière irrationnelle. D'un point de vue épistémologique, ce concept redéfinit la nature du choix.

En effet, la rationalité ne consiste plus à optimiser ses choix, instant après instant, dans un vide social. Au contraire, elle consiste à ériger des barrières, des règles, des institutions ou des constitutions politiques pour se protéger de ses propres passions futures. L'explication des comportements et des institutions humaines, comme par exemple les lois constitutionnelles qui empêchent un gouvernement de modifier les règles électorales à son avantage en période de crise, doit donc être lue comme une tentative de s'auto-contraindre. La science sociale elstérienne intègre ainsi cette réflexivité de l'agent : l'homme est un animal politique précisément, du fait qu'il se sait irrationnel et qu'il

invente des prothèses institutionnelles pour feindre la rationalité. Le cœur de la nouvelle scientificité elstérienne repose dès lors sur la modélisation de l'esprit humain, non pas comme une machine linéaire réglée par des lois générales, mais comme un réservoir de mécanismes psychologiques qui s'activent souvent par paires antagonistes. Face à une même situation initiale, l'esprit humain ne réagit pas de manière uniforme. Il peut déclencher deux processus causaux parfaitement inverses, sans que le chercheur puisse déterminer à l'avance lequel l'emportera.

Jon Elster illustre cette thèse par de splendides analyses psychologiques issues de la littérature classique, de Montaigne à Stendhal, et de la psychologie cognitive. Prenons le cas de la frustration face à l'inaccessible. Un agent qui ne peut obtenir un bien peut réagir selon le mécanisme des « raisins secs » ou des « raisins verts », issu de la fable d'Ésope : il va dénigrer l'objet interdit pour réduire sa dissonance cognitive (*« ce raisin est trop vert, il n'est bon que pour les goujats »*). C'est le mécanisme de la formation des préférences adaptatives. Mais l'agent peut tout aussi bien réagir selon le mécanisme inverse, celui des « souris vertes » ou du « fruit défendu » : le fait qu'un objet soit inaccessible le rend démesurément attractif, amplifiant le désir au-delà de toute rationalité. Le modèle nomologique classique de C. Hempel (1965) s'effondre. Une loi scientifique exige la régularité stricte de la cause et de l'effet. Or, ici, la cause, c'est-à-dire l'inaccessibilité, produit parfois le résultat A, et parfois le résultat B. J. Elster soutient que cette indétermination n'est pas un aveu d'échec pour la science sociale, mais le reflet exact de la complexité de la psychologie humaine.

Le mécanisme permet d'expliquer l'action une fois qu'elle s'est produite, le chercheur peut retracer le chemin causal qui a mené aux « raisins verts », mais il interdit toute prédiction certaine. La scientificité change de nature : elle abandonne l'idéal de la prévision divinatoire pour devenir une science de la reconstruction des processus. Cette refonte épistémologique, bien qu'extrêmement séduisante par son réalisme psychologique et son refus des abstractions mathématiques déconnectées du réel, suscite de profondes réserves critiques quant à sa viabilité méthodologique à long terme. La principale objection que l'on peut adresser au paradigme des mécanismes de Jon Elster concerne le statut de la réfutabilité popperienne.

Si une théorie affirme qu'un événement X peut produire l'effet Y (par le mécanisme A) ou l'effet inverse Z (par le mécanisme B), comment peut-on jamais la prendre en défaut ? Quels que soient les faits observés *a posteriori* (que l'agent dénigre le bien ou qu'il en devienne obsédé), la boîte à outils de Jon Elster disposera toujours d'un mécanisme pour justifier le résultat. Une telle approche ne court-elle pas le risque de glisser vers une forme de descriptivisme *ad hoc*, où le chercheur se contente de raconter une histoire causale plausible après coup, sans jamais mettre sa théorie en péril face à l'expérience future ? En renonçant au pouvoir prédictif, Jon Elster prive les sciences sociales de leur outil de contrôle le plus rigoureux : le test de la prédiction erronée. De surcroît, le remplacement des lois universelles par une multitude de micro-mécanismes fragmentarise la connaissance.

La science sociale elstérienne ressemble à un catalogue infini de réactions psychologiques locales plutôt qu'à un édifice théorique unifié. On peut légitimement se demander si cette démarche ne dissout pas l'ambition même des sciences de la société, qui est de dégager des régularités macro-sociales au-delà des singularités psychologiques des individus. En se focalisant exclusivement sur les rouages internes de l'esprit, Jon Elster tend à psychologiser les phénomènes sociaux, oubliant parfois que les structures économiques, politiques et matérielles exercent des contraintes qui dépassent la simple mécanique des passions individuelles. L'intelligibilité rétrospective est certes un gain précieux pour le philosophe de l'action, mais elle laisse le sociologue et le décideur public désarmés face à l'avenir, confirmant le coût épistémologique exorbitant de cette cure de modestie scientifique.

Conclusion

Au terme de cette enquête épistémologique, il apparaît avec évidence que la déconstruction de l'*homo economicus* opérée par Jon Elster ne constitue pas une charge nihiliste contre la possibilité d'une science sociale rigoureuse. Bien au contraire, en affrontant les deux grandes anomalies théoriques que sont le désintéressement et l'irrationalité, Jon Elster accomplit un geste de salubrité méthodologique indispensable. Il arrache la

philosophie de l'action à l'illusion d'une harmonie mathématique universelle pour la confronter au principe de réalité de la psychologie humaine. L'enjeu de son œuvre n'est pas d'exclure la rationalité du champ de la connaissance, mais de la destituer de son statut infondé de description universelle du monde pour lui restituer sa véritable fonction : celle d'une norme fragile, un idéal régulateur que l'homme s'efforce péniblement d'atteindre.

La trajectoire de notre réflexion a permis de mesurer l'ampleur de cette rupture paradigmatique. En premier lieu, l'analyse du désintéressement a mis au jour le caractère tautologique et ultimement infalsifiable de l'impérialisme économique standard. En démontrant que le dévouement normatif et l'altruisme pur ne se laissent pas réduire à des calculs égoïstes déguisés, Jon Elster a brisé le dogme de l'unification motivationnelle, imposant un pluralisme explicatif qui redonne droit de cité aux valeurs et aux normes sociales. En second lieu, l'anatomie de l'irrationalité a révélé que les déraillages de l'esprit, qu'il s'agisse du court-circuit des croyances motivées ou de la fracture interne de la volonté, ne relèvent pas d'un chaos inintelligible. L'irrationalité elstérienne dispose d'une structure interne, d'une causalité propre et de micro-mécanismes réguliers que le chercheur peut rigoureusement modéliser. Le pivot de cette révolution épistémologique réside cependant dans la substitution du modèle des « mécanismes » au modèle classique des « lois générales ». En affirmant qu'un mécanisme est un déclencheur causal récurrent, mais dont les conséquences sont indéterminées et dépendent de paires antagonistes, à l'instar, par exemple, du conflit entre les « raisins verts » et le « fruit défendu », Jon Elster opère un divorce salutaire entre l'explication et la prédiction.

Les sciences humaines doivent faire le deuil de la prévisibilité divinatoire au profit d'une intelligibilité rétrospective. Malgré le risque de fragmentation psychologique et de perte de réfutabilité a priori, ce réalisme méthodologique offre le seul sol scientifique ferme. La pensée de Jon Elster fonde ainsi une anthropologie renouvelée : l'homme y est un être divisé, imparfaitement rationnel, mais capable d'anticiper ses failles par des stratégies de contrainte préalable. En troquant l'élégance formelle des lois universelles contre la pertinence concrète d'une boîte à outils de mécanismes, Elster restitue toute sa dignité aux sciences sociales pour saisir la complexité tragique de l'action. Mais cette destitution de la certitude

ne nous invite-t-elle pas à reconsidérer la liberté humaine elle-même, non plus comme une souveraineté de la raison, mais comme la capacité tragique et créatrice d'agir au cœur même de nos propres paradoxes ?

Références bibliographiques

ELSTER Jon, 2009. *Le Désintéressement. Traité critique de l'homme économique I*, Paris, Éditions du Seuil, coll. « Les livres du nouveau monde », 384 p.
ELSTER Jon, 2010. *L'Irrationalité. Traité critique de l'homme économique II*, Paris, Éditions du Seuil, coll. « Les livres du nouveau monde », 352 p.
ELSTER Jon, 1984. *Ulysse et les Sirènes. Études sur la rationalité et l'irrationalité*, trad.fr. C. Malamoud, Paris, Éditions de la Maison des sciences de l'homme, 248 p.
ELSTER Jon, 2003. *Le Laboureur et ses enfants. Deux essais sur les limites de la rationalité*, trad.fr. M. Saint-Upéry, Paris, Éditions de minuit, coll. « Arguments », 208 p.
ELSTER Jon, 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*, Cambridge, Cambridge University Press, 496 p.
DAVIDSON Donald, 1993. *Actions et événements*, trad.fr. P. Engel, Paris, Presses Universitaires de France, coll. « Épiméthée », 416 p.
HEMPEL G. Carl, 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York, The Free Press, 504 p.
KUHN Thomas, 2008. *La Structure des révolutions scientifiques*, trad. de l'américain par L. Meyer, Paris, Flammarion, coll. « Champs », 284 p.
POPPER Karl, 2007. *La Logique de la découverte scientifique*, trad.fr. N. Thyssen-Rutten et P. Devaux, Paris, Payot, 480 p.
SEN Amartya, 1977, « Rational Fools: A Critique of the Behavioral Foundations of Economic Theory », *Philosophy & Public Affairs*, vol. 6, n° 4, p. 317-344.